SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

A comparative analysis of artificial neural networks and wavelet hybrid approaches to long-term toxic heavy metal prediction

Peifeng Li[1], Pei Hua[2,3], Dongwei Gui[4], Jie Niu[5], Peng Pei[6], Jin Zhang[5]✉ & Peter Krebs[1]

The occurrence of toxic metals in the aquatic environment is as caused by a variety of contaminations which makes difficulty in the concentration prediction. In this study, conventional methods of back-propagation neural network (BPNN) and nonlinear autoregressive network with exogenous inputs (NARX) were applied as benchmark models. Explanatory variables of Fe, pH, electrical conductivity, water temperature, river flow, nitrate nitrogen, and dissolved oxygen were used as different input combinations to forecast the long-term concentrations of As, Pb, and Zn. The wavelet transformation was applied to decompose the time series data, and then was integrated with conventional methods (as WNN and WNARX). The modelling performances of the hybrid models of WNN and WNARX were compared with the conventional models. All the given models were trained, validated, and tested by an 18-year data set and demonstrated based on the simulation results of a 2-year data set. Results revealed that the given models showed general good performances for the long-term prediction of the toxic metals of As, Pb, and Zn. The wavelet transform could enhance the long-term concentration predictions. However, it is not necessarily useful for each metal prediction. Therefore, different models with different inputs should be used for different metals predictions to achieve the best predictions.

Due to the nature of ubiquity, toxicity at a trace level, and hard biodegradation, elevated metals in aquatic environments are a global concern[1,2]. A long-term exposure of toxic metals by the ingestion of the contaminated water and fish can cause chronic diseases[3,4]. For example, Arsenic (As) destroys the redox capacity of cells, affects the normal metabolism, causes tissue damage and body disorders, and even directly causes poisoning death when ingested in small quantities[5]. Lead (Pb) affects nerves, digestion, urinary, reproductive and developmental, cardiovascular, endocrine, immune, bone, and other organ systems. More serious is that Pb affects the growth and mental development of infants and young children, impairs brain function such as cognition[6]. In addition, a high level of Zinc (Zn) weakens immune function, leads to iron deficiency anaemia, affects the function of the digestive system, and causes damage to blood vessels[7]. Due to the special significance to water quality, As, Pb, and Zn are included in the priority pollutant list by the United States Environmental Protection Agency. Therefore, it is essential to understand the environmental behaviours of toxic metal in rivers to protect the drinking water intake.

Traditionally, the toxic metals in trace levels were required to be routinely sampled and determined in the laboratory. However, there are some constraints for the environmental managers to adequately and timely receive the metal contents and respond to the metal pollution, such as (i) expense of field monitoring, (ii) staffs availability and resources, (iii) field safety issues, and (iv) large time intervals between data collection, reporting and public notification[8]. Therefore, to decrease the cost of aquatic environmental monitoring and provide an early-warning proactive approach to metal pollutions, a forecasting approach is essential.

[1]Institute of Urban and Industrial Water Management, Technische Universität Dresden, 01062 Dresden, Germany. [2]Environmental Research Institute, Guangdong Provincial Key Laboratory of Chemical Pollution and Environmental Safety and MOE Key Laboratory of Theoretical Chemistry of Environment, South China Normal University, Guangzhou 510006, China. [3]School of Environment, South China Normal University, University Town, Guangzhou 510006, China. [4]State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China. [5]Institute of Groundwater and Earth Sciences, Jinan University, Guangzhou 510632, China. [6]College of Mines, Guizhou University, Guiyang 550025, China. ✉email: jzhang@jnu.edu.cn

Regarding the modelling approach, a variety of models have been used to forecast the levels of toxic metals. On two categories of models were commonly classified as physical principles based on mechanical models[9], and historical data based numerical models[10]. Mechanical models of water quality require detailed information and principles about the processes[11]. Considerable parameters for model setup, simulation, and post-processing gained from a large number of trials are needed[12]. However, to forecast the fate of metals in surface water, mechanical models require water, sediment, and related data to describe the complex biological, physical, and chemical processes that influence metals' behaviours, which are not generally available for most surface waters[13]. Therefore, machine learning based numerical models could be more beneficial because the models could be transferred and applied from one point to another with relative ease and convenience[14].

In terms of machine learning technology, artificial neural networks (ANNs) have received wide attention in recent years, being implemented and popularized with the development of the computer age[15]. ANNs showed advantages over traditional multiple regression models especially when the underlying functions and data sets are highly complex and nonstationary[16]. Besides, back-propagation neural network (BPNN) and nonlinear autoregressive exogenous (NARX) models are typical time series prediction approaches. They were successfully applied to forecast the environmental factors of rainfall patterns[17], river flows[18], suspended sediment concentrations[19], river levels[20], and dissolved oxygen (DO)[21], etc. It was applied to predict heavy metal concentrations in the aquatic environment, and even works when the underlying function cannot be expressed in terms of any known mathematical functions. More explicit, Alizamir[22] developed and employed the feedforward ANN to forecast Pb and Zn concentrations in groundwater of Asadabad plain. The models were trained with the data collected from the field and then utilized as prediction tools. Ke[23] established a non-linear regression-based model to forecast the contents of Cd, Pb, Cu, Zn, As, and Cr in Xiangjiang River, China. Verification showed that this model had high precision, and the spatial variation of the predicted metal content was consistent with the actual conditions. Although the success of these earlier studies shows the beneficial of ANN modelling to the short- and long-term forecasts in many areas, it has certain limitations and problems in dealing with non-stationary data sets (i.e., statistical properties fluctuate over time).

Wavelet transform is an effectual tool for handling non-stationary data sets, which has been spread for time series and spatial data analysis over a few past decades[24]. A necessary feature of wavelet analysis is the ability to decompose the original data sets into high- and low-frequency contributions (i.e., fine and coarse features in the data) for further analysis[25]. Therefore, the hybrid methods of wavelet-ANN (WANN), including wavelet-BPNN and wavelet-NARX, have been reported by recent studies for the occurrence forecasting of daily river discharge[26], suspended sediment[27], rainfall runoff[28], and droughts[29], etc. However, these methods have less application in the metal concentrations prediction for surface water management.

Therefore, to mitigate the occurrence of metal pollutions, and eventually facilitate the minimization of the adverse effect of toxic metal to the aquatic environment, this study examines and compares the performance of conventional and hybrid neural network models for characterizing the toxic. Specific questions would be addressed in this study: (1) the effect of inputs selection and division on the performance of conventional and hybrid models of BPNN, NARX, WNN, and WNARX with time-series data; (2) the evaluation for long-term forecasts with a high degree of confidence as quantified by standard metrics such as the coefficient of determination and root-mean-square errors; and (3) the selection of the optimum approaches and inputs for each metal prediction with the best performance.

Materials and methods

Study area and water quality data. The Elbe River is one of the most important rivers in Europe. It crosses Germany (65.5% of the total length), the Czech Republic (33.7%), Austria (0.6%), and Poland (0.2%). It takes tasks including flood management, urban water supply, and navigation. The Elbe River basin, comprised of the Elbe River and its tributaries, has an area of 148,268 km^2 and sustains the consumption of about 25 million people[30].

The water quality data was recorded from 1998 to 2017 at Schmilka station (50°53′ N 14°13′ E) in the Elbe River as shown in Fig. 1. Weekly time-series data of Fe, Pb, Zn, and As were measured from one-week mixture samples by River Basin Community Elbe office (Flussgebietsgemeinschaft Elbe Geschäftsstelle). The daily time series data of pH, electrical conductivity (EC), water temperature (WT), river flow, nitrate nitrogen (NO_3–N), and DO were transferred to the weekly time series by mean of the seven consecutive days. The methods mentioned in the following sections were implemented based on the wavelet analysis and neural network toolboxes in MATLAB 2019a (The Mathworks Inc., Natick, MA).

The recorded data was classified into computing and simulating categories. The data from 1998 to 2015 was classified into the computing part. While the data between 2016 to 2017 was used for simulating the training networks. The computing part was categorized into three sets of training, validation, and testing sets. Besides, to evaluate the effectiveness of different combinations of data sets, the raw data was distributed as different modes shown in Table 1.

Input identification. Due to the complexities of metals' behaviour, a larger input data size and more input parameters may not necessarily ensure fewer errors at the test phase, though it may perform less error at the training phase[31]. Therefore, identifying the best input combination is the first step of the model establishment. Iron (Fe) is the most abundant element in Earth and the environment levels of Fe were usually regarded as non-toxic. It is usually combined with the other elements in hundreds of minerals. In other words, the occurrence of Fe is strongly linked to the other metals[32]. Therefore, for a better modelling effect, Fe was selected as an input parameter. Besides, the values of pH, EC, WT, flow, NO_3–N, and DO of the given river were considered as the candidates of the input parameters.

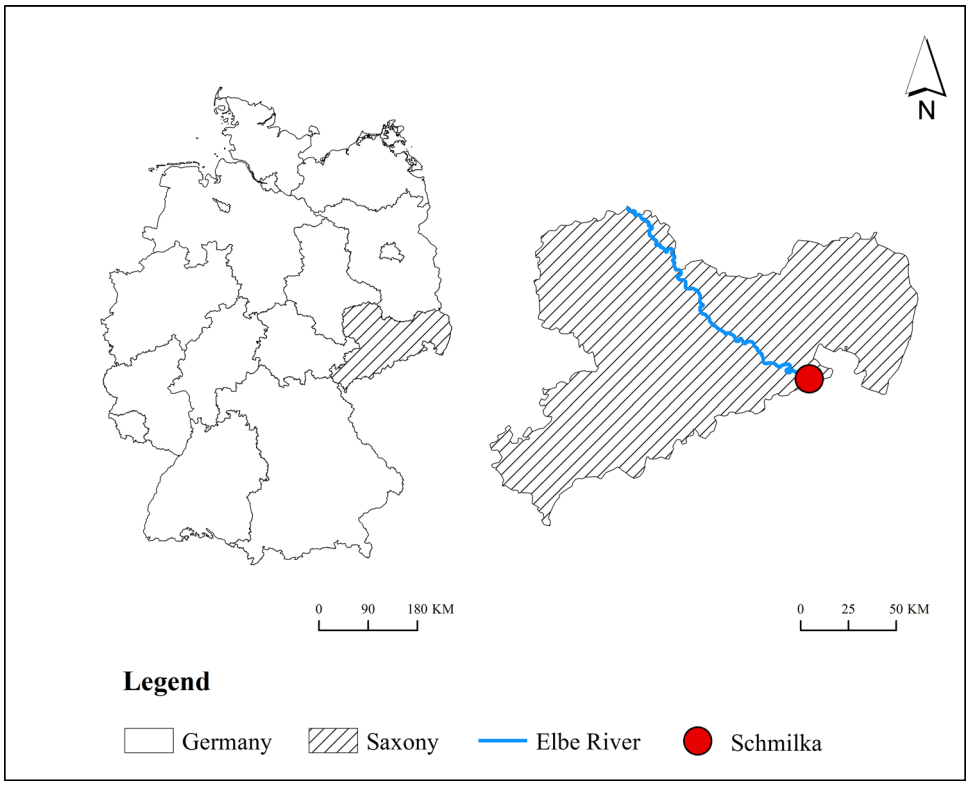

Figure 1. Location of Schmilka station. This figure shows the location of the measuring point. Environmental Systems Research Institute (ESRI). (2018). ArcGIS Release 10.6. Redlands,CA.

Mode	Train volume (years)	Validation volume (years)	Test volume (years)
1	16	1	1
2	15	2	1
3	14	3	1
4	13	4	1

Table 1. Different modes for data distribution.

As shown in Table 2, considering the statistical analysis of the Person correlation coefficients and significance analysis, several optimal input combinations were chosen to estimate the toxic metals according to the following conditions: (1) the p-value less than 0.05 indicating the relatively strong relationship between the inputs and targets, (2) the absolute correlation coefficients between the inputs and studied variables are relatively higher.

Wavelet transform. Wavelet transform method is commonly used to perform time-localized filtering in both time and frequency domains[33]. It expresses the asymmetric and unstable input time-series signals as the sum of the sub-signals and characterized as the continuous and discrete wavelet transforms (CWT and DWT)[34]. In this study, the wavelet transform was used for decomposing the time series data. It is based on a mother wavelet function that constructs a family of wavelets of a finite interval shown as below:

$$\mathrm{W(a,b)} = \frac{1}{\sqrt{a}} \int_{+\infty}^{-\infty} f(t)\overline{\psi}\left(\frac{t-b}{a}\right) dt \tag{1}$$

where a is a scale or frequency parameter, b is the shift parameter, $f(t)$ is the time series, and $\overline{\psi}(t)$ is the complex conjugate function of mother wavelet $\psi(t)$.

According to the discretization of the hydrometeorological time series data, the DWT is preferred in most hydrological forecasting problems. The DWT operates on two sets of functions viewed as high- and low-pass filters to produce discrete wavelet coefficients (DWC). For an input signal x, the first step produces two sets of DWCs: high pass approximation coefficients, $A1$ (low frequency), and low pass detail coefficients, $D1$ (high frequency). The next step splits the approximation coefficient $A1$ into two parts using the same scheme, replacing

		As	Pb	Zn	Fe	Flow	pH	WT	NO_3–N	EC	Do
As	Pearson Corr	1									
	p-value	–									
Pb	Pearson Corr	–	1								
	p-value	–	–								
Zn	Pearson Corr	–	–	1							
	p-value	–	–	–							
Fe	Pearson Corr	0.755	0.818	0.614	1						
	p-value	0	0	0	–						
Flow	Pearson Corr	0.161	0.418	0.174	0.534	1					
	p-value	1.32E−06	0	1.72E−07	0	–					
pH	Pearson Corr	− 0.260	− 0.084	− 0.118	− 0.091	− 0.048	1				
	p-value	2.66E−15	0.01	4.15E−04	0.01	0.15	–				
WT	Pearson Corr	0.172	− 0.030	0.031	− 0.102	− 0.420	0.064	1			
	p-value	2.33E−07	0.38	0.35	0	0	0.06	–			
NO_3–N	Pearson Corr	0.052	0.177	0.279	0.209	0.441	− 0.089	− 0.582	1		
	p-value	0.13	1.68E−07	0	6.06E−10	0	0.01	0	–		
EC	Pearson Corr	0.048	− 0.158	0.084	− 0.193	− 0.498	− 0.173	− 0.111	0.184	1	
	p-value	0.15	2.13E−06	0.01	6.70E−09	0	1.95E−07	9.59E−04	5.47E−08	–	
Do	Pearson Corr	− 0.246	0.020	− 0.059	0.088	0.442	0.342	− 0.881	0.538	−0.033	1
	p-value	8.75E−14	0.54	0.08	0.01	0	0	0	0	0.32	–

Table 2. Person correlation coefficients and significance analysis.

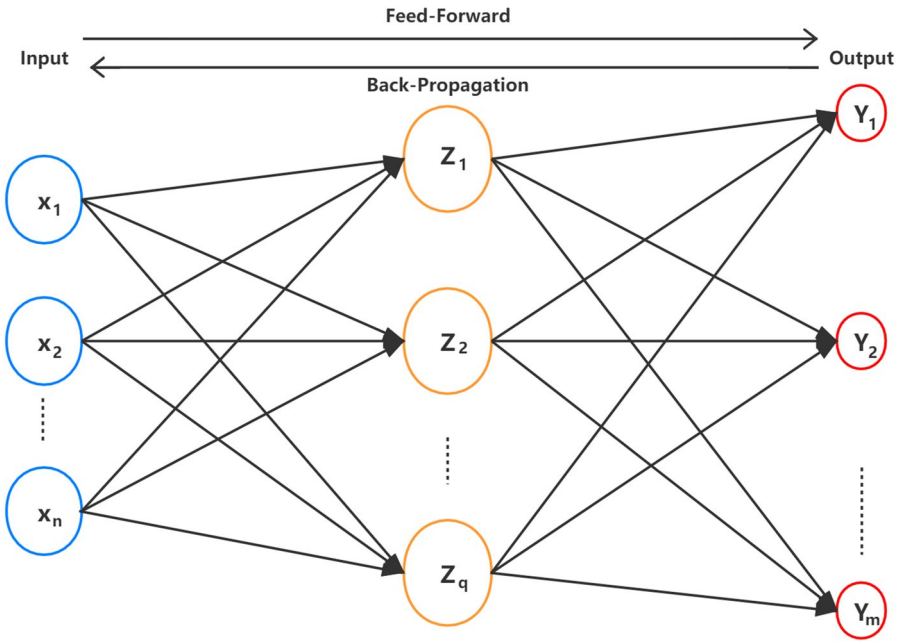

Figure 2. Topology of neural network.Fig. 1. This figure shows the basic structure of the neural network model.

x by $A1$, and produces $A2$ and $D2$, and so on. The wavelet decomposition of the input x analysed at level n has the structure of $[A_n, D_n, D_{n-1}, \ldots, D_2, D_1]$[35].

Back-propagation neural network (BPNN) model. BPNN is a supervised self-learning algorithm designed to minimize the mean square error between the computed output of the network and desired output[36]. As shown in Fig. 2, BPNN was formed by one input layer, one or more hidden layers, and one output layer. In this research, BPNN was trained with the Levenberg–Marquardt (LM) algorithm. It is a classic back-propagation algorithm that uses heuristics, relies on numerical optimization techniques to minimize and accelerate the calculation process, leading to a faster training[37]. The optimal number of hidden neurons for BPNN was determined by trial and error procedures.

In the feed-forward process, it was supposed that the input layer of the BP network has n nodes, the hidden layer has q nodes, the output layer has m nodes, the weight between the input and hidden layers is v_{ki}, and the weight between the hidden and output layers is w_{kj}. The transfer function of the hidden layer is $f_1(\cdot)$, and the transfer function of the output layer is $f_2(\cdot)$. Then the output of the hidden layer node z_k is:

$$Z_k = f_1\left(\sum_{i=0}^{n} v_{ki}x_i\right), i = 1, 2, \ldots q \tag{2}$$

The output of the output layer node y_j is:

$$y_j = f_2\left(\sum_{k=0}^{q} w_{ki}z_k\right), j = 1, 2, \ldots m \tag{3}$$

The function could be chosen by tansig, logsig, and purelin in MATLAB 2019a.

In the back-propagation process, using the squared error function, the error E_p of the Pth sample is obtained:

$$E_p = \frac{1}{2}\sum_{j=1}^{m}\left(t_j^P - y_j^P\right)^2 \tag{4}$$

where t_j^P is the expected output value.

The Levenberg–Marquardt algorithm uses this approximation to the Hessian matrix in the following Newton-like update:

$$\Delta w = \left(J^T J + \mu I\right)^{-1} J^T E \tag{5}$$

where J is the Jacobian matrix that contains first derivatives of the network errors for the weights and biases; E is a vector of network errors; μ is a scalar and its initial value is 0.001; I is the identity matrix; and Δw represents the adjustment of current weight value.

In this study, an adopted one-hidden-layer network was applied. The number in the hidden layer was estimated by the empirical formula given in Eq. (6)[38]:

$$N = \sqrt{n+m} + a \tag{6}$$

where N is the number of neurons in the hidden layer; n is the number of input variables; m is the number of output variables; and a is a number between 0 to 10. The optimal value of a is determined by trial and error. The optimum neuron number of the hidden layer was determined by gradually varying the number of nodes in the hidden layer through trial and error.

Nonlinear autoregressive exogenous (NARX) model. NARX is a nonlinear autoregressive model with exogenous inputs developed to predict the indicators. The model studies the relationship of the target value of a time series as well as current and past values of the exogenous series which influence the series of interest. It can be defined algebraically by:

$$y(t) = f\left(y(t-1), y(t-2), \ldots, y\left(t-n_y\right), u(t-1), u(t-2), \ldots, u(t-n_u)\right) \tag{7}$$

where, y(t) is the target values, and u is the externally related variables. In this scheme, $y(t-1)$ to $y\left(t-n_y\right)$ represent the past time series of the target. $u(t-1)$ to $u(t-n_u)$ denotes the past information about u, which helps predict the target values. f represents the nonlinear function approximated based on a feed-forward neural network.

Toto determine the inputs lag of the NARX model from the metal factors measured at the station, the cross-correlation between the factors and the autocorrelation of toxic metal was examined and checked. The lag for the U inputs was defaulted to one-week before the target according to the lag analysis based on the auto- and cross-correlations of the metal variables. The feedback loop performs multi-step-ahead prediction after the training process of the model. The closed-loop of the NARX network is established in the simulation process.

Wavelet and BPNN (WNN) hybrid model. WNN is an advanced neural network proposed in 1992[39], with the combination of wavelet and traditional BP neural network (n input nodes, q hidden nodes, m output nodes) which replaces the activation function of BP neural network hidden layer by wavelet basis function $\psi(x)\left(i = 1, 2, ..., q\right)$. The basic model of WNN[40] is:

$$y_i = \sum_{j=1}^{q} C_{ji}\psi\left(A_j x - b_j^{'}\right) \tag{8}$$

where, $1 \le j \le q, 1 \le i \le m$, $A_j = diag\left(a_{1j}^{-1}, L, a_{nj}^{-1}\right)$; $b_j^{'} = \left[a_{1j}^{-1}b_{1j}, L, a_{nj}^{-1}b_{nj}\right]^T$; a is the scale parameter; and b is the translation parameter.

This combination maintains the advantages of the BP neural network and overcomes insufficient accuracy of prediction results due to local extremum based on the ability of the wavelet transform to extract local information by amplifying the signal to optimize the weight and threshold of the BP neural network.

Wavelet and NARX (WNARX) hybrid model. A WNARX model is an integrated model combining two algorithms of the NARX and the wavelet transform. The wavelet decomposition coefficients of the water quality data are transported into the NARX model to set up a forecast hybrid model. For the WNARX model inputs, the original water quality time series is decomposed into various detail components at different resolution levels using the high- and low-pass filtering approaches. The prediction results are summarised as:

$$\mathrm{y(t)} = \sum_{i=1}^{L} f_i \begin{pmatrix} y_i(t-1), y_i(t-2), \ldots, y_i\left(t-n_y\right), \\ x_i(t-1), x_i(t-2), \ldots, x_i(t-n_u) \end{pmatrix} \tag{9}$$

where, y_i and x_i represents the divided signal of the separated input. It is recommended that the number of wavelet levels $\mathrm{L} = \mathrm{int}\left[\log_{10}(\mathrm{N})\right]$ levels are needed for transformation[41], where N denotes the number of transformed data.

Model performance evaluation. The models' performance was evaluated by error evaluation measurements of the coefficient of determination (R^2) and the root-mean-square errors (*RMSE*). R^2 was used to assess the predictive ability and accuracy of the model as expressed in Eq. (10).

$$R^2 = 1 - \frac{\sum \left(X_{forecast} - X_{measured}\right)^2}{\sum \left(X_{measured} - X_{meanmeasured}\right)^2} \tag{10}$$

RMSE is the measure of the difference between the measured and forecast values expressed in Eq. (11).

$$RMSE = \sqrt{\sum_{i=1}^{N} \left(X_{forecast} - X_{measured}\right)^2 \Big/ N} \tag{11}$$

A higher value of R^2 and a lower value of *RMSE* indicates better fitness and a smaller discrepancy between the observation and prediction. Generally, the R^2 greater than 0.6 and *RMSE* less than 10% of the range of target values are considered as the acceptable fitness between both series[42].

Results

Model establishment. *Inputs selection.* According to the previous studies, the occurrences of Fe, As, Pb, and Zn were usually linked with the industries activities[43]. Therefore, Fe could exist as associated emissions with As, Pb, and Zn. Flow, as a common hydrological monitor parameter to study the water environment capacity[44], affects the accumulation speed of metals and the solubility of metals as the solvent directly[45–47]. The pH value could influence the concentration of dissolved As, Pb, and Zn in water. Furthermore, pH and DO affect the form of As in water[48]. The water temperature affects the dissolution of As, but no consistent effects were observed on Pb, Zn concentration[49]. Pb and Zn, which could conduct electric current as metals, present intrinsic EC values respectively[50]. Hence, EC could have a partial relationship with the concentrations of Pb and Zn. Therefore, pH, EC, WT, flow, NO_3–N, and DO, and Fe were considered as the candidates of the input parameters. As shown in Table 2, the input combinations considered for each metal perdition were: As forecasted with inputs of (1) Fe, flow, pH, WT, and DO; (2) Fe, pH, and DO; (3) Fe; Pb forecasted with inputs of (4) Fe, flow, pH, NO_3–N, and EC; (5) Fe and flow, (6) Fe; and Zn forecasted with inputs of (7) Fe, flow, pH, NO_3-N, and EC, (8) Fe and NO_3-N, and (9) Fe.

Model structure. The optimal architecture of the different models and their parameter variation were determined based on their characteristics after testing the different data sets. BPNN, NARX, WNN, and WNARX models with different inputs were compared in the simulation phase. The R2 and RMSE values for the simulation processes of all given models were denoted. It was apparent that all the performances of these scenarios show a range of differences because of the different inputs or model structures. To get an effective evaluation of BPNN, NARX, WNN, and WNARX models' performance, the statistical results have been used as the criteria.

For BPNN, Figure S1-9 described the R^2 range of the separated scenarios with different parameter settings. As for BPNN, different data distribution and built-in parameter settings caused larger changes in the simulation process. Among them, the data distribution in BPNN has no obvious effect on the setting of purelin-purelin's activation function, impossible to be further adjusted more accurately. The function logsig as the output layer leads to a lower prediction ability for the BPNN structure. The settings of tansig-tansig ensured the impact of data distribution on the fitting results, indicating that the structure of BPNN was relatively stable. These structures present a further potential for parameter adjustment. The $BPNN_1$ with 5 hidden neurons and mode 3 were respectively chosen to be the optimal structure for the As prediction. The $BPNN_4$ with 4 hidden neurons and mode 3, the $BPNN_7$ with 4 hidden neurons and mode 2 were set for Pb and Zn predictions, respectively. As for NARX, the close-loop structure was established to iterative forecasting in scenarios. Shown in Figure S10, $NARX_1$ with mode 3mode3, $NARX_5$ with mode 3 and $NARX_9$ mode 4, these different data allocations, were applied to As, Pb, and Zn predictions by comparing the different modes. However, different input contaminants were chosen for different metal predictions.

Metal	Model	Scenario	Input	R^2	RMSE
As	BPNN	$BPNN_1$	Fe, Flow, pH, WT, DO	0.550	0.383
		$BPNN_2$	Fe, pH, DO	0.415	0.376
		$BPNN_3$	Fe	0.442	0.163
	NARX	$NARX_1$	Fe, Flow, pH, WT, DO	0.537	0.512
		$NARX_2$	Fe, pH, DO	0.468	0.499
		$NARX_3$	Fe	0.279	0.255
	WNN	WNN_1	Fe, Flow, pH, WT, DO	0.122	0.279
		WNN_2	Fe, pH, DO	0.101	0.026
		WNN_3	Fe	0.439	0.178
	WNARX	$WNARX_1$	Fe, Flow, pH, WT, DO	0.321	0.475
		$WNARX_2$	Fe, pH, DO	0.631	0.300
		$WNARX_3$	Fe	0.335	0.278
Pb	BPNN	$BPNN_4$	Fe, Flow, pH, NO_3–N, EC	0.703	1.290
		$BPNN_5$	Fe, Flow	0.666	0.807
		$BPNN_6$	Fe	0.632	0.794
	NARX	$NARX_4$	Fe, Flow, pH, NO_3–N, EC	0.621	1.006
		$NARX_5$	Fe, Flow	0.622	0.919
		$NARX_6$	Fe	0.611	0.777
	WNN	WNN_4	Fe, Flow, pH, NO_3–N, EC	0.648	0.764
		WNN_5	Fe, Flow	0.691	0.714
		WNN_6	Fe	0.614	0.816
	WNARX	$WNARX_4$	Fe, Flow, pH, NO_3–N, EC	0.013	3.306
		$WNARX_5$	Fe, Flow	0.039	1.085
		$WNARX_6$	Fe	0.602	0.761
Zn	BPNN	$BPNN_7$	Fe, Flow, pH, NO_3–N, EC	0.780	6.702
		$BPNN_8$	Fe, NO_3–N	0.714	5.033
		$BPNN_9$	Fe	0.632	3.499
	NARX	$NARX_7$	Fe, Flow, pH, NO_3–N, EC	0.385	9.280
		$NARX_8$	Fe, NO_3–N	0.345	5.538
		$NARX_9$	Fe	0.575	4.067
	WNN	WNN_7	Fe, Flow, pH, NO_3–N, EC	0.768	3.428
		WNN_8	Fe, NO_3–N	0.700	3.425
		WNN_9	Fe	0.613	2.884
	WNARX	$WNARX_7$	Fe, Flow, pH, NO_3–N, EC	0.034	10.188
		$WNARX_8$	Fe, NO_3–N	0.006	13.727
		$WNARX_9$	Fe	0.637	3.407

Table 3. The structure and the performance statistics prediction.

As for WNN, the structures of the model mimicked the final scheme of the BPNN, with the morlet wavelet replacing the tansig, logsig and purelin function as the activation function. In this model, the different number of hidden neurons present little effect on the optimal results judged by R^2 described by Figure S11. The highest R2 for As was WNN_3 under mode 3 with 4 hidden neurons; WNN_5 mode 2 with 7 hidden neurons for Pb; and WNN_7 mode 2 with 6 hidden neurons for Zn. For WNARX, the Daubechies (db3) function with three-level decomposition was found to be the optimal wavelet for series analysis. Actually, the computing process of WNARX model was first to decompose the time series of the inputs data and then integrate it into the NARX calculation. As shown in Figure S12, the model $WNARX_6$ and $WNARX_9$ with mode 3 had the best simulations for Pb and Zn. Besides, it has a relatively good simulation for As under $WNARX_2$ with mode 4, with the optimal value of R^2 in mode 4.

As given in Table 3, the optimal model with the best combination of inputs in this study considering the values of R^2 and RMSE are : (1) WNARX with inputs of Fe, pH, and DO for the prediction of As ($WNARX_2$), (2) WNN with inputs of Fe, Flow, pH, NO3-N, and EC for the prediction of Pb (WNN_5), (3) WNN with inputs of Fe, Flow, pH, NO_3–N, and EC for the prediction of Zn (WNN_7). the values of R^2 and RMSE of $WNARX_2$, WNN_5, WNN_7 larger than 0.63 and less than 10% of the fluctuation ranges (with 5.1 μg/L for As, 14.7 μg/L for Pb, and 86.2 μg/L for Zn respectively).

Performance analysis of the optimal scenarios. *Trend analysis.* As shown in Fig. 3, all scenarios presented a certain fitting ability in the trend prediction. It suggests that in the prediction of As, the input data selection was more appropriate. Compared with other models, the prediction curves of WNN models were more

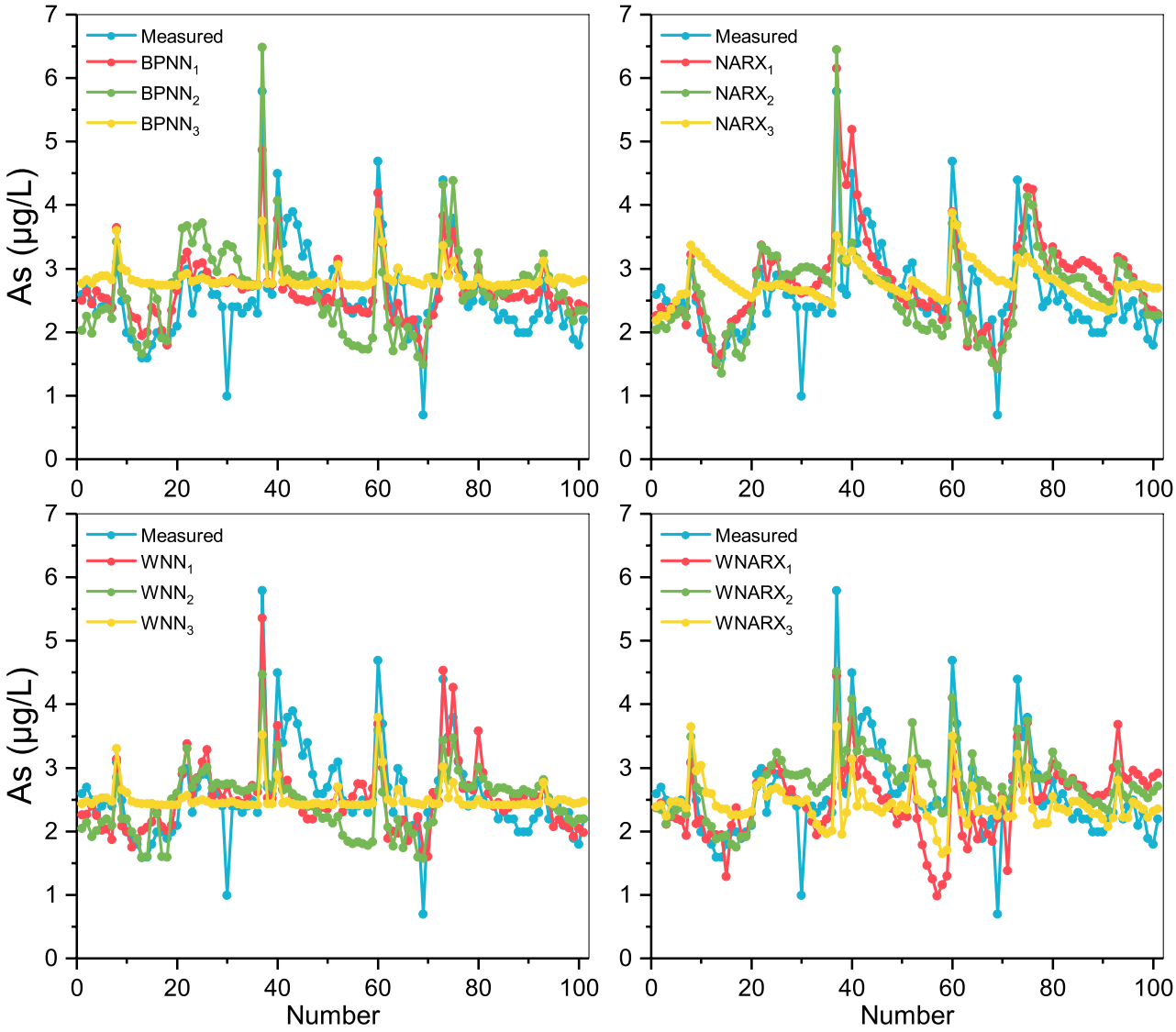

Figure 3. Comparison of the forecasting values and the measured values for As. This figure shows the different prediction results for As of the settings of the best model parameters in each model algorithm, indicating the concentrations of As changes within 2 years' measurement. The red represents the situation with inputs: Fe, flow, pH, WT, and DO, the green represents the situation with inputs: Fe, pH, and DO, and the yellow represents the situation with inputs: Fe.

stable, which showed that the wavelet analysis could reduce the noise after processing the input data series. The NARX model showed a better fitting effect in the downtrend stage. It indicates that the iterative prediction model has a certain prediction effect on the downward fluctuation trend in As compared with BPNN. The optimized model was set up based on a combination of wavelet and NARX algorithm, which retained the advantages of NARX and the stabilized ability of wavelet. In addition, the selection of input data by $WNARX_2$ not only avoided the interference of too much data in $WNARX_1$ on the As results (No. 15-20 and No. 55-71). It also avoided the disadvantage of single data in $WNARX_3$, allowing the result to predict the fluctuation trend relatively smoothly in As concentration.

As shown in Fig. 4, the prediction effects of BPNN, NRRX, and WNN models were similar, but the fitting deviations of WNARX were relatively large. It indicates that the amplification effect of the input wavelet was not suitable for NARX model prediction, even leading to negative effects. Besides, the red trend line represented by scenario 4 was significantly more volatile than the other trend lines. It shows that for toxic metal Pb prediction, 5 inputs could affect the accuracy of prediction. In fact, the optimal BPNN and WNN models showed similar effects on the prediction of Pb with R^2 values around 0.7. It might be due to the operation of WNN model had a better performance in RMSE values. Then, WNN_5 was selected as the optimal model. Besides, regarding the extreme points, the prediction effect of Pb was the best. It shows that the discharge, distribution, and degradation of Fe and Pb in rivers were similar. Therefore, Fe was a good reference value for Pb prediction in rivers.

As shown in Fig. 5, Zn fluctuated mostly in long sequences among the three heavy metals. BPNN model had the best performance in the simulating process judged by R^2 values, reaching the R^2 as 0.78. However, the stability

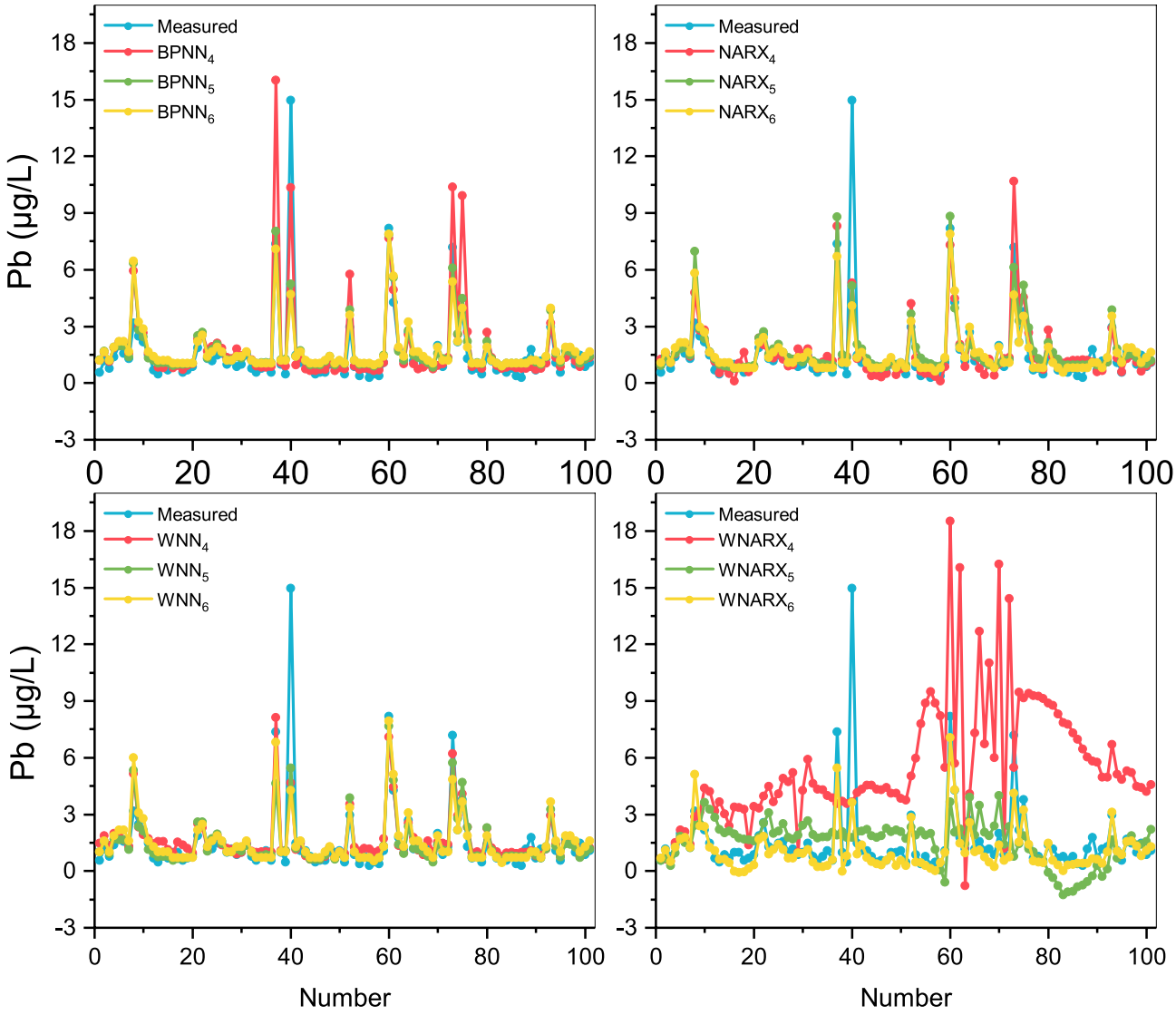

Figure 4. Comparison of the forecasting values and the measured values for Pb. This figure shows the different prediction results for Pb of the settings of the best model parameters in each model algorithm, indicating the concentration of Pb changes within 2 years' measurement. The red represents the situation with inputs: Fe, flow, pH, NO_3–N, and EC, the green represents the situation with inputs: Fe and flow, and the yellow represents the situation with inputs: Fe.

ability of WNN could achieve better converge the divergence effect brought by multiple inputs. In other words, the wavelet function as the activation function does not lose the sensitivity for its prediction of extreme values but show better adjustment for daily values. Therefore, the best WNN model could obtain lower RMSE values by keeping the R^2 values of 0.77. NARX and WNARX models showed a large difference in the predictions and both showed negative values in scenario 7. It indicates that both multi-input and NARX models were not suitable for the prediction of Zn. Therefore, WNN_7 was chosen as the most optimized model.

Scatter analysis. It can be seen from Fig. 6 that the yellow spots represent that the distribution with only Fe concentration as input had a significantly lower slope than the other two-liner fits. In other words, a single concentration input for the prediction of As was not desirable. The distributions of spots of $WNARX_1$ and $WNARX_3$ were more scattered than those of $WNARX_2$. The slopes of the linear fits of $WNARX_1$ and $WNARX_3$ were lower than those that of $WNARX_2$. Thus, $WNARX_2$ was still the optimal choice.

From Fig. 7, the slopes of the linear fits of $WNARX_4$ and $WNARX_5$ were close to 0, indicating that there is almost no correlation between the data. The slope of the fitted line of $BPNN_4$ performed better than the other two, closest to bisector. For the other models, the linear fits of the WNN models processed by wavelet functions keep almost the same slopes. Considering the concentrated scatters of WNN ranged from 0 to 3 20 µg/L showing the best performance in predicting the concentration of Pb. Therefore, WNN_5 was regarded as an acceptable scenario.

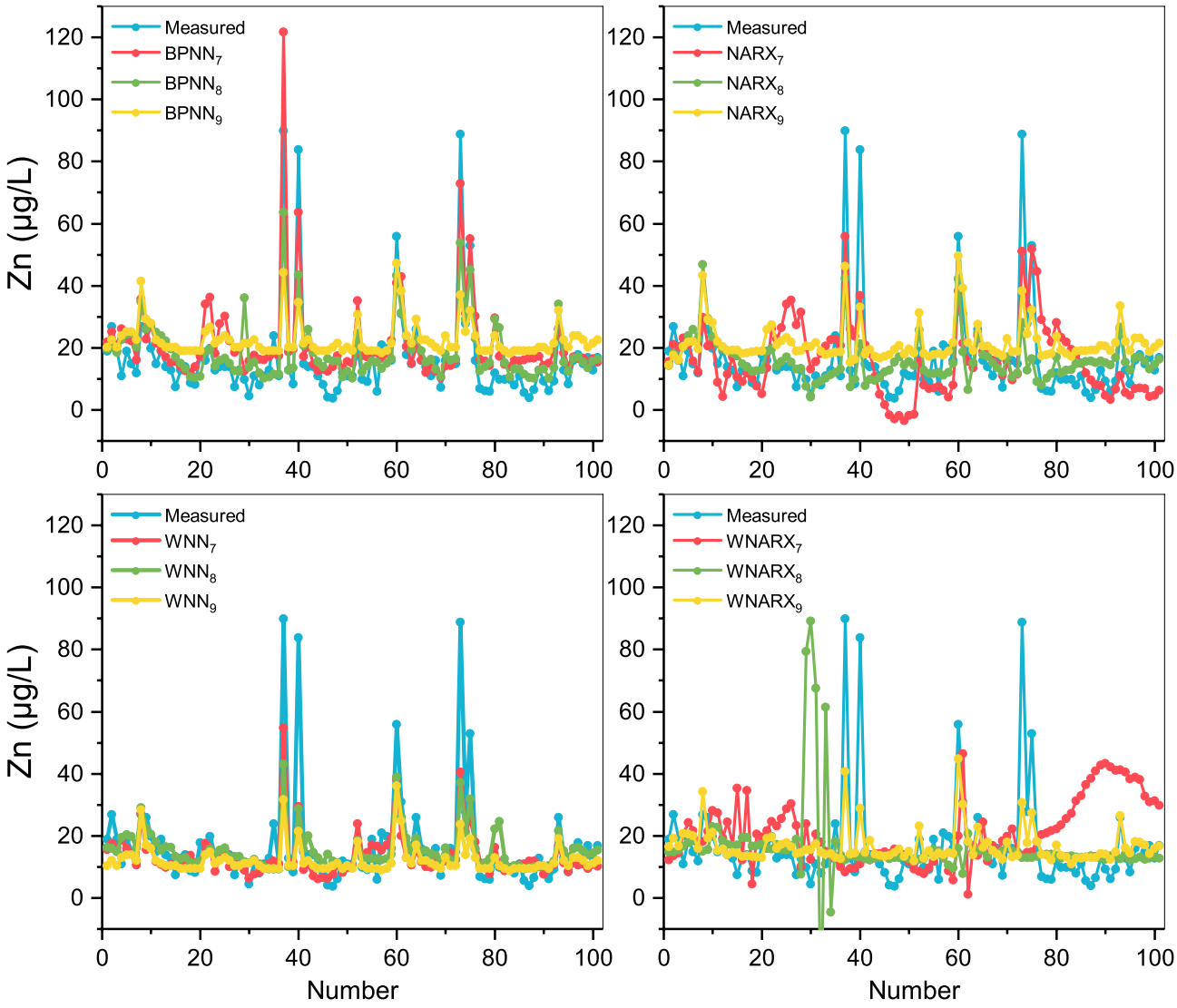

Figure 5. Comparison of the forecasting values and the measured values for Zn. This figure shows the different prediction results for Zn of the settings of the best model parameters in each model algorithm, indicating the concentration of Zn changes within 2 years' measurement. The red represents the situation with inputs: Fe, flow, pH, NO_3–N, and EC, the green represents the situation with inputs: Fe and NO_3–N, and the yellow represents the situation with inputs: Fe.

From the display in Fig. 8, the fitting effect of WNARX models was relatively poor. Considering the effect of BPNN, more input data could adjust the data close to the bisector. However, the WNN model could process the same input data more concentratedly. It shows that WNN is the model that could decrease the range of the prediction of Zn concentrated between 0 and 20 μg/L. Thus, WNN_7 used the most data input and was the most suitable model for daily prediction.

Discussion

In terms of As prediction models, when Fe was included as an input, their R^2 is significantly smaller than the Pb and Zn models of the same input, but WNARX has a significant improvement over NARX. Moreover, Fe, pH, and DO were the best inputs for WNARX models, the model also shows a better regression effect compared to the others. It shows that the wavelet decomposition can extract the division signals of the inputs' series, and these signals have a positive effect on the long-term prediction of As content in water. As for Pb, its performances in $BPNN_4$ and WNN_5 were similar. But due to the relatively lower RMSE value of the WNN model, this study selected WNN as the best model for Pb prediction. At the same time, these two approaches performed better predication when including Fe and flow as input data. It indicates that the higher correlation coefficient of Fe and flow with Pb will have a better performance. Considering the results of Zn, the BPNN and WNN model showed the optimal regression. According to the lower RMSE values, it denotes that wavelet, as the activation function in the neural network, could extract local information by processing the signal. As for WNARX, multiple inputs caused anomalies in the fitting effect. It implies that the wavelet divided signals were not suitable for the predictions of Pb and Zn.

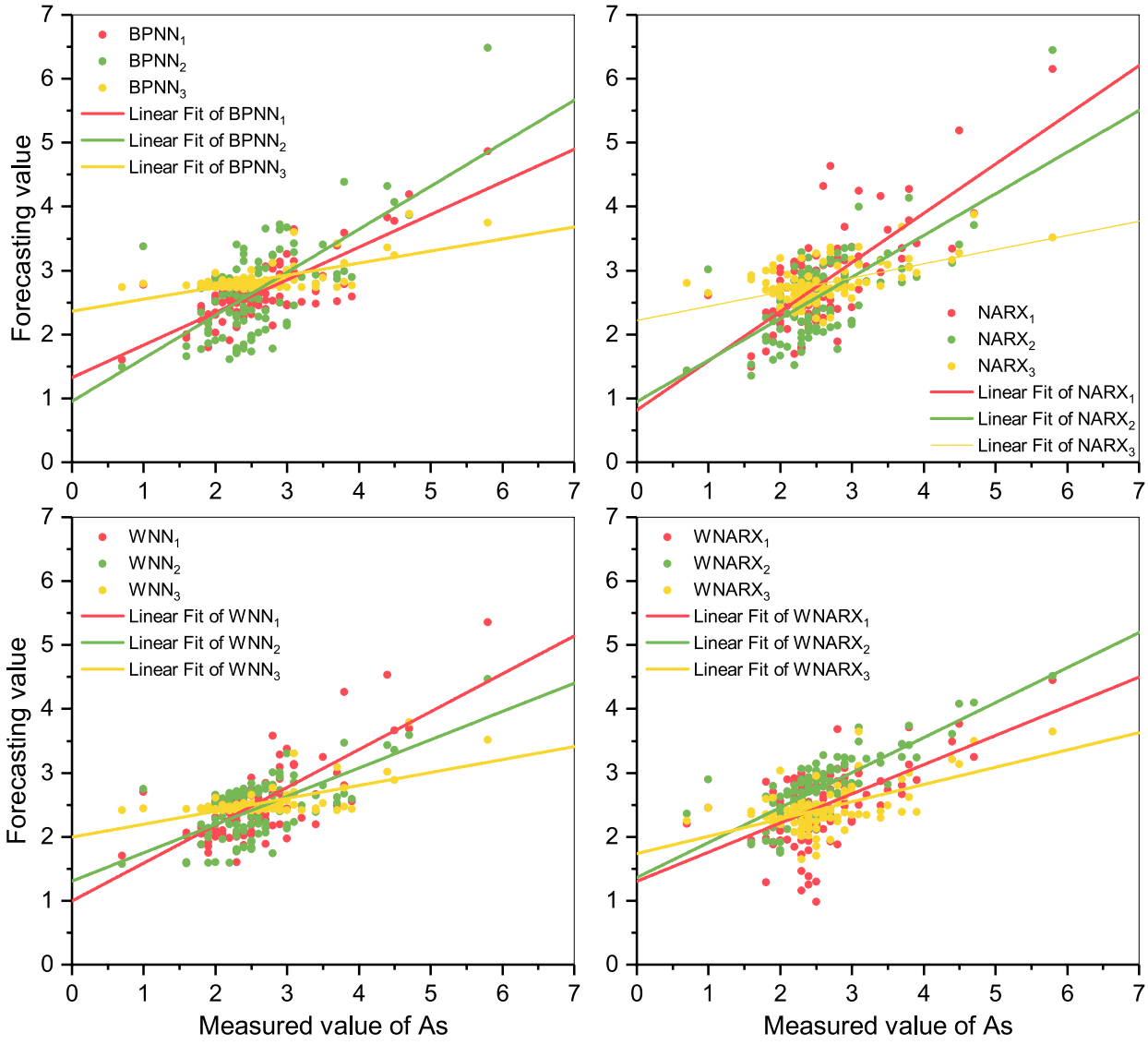

Figure 6. Scatter plots of As. The figure shows the scatters distributions under different scenarios for As and the linear fit of different scenarios. Meanwhile, the color choice keeps the same as Fig. 4.

Consequently, according to the models (BPNN, NARX, WNN, and WNARX) used as for the time series predictions for different targets, different combinations of inputs and models were considered as the algorithm to forecast the different metals concentrations in the river. These findings provide a new perspective for the long-term prediction of heavy metals in natural rivers. In addition, the improvement noted in our study was that the results analysed based on the supernumerary 2-years simulation instead of the test series of the training process. This study, therefore, indicates that the optimal models have practical application value and generalization. Our results provide compelling evidence for long-term prediction of As, Pb, and Zn concentrations and suggest that this approach appears to be effective in fields of water quality prediction.

Although our hypotheses were supported statistically, the results present a certain fitness between the forecasted and measured values. There are still many unanswered questions about the specific extreme value content prediction. Future work should, therefore, include follow-up work designed to find out other factors that influence the drastic changes in metal concentrations and whether they continue to be useful to improve accuracy. Besides, according to the previous studies, the physical model of WASP has been applied to simulate the spatial distribution of heavy metals in estuary[51], another physical model Delft3D-WAQ also could be used to simulate the heavy metal[52]. Although these models require detailed information, and a large number of data set were needed for the model set up and validation, a further comparison between physical models and our proposed models is useful and valuable for a deep understanding of the metals' behaviors in the aquatic environment.

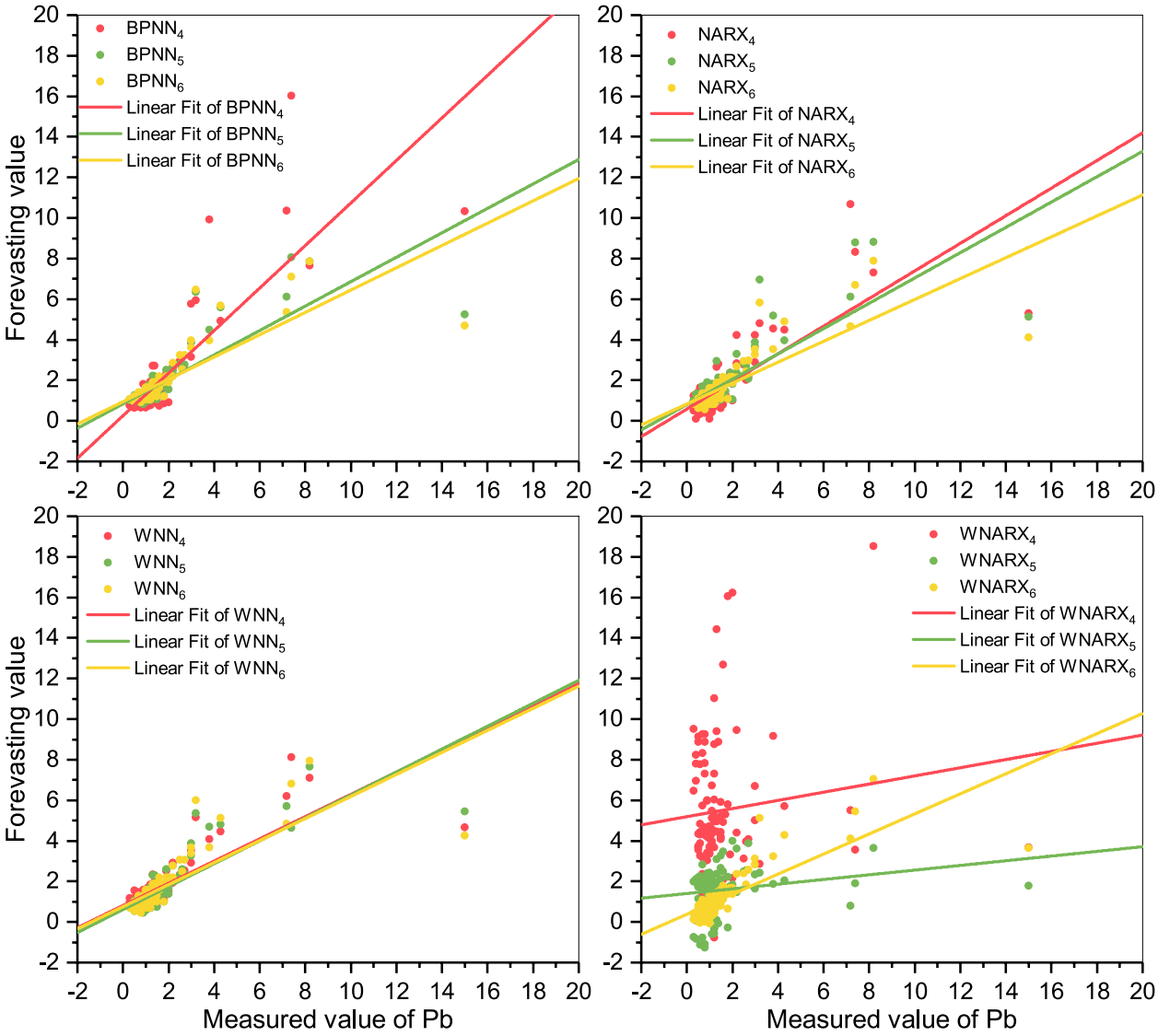

Figure 7. Scatter plots of Pb. The figure shows the scatters distributions under different scenarios for Pb and the linear fit of different scenarios. Meanwhile, the color choice keeps the same as Fig. 5.

Conclusion

To address the issues of long-term toxic metals' prediction, models of BPNN, NARX, WNN, and WNARX were employed in this study using the hybrid concepts of wavelet transform and artificial neural networks. The efficacy and fitness of the models were evaluated for their application in surface waters. The results revealed that: (1) the given models showed good performances for the long-term prediction of the toxic metals of As, Pb, and Zn; (2) the wavelet transform can enhance the long-term concentration prediction of As, Pb, and Zn especially in the daily conditions; and therefore (3) different models and inputs were required for different metal predictions to guarantee the optimum results.

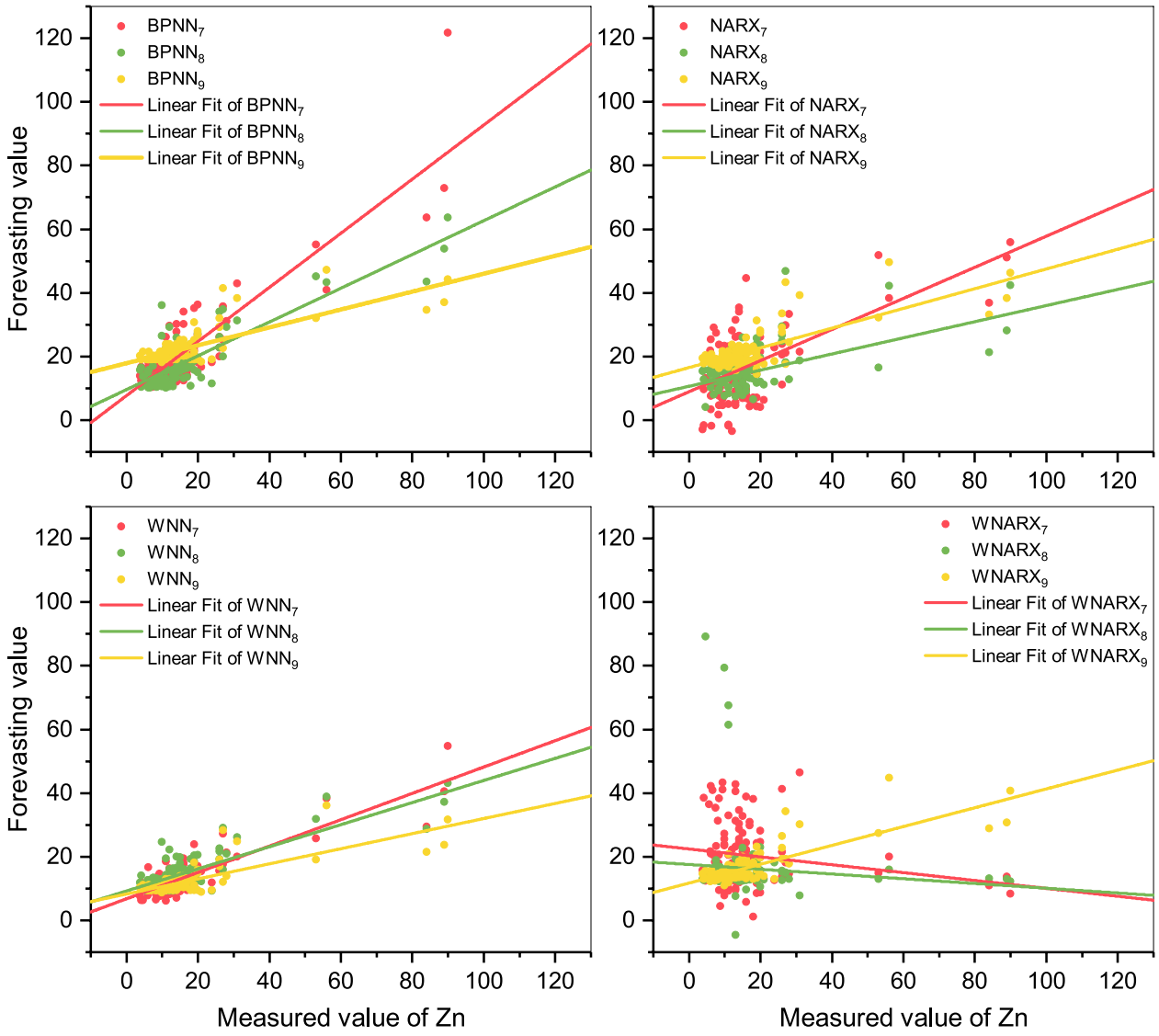

Figure 8. Scatter plots of Zn. The figure shows the scatters distributions under different scenarios for Zn and the linear fit of different scenarios. Meanwhile, the color choice keeps the same as Fig. 6.

Received: 13 May 2020; Accepted: 10 July 2020
Published online: 10 August 2020

References

1. Zhang, J., Hua, P. & Krebs, P. The build-up dynamic and chemical fractionation of Cu, Zn and Cd in road-deposited sediment. *Sci. Total Environ.* **532**, 723–732 (2015).
2. Wang, Z. *et al.* Concentration decline in response to source shift of trace metals in Elbe River, Germany: a long-term trend analysis during 1998–2016. *Environ. Pollut.* **250**, 511–519. https://doi.org/10.1016/j.envpol.2019.04.062 (2019).
3. Albering, H. J., Van Leusen, S. M., Moonen, E., Hoogewerff, J. A. & Kleinjans, J. Human health risk assessment: a case study involving heavy metal soil contamination after the flooding of the river Meuse during the winter of 1993–1994. *Environ. Health Perspect.* **107**, 37–43 (1999).
4. Enitan, I. T., Enitan, A. M., Odiyo, J. O. & Alhassan, M. M. Human health risk assessment of trace metals in surface water due to leachate from the municipal dumpsite by pollution index: a case study from Ndawuse River, Abuja, Nigeria. *Open Chem.* **16**, 214–227 (2018).
5. Graeme, K. A. & Pollack, C. V. Jr. Heavy metal toxicity, part I: arsenic and mercury. *J. Emerg. Med.* **16**, 45–56 (1998).
6. Malar, S., Vikram, S. S., Favas, P. J. & Perumal, V. Lead heavy metal toxicity induced changes on growth and antioxidative enzymes level in water hyacinths [*Eichhornia crassipes* (Mart.)]. *Bot. Stud.* **55**, 54 (2016).
7. Fu, F. & Wang, Q. Removal of heavy metal ions from wastewaters: a review. *J. Environ. Manag.* **92**, 407–418 (2011).
8. Tian, W., Liao, Z. & Zhang, J. An optimization of artificial neural network model for predicting chlorophyll dynamics. *Ecol. Model.* **364**, 42–52 (2017).
9. Wool, T. A., Ambrose, R. B., Martin, J. L., Comer, E. A. & Tech, T. Water quality analysis simulation program (WASP). *User's Manual, Version* **6** (2006).
10. Scharf, L. L. *Statistical Signal Processing* Vol. 98 (Addison-Wesley Reading, MA, 1991).
11. Hua, P., Vasyukova, E. & Uhl, W. A variable reaction rate model for chlorine decay in drinking water due to the reaction with dissolved organic matter. *Water Res.* **75**, 109–122 (2015).

12. Frostick, L. E., McLelland, S. J. & Mercer, T. G. *Users Guide to Physical Modelling and Experimentation: Experience of the HYDRALAB Network* (CRC Press, Boca Raton, 2011).
13. Allen, H. E., Luther, G. W. & Garrison, W. *Metals in Surface Waters* (CRC Press, Boca Raton, 1997).
14. Tu, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **49**, 1225–1231 (1996).
15. Mosavi, A., Dehghani, M. & Várkonyi-Kóczy, A. R. Deep learning and machine learning in hydrological processes, climate change and earth systems: a systematic review (2019)
16. Bejou, D., Wray, B. & Ingram, T. N. Determinants of relationship quality: an artificial neural network analysis. *J. Bus. Res.* **36**, 137–143 (1996).
17. Tokar, A. S. & Johnson, P. A. Rainfall-runoff modeling using artificial neural networks. *J. Hydrol. Eng.* **4**, 232–239 (1999).
18. Kişi, Ö. River flow modeling using artificial neural networks. *J. Hydrol. Eng.* **9**, 60–63 (2004).
19. Rajaee, T., Mirbagheri, S. A., Zounemat-Kermani, M. & Nourani, V. Daily suspended sediment concentration simulation using ANN and neuro-fuzzy models. *Sci. Total Environ.* **407**, 4916–4927 (2009).
20. Leahy, P., Kiely, G. & Corcoran, G. Structural optimisation and input selection of an artificial neural network for river level prediction. *J. Hydrol.* **355**, 192–201 (2008).
21. Ranković, V., Radulović, J., Radojević, I., Ostojić, A. & Čomić, L. Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia. *Ecol. Model.* **221**, 1239–1244 (2010).
22. Alizamir, M. & Sobhanardakani, S. Forecasting of heavy metals concentration in groundwater resources of Asadabad plain using artificial neural network approach. *J. Adv. Environ. Health Res.* **4**, 68–77 (2016).
23. Ke, N. The prediction model of heavy metal pollution in Xiangjiang River based on matlab. *J. Anhui Agric. Sci.* **9** (2012).
24. Wang, W. & Ding, J. Wavelet network model and its application to the prediction of hydrology. *Nat. Sci.* **1**, 67–71 (2003).
25. Torrence, C. & Compo, G. P. A practical guide to wavelet analysis. *Bull. Am. Meteor. Soc.* **79**, 61–78 (1998).
26. Sehgal, V., Tiwari, M. K. & Chatterjee, C. Wavelet bootstrap multiple linear regression based hybrid modeling for daily river discharge forecasting. *Water Resour. Manag.* **28**, 2793–2811 (2014).
27. Rajaee, T. Wavelet and ANN combination model for prediction of daily suspended sediment load in rivers. *Sci. Total Environ.* **409**, 2917–2928 (2011).
28. Nourani, V., Komasi, M. & Mano, A. A multivariate ANN-wavelet approach for rainfall–runoff modeling. *Water Resour. Manag.* **23**, 2877 (2009).
29. Mehr, A. D., Kahya, E. & Özger, M. A gene–wavelet model for long lead time drought forecasting. *J. Hydrol.* **517**, 691–699 (2014).
30. Matthies, M., Berlekamp, J., Lautenbach, S., Graf, N. & Reimer, S. System analysis of water quality management for the Elbe river basin. *Environ. Model. Softw.* **21**, 1309–1318 (2006).
31. Bray, M. & Han, D. Identification of support vector machines for runoff modelling. *J. Hydroinf.* **6**, 265–280 (2004).
32. Hong, N. *et al.* Quantitative source tracking of heavy metals contained in urban road deposited sediments. *J. Hazardous Mater.* **393**, 122362 (2020).
33. Morlet, J., Arens, G., Fourgeau, E. & Glard, D. Wave propagation and sampling theory—part I: complex signal and scattering in multilayered media. *Geophysics* **47**, 203–221 (1982).
34. Cannas, B., Fanni, A., See, L. & Sias, G. Data preprocessing for river flow forecasting using neural networks: wavelet transforms and data partitioning. *Phys. Chem. Earth A/B/C* **31**, 1164–1171 (2006).
35. Nanda, T., Sahoo, B., Beria, H. & Chatterjee, C. A wavelet-based non-linear autoregressive with exogenous inputs (WNARX) dynamic neural network model for real-time flood forecasting using satellite-based rainfall products. *J. Hydrol.* **539**, 57–73 (2016).
36. Goh, A. T. Back-propagation neural networks for modeling complex systems. *Artif. Intell. Eng.* **9**, 143–151 (1995).
37. Beale, M. H., Hagan, M. T. & Demuth, H. B. Neural network toolbox user's guide. *The MathWorks Inc* (1992).
38. Fan, J., Wang, Z. & Qian, F. Research progress structural design of hidden layer in BP artificial neural networks. *Control Eng. China* **12**, 105–109 (2005).
39. Zhang, Q. & Benveniste, A. Wavelet networks. *IEEE Trans. Neural Netw.* **3**, 889–898 (1992).
40. Alexandridis, A. K. & Zapranis, A. D. Wavelet neural networks: a practical guide. *Neural Netw.* **42**, 1–27 (2013).
41. Tiwari, M. K. & Chatterjee, C. Development of an accurate and reliable hourly flood forecasting model using wavelet–bootstrap–ANN (WBANN) hybrid approach. *J. Hydrol.* **394**, 458–470 (2010).
42. Alexander, D. L., Tropsha, A. & Winkler, D. A. Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *J. Chem. Inf. Model.* **55**, 1316–1322 (2015).
43. World Health Organization. Guidelines for drinking-water quality. *WHO chronicle* (2011).
44. Chen, L. *et al.* Water environmental capacity calculated based on point and non-point source pollution emission intensity under water quality assurance rates in a tidal river network area. *Int. J. Environ. Res. Public Health* **16**, 428 (2019).
45. Slooff, W. *et al.* Integrated criteria document arsenic. *RIVM Rapport 710401004* (1990).
46. Lenntech. *Zinc (Zn) and water*, <https://www.lenntech.com/periodic/water/zinc/zinc-and-water.htm> (2020).
47. Kim, E. J., Herrera, J. E., Huggins, D., Braam, J. & Koshowski, S. Effect of pH on the concentrations of lead and trace contaminants in drinking water: a combined batch, pipe loop and sentinel home study. *Water Res.* **45**, 2763–2774 (2011).
48. Association, T. W. Q. *Arsenic Fact Sheet*, <https://www.lenntech.com/periodic/water/zinc/zinc-and-water.htm> (2013).
49. Bonte, M., van Breukelen, B. M. & Stuyfzand, P. J. Temperature-induced impacts on groundwater quality and arsenic mobility in anoxic aquifer sediments used for both drinking water and shallow geothermal energy production. *Water Res.* **47**, 5088–5100 (2013).
50. Helmenstine, A. M. Table of electrical resistivity and conductivity. *ThoughtCo. Sep.* **24**, 2018 (2018).
51. De Smedt, F., Vuksanovic, V., Van Meerbeeck, S. & Reyns, D. *in Trace Metals in the Westerschelde Estuary: A Case-Study of a Polluted, Partially Anoxic Estuary 143–155* (Springer, Berlin, 1998).
52. Negm, A. M., Bek, M. A. & Abdel-Fattah, S. *Egyptian Coastal Lakes and Wetlands: Part II: Climate Change and Biodiversity, 72* (Springer, Berlin, 2018).

Acknowledgements

The authors would like to gratefully thank the Saxony State Office for Environment, Agriculture, and Geology (Landesamt für Umwelt, Landwirtschaft und Geologie, LfULG) for providing the data and the Public Operating Company for Environment and Agriculture (Staatliche Betriebsgesellschaft für Umwelt und Landwirtschaft, BfUL) for measuring the data. This work was supported by the state-sponsored scholarship program provided by the China Scholarship Council (CSC) (No.: 201908080087) and Guangdong Basic and Applied Basic Research Foundation (No.: 2020A1515011130). The mentioning of trade names or commercial products does not constitute endorsements or recommendations for use.

Author contributions

P.L.: Writing- Original draft preparation, Methodology, Software. P.H.; D.G.; J.N.; P.P.: Writing- Reviewing and Editing. J.Z.: Conceptualization, Writing- Reviewing and Editing, Project administration. P.K.: Resources, Supervision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-70438-8.

Correspondence and requests for materials should be addressed to J.Z.

Reprints and permissions information is available at www.nature.com/reprints.

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