Journal of
Hydroinformatics

© 2021 The Authors

Journal of Hydroinformatics Vol 23 No 4, 864 doi: 10.2166/hydro.2021.023

A hybrid wavelet-Lyapunov exponent model for river water quality forecast

Jiping Jiang[a,b,*], Sijie Tang[b], Rentao Liu[c], Bellie Sivakumar[d,e], Xiaoye Wu[f] and Tianrui Pang[a]

[a] State Key Laboratory of Urban Water Resource and Environment, School of Environment, Harbin Institute of Technology, Harbin 150090, China
[b] Shenzhen Municipal Engineering Lab of Environmental IoT Technologies, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[c] School of Municipal Engineering Technology, Heilongjiang College of Construction, Harbin 150025, China
[d] Department of Civil Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India
[e] State Key Laboratory of Hydroscience and Engineering, Tsinghua University, Beijing 100084, China
[f] Faculty of Civil Engineering, RWTH Aachen University, Aachen 52062, Germany
*Corresponding author. E-mail: jiangjp@sustech.edu.cn

ABSTRACT

The use of spectral theory and chaos theory on river water quality modeling is reported in a very limited way. This study proposes a wavelet-maximum Lyapunov exponent (WMLE) hybrid model for river water quality dynamics, combining spectral theory and chaos theory. The methodology involves the following major steps: (1) use of wavelet transformation to filter the noisy signal in the water quality time series; (2) reconstruction of phase space to embed the water quality time series and determine the trajectory of the underlying dynamics; and (3) identification of the presence/absence of chaos and prediction using the largest Lyapunov exponent value. Case studies on the Huaihe River in China and the Potomac River in the United States, as representatives of low-frequency and high-frequency forecast, show average relative errors on weekly dissolved oxygen (DO), chemical oxygen demand (COD), and ammonia nitrogen (NH_3-N) data are 2.35%, 4.53%, and 18.85%, and on 15-minute based DO data are 1.185%. It also indicates that the hybrid model performs better to some extent when compared to the purely Lyapunov exponent model, ARMA model, and ANN model. This study is a proof that the combination of spectral theory and chaos theory is promising to describe and predict fluctuation of particular water quality indicators in rivers.

Key words: chaos theory, hybrid model, Lyapunov exponent, nonlinearity, water quality, wavelet

HIGHLIGHTS

- River water quality dynamics of DO, COD and NH_3-N fluctuation present chaos behavior.
- A hybrid Wavelet-Lyapunov exponent model for water quality forecast is newly proposed.
- WMLE model performs well for weekly DO and COD forecast but relative poor for NH_3-N.
- Forecast for 15-minute based DO time series achieve high accuracy in Potomac river.
- WMLE model slightly overweight ANN, ARMA and pure Lyapunov model in the study cases.

INTRODUCTION

Forecasting the river water quality fluctuation is one of the key elements of watershed water quality management. It provides an early warning for chemical release incidents, risk assessment for water usage, guiding the operation of water supply and drainage system. The existing modeling or forecast tools are normally grouped as mechanistic models and data-driven models. Due to the ability to represent physical processes, the mechanistic models are commonly used in pollution control planning of water environment or situational analysis of long-term transition (Gordillo *et al.* 2019). However, those models generally require data of variables, which are hard to collect in reality.

Data-driven models, due to their ability to identify patterns in the system dynamics often based on a single time series alone, have clear advantages, especially in short-term prediction (Babovic 2005; Solomatine & Ostfeld 2008; Tongal & Booij 2018). Although the types of data used in the data-driven models are much less when compared to those in the mechanistic models, the data-driven models often require long-term observations of water quality (Sun *et al.* 2010). With growing infrastructure and analysis methods for environmental monitoring and data accumulation around the world in recent years, the arrival of the era of data-intensive innovation provides new opportunities for the development and use of more sophisticated data-driven models. In this regard, the rapid development of many machine learning techniques certainly provides new avenues for modeling and prediction of water quality dynamics (Meszaros & El Serafy 2018; Yajima & Derot 2018).

River systems are open, complex, and nonlinear systems. The dynamics of water quality (and quantity) in rivers are often governed by complex and nonlinear interactions among the variables influencing river systems (Lintern *et al.* 2018). Such dynamics are also often subjected to sensitive dependence on initial conditions, which limits our ability to predict the dynamics in the long term but, nevertheless, allows reliable short-term predictions. Therefore, in the context of complex and nonlinear interactions with sensitive dependence on initial conditions, the dynamics of water quality (and quantity) in rivers can be studied using the concepts of nonlinear dynamic and chaos theories; see the monograph by Sivakumar (2017) for a comprehensive account of chaos theory applications to hydrological systems.

Chaos theory was discovered in 1963 by the famous American meteorologist Edward Lorenz through numerical experiments of the atmosphere (Lorenz 1963). The term 'chaos' was coined, and has been widely used to refer to situations where complex and 'random-looking' behaviors arise from simple deterministic systems with sensitive dependence on initial conditions. The development of many methods for identification of chaos in time series since the 1980s has led to their applications in many different scientific and engineering fields, such as meteorology, hydrology, geology, biology, medical science, electronics, transportation, and astronomy. In the field of hydrology, there have been numerous applications of the concepts of chaos theory to study many different processes, data, and problems associated with river systems. These studies include analysis of rainfall, river flow, and sediment transport processes (Yu *et al.* 2004; Sivakumar 2005; Huang *et al.* 2017), addressing system characterization, prediction, missing data estimation, and data disaggregation, among others.

The outcomes of such studies are certainly encouraging, especially for more reliable short-term predictions of river system dynamics when compared to the predictions achieved using stochastic and other data-based approaches. A particular advantage of the concepts of chaos theory is that they can offer useful clues regarding the predictability of system dynamics. For instance, the Lyapunov exponents provide important information on the predictability horizon for a system (Huang *et al.* 2017), as the Lyapunov exponents are essentially the average exponential rates of divergence (expansion) or convergence (contraction) of nearby orbits in the phase space which is essentially a graph or a coordinate diagram, whose coordinates represent the variables necessary to completely describe the state of the system at any moment. Recently, some studies have reported forecasting high-dimensional chaotic systems by long short-term memory networks (Lu *et al.* 2018; Vlachas Pantelis *et al.* 2018).

While the chaos theory has found widespread applications in river system studies, the applications have largely been to study water quantity, such as river flow discharge and water level (Tongal 2013, 2020; Tongal & Berndtsson 2017). To our knowledge, chaos applications to study river water quality dynamics have, thus far, been very limited. One possible reason for this situation may be the general belief that chaos identification and prediction methods require infinite (or at least very long) and noise-free time series, while real water quality time series are often very short and always contaminated with noise. However, studies have shown that the data size issue for chaos identification and prediction methods is not as serious as it is generally believed to be and that the methods can provide reliable results even when the data size is small; see, for example, Sivakumar *et al.* (2002), Sivakumar (2005), Siek & Solomatine (2010), and Zhang (2013) for some details on the issue of data size in the application of chaos identification and prediction methods for river system-related time series.

In recent years, there has been increasing attention on developing hybrid models to increase the predictability in data-driven modeling (Huang *et al.* 2017; Li *et al.* 2017). A good understanding of the specific advantages and disadvantages of the existing models allows selection of two (or more) models that are suitable for the problem/data at hand for more reliable modeling and predictions. Wavelet technology has the feature of good localization performance of time frequency, and it has strong adaptability to represent the local feature of signal in the time-frequency domain (Sang 2013; Alizadeh *et al.* 2018). There have been many studies that have successfully combined wavelets with other machine learning methods, such as artificial neural networks (ANNs), support vector machine (SVM) (Yu *et al.* 2004; Lin *et al.* 2013), and extreme learning machine (Roushangar *et al.* 2018). For example, Shi *et al.* (2017) proposed a coupled wavelet-ANN model and applied it for early warning of water quality conditions. Barzegar *et al.* (2018) combined wavelets and extreme learning machine to predict electrical conductivity. Ren *et al.* (2013) proposed a combination of adaptive neural-fuzzy inference system and wavelet analysis to predict monthly runoff and reported very good prediction. With the usefulness and advantages of the concepts of chaos theory, there seems great potential to couple chaos methods and wavelets to enhance our analysis, modeling, and prediction of time series. To our knowledge, no study has attempted to study the effectiveness of the hybrid model for water quality time series. The present study makes an effort to further advance research in this direction.

In the present study, a hybrid prediction model is developed by coupling chaos theory and wavelets for water quality prediction. In particular, a coupled wavelet-Lyapunov exponent model is proposed. The wavelet is used to effectively extract feature information of the observed time series through decomposition and reconstruction. It can be applied in short time series and greatly simplifies the calculation complexity of prediction model of the maximal Lyapunov exponent (MLE); meanwhile, it increases flexibility and improves the forecast precision of the MLE model when applying on larger fluctuations of time-sequence data. Case studies are carried out based on long-term water quality monitoring on Huaihe River, China and high-frequency monitoring on Potomac River, United States (US). Based on the data availability, chemical oxygen demand (COD), dissolved oxygen (DO), and ammonia nitrogen (NH_3-N), nitrite-nitrogen (NO_2-N) were the main water quality indicators under investigation.

As well, this work discussed the following hypotheses or questions: (1) Is there presence/absence of chaos on the river water quality variation? (Does the chaos exist in the river water quality variation?) (2) Can the chaos theory-based modeling approach be qualified for predicting different water quality parameters? (3) Does the hybrid method combining with spectral theory improve the performance of purely chaos-based predication? (4) How is the performance of the hybrid method compared with traditional data-driven models, i.e., ARMA, ANN, etc. The following sections include illustration of methodology, descriptions of study area and monitoring data, results and discussion on chaos identification, model performance and comparison.

METHODOLOGY

Decomposition and reconstruction of wavelet

Wavelet transformation has good localized performance of time frequency and it is one kind of conversion tool of time frequency that is widely used (Sang 2013). It has high frequency resolution and low time resolution in the low-frequency part and low frequency resolution and high time resolution in the high-frequency part. Therefore, it has strong adaptability to represent the localized feature of signal in the time-frequency domain. Wavelet function in wavelet analysis can be diverse, and different wavelet functions are often needed in different situations in applications. The *db*-wavelet has reliable time-frequency resolution, and it can match the original signal to the greatest degree. The *db5* wavelet has been found to provide optimal effect upon repeated calculations and analysis in many applications (Sang 2013; Alizadeh *et al.* 2018; Roushangar *et al.* 2018), and so this wavelet function is selected here. The specific steps are as follows (Kim *et al.* 1999):

Step 1 – Wavelet decomposition: the wavelet decomposition is done according to:

$$c_j(t) = \sum_l h(l)c_{j-1}(t + 2^{j-1}l), j = 1, 2, \cdots, k \tag{1}$$

$$\omega_j(t) = c_{j-1}(t) - c_j(t), j = 1, 2, \cdots, k \tag{2}$$

where $c_j(t)$ is the decomposition scale coefficient corresponding to decomposition scale j, $h(l)$ is the low pass filter, and $\omega_j(t)$ is wavelet coefficient corresponding to decomposition scale j.

Step 2 – Wavelet reconstruction: the wavelet reconstruction is done according to:

$$c_0(t) = \sum_{j=1}^{k} \omega_j(t) + c_k(t) \tag{3}$$

where $c_0(t)$ is the time series upon reconstruction, $c_k(t)$ is an approximate series, and $\sum_{j=1}^{k} \omega_j(t)$ is a detailed series.

Identification of chaos

There exist many methods for identification and prediction of chaos in a time series. These include correlation dimension method, Lyapunov exponent method, false nearest neighbor algorithm, and nonlinear local approximation prediction method, among others; see, for example, Kantz & Schreiber (2004) and Sivakumar (2017) for details. Almost all of the chaos identification and prediction methods use the concept of phase space for embedding (reconstruction) of the time series (e.g., Packard *et al.* 1980; Takens 1981). In the present study, the correlation dimension method and the Lyapunov exponent method are used for identification of chaos in the water quality time series and, therefore, they are briefly described here. Since it is common to both methods, it is presented first.

Let us consider a water quality time series $x(t_i)$. The phase space for this time series can be reconstructed, as follows:

$$
\begin{aligned}
&Y(t_1) = [x(t_1), x(t_1+\tau), x(t_1+2\tau), \cdots, x(t_1+(m-1)\tau)]\\
&Y(t_2) = [x(t_2), x(t_2+\tau), x(t_2+2\tau), \cdots, x(t_2+(m-1)\tau)]\\
&Y(t_i) = [x(t_i), x(t_i+\tau), x(t_i+2\tau), \cdots, x(t_i+(m-1)\tau)]\\
&Y(t_M) = [x(t_M), x(t_M+\tau), x(t_M+2\tau), \cdots, x(t_M+(m-1)\tau)]
\end{aligned}
\tag{4}
$$

where $Y(t_i)$ are a series of phase points (vectors) embedded in phase space of m dimensions, $i = 1, 2, \cdots M$, $M = n - (m-1)\tau$, and τ is delay time. The selection of the embedded dimension m and time delay τ is of critical importance in the process of reconstructing the phase space for an appropriate embedding. While there are specific methods to select an appropriate delay time (albeit some potential pitfalls), the most suitable embedding dimension is normally obtained by increasing the embedding dimension and identifying the best results. A brief discussion about these is offered next.

Determination of delay time τ

The selection of an appropriate delay time τ is significant for an appropriate embedding, but as such, is also oftentimes tricky. If the value of τ is too small, there is not much difference between neighboring coordinates in the reconstructed phase space (i.e., redundancy) and so it cannot reflect the system dynamics well. Conversely, if the value of τ is too big, then there is not much relation between the two coordinates and they are too independent (i.e., irrelevance) to reflect the changes in the system dynamics. Therefore, τ should be appropriate and the selection should avoid both redundancy and irrelevance. The common methods to select an appropriate delay time τ or even delay time window include the autocorrelation function (ACF) (Li *et al.* 2016), mutual information (MI) (Ren *et al.* 2013), and correlation integral and its variants, including the C-C method (Liu *et al.* 2011), among others. In this study, the non-partial multiple autocorrelation coefficient method is adopted to select the delay time τ, and is given by:

$$
C_{xx}^{m}(\tau) = \frac{2}{N(N-1)} \sum_{i=0}^{N-1} \sum_{j=1}^{m-1} (x_i - \bar{x})(x_{i+j\tau} - \bar{x})
\tag{5}
$$

where $\bar{x}$ is the mean of the time series. The delay time is chosen when the time of value $C_{xx}^{m}(\tau)$ is reduced to $(1 - 1/\mathrm{e})$ of the initial value for the first time.

Determination of optimal embedded dimension *m* and correlation dimension D

The correlation dimension of a time series is a reliable indicator of the number of dominant variables governing the underlying system dynamics. For estimation of the correlation dimension method, the Grassberger-Procaccia (GP) algorithm (Li *et al.* 2013) is widely used. The main steps of the GP algorithm are as follows:

1. A small m_0 (say, $\mathrm{m}_0 = 2$) is first offered to reconstruct the phase space with time series $x_1, x_2, \cdots, x_n$, as shown in Equation (1).
2. The correlation function or integral is calculated according to:

$$
C(r) = \lim_{N\to\infty} \frac{1}{N} \sum_{i,j=1}^{N} \theta(r - |\mathrm{Y_i} - \mathrm{Y_j}|)
\tag{6}
$$

   where $|Y_\mathrm{i} - Y_\mathrm{j}|$ indicates the distance between phase space points $\mathrm{Y_i}$ and $\mathrm{Y_j}$, and $\theta(z)$ is the Heaviside step function, and $C(r)$ indicates probability that the distance between two points on the attractor in phase space is less than r.
3. The dimension d and cumulative distribution function $C(r)$ of the attractor should satisfy $d(m) = \ln C(r)/\ln r$ for certain proper scope of r. The estimated value $d(m_0)$ of correlation exponent corresponding to m_0 can be obtained by fitting.
4. Then, an embedded dimension $m_1 > m_0$ is added, and steps (2) and (3) are repeated until the corresponding estimated value $d(m)$ of dimension is convergent to a stable value. The value d obtained in this way is the correlation dimension

of the attractor, and the embedding dimension that is just above the correlation dimension value is considered to be equal to the number of variables dominantly governing the underlying system dynamics.

Lyapunov exponents

Lyapunov exponents are the average exponential rates of divergence (expansion) or convergence (contraction) of nearby orbits in the phase space. In general, the presence of a positive Lyapunov exponent in a time series indicates the presence of chaos, i.e., at least the MLE λ_1 is above zero. There are many different methods to calculate the Lyapunov exponent of a time series (Zhang 2013). However, the algorithm by Wolf *et al.* (1985) is widely used, and so is used in the present study (Huang *et al.* 2017). The main calculation process is as follows:

Suppose the Euclidean distance between two phase points Y_i and Y_j is indicated with $\|Y_\mathrm{i} - Y_\mathrm{j}\|$ in phase space of dimension d and Y_m is selected as the reference phase point. The relation between Y_{nbt} and Y_m can be indicated as $Y_{nbt} = \min[\|Y_\mathrm{m} - Y_\mathrm{j}\|]$ $(\mathrm{j} = 1,\ 2,\ \ldots,\ \mathrm{m} - 1)$, where Y_{nbt} is the nearest-neighbor phase point. The nearest-neighbor phase point of each phase point Y_j in the phase space can be calculated and sought according to this method upon reconstructing phase space with time series. Suppose L_{k-1} is the distance of the nearest-neighbor phase points in k^{th} step and it evolves into L_k upon step length k. Then, the value of the maximum Lyapunov exponent λ_1 can be obtained according to:

$$\lambda_1 = \frac{1}{k\Delta t}\log_2\frac{L_k}{L_{k-1}} \tag{7}$$

where Δt is the time interval of time sequence and $k\Delta t$ is the time needed for L_{k-1} to develop to L_k.

Prediction model of MLE

The prediction model of a chaotic time series for the MLE is as follows (Zhang 2013):

$$2^{\lambda_1 k\Delta t} = \frac{\|Y(t_M + T) - Y_{nbt}(t + T)\|}{\|Y(t_M) - Y_{nbt}(t)\|} \tag{8}$$

where $M = N - (m - 1)\tau$, k is time step of observation, T is the forecasted time of the system in advance (i.e., lead time) with $\mathrm{T} = k\Delta t$, $Y(t_M + T)$ is the evolution value of the center point $Y(t_M)$ upon time T, $Y_{nbt}(t_i)$ is the nearest-neighbor phase point of $Y(t_M)$, and $Y_{nbt}(t_i + \mathrm{T})$ is the evolution value of $Y_{nbt}(t_i)$ upon time T.

In Equation (8), only the last element in $Y(t_M + T)$ is unknown, i.e., $x(t_M + K\Delta t)$ only if $\mathrm{T} \le \tau$. Thus, the calculation equation of prediction value can be obtained as:

$$Y_{(t_M+\mathrm{k}\Delta t)}(M) = Y_{nbt(t+\Delta t)}(M) \pm \sqrt{(d2^{\lambda_1\Delta t})^2 - \sum_{j=1}^{m-1}(Y_{(t_M+\mathrm{k}\Delta t)}(j) - Y_{nbt(t+k\Delta t)}(j))^2} \tag{9}$$

Wavelet-Lyapunov exponent model

As long as the presence of chaos in a time series is identified, the hybrid WLME model can be used for decomposition of the time series and subsequent predictions. It is assembled in three steps. First, wavelet decomposition is conducted on the original time series, to obtain the low-frequency signal and high-frequency signal. Second, the low-frequency signal is predicted in the application of the MLE model and the high-frequency signals are all set as zero. Third, wavelet reconstruction is conducted in combination with low-frequency prediction result and the reset high-frequency signal to obtain the final prediction result. A flow chart of the WLE model is shown in Figure 1.

STUDY AREA AND MONITORING DATA

For watershed water quality management, routine monitoring is typically conducted on a monthly basis or weekly basis. In many cases, online high-frequency monitoring (e.g., 5 to 60 min interval) is available for physical (e.g., temperature, turbidity) and chemical (e.g., conductivity, dissolved oxygen) indicators due to the development of *in-situ* water quality sensors (Rode *et al.* 2016). This is of particular importance in monitoring drinking water resources. Considering these, in this study, we

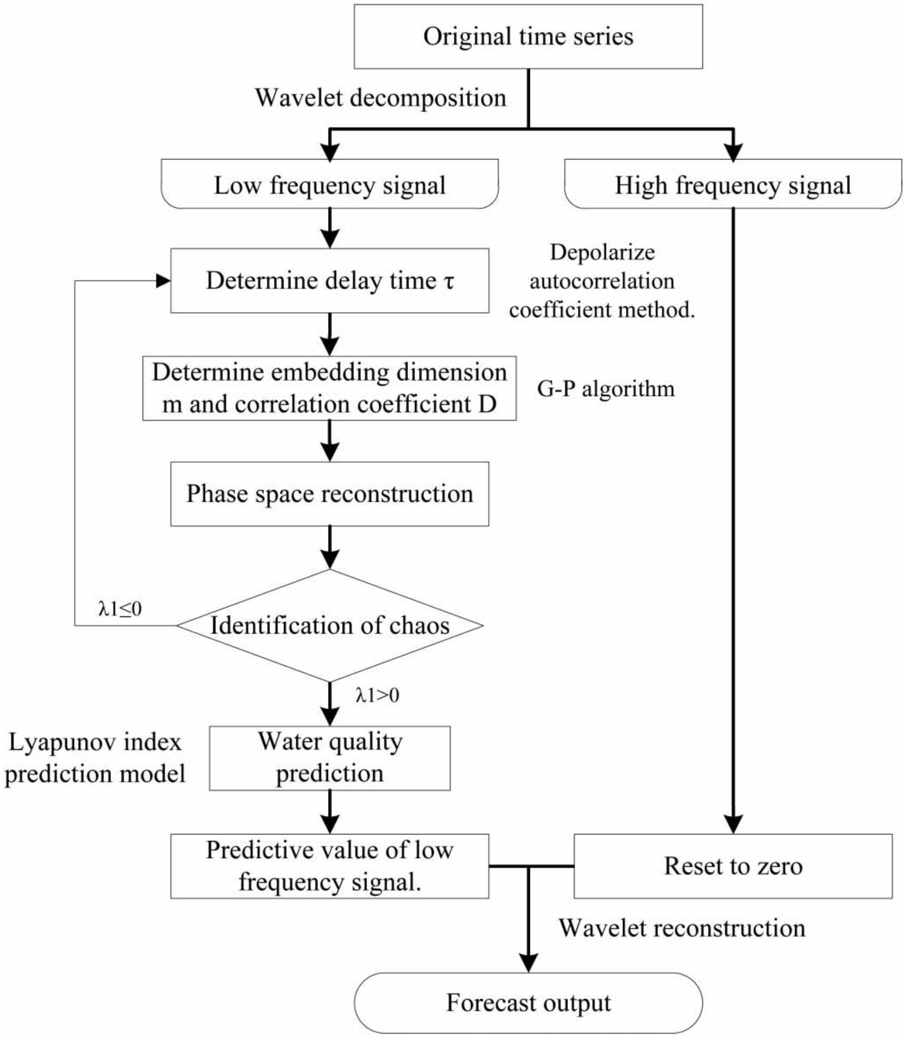

Figure 1 | Flowchart of wavelet-Lyapunov model.

select water quality data observed at two different time resolutions for investigation, to represent low-resolution monitoring and high-resolution monitoring: (1) weekly water quality data observed at the Huaihe River in China and (2) 15-minute water quality data observed at the Potomac River in the US. In both cases, three water quality parameters are analyzed: COD, DO, and NH_3-N. One-step ahead and multi-step ahead forecasts are made. The two study regions and the data considered are described next.

Weekly monitoring at Huaihe River, China

The Huaihe River Basin is located in eastern China between the Yellow River Basin and the Yangtze River Basin (Figure 2). The Huaihe River flows through five provinces, including Hubei, Henan, Anhui, Shandong, and Jiangsu. The area of the entire basin is about 270,000 km^3, and the basin population is about 165 million. The average annual rainfall in the basin is about 900 mm, over 70% of which is concentrated in the rainy season between June and October. In the Huaihe River Basin, there are 27 automatic water quality monitoring stations, and water quality data have been monitored at weekly intervals, through the national automatic monitoring network.

In the present study, water quality data observed at the Aishanxi Bridge monitoring station are considered for analysis. This station is located in Picang, Pizhou in Jiangsu province. The specific location of this station is westward drift of the floodway in Picang in the junction of Shandong and Jiangsu provinces. Weekly water quality data from January 2009 to June 2014 are

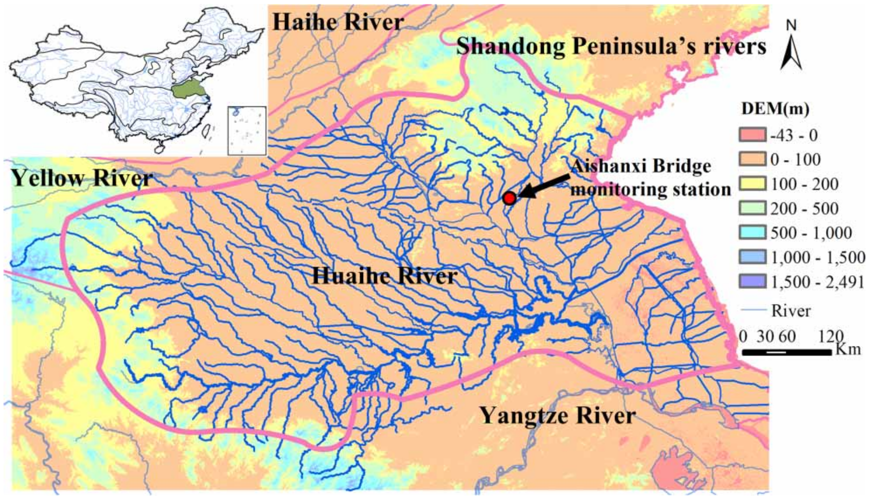

Figure 2 | Location of Huaihe River Basin and Aishanxi Bridge station.

considered for analysis, for a total of 281 data (weeks). The water quality variables considered for study are COD, DO, and NH_3-N. The time series of these three water quality variables over the 281-week period are shown in Figure 3.

Sub-hourly monitoring in Potomac River, US

The Potomac River is the fourth largest river in the US and is located along the coast of the Atlantic Ocean (Figure 4). The river originates in the Appalachian Mountains of West Virginia and runs into Chesapeake Bay. Approximately five million people live in this basin. The average flow in this river is 306 m^3/s (USEPA 2017).

The United States Geological Survey (USGS) (https://waterdata.usgs.gov/nwis/sw) has installed more than 850,000 monitoring stations to collect surface-water data, including water levels, streamflow (discharge), surface-water quality, rainfall, etc. They are mainly detected by sensors at stations with a fixed interval of 15 to 60 minutes and transmitted to the USGS every hour. The surface water quality parameters include water temperature (WT), specific conductivity (SC), DO, pH, turbidity (TURB), and nitrate + nitrite nitrogen (NOx). Among the variables, only DO is common for both the river basins considered in this study. Long-term COD and NH_3-N for the Potomac River are not available.

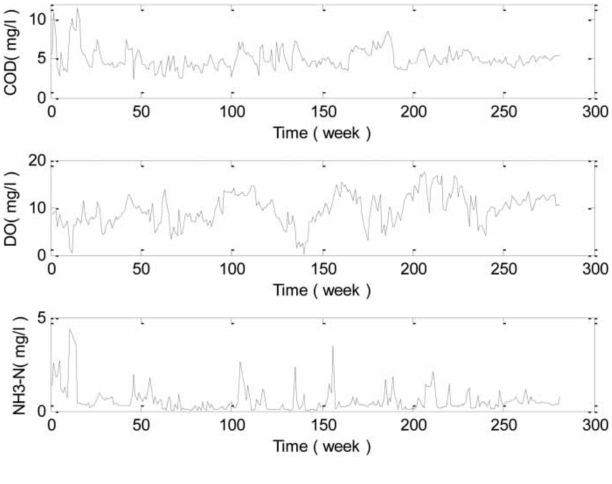

Figure 3 | Time series of observed water quality data in the Aishanxi Bridge station of the Huaihe River Basin.

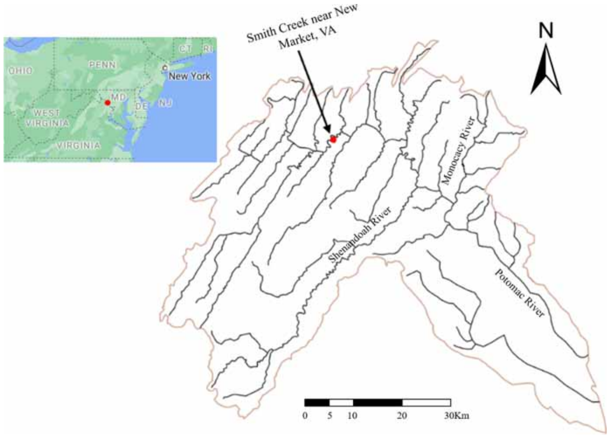

Figure 4 | Location of Potomac River and Smith Creek station.

For the present study, water quality data observed from Smith Creek close to New Market in Virginia (USGS Station #01632900) are considered for analysis. The data are available at a very high resolution of 15 minutes as shown in Figure 5.

RESULTS AND DISCUSSION

The COD dataset in the Huaihe River is first used for model verification as a baseline scenario. The WMLE model is then compared with pure Lyapunov exponent model (LE model) and two classical data-driven prediction models, ARMA (Shi *et al.* 2017) and ANN prediction model (Du *et al.* 2017). WMLE model performance of COD dynamics is compared on the application on DO and NH_3-N dynamics and, finally, the performance on high-resolution water quality dynamics is compared.

COD monitoring records are divided into two groups: calibration period or ‘warm-up’ period (total 250 data from the first week in 2009 to the 41st week in 2013) and verification period (total 30 data from the 42nd week in 2013 to the 19th week in 2014) (Grassberge & Procaccia 1983; Zhang *et al.* 2016). Warm-up means the chaos feature of the low-frequency signal of decomposed surface water quality time series is identified at this period. Tests on DO and NH_3-N dynamics followed the same routine.

Chaos identification of low-frequency signal of surface water quality

Wavelet decomposition is first conducted on COD monitoring data from the first 250 weeks of the original time series with *db5* wavelet function and the decomposition layer is two. Thus, the low-frequency signal presented the main information in

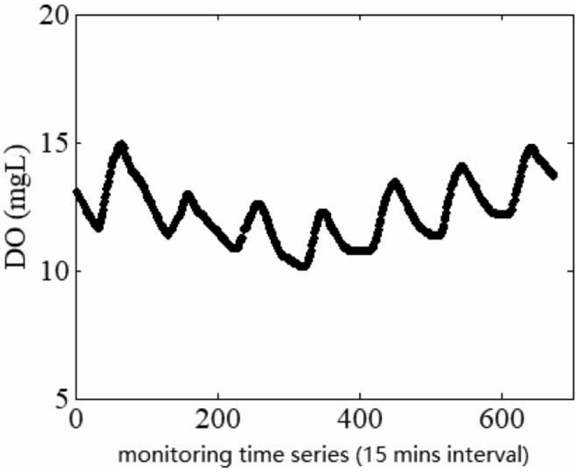

Figure 5 | Time series of observed water quality data in Smith Creek station of the Potomac River.

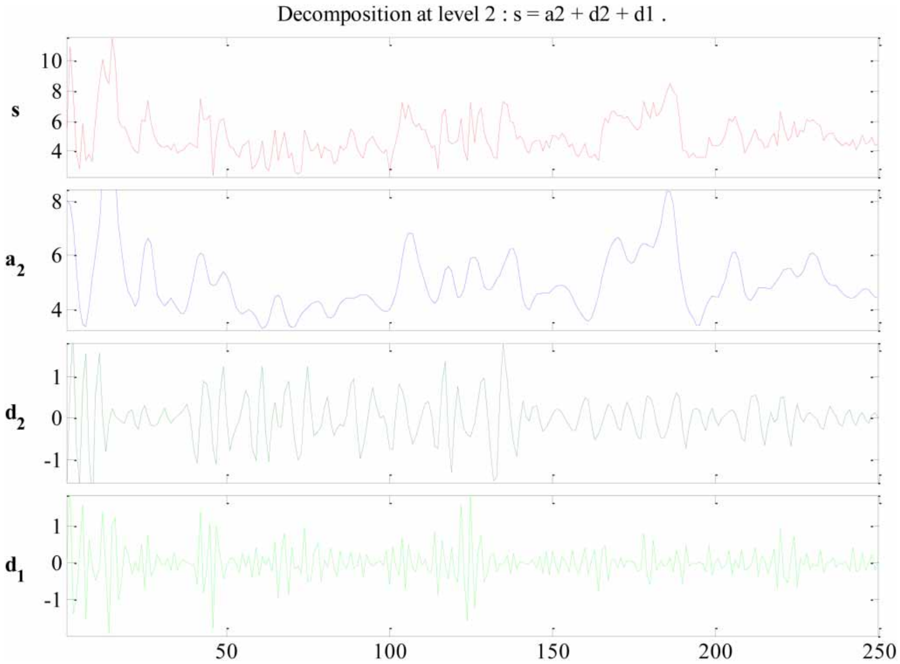

Figure 6 | Decomposition of wavelet signal for COD records at Huaihe River.

monitoring data of water quality and the high-frequency signal is composed of main noise. Wavelet decomposition waveform of the original time series can be seen in Figure 6, and S is the original signal. a_2 is low-frequency signal and d_2 and d_1 are high-frequency signals.

To determine the delay time τ, non-partial/multiple autocorrelation coefficient $C_{xx}^{m}(\tau)$ is calculated according to the method hereinbefore and the result can be seen in Supplementary Material, Table S1. Time of value $C_{xx}^{m}(\tau)$ reduced to $(1 - 1/e)$ of the initial value for the first time is selected as time delay τ, namely, optimal delay time of reconstructed phase space. It can be found that $\tau = 3$ from Figure 7 and Supplementary Material, Table S1.

Phase space of the low-frequency signal is then reconstructed. The projection shown in Figure 8(b) corresponds to a delay time value $\tau = 3$. The reconstruction yields relatively well-defined dynamics. Figure 8(c) shows correlation integral, C(r), and the radius, r, for embedding dimensions, m, from 2 to 13. Fitting should be conducted on the parts of all the curves mostly close to the straight line with least squares method and the slope of the fitted straight line is, namely, correlation exponent D. It can be found from Supplementary Material, Table S2 and Figure 8(d) that correlation coefficient D tends to be stable when

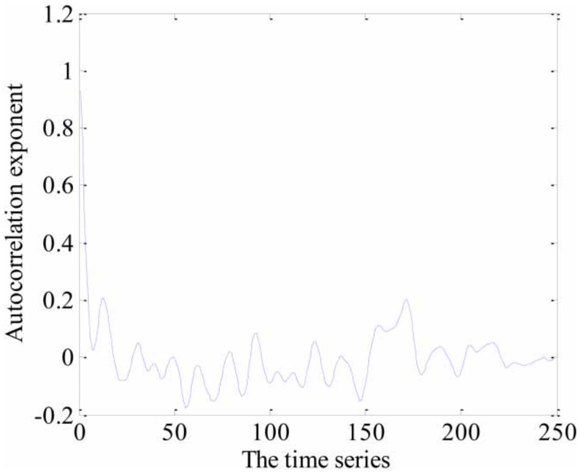

Figure 7 | Variation of autocorrelation coefficients with time.

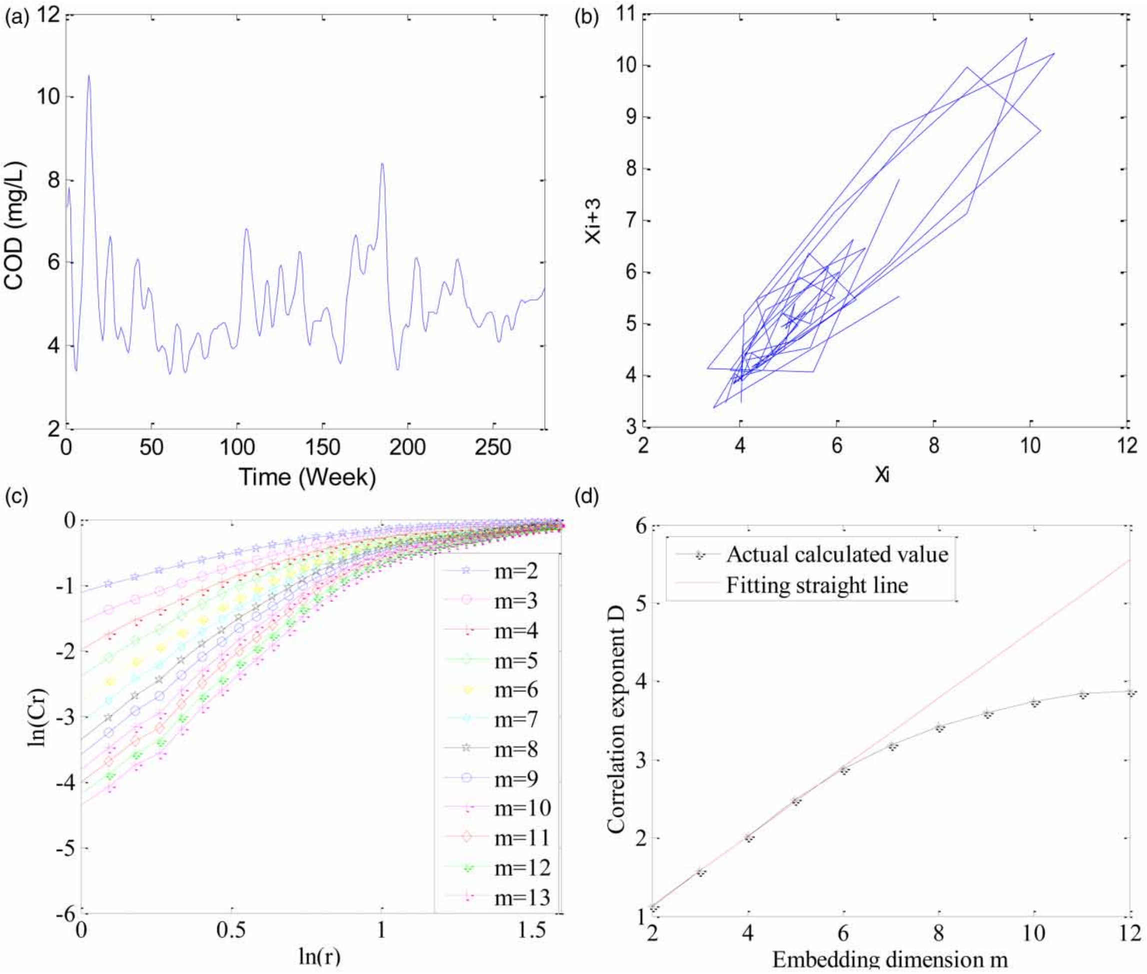

Figure 8 | Chaos characteristics of weekly COD time series at Huaihe River: (a) time series, (b) phase space, (c) *LnC(r)* versus *Ln(r)*, and (d) relationship between correlation exponent and embedding dimension.

embedded dimension $m = 11$, which is much larger than the case for runoff modeling (Sivakumar 2017). Thus, it is determined that embedded dimension $m = 11$ and corresponding correlation coefficient $D = 3.8415$.

Correlation exponent has a platform as the embedding dimension goes higher in Figure 8(d), which is a proof of chaos absence. Further, it can be obtained that MLE $\lambda_1 = 1.1710$ (low-frequency signal) and 1.2413 (original data) through calculation according to Equation (4) and delay time $\tau = 3$ and embedded dimension $m = 11$ determined in step 1 and step 2. MLE λ_1 is positive, also denoting that the low-frequency signal of time series of river water quality has chaotic property and can be predicted with chaos theory.

Performance of WMLE model and comparison with classical data-driven approaches

COD dataset was first used for investigation. The one-step prediction results of the WMLE model on the upcoming 30 weeks have maximum relative error (MRE) 14.61% and average relative error (ARE) 4.53% (Figure 9).

Prediction can be conducted on the same data and scenario with pure Lyapunov exponent model (LE model) without wavelet filtering (Equation (4)). It can be obtained that $\tau = 3$, $m = 15$ and corresponding correlation exponent $D = 4.5000$ and MLE $\lambda_1 = 1.2413$ according to the mentioned method hereinbefore. LE model results can be seen in Figure 9. MRE is 17.17% and ARE is 6.72% in this case.

ARMA was proposed by American statisticians Jenkins and Box in the 1970s and it has become the benchmark model for time series prediction. The basic principle is: time series can be regarded as a random process and it can be described or

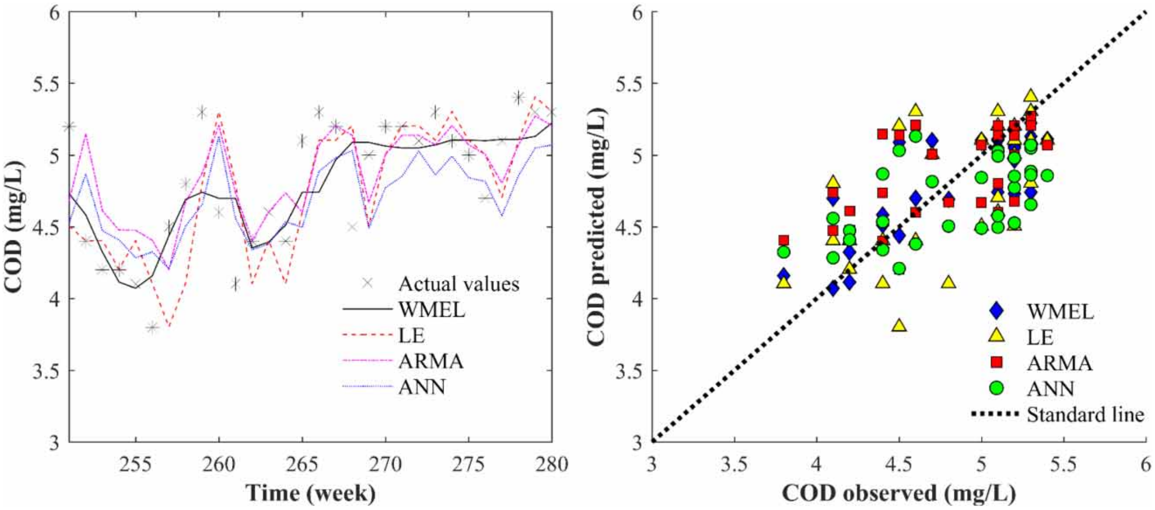

Figure 9 | Prediction values by different models and actual measured value at Huaihe River station.

simulated with a mathematical model (Thornton & Chambers 2017). ARMA model was constructed here and results of COD at Aishanxi Bridge monitoring station are reported in Figure 9 as well. MRE is 19.82% and ARE is 7.11%.

ANN model is another benchmark gray-box model on the scope of machine learning. A back-propagation ANN was used for comparison (Tongal 2013). The log-sigmoid transfer function *logsig* was used in the hidden layer, and the pure linear function *purelin* in the output layer. After training test and optimization (see Supplementary information), a structure of 10 × 8 × 1 was selected. During the training, maximum step of 1,000, minimum error with 0.0001, and learning rate 0.05 were set. Results show MRE is 13.80% and the ARE is 7.25% in the case of Huaihe River. Supplementary Material, Figures S1 and S2 with corresponding interpretation give the information on the construction of the ANN model. The detailed steps of the ARMA model are shown in Figures S3 and S4 and the accompanying text.

Table 1 lists the MRE and ARE of the above four methods for comparison. The hybrid WMLE model is superior to almost all other models in this case. The wavelet process largely improved the performance of the purely Lyapunov model which denotes the merit of the hybrid model. As a benchmark model, ARMA interestingly performs well in this case (see Supplementary information for more details). The ANN model performs worst in ARE, however, with a minimum MRE. This is probably due to the noise and input neurons. ANN is normally used as a 'half-mechanism' model, and relative impact factors of COD, like temperature, DO, etc., are necessary to be considered in an input layer besides historical COD data (Wang *et al.* 2013). As a nonlinear local approximation approach, it was found that although the pure Lyapunov model is significantly superior to the ANN model for river flow forecast (Sivakumar *et al.* 2002), the ANN model performs slightly better on COD forecast in this study. Performance on long-term forecast, and multi-step, was not investigated here since the time scale in this case is coarse.

The combination of chaos system with machine learning is also promising. One pathway is chaos evidence improving machine learning models (Sun *et al.* 2010; Huang *et al.* 2017). The other is machine learning models reconstructing chaos and improve the forecasting performance (Pathak *et al.* 2018). No studies were reported on riverine water quality dynamics.

Table 1 | Prediction precision contrast of four methods based on COD time series at Huaihe River

Models	Maximum relative error (%)	Average relative error (%)
Wavelet-Lyapunov	14.61	4.53
Lyapunov	17.17	6.72
ARMA	17.01	6.39
ANN	13.80	7.25

Performance of WMLE model on different water quality parameters

The same process of prediction, verification, and analysis is also conducted on DO and NH_3-N time series as COD. It was found that DO has delay time $\tau = 5$ and embedding dimension $m = 4$ while NH_3-N has $t = 3$ and $m > 20$ (misconvergence). Results shown MRE of DO prediction is 6.11% and ARE is 2.35%; MRE of NH_3-N prediction is 55.20% and ARE is 18.85%.

NH_3-N prediction is most challenged here compared with DO and COD. This is reasonable, since sources and dynamics of the in-stream NH_3-N process are more complex than those of COD and DO. The many big jumps in the original datasets of NH_3-N also offer some preliminary observations on this (see Figure 3). Such jumps probably result from storm water or pollutant discharge and they can be identified as anomalies, in practice. For example, NH_3-N abruptly jumped to 3.5 mg/Lat the 156th week, six times higher than the average value 0.57 mg/L of the whole series. Such abnormalities surely exceed the predication capacity of models, hence the increase of both MRE and ARE to a certain degree.

Performance of WMLE model on high-resolution water quality dynamic

High-resolution observation obviously leads to a different pattern of water quality dynamics from that of weekly observation. In the Potomac River case, DO and NO_2^--N were investigated due to data availability. The previous 5 days, from January 1 to January 5, is set as the warm-up period which contains 480 data on a 15 minute basis. Ninety-six data on the 6th day are used as model verification. Delay time τ is selected according to non-partial multiple autocorrelation coefficient method, as shown in the Methodology section. As presented in Table S1 and Figure 10(a), $\tau = 18$ is determined, i.e., 4.5 hours. Embedded dimension m is determined by *ln C*(*r*)-*ln r* relationship (Figure 10(a)) and $m = 6$ is selected. Further, Lyapunov exponent $\lambda_1 = 0.0042$

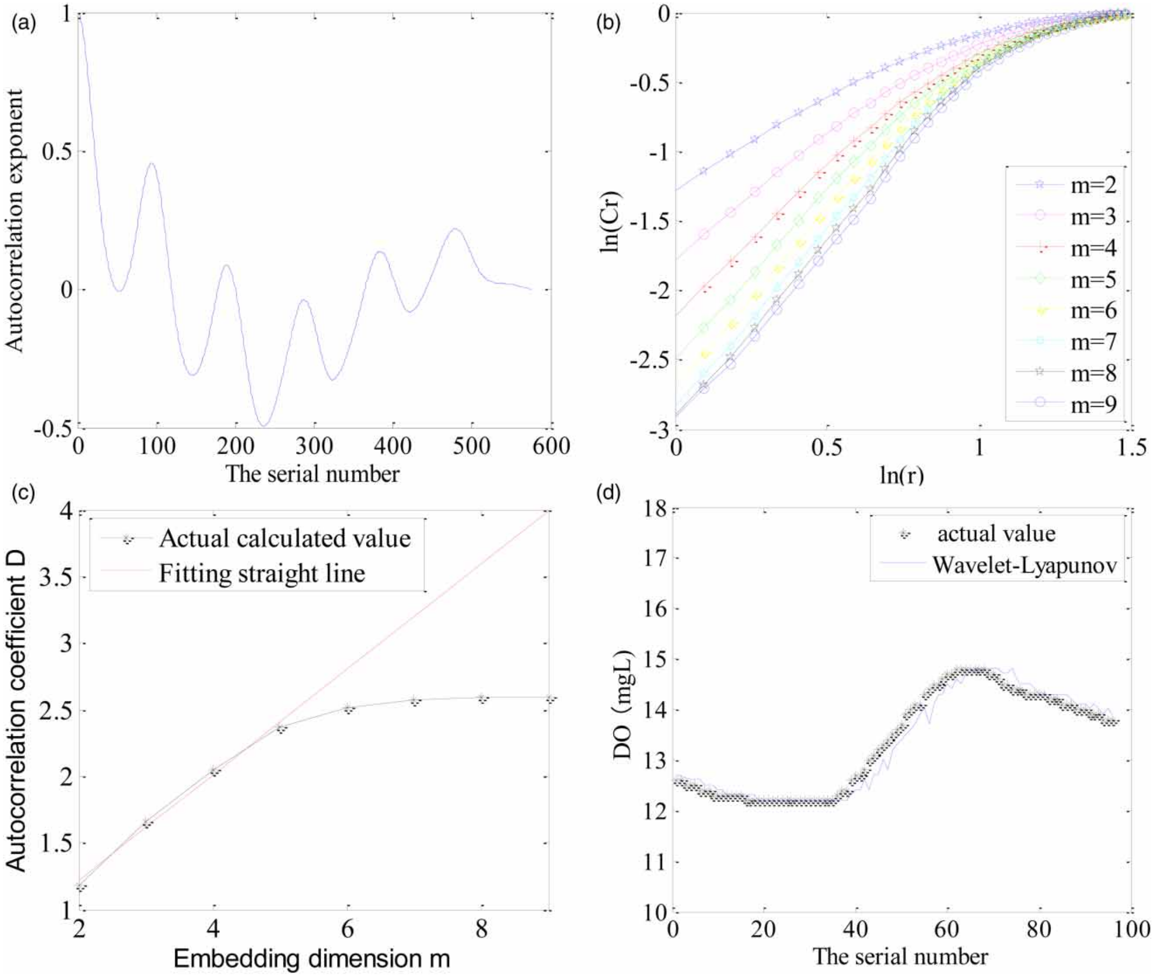

Figure 10 | Chaos characteristics of high-frequency DO time series: (a) variation of autocorrelation, (b) *lnC*(*r*) versus *ln*(*r*), (c) relationship between correlation exponent and embedding dimension, and (d) forecast results.

is obtained which indicates the presence of chaos of the high-resolution DO system. NO_2^--N dataset did not present a chaotic behavior after the same analysis.

Figure 10(d) shows the forecast results of the WMLE model of high-resolution DO time series. The MRE is 4.76% and ARE is 1.18%. It is a good enough result. A smooth and narrow diurnal fluctuation of DO accounts for this. Compared with the weekly DO forecast, short-term forecast succeeded better since the Lyapunov method is a local approximation approach.

Interestingly, the results also reveal that the embedded dimension m for weekly time series (Huaihe River case) is obviously larger than the high-resolution monitoring time series (Potomac River) while the characteristic of delay time τ is reversed. For delay time, it is reasonable for high-resolution observation to have a larger time lag. Embedded dimension m is a feature that denotes complexity. The larger m the more complex water quality dynamics. Here, the weekly monitoring data need a higher dimension to present attractors – that is, only in higher dimensional space can the chaos emerge.

CONCLUSIONS

The following conclusions can be drawn in this study: (1) Chaos phenomenon surely existed in river water quality dynamic systems, but not all the water quality indexes presented chaos behavior, possibly due to the data properties. Some stations in Huaihe River Basin and NO_2^- in the Potomac River did not present chaotic dynamics. Therefore, chaos behavior recognition of water quality dynamics should be first conducted before utilizing a chaos theory-based modeling approach. (2) A hybrid model wavelet–maximal Lyapunov exponent method for river water quality prediction was successfully established. The WMLE model is superior to traditional data-driven models such as ARMA and ANN, and outweighs the pure MLE model in this investigation. It also has good performance on weekly forecasting of DO and COD time series with average relative error 2.35% and 4.53%, respectively. NH_3-N forecasting is challenged with the average relative error of 18.85%. (3) High-frequency DO time series can be more accurately forecasted than weekly ones with a relative average error of 1.18%. This indicates the developed hybrid model WMLE combined with sensors and abnormal detection algorithm can provide an eligible alternative for river water quality early warning. (4) Data-driven models are experiencing a renaissance. Further studies will be meaningful on the chaos characteristics of water quality dynamics based on large-scale datasets, and their fundamental mechanism of generic chaos dynamics can be discussed since chaos bridges the deterministic and stochastic.

ACKNOWLEDGEMENTS

This work was supported by Key-Area Research and Development Program of Guangdong Province (Grant No. 2019B110205005), National Natural Science Foundation of China (Grant No. 51979136), and Open Project of State Key Laboratory of Urban Water Resource and Environment, Harbin Institute of Technology (Grant No. ESK201901), Science, Technology and Innovation Commission of Shenzhen (JSGG20180508151852303). We are grateful to Dr Bin Shi from Research Center for Eco-Environmental Sciences, CAS for assistance in coding and data collection.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories (https://github.com/nantekoto/WQprediction).

REFERENCES

Alizadeh, M. J., Nourani, V., Mousavimehr, M. & Kavianpour, M. R. 2018 Wavelet-IANN model for predicting flow discharge up to several days and months ahead. *Journal of Hydroinformatics* **20** (1), 134–148.

Babovic, V. 2005 Data mining in hydrology. *Hydrological Processes* **19** (7), 1511–1515.

Barzegar, R., Moghaddam, A. A., Adamowski, J. & Ozga-Zielinski, B. 2018 Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. *Stochastic Environmental Research and Risk Assessment* **32** (3), 799–813.

Du, K., Zhao, Y. & Lei, J. 2017 The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series. *Journal of Hydrology* **552**, 44–51.

Gordillo, G., Morales-Hernández, M. & García-Navarro, P. 2019 Finite volume model for the simulation of 1D unsteady river flow and water quality based on the WASP. *Journal of Hydroinformatics* **22**, 327–345.

Grassberge, P. & Procaccia, I. 1983 Characterization of strange attractor. *Physical Review Letters* **50**, 346–349.

Huang, F., Huang, J., Jiang, S.-H. & Zhou, C. 2017 Prediction of groundwater levels using evidence of chaos and support vector machine. *Journal of Hydroinformatics* **19** (4), 586–606.

Kantz, H. & Schreiber, T. 2004 *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, UK.

Kim, H. S., Eykholt, R. & Salas, J. D. 1999 Nonlinear dynamics, delay times, and embedding windows. *Physica D: Nonlinear Phenomena* **127**, 48–60.

Li, Z.-m., Cui, L.-g., Xu, S.-w., Weng, L.-y., Dong, X.-x., Li, G.-q. & Yu, H.-p. 2013 Prediction model of weekly retail price for eggs based on chaotic neural network. *Journal of Integrative Agriculture* **12** (12), 2292–2299.

Li, Y., Yang, J., Zhang, N., Yang, J., Zhou, H. & He, J. 2016 Study on sensor fault instability prediction for the Internet of agricultural things based on largest Lyapunov exponent. *Tehnicki Vjesnik – Technical Gazette* **23**, 1.

Li, X., Sha, J., Li, Y.-m. & Wang, Z.-L. 2017 Comparison of hybrid models for daily streamflow prediction in a forested basin. *Journal of Hydroinformatics* **20** (1), 191–205.

Lin, G.-F., Lin, H.-Y. & Chou, Y.-C. 2013 Development of a real-time regional-inundation forecasting model for the inundation warning system. *Journal of Hydroinformatics* **15** (4), 1391–1407.

Lintern, A., Webb, J. A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., Leahy, P., Wilson, P. & Western, A. W. 2018 Key factors influencing differences in stream water quality across space. *Wiley Interdisciplinary Reviews-Water* **5** (1), 1–31.

Liu, Y., Lei, S., Sun, C., Zhou, Q. & Ren, H. 2011 A multivariate forecasting method for short-term load using chaotic features and RBF neural network. *European Transactions on Electrical Power* **21** (3), 1376–1391.

Lorenz, E. N. 1963 Deterministic nonpcriodic flow. *Journal of the Atmospheric Sciences* **20**, 130–141.

Lu, Z., Hunt, B. R. & Ott, E. 2018 Attractor reconstruction by machine learning. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **28** (6), 061104.

Meszaros, L. & El Serafy, G. 2018 Setting up a water quality ensemble forecast for coastal ecosystems: a case study of the southern North Sea. *Journal of Hydroinformatics* **20** (4), 846–863.

Packard, N. H., Crutchfield, J. P., Farmer, J. D. & Shaw, R. S. 1980 Geometry from a time series. *Physical Review Letters* **45** (9), 712.

Pathak, J., Hunt, B., Girvan, M., Lu, Z. & Ott, E. 2018 Model-free prediction of large spatiotemporally chaotic systems from data: a reservoir computing approach. *Physical Review Letters* **120** (2), 024102.

Ren, L., Xiang, X.-Y. & Ni, J.-J. 2013 Forecast modeling of monthly runoff with adaptive neural fuzzy inference system and wavelet analysis. *Journal of Hydrologic Engineering* **18** (9), 1133–1139.

Rode, M., Wade, A. J., Cohen, M. J., Hensley, R. T., Bowes, M. J., Kirchner, J. W., Arhonditsis, G. B., Jordan, P., Kronvang, B., Halliday, S. J., Skeffington, R. A., Rozemeijer, J. C., Aubert, A. H., Rinke, K. & Jomaa, S. 2016 Sensors in the stream: the high-frequency wave of the present. *Environmental Science & Technology* **50** (19), 10297–10307.

Roushangar, K., Alizadeh, F. & Nourani, V. 2018 Improving capability of conceptual modeling of watershed rainfall-runoff using hybrid wavelet-extreme learning machine approach. *Journal of Hydroinformatics* **20** (1), 69–87.

Sang, Y.-F. 2013 A review on the applications of wavelet transform in hydrology time series analysis. *Atmospheric Research* **122**, 8–15.

Shi, B., Wang, P., Jiang, J. & Liu, R. 2017 Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies. *Science of the Total Environment* **610–611**, 1390–1399.

Siek, M. & Solomatine, D. P. 2010 Nonlinear chaotic model for predicting storm surges. *Nonlinear Processes in Geophysics* **17** (5), 405–420.

Sivakumar, B. 2005 Chaos in rainfall: variability, temporal scale and zeros. *Journal of Hydroinformatics* **7** (3), 175–184.

Sivakumar, B. 2017 *Chaos in Hydrology: Bridging Determinism and Stochasticity*. Springer, Dordrecht, the Netherlands.

Sivakumar, B., Jayawardena, A. W. & Fernando, T. M. K. G. 2002 River flow forecasting: use of phase-space reconstruction and artificial neural networks approaches. *Journal of Hydrology* **265** (1), 225–245.

Solomatine, D. P. & Ostfeld, A. 2008 Data-driven modelling: some past experiences and new approaches. *Journal of Hydroinformatics* **10** (1), 3–22.

Sun, Y., Babovic, V. & Chan, E. S. 2010 Multi-step-ahead model error prediction using time-delay neural networks combined with chaos theory. *Journal of Hydrology* **395** (1–2), 109–116.

Takens, F. 1981 Detecting strange attractors in turbulence. In: *Dynamical Systems and Turbulence* (Rand, D. & Young, L. S., eds). Springer Verlag, Berlin, Heidelberg, pp. 366–381.

Thornton, M. A. & Chambers, M. J. 2017 Continuous time ARMA processes: discrete time representation and likelihood evaluation. *Journal of Economic Dynamics and Control* **79**, 48–65.

Tongal, H. 2013 Nonlinear forecasting of stream flows using a chaotic approach and artificial neural networks. *Earth Sciences Research Journal* **17**, 119–126.

Tongal, H. 2020 Comparison of local and global approximators in multivariate chaotic forecasting of daily streamflow. *Hydrological Sciences Journal* **65** (7), 1129–1144.

Tongal, H. & Berndtsson, R. 2017 Impact of complexity on daily and multi-step forecasting of streamflow with chaotic, stochastic, and black-box models. *Stochastic Environmental Research and Risk Assessment* **31** (3), 661–682.

Tongal, H. & Booij, M. J. 2018 Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *Journal of Hydrology* **564**, 266–282.

USEPA 2017 National hydrography dataset high-resolution flowline data. The national map. Available from: https://www.data.gov/, Accessed 1st June 2021.

Vlachas Pantelis, R., Byeon, W., Wan Zhong, Y., Sapsis Themistoklis, P. & Koumoutsakos, P. 2018 Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **474** (2213), 20170844.

Wang, Y., Zheng, T., Zhao, Y., Jiang, J., Wang, Y., Guo, L. & Wang, P. 2013 Monthly water quality forecasting and uncertainty assessment via bootstrapped wavelet neural networks under missing data for Harbin, China. *Environmental Science and Pollution Research* **20** (12), 8909–8923.

Wolf, A., Swift, J. B., Swinney, H. L. & Vastano, J. A. 1985 Determining Lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena* **16** (3), 285–317.

Yajima, H. & Derot, J. 2018 Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. *Journal of Hydroinformatics* **20** (1), 206–220.

Yu, X., Liong, S.-Y. & Babovic, V. 2004 EC-SVM approach for real-time hydrologic forecasting. *Journal of Hydroinformatics* **6** (3), 209–223.

Zhang, Y. 2013 New prediction of chaotic time series based on local Lyapunov exponent. *Chinese Physics B* **22** (5), 050502.

Zhang, L., Zou, Z. H. & Zhao, Y. F. 2016 Application of chaotic prediction model based on wavelet transform on water quality prediction. *IOP Conference Series: Earth and Environmental Science* **39**, 012001.

First received 17 February 2021; accepted in revised form 25 April 2021. Available online 28 May 2021