



Statistical comparison between SARIMA and ANN's performance for surface water quality time series prediction

Xuan Wang¹ · Wenchong Tian¹ · Zhenliang Liao^{1,2} 

Received: 23 November 2020 / Accepted: 17 February 2021 / Published online: 27 February 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract

The performance comparison studies of the autoregressive integrated moving average model (ARIMA) and the artificial neural network (ANN) were mostly carried out between the selected model structures through trial-and-error, strongly influenced by model structure uncertainty. This research aims to make up for this inadequacy. First, a surface water quality prediction case study including eight monitoring sites in China was introduced. Second, the ARIMA and ANN's performance was compared statistically between 6912 Seasonal ARIMA (SARIMA) and 110,592 feedforward ANN with different model structures, based on the mean square error (MSE) distributions depicted by boxplots. In a statistical view, the ANN models obtained a significantly lower median value and a more concentrated distribution of validation MSEs, which indicated lighter overfitting and better generalization ability. Furthermore, the optimal SARIMA models' performance is inferior to even the median of the ANN models in the case study. In contrast with the previous comparisons among selected models, the statistical comparison in this study shows lower uncertainty.

Keywords ANN · ARIMA · Surface water quality · Time series prediction · Statistical comparison · Grid sampling

Introduction

Surface water quality prediction is crucial in the water planning and management process. With the help of prediction models, the degree and trend of water pollution could be forecasted, providing timely and effective decision support for the administrators (Zhou et al. 2018). Generally, surface water quality prediction models are classified as theory-driven models and data-driven models (Hunter et al. 2018). In recent decades, data-driven models have been more widely applied in water quality prediction tasks (Mount et al. 2016), owing to the accumulation of surface water monitoring data, the improvement of computing power, and algorithm optimization (Kang et al. 2017).

There are mainly three types of data-driven models for water quality prediction. The first is the traditional statistical method-based models such as multiple linear regression (MLR), autoregression (AR), and autoregressive integrated moving average model (ARIMA) (Monteiro and Costa 2018; Khairuddin et al. 2019). Among them, ARIMA proposed by Box and Jenkins (1976) has become one of the most widely used techniques (Khairuddin et al. 2019) and has been proven effective for water quality predictions. For example, Ahmad et al. (2001) identified the ARIMA as the appropriate model for modeling water quality data of the Ganges River in India. Salmani and Salmani Jajaei (2016) proved that SARIMA models satisfied the necessary tests and conditions for water quality forecast in the Karoun River in Iran. Sheikhy Narany et al. (2017) predicted nitrate contamination inside water resources in Malaysia based on an ARIMA model. The second type of data-driven model is the machine learning-based models (Ansari et al. 2018; Garcia Nieto et al. 2019; Hanson et al. 2020), such as artificial neural networks (ANNs), support vector machine (SVM), and adaptive neural fuzzy interference (ANFIS). ANN is the predominant among these models (Bhagat et al. 2020). Increasing studies of ANN for water quality predictions have been done over these years and exerted satisfactory performance (Maier et al. 2010). For example, Haghiabi et al. (2018) demonstrated

Responsible Editor: Xianliang Yi

✉ Zhenliang Liao
zl_liao@tongji.edu.cn

¹ College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China

² College of Civil Engineering and Architecture, Xinjiang University, Urumqi 830046, China

ANN had suitable performance for predicting water quality variables (WQ variables) (Tiyasha et al. 2020) in the Tيره River. Hameed et al. (2017) found ANN exhibited a remarkable ability to capture the nonlinearity pattern from the tropical rivers' water quality data. Shi et al. (2018) proved that ANN was effective for high-frequency surface water quality prediction. The last type is hybrid models, such as ANN-ARIMA, wavelet-neural networks (WANN), and extreme gradient boosting-ANN (XGBoost-ANN) (Bhagat et al. 2021).

Many comparison studies of various data-driven models' performance in the field of water resources have been conducted. Among them, comparisons between ARIMA and ANN are common. Raman and Sunilkumar (1995) used ARIMA and ANN to predict reservoir inflow. The trial-and-error method was used to select the "optimal" model structure. The results showed ANN provided a better fit in the absence of data preprocessing. Landeras et al. (2009) compared ARIMA and ANN in evapotranspiration forecast tasks. They selected the best two from the 18 ARIMA models and the 12 ANN models, respectively. The performance comparison results showed that ANN performed better for the summer months while ARIMA performed better from September to November. In the comparison study carried out by Ömer Faruk (2010), the optimal ARIMA structure was selected via AIC (Akaike information criterion), while the ANN structure was determined on an ad-hoc basis (Maier et al. 2010). The results showed that ANN had better performance than ARIMA for water quality prediction. Similarly, Valipour et al. (2013) also selected the optimal ARIMA structure via AIC, while the neurons' number of the single-hidden layer ANN was determined via trial-and-error. Results showed that the performance of the two optimal models was acceptable in forecasting the reservoir inflow. Elkiran et al. (2019) concluded that machine learning models were more robust than ARIMA in the river dissolved oxygen prediction research. The comparisons were conducted among the selected ARIMA and machine learning models, of which the model structures were also determined through trial-and-error. Two inadequacies could be found throughout the above comparison studies: (i) although the comparison studies of ARIMA and ANN in the field of water resources were conducted a lot, the majority were water-quantity-related. In contrast, research on surface water quality prediction issues was rare. (ii) Mostly, comparisons were carried out among the "optimal" model structures selected via trial-and-error or in an ad hoc way. However, given the high uncertainty associated with the model structure (Zhang et al. 2011), the comparison results among a couple of selected models seem inconclusive.

This study seeks to make up for the inadequacies. First, a surface water quality time series prediction case study of eight monitoring sites in China was introduced. Second, ARIMA and ANN's prediction performance was compared statistically between thousands of models of different model structures.

Specifically, the training and validation samples are weekly automatic monitoring data, published by the China National Environmental Monitoring Centre. Based on data analysis and previous experiments, the value sets of the model structure hyperparameters were determined. Then, 6912 SARIMA models and 110,592 feedforward ANN models with grid-sampled model structure hyperparameters sets were developed and trained. Moreover, the performance metric, mean square error (MSE), was calculated for each model on the training and validation datasets. Afterward, the two types of models were compared statistically based on the MSE distributions. The distributions were depicted by boxplots, where the median and percentile MSE values were given.

The paper's remaining part was organized as follows: ARIMA and ANN's brief reviews were presented in the "Methods" section, followed by the comparison case study for surface water quality time series prediction in the "Case study" section. Then, the ARIMA and ANN models' performance assessment results were illustrated separately and then compared and discussed in the "Results and discussion" section. Finally, the summary and conclusions were presented in the "Summary and conclusions" section.

Methods

Overall, the ARIMA and ANN models with different structures need to be developed and trained first. Then, the performance metric of each model on both train and validation dataset needs to be calculated. Afterward, the two types of models' performance should be statistically compared based on the median and percentile metric values. The model structures value sets of ARIMA were specified via the auto correlation function (ACF) and partial auto correlation function (PACF) curves, while the ANNs were determined based on literature review.

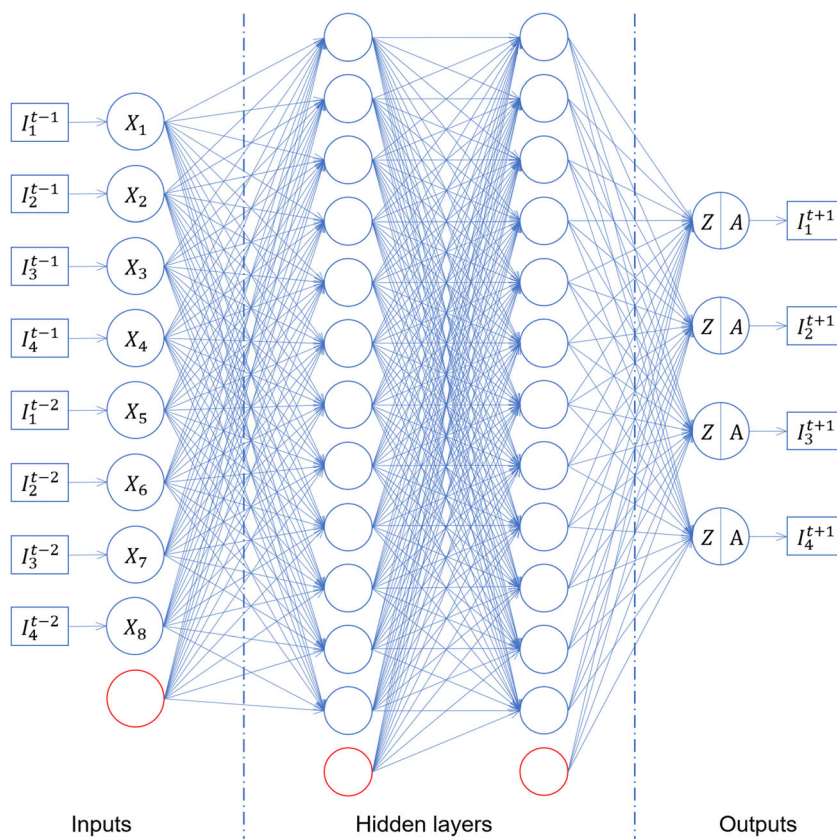
ARIMA

Autoregressive integrated moving average model (ARIMA) contains three parts: (i) autoregressive (AR) module, which describes the memory of the system's former state. (ii) Moving average (MA) module, which describes the memory of the noise that entered the system in the past. (iii) The integration procedure indicates the number of differences required to guarantee the series's stationarity (Rafael et al. 2019). Since stability is a prerequisite for the AR and MA modules, stabilization is required. Equations are as follows:

$$I_t' = \Delta^d I_t = \Delta(\Delta^{d-1} I_t) \quad (1)$$

$$I_t' = \delta + \sum_{i=1}^p \phi_i I_{t-i}' + \sum_{i=1}^q \theta_i e_{t-i} + e_t \quad (2)$$

Fig. 1 Schematic view of an example feedforward ANN with eight inputs ($n=2$) and a bias unit, added with two hidden layers with 12 neurons and a bias unit in each layer. The symbol I_1^{t-1} represents input or output data, where the superscript $(t-1)$ represents data one week lagged, and the subscript refers to the corresponding WQ variables: 1 - pH , 2 - DO , 3 - COD , 4 - NH_3-N



In Eq. (1), the d -order non-seasonal difference is made to series I_t . In Eq. (2), $\sum_{i=1}^p \phi_i I_{t-i}'$ is a p -order AR model, in which p is the autoregression order. $\sum_{i=1}^q \theta_i e_{t-i}$ is a q -order MA model, in which q is the moving average order. ϕ_i and θ_i are the model parameters that are to be optimized via the least square algorithm. δ is the intercept, e_t is the white noise that obeys the independent identical distribution: $e_t \sim N(0, \sigma_e^2)$.

Furthermore, given the seasonal fluctuation of the original surface water quality time series (see Fig. 3), the SARIMA, an extension of the ARIMA, is more likely to fit the data well. A seasonal component was considered in the SARIMA so that a seasonal difference step was added (Eq. (3)), and seasonal AR and MA modules were introduced (Eq. (4)).

$$I_t^\lambda = \Delta^D I_t' = \Delta \left(\Delta^{D-1} I_t' \right) \tag{3}$$

$$I_t^\lambda = \delta \sum_{i=1}^p \phi_i I_{t-i}^\lambda + \sum_{i=1}^q \theta_i e_{t-i} + \sum_{i=1}^P \Phi^{is} I_{t-is}^\lambda + \sum_{i=1}^Q \Theta_{is} e_{t-is} + e_t \tag{4}$$

In Eq. (4), $\sum_{i=1}^P \Phi^{is} I_{t-is}^\lambda$ is a P -order seasonal AR model, $\sum_{i=1}^Q \Theta_{is} e_{t-is}$ is a Q -order seasonal MA model. Φ_i and Θ_i are the parameters to be optimized via the least square algorithm.

To sum up, the model structure of a SARIMA in this case study was determined by seven hyperparameters:

$$(p, d, q), (P, D, Q, S) \tag{5}$$

where the orders of the non-seasonal and seasonal difference, d and D , along with the seasonal period, S , could be determined based on the ACF curves (see Fig. 4). After the difference step, the orders of non-seasonal and seasonal AR and MA, (p, q) and (P, Q) could be determined based on the ACF and PACF curves of the stabilized series.

ANN

Artificial neural network (ANN) can be classified as feedforward ANN, recurrent ANN, and convolution ANN. Among them, feedforward ANN is the first and simplest architecture (Schmidhuber 2015), and was used in this study. The model structure of a feedforward ANN is determined by an input layer, a hidden layer(s), and an output layer (Schmidhuber 2015).

In the surface water quality time series prediction, the inputs of a feedforward ANN model were lagged water quality data. Thus, the hyperparameter “inputs” was appointed to represent lagged time steps. The number of hidden layers and neurons in each hidden layer were set as hyperparameter “layers” and “neurons.” The two determine the depth and

Fig. 2 Locations of the eight monitoring sites



width of the network as well as the ANN's generalization capability and complexity. An example was given in Fig. 1 to illustrate the feedforward ANN model structure. Hyperparameters set in Eq. (6) represent that two weeks of lagged data for each of the four WQ variables are used as model inputs. The number of hidden layers is two, with 12 neurons in each.

$$\begin{cases} \text{inputs} = 2, \\ \text{layers} = 2, \\ \text{neurons} = 12. \end{cases} \quad (6)$$

Then, a specific activation function needs to be added to each neuron to introduce nonlinear informational transformation to the ANN. The two recommended activation functions were *Tanh* and *ReLU* (Eq. (6)-(7)).

$$g_{\text{Tanh}}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (7)$$

$$g_{\text{ReLU}}(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \quad (8)$$

In the training process, the adaptive moment estimation (Adam) algorithm was used. Thus, the initial learning rate value needed to be set. We appointed it as the hyperparameter “*lr*.” Then, we adopted the mini-batch gradient descent strategy, so the hyperparameter “*batch size*” was set to represent the number of samples used by each epoch to update parameters.

To sum up, the model structure of an ANN in this case study was determined by the following six hyperparameters:

$$\{\text{inputs}, \text{layers}, \text{neurons}, \text{activation}, \text{lr}, \text{batch size}\} \quad (9)$$

Statistical comparison

The performance metric measures the model's accuracy by judging the similarity between the predicted output(s) and the real one(s). The most commonly used metric in water resources models is the mean square error (MSE, Eq. (10)).

$$MSE = \frac{1}{n} \sum_{k=1}^n [y(t_k) - y_p(t_k)]^2 \quad (10)$$

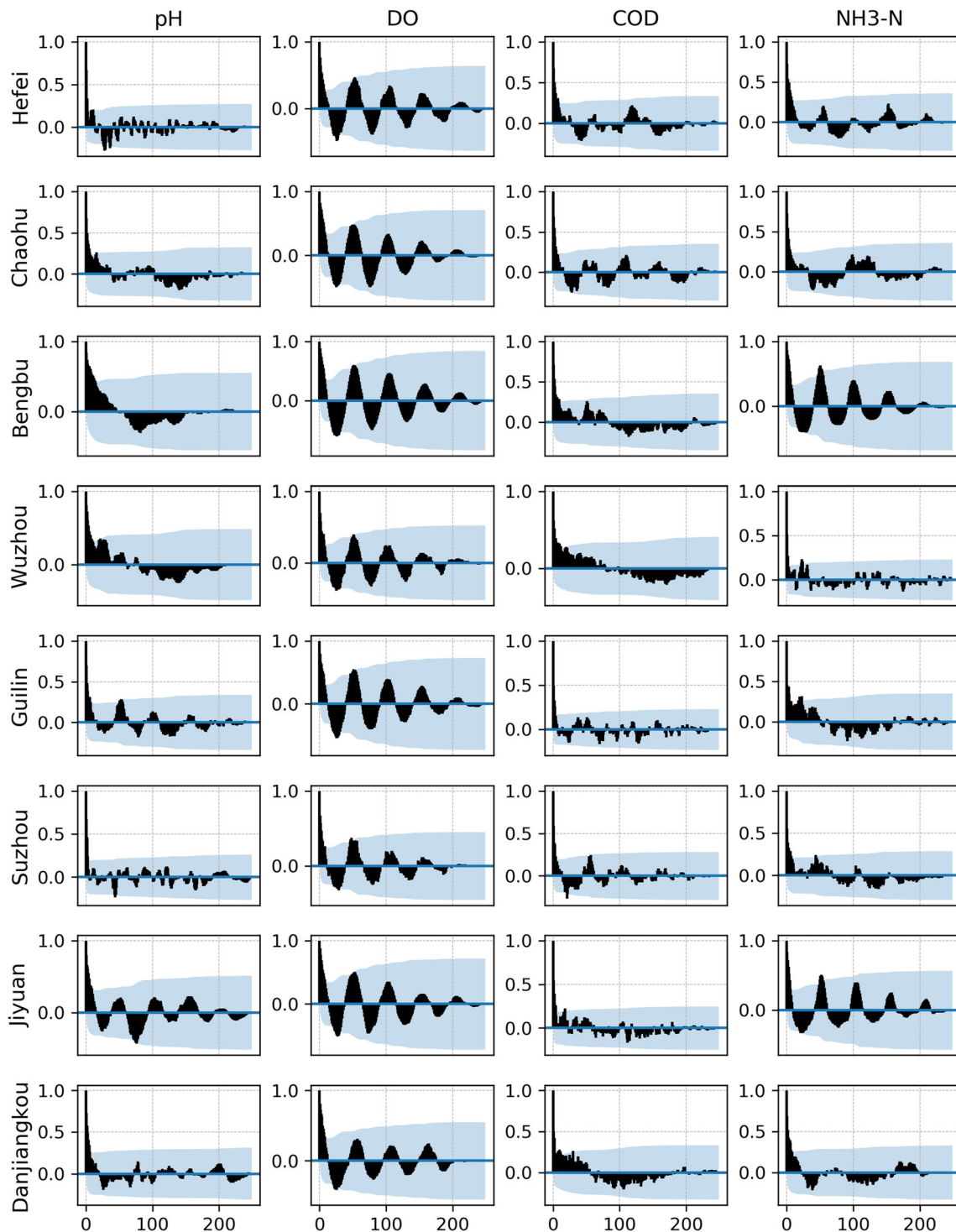


Fig. 3 ACF curves of the original series

where $y_p(t_k)$ and $y(t_k)$ are the predicted output(s) and the real one(s) at the time of t_k , respectively. The optimal value is 0 for MSE.

It should be noted that the SARIMA models for the four WQ variables: potential of hydrogen (*pH*), chemical oxygen demand (*COD*), dissolved oxygen (*DO*), and ammonia nitrogen (*NH₃-N*) were developed separately. Thus, the MSEs were first calculated for the four WQ variables, respectively.

And then, the average MSEs distributions were obtained for each of the eight sites. Otherwise, the ANN models integrated all the four WQ variables as outputs. Thus, the average MSEs were calculated directly for each site.

To statistically compare the performance of ARIMA and ANN, we used boxplots to describe the MSE distributions of the SARIMA and ANN models with different structures. The

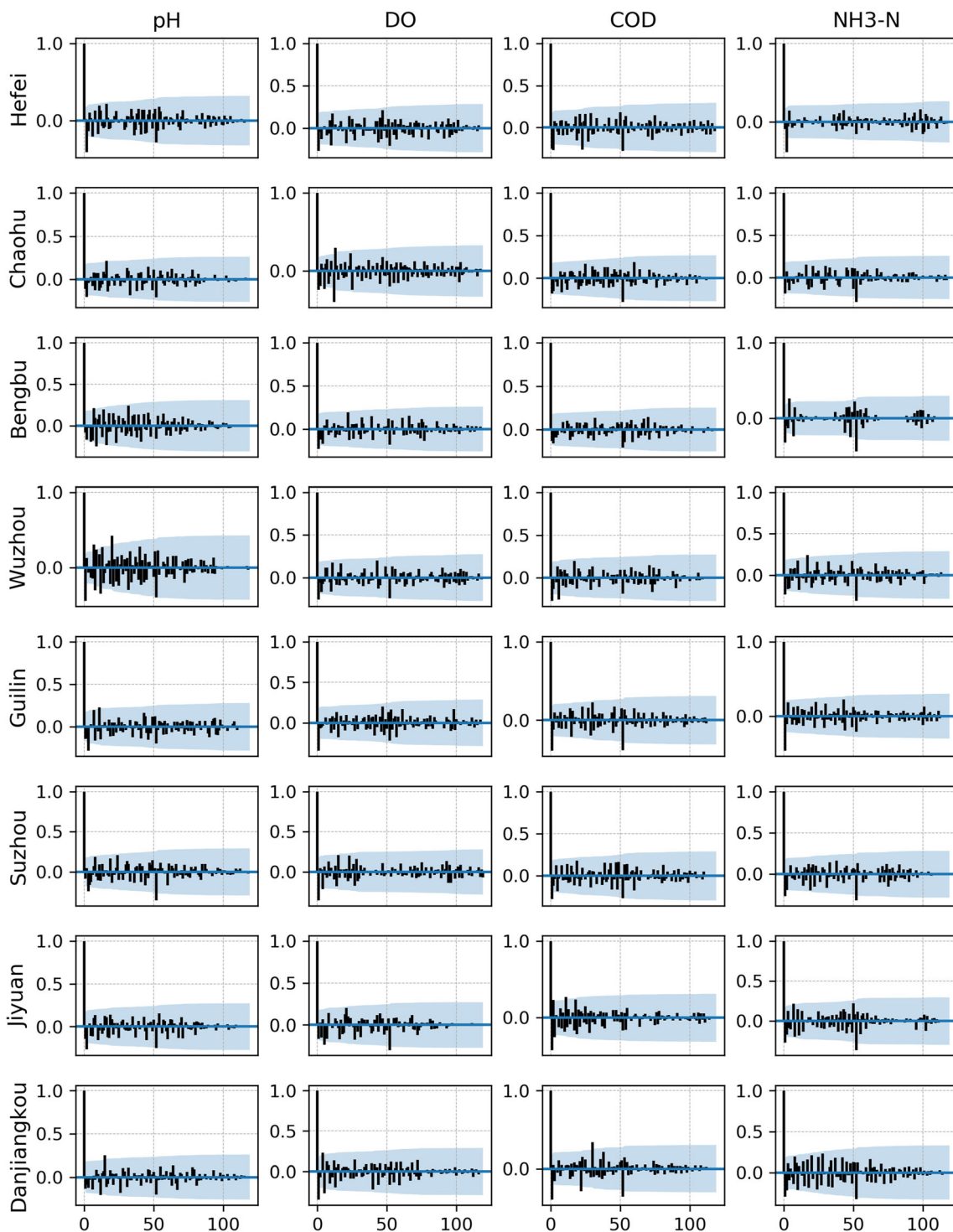


Fig. 4 ACF curves of the seasonal and first-order differences

Table 1 Value sets of SARIMA’s model structure hyperparameters

Hyperparameters	Implications	Value set
d, D	(Seasonal) differential order	$\{0, 1\}$
p, P	(Seasonal) autoregressive order	$\{0, 1, 2\}$
q	Moving average order	$\{0, 1, 2\}$
Q	Seasonal moving average order	$\{0, 1\}$

boxplots in this study illustrated the median and percentile values (5th, 25th, 75th, 95th percentiles), little affected by outliers. The boxplots can describe the central tendency (median value) of the MSEs, and depict the dispersion feature accurately and stably. Moreover, the boxplots can provide a simple but effective approach to compare distributions among variables.

Table 2 Value sets of ANN’s model structure hyperparameters

Hyperparameters	Implications	Value set
<i>inputs</i>	Lagged steps of data used as inputs	{1, 2, 3, 4}
<i>layers</i>	Number of hidden layers	{1, 2, 3, 4, 5, 6, 7, 8}
<i>neurons</i>	Number of neurons in each hidden layer	{4, 8, 12, 16, 20, 24, 28, 32, 36}
<i>activation</i>	Activation function	{Tanh, ReLU}
<i>lr</i>	Initial learning rate of Adam	{0.001, 0.003, 0.01, 0.03}
<i>batch size</i>	Number of samples used in each epoch	{2, 4, 8, 16, 32, 64}

Case study

Surface water quality time series data

The two types of data-driven models were applied to a case study for surface water quality time series prediction. The weekly automatic monitoring data were published by the China National Environmental Monitoring Centre. The data validity was guaranteed by periodic equipment calibration and replacement of electrodes. Most of the published data started in 2004, including four WQ variables: *pH*, *COD*, *DO*, *NH₃-N*.

Considering data integrity and site distribution, we selected eight monitoring sites located in eight cities in China: Hefei, Chaohu, Bengbu, Wuzhou, Guilin, Suzhou, Jiyuan, Danjiangkou, respectively (Fig. 2). Then 300 samples of each site, monitored from 2005/01/04 to 2009/09/27, were used to train the models, while the next 200 samples, monitored from 2009/10/04 to 2013/07/28, were used to validate the models. Since the data set with missing values was not valid for training ANN models, all the time series data were complemented with mean values. The time series diagrams of the eight sites after data complementation are available in the [supplementary document](#).

All the sample datasets for the SARIMA and ANN models were divided into training sets and validation sets.

Considering temporal order, we directly set the first 60% as training samples and the last 40% as validation samples (Wu et al. 2013). Since data normalization is helpful to improve the convergence rate of training, the Min-Max method (Eq. (11)-(12)) was used. The ahead time step to prediction was set as one for both the SARIMA and ANN models.

$$x'_i = \frac{x_i - \max(x)}{\max(x) - \min(x)} \tag{11}$$

$$y'_i = \frac{y_i - \max(y)}{\max(y) - \min(y)} \tag{12}$$

Development of the SARIMA models

According to the basic knowledge of ARIMA in the “ARIMA” section, model structure hyperparameters (*p*, *d*, *q*) and (*P*, *D*, *Q*, *S*) need to be determined before model training. Figure 3 shows the ACF curves of the original series. Apparent seasonality feature could be captured from the DO series in all the eight sites, the pH series in the Guilin and Jiyuan sites, the COD series in the Chaohu and Suzhou sites, the NH₃-N series in the Hefei, Bengbu, and Jiyuan sites. A seasonal cycle of 52 weeks (one year)

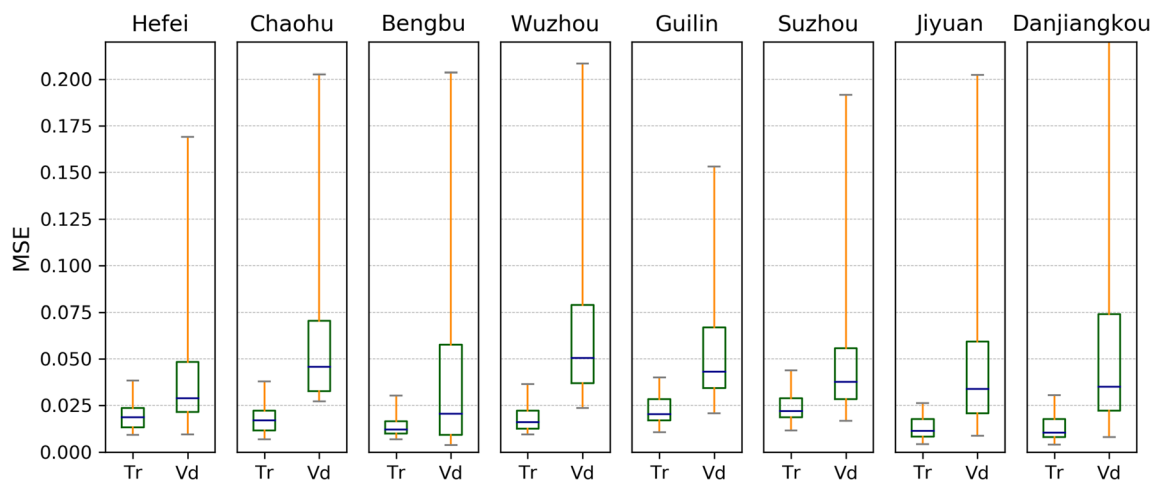


Fig. 5 MSE boxplots of the SARIMA models for the eight sites (captions: lower and upper limit of the whisker refer to the 5th to 95th percentiles of the data, Tr—training datasets, Vd—validation datasets, the same below)

Table 3 The optimal model structure hyperparameters for the SARIMA

Hyperparameter	pH		DO		COD		NH ₃ -N	
	(p, d, q)	(P, D, Q)	(p, d, q)	(P, D, Q)	(p, d, q)	(P, D, Q)	(p, d, q)	(P, D, Q)
Hefei	(0, 0, 2)	(1, 0, 1)	(1, 1, 1)	(2, 0, 0)	(2, 0, 1)	(0, 0, 0)	(2, 1, 1)	(1, 0, 1)
Chaohu	(2, 0, 2)	(1, 0, 1)	(0, 0, 2)	(0, 1, 1)	(2, 1, 2)	(0, 0, 0)	(0, 0, 2)	(1, 1, 0)
Bengbu	(2, 1, 2)	(2, 1, 1)	(0, 0, 2)	(1, 0, 1)	(0, 1, 2)	(2, 0, 0)	(0, 0, 0)	(1, 0, 1)
Wuzhou	(1, 1, 1)	(2, 0, 0)	(0, 1, 2)	(0, 1, 1)	(1, 0, 1)	(0, 1, 1)	(2, 0, 2)	(1, 0, 1)
Guilin	(1, 0, 2)	(2, 0, 1)	(2, 1, 0)	(0, 1, 1)	(0, 1, 0)	(0, 0, 1)	(1, 1, 2)	(1, 0, 1)
Suzhou	(2, 1, 2)	(0, 0, 0)	(0, 0, 2)	(1, 0, 0)	(2, 0, 1)	(1, 0, 0)	(0, 0, 1)	(1, 0, 1)
Jiyuan	(2, 1, 2)	(0, 1, 0)	(1, 1, 1)	(0, 1, 1)	(2, 1, 0)	(2, 1, 0)	(1, 0, 0)	(1, 0, 1)
Danjiangkou	(2, 0, 2)	(1, 0, 1)	(2, 0, 2)	(2, 1, 1)	(2, 0, 2)	(2, 0, 0)	(2, 0, 2)	(0, 0, 1)

was found. Besides, the autocorrelation coefficient declines with the increase of the lag steps, which indicates a tendency. Therefore, the value sets of hyperparameter *d* and *D* were both specified as {0, 1}, and the seasonal period *S* was set as 52.

Figure 4 presents the ACF curves of the series after the 1-order seasonal and 1-order non-seasonal difference. It can be seen that the absolute ACF values of most series are high in the lags 1, 2, and 52, while not statistically significant in the other lags. Therefore, the value sets of hyperparameters *q* and *Q* should be specified as {0, 1, 2} and {0, 1}, respectively. Similarly, according to the PACF curves (see the supplementary document), the value sets of hyperparameters *p* and *P* were specified as {0, 1, 2}.

The SARIMA models were to be developed for the 32 original time series (eight sites with four WQ variables for each). Table 1 lists all the model structure hyperparameters aforementioned and their value sets. In line with the sets, 216 SARIMA models with different model structures (grid method) were constructed for each series. The models were then optimized via the least square algorithm, and the residual series were tested for

independence and normality. The model performance was estimated both on training and validation datasets based on the MSE values.

Development of the ANN models

The model structure hyperparameters of feedforward ANN (see the “ANN” section) should be determined before model training. Based on our previous experiments, the six hyperparameters' value sets were specified in Table 2.

Based on the development protocol proposed by Wu et al. (2014) and the surface water quality time series data, 13,824 surface water quality prediction ANN models with different model structures (grid sampling method) were constructed for each of the eight monitoring sites. The models were trained via Adam algorithm, and the number of training epochs was fixed at 100, which could ensure training convergence of almost all the models. In line with the SARIMA, the MSE values were also calculated both on training and validation datasets to estimate the ANN’s model performance.

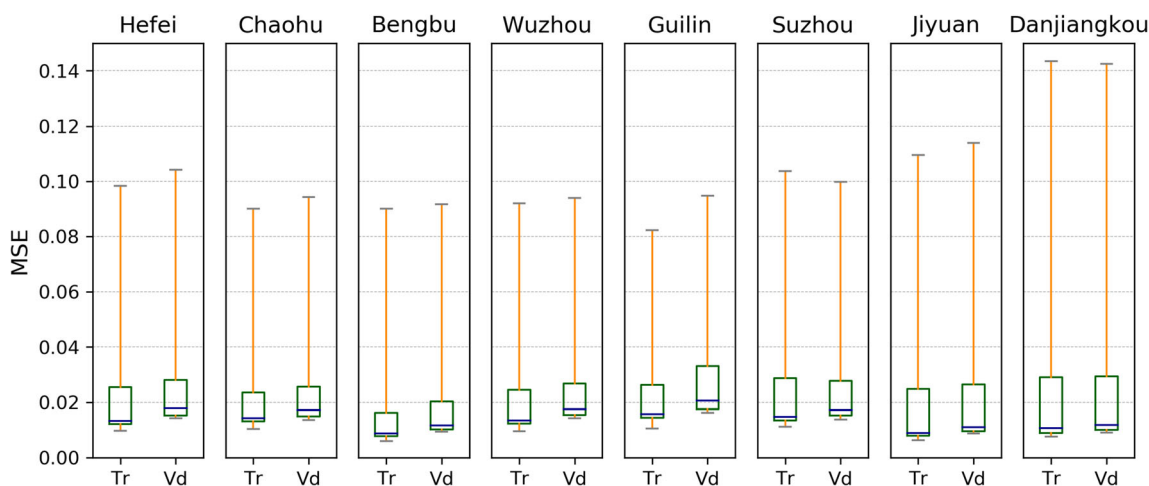


Fig. 6 MSE boxplots of the ANN models for the eight sites

Table 4 The optimal model structure hyperparameters for the ANN

Hyperparameter	Inputs	Layers	Neurons	Activation	lr	Batch size	MSE
Hefei	4	6	32	Tanh	0.001	64	0.013
Chaohu	3	2	36	Tanh	0.001	64	0.013
Bengbu	2	2	28	ReLU	0.001	32	0.009
Wuzhou	2	1	8	Tanh	0.003	4	0.014
Guilin	3	1	24	Tanh	0.001	64	0.015
Suzhou	2	7	8	Tanh	0.001	16	0.013
Jiyuan	3	1	28	ReLU	0.001	16	0.008
Danjiangkou	3	1	8	ReLU	0.001	32	0.008

Results and discussion

Performance assessment of the SARIMA models

Model performance was estimated by MSE both on training and validation datasets. Figure 5 shows the MSE boxplot of the SARIMA models for each of the eight sites. Note that it was the distributions of the average MSE values of the four WQ variables for each site. From the central tendency perspective, significant gaps can be seen between the training median MSEs and the validation ones. Besides, in contrast with the MSEs on the validation sets, the ones on the training sets were in much more concentrated distributions. In summary, the SARIMA models performed significantly worse on validation sets, which indicates the models were overfitted, resulted in poor generalization ability. Furthermore, different prediction performance can be found for the eight sites. From the central tendency’s perspective, the SARIMA models for site Bengbu exerted the best

performance. Otherwise, as to the dispersion, the models for site Hefei performed better than the others (obtained the lowest 75th percentile MSE, 0.0484).

Table 3 shows the optimal SARIMA structures for each of the 32 series selected based on the MSE values on the validation datasets. The seasonal period S identically equals 52. It can be inferred that the seasonal periodicity was significant for the DO series, whereas no necessity for the pH, COD, NH₃-N series to be stabilized by seasonal difference. Furthermore, the NH₃-N series showed a weak tendency, while the tendency feature for the pH, DO, COD series varied with each site.

Performance assessment of the ANN models

Figure 6 shows the MSE boxplot of the ANN models for each of the eight sites. From the perspective of central tendency, gaps between the training median MSEs and the validation ones existed, especially in Hefei and Wuzhou. However, the gaps were much smaller than the SARIMA models. On the other hand, few differences existed between the dispersion feature of the MSE distributions on the training and validation sets. These results suggest that, in contrast with the SARIMA, the ANN models exerted better generalization ability. Furthermore, Fig. 6 illustrates different prediction performance for the eight sites as well. From the central tendency’s perspective, the ANN models for site Bengbu exerted the best performance, consistent with the SARIMA. Otherwise, as to the dispersion, better performance was also found in site Bengbu, which was different from the SARIMA. Performance differences among the sites are mainly related to the original data. For example, apparent autocorrelation and seasonal periodicity could be found in the pH, DO, and NH₃-N series at site Bengbu (see Fig. 3), contributing to the best prediction performance.

Table 4 shows the optimal ANN structures selected based on the MSE values (the average of the four WQ variables) on the validation datasets for each of the eight sites. The results show little similarity among the model structure hyperparameters for the eight sites except for the learning rate.

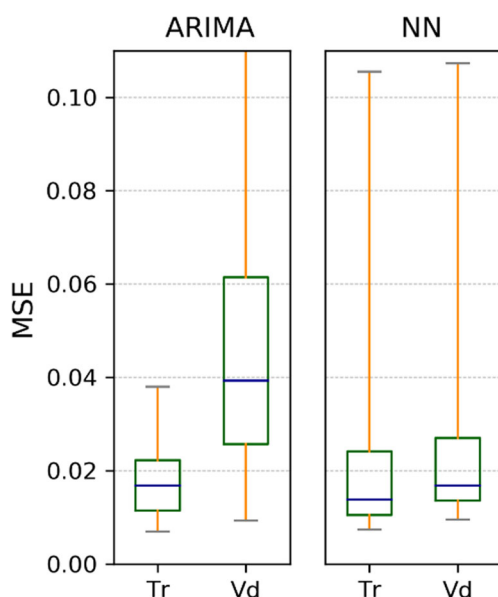
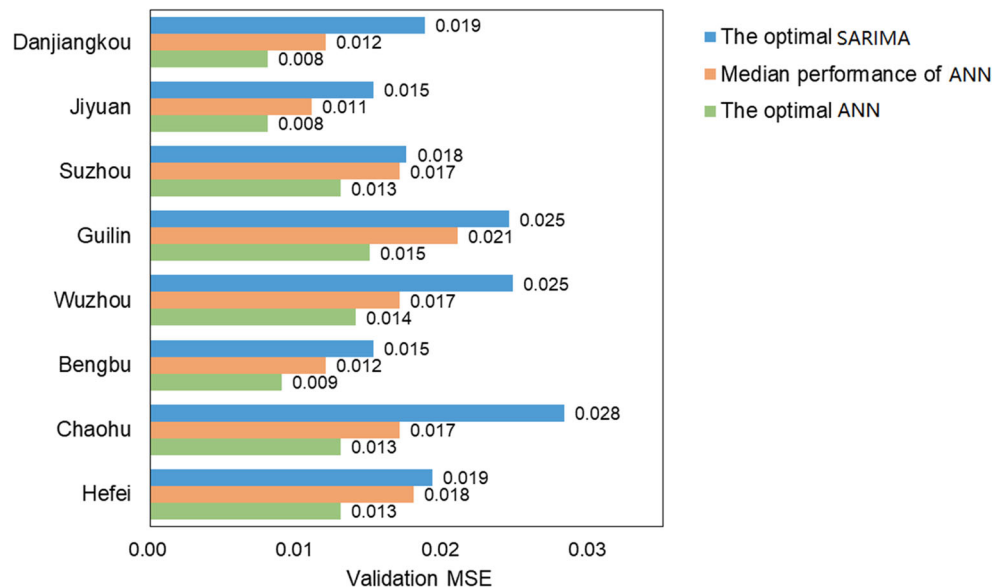


Fig. 7 MSE boxplots for comparison of the ARIMA and ANN models

Fig. 8 Comparison between the optimal ARIMA and ANN models



Statistical comparison between SARIMA and ANN's performance

As mentioned in the “Statistical comparison” section, for the SARIMA models, the average MSEs of the four WQ variables were calculated. Then the MSE distributions of all the eight sites were compared between the SARIMA and ANN models. As shown in Fig. 7, the median training MSE of the ANN models (0.014) was slightly lower than that of the SARIMA models (0.017). Simultaneously, for the validation MSEs, the ANN obtained a significantly lower median value (0.017) than the SARIMA (0.039), which indicates lighter overfitting and better generalization ability of the ANN. Considering dispersion, the training MSEs of the SARIMA were in the most concentrated distribution (5th to 95th), but the distribution of the SARIMA's validation MSEs spread most broadly.

Additionally, the MSEs distributions between the 75th and 95th percentiles of the ANN models were in apparently high dispersion, which stresses the importance of model structure selection. Furthermore, comparisons of the SARIMA and ANN were made in terms of different

sites (see Figs. 5, 6). The results show that the ANN models achieved better performance in all the eight sites, especially for the site Wuzhou, where the median validation MSE of the SARIMA was almost twice higher than the ANN.

In summary, the ANN models statistically performed better than the SARIMA in this surface water quality time series prediction case study.

Optimal performance comparison between SARIMA and ANN

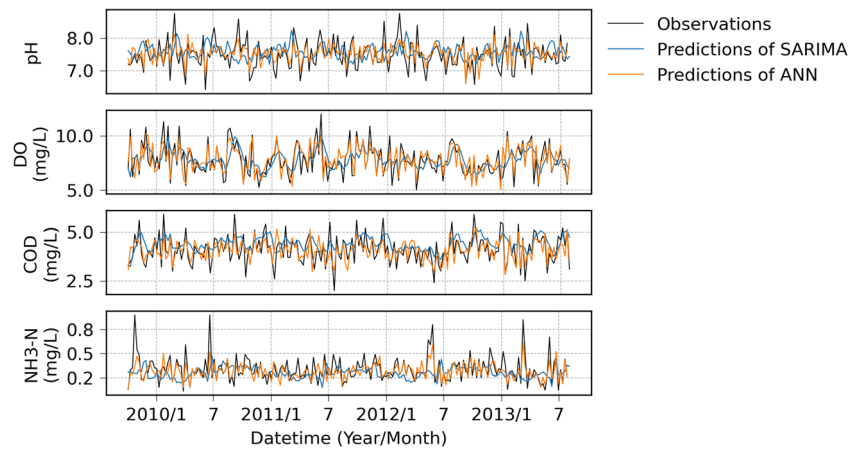
Figure 8 illustrates that for each of the eight sites, the optimal SARIMA's prediction performance on the validation dataset was inferior to the median performance of the ANN, not to mention the optimal ANN. Especially for the sites Danjiangkou, Jiyuan, and Chaohu, where the validation MSE of the optimal SARIMA was around twice as large as the optimal ANN's.

Figures 9 and 10 compare the observed WQ data at Chaohu with the optimal SARIMA and ANN models' predictions. Table 5 lists three performance metrics, root mean square error (RMSE), Nash Sutcliffe efficiency (NSE), and mean absolute percentage error (MAPE), to further evaluate the models. For the calculation formulas of these metrics, please refer to Faruk's work (Ömer Faruk 2010). Good performance of the ANN was shown in the predictions for DO and COD, while the performance for pH and NH₃-N predictions was slightly inferior. Similar phenomena were found for the optimal SARIMA. As Fig. 3 shows, the DO and COD series of site Chaohu are of apparent autocorrelation and seasonal periodicity, leading to better prediction performance compared to pH and NH₃-N.

Table 5 Performance metrics of the optimal SARIMA and ANN at Chaohu

WQ variables	RMSE		NSE		MAPE	
	SARIMA	ANN	SARIMA	ANN	SARIMA	ANN
pH	0.602	0.509	0.355	0.832	12.5	10.0
DO (mg/L)	1.22	0.806	0.545	0.967	12.4	7.82
COD (mg/L)	0.492	0.288	0.556	0.867	5.31	3.02
NH ₃ -N (mg/L)	0.140	0.089	0.394	0.810	30.7	17.7

Fig. 9 WQ variables series predicted by the optimal SARIMA and ANN model at Chaohu



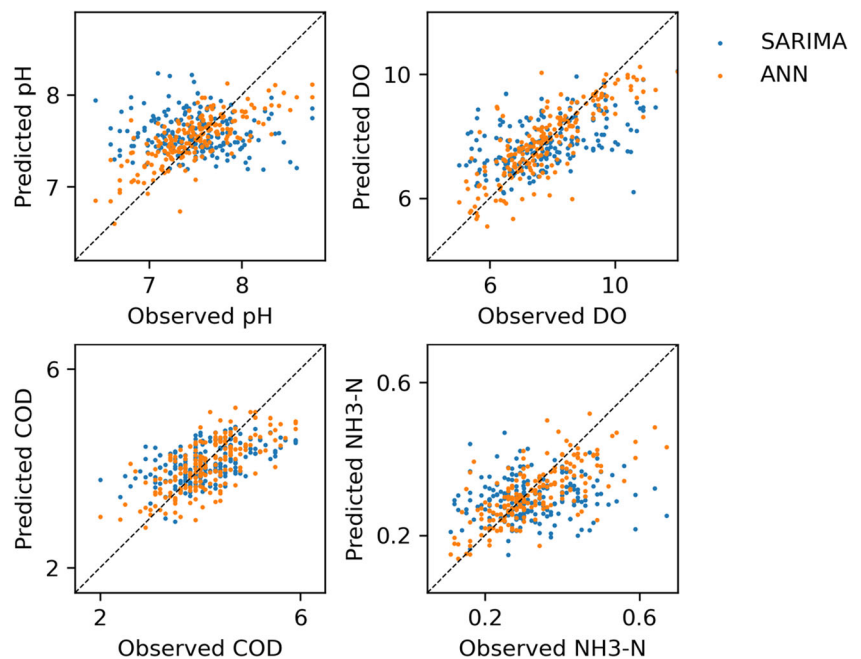
Discussion of the generality

The statistical comparison results could be explained as follows. In principle, ARIMA describes the linear correlations between the system's future and the past state (Edwin and Martins 2014). Therefore, the complex and nonlinear dynamics in water systems could not be described well. ANN has a strong nonlinear mapping ability (García-Alba et al. 2018), so that we believe it can be more qualified than ARIMA for water systems related problems.

However, ANN has shortcomings, of which “poor interpretability” is the leading one (Doshi-Velez and Kim 2017). In contrast with ARIMA, model structure hyperparameters of ANN are not aligned with exact physical meanings. Besides, there are many model structure hyperparameters for ANN to be determined, and the selection space is vast. It can be

concluded from this case study that the selection of the model structure has a significant impact on ANN’s prediction performance (Diez-Sierra and del Jesus 2020). Therefore, the model structure optimization process is essential but challenging, and with high uncertainty (Zhang et al. 2011). As can be inferred from Fig. 7, there were many SARIMA models obtaining lower MSE values than part of the ANN models. In this regard, it is likely to conclude that ARIMA performs better than ANN when the comparison is merely carried out among a couple of selected models, and the ANN model structure is determined via trial-and-error or other ad-hoc ways. In contrast, the statistical comparison in this study is of lower uncertainty and higher replicability. Figure 11 shows the model structure hyperparameters count of the ANN models which performed inferior to the optimal SARIMA for each monitoring site. Similar model structures of ANN that more likely led to inferior performance could be found for the eight sites.

Fig. 10 Observed versus predicted WQ variables of the optimal SARIMA and ANN at Chaohu



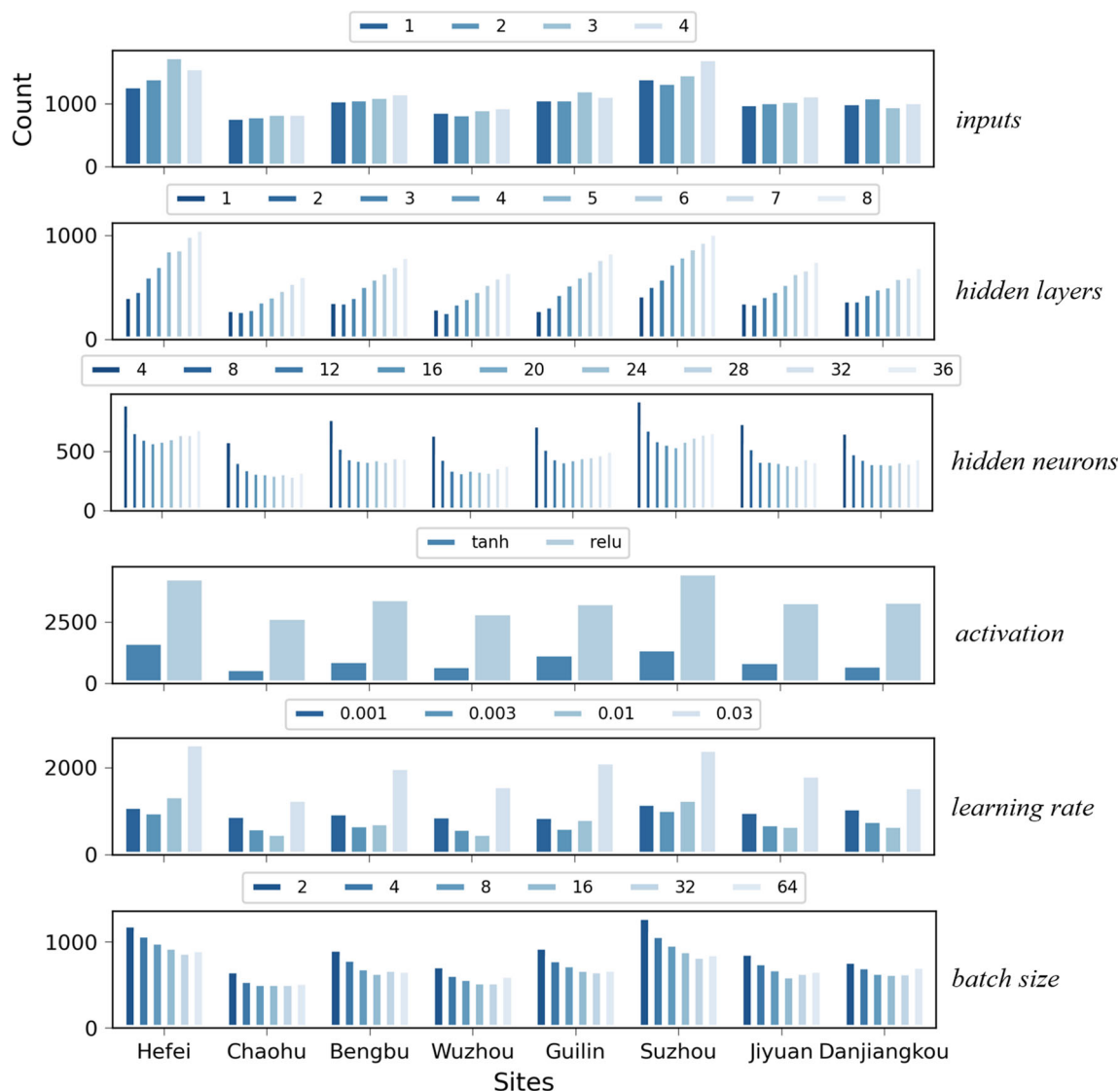


Fig. 11 Model structure hyperparameters count of the ANNs performed inferior to the optimal SARIMA for each site

In addition, we believe the performance (estimated by metrics like MSE) of the optimal ARIMA can be future used as a benchmark to screen out ANN models with better performance, thereby saving computing and memory resources.

Summary and conclusions

Overall, 6912 SARIMA models and 110,592 feedforward ANN models with grid-sampled model structure hyperparameters sets were developed and trained, and the performance metric, MSE, was calculated for each model both on the training and validation datasets. Then the two types of models were compared based on the MSE distributions depicted by boxplots. In a statistical view, the main findings and comparison results of this study are as follows:

- (i) For the SARIMA models, significant gaps existed between the training median MSEs and the validation ones. Besides, the validation MSE distributions were in much higher dispersion. These results indicate the SARIMA models can be easily overfitted and have poor generalization ability.
- (ii) For the ANN models, smaller gaps were found between the training and validation median MSEs. Little differences were revealed between the dispersion feature of the MSE distributions. These results suggest the ANN models exert better performance in generalization ability.
- (iii) In contrast with the SARIMA models, the ANN models obtained a significantly lower median value and a more concentrated distribution of validation MSEs, which indicates the ANN models statistically performed better in this surface water quality time series prediction case study.

- (iv) The optimal SARIMA models' prediction performance is inferior to the median of the ANN models, not to mention the optimal ones.
- (v) In contrast with the previous comparison studies among a couple of selected models, the statistical comparison in this study is of lower uncertainty.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11356-021-13086-3>.

Acknowledgements This study was financially supported by the Major Science and Technology Project of Water Pollution Control and Management in China (grant no. 2018ZX07208006) and the National Natural Science Foundation of China (grant no. 51778451). We also thank the 111 Project (B13017) of Tongji University.

Author contribution Xuan Wang: conceptualization, methodology, software, data curation, writing—original draft preparation.

Wenchong Tian: methodology, investigation, writing—reviewing, and editing.

Zhenliang Liao: supervision, writing—reviewing, and editing.

Funding Major Science and Technology Project of Water Pollution Control and Management in China (grant no. 2018ZX07208006). National Natural Science Foundation of China (grant no. 51778451). 111 Project (B13017) of Tongji University.

Data availability The datasets and codes are available in our GitHub repository: <https://github.com/MrBrenda/WaterResourcesFNNModels.git>.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare that they have no competing interests.

References

- Ahmad S, Khan IH, Parida B (2001) Performance of stochastic approaches for forecasting river water quality. *Water Res* 35:4261–4266. [https://doi.org/10.1016/S0043-1354\(01\)00167-1](https://doi.org/10.1016/S0043-1354(01)00167-1)
- Ansari M, Othman F, Abunama T, El-Shafie A (2018) Analysing the accuracy of machine learning techniques to develop an integrated influent time series model: case study of a sewage treatment plant, Malaysia. *Environ Sci Pollut Res* 25:12139–12149. <https://doi.org/10.1007/s11356-018-1438-z>
- Bhagat SK, Tung TM, Yaseen ZM (2020) Development of artificial intelligence for modeling wastewater heavy metal removal: State of the art, application assessment and possible future research. *J Clean Prod* 250:119473. <https://doi.org/10.1016/j.jclepro.2019.119473>
- Bhagat SK, Tiyyasha T, Awadh SM, Tung TM, Jawad AH, Yaseen ZM (2021) Prediction of sediment heavy metal at the Australian Bays using newly developed hybrid artificial intelligence models. *Environ Pollut* 268:115663. <https://doi.org/10.1016/j.envpol.2020.115663>
- Box GE, Jenkins GM (1976) *Time series analysis: forecasting and control*, vol 31, third edn. Holden Day, Oakland, p 303
- Diez-Sierra J, del Jesus M (2020) Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods. *J Hydrol* 586:124789. <https://doi.org/10.1016/j.jhydrol.2020.124789>
- Doshi-Velez F, Kim B (2017) Towards A Rigorous Science of Interpretable Machine Learning 1–13. <https://arxiv.org/abs/1702.08608v2>.
- Edwin AI, Martins OY (2014) Stochastic Characteristics and Modelling of Monthly Rainfall Time Series of Ilorin, Nigeria. *Open J Mod Hydrol* 04:67–79. <https://doi.org/10.4236/ojmh.2014.43006>
- Elkiran G, Nourani V, Abba SI (2019) Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. *J Hydrol* 577:123962. <https://doi.org/10.1016/j.jhydrol.2019.123962>
- García Nieto PJ, García-Gonzalo E, Alonso Fernández JR, Díaz Muñoz C (2019) Water eutrophication assessment relied on various machine learning techniques: A case study in the Englishmen Lake (Northern Spain). *Ecol Model* 404:91–102. <https://doi.org/10.1016/j.ecolmodel.2019.03.009>
- García-Alba J, Bárcena JF, Ugarteburu C, García A (2018) Artificial neural networks as emulators of process-based models to analyse bathing water quality in estuaries. *Water Res* 150:283–295. <https://doi.org/10.1016/j.watres.2018.11.063>
- Haghiabi AH, Nasrolahi AH, Parsaie A (2018) Water quality prediction using machine learning methods. *Water Qual Res J Can* 53:3–13. <https://doi.org/10.2166/wqrj.2018.025>
- Hameed M, Sharqi SS, Yaseen ZM, Afan HA, Hussain A, Elshafie A (2017) Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. *Neural Comput & Applic* 28:893–905. <https://doi.org/10.1007/s00521-016-2404-7>
- Hanson PC, Stillman AB, Jia X, Karpatne A, Dugan HA, Carey CC, Stachelek J, Ward NK, Zhang Y, Read JS, Kumar V (2020) Predicting lake surface water phosphorus dynamics using process-guided machine learning. *Ecol Model* 430:109136. <https://doi.org/10.1016/j.ecolmodel.2020.109136>
- Hunter JM, Maier HR, Gibbs MS, Foale ER, Grosvenor NA, Harders NP, Kikuchi-Miller TC (2018) Framework for developing hybrid process-driven, artificial neural network and regression models for salinity prediction in river systems. *Hydrol Earth Syst Sci* 22:2987–3006. <https://doi.org/10.5194/hess-22-2987-2018>
- Kang G, Gao JZ, Xie G (2017) Data-driven water quality analysis and prediction: A survey. *Proc - 3rd IEEE Int Conf Big Data Comput Serv Appl BigDataService 2017* 224–232. <https://doi.org/10.1109/BigDataService.2017.40>
- Khairuddin N, Aris AZ, Elshafie A, Sheikhy Narany T, Ishak MY, Isa NM (2019) Efficient forecasting model technique for river stream flow in tropical environment. *Urban Water J* 16:1–10. <https://doi.org/10.1080/1573062x.2019.1637906>
- Landeras G, Ortiz-Barredo A, López JJ (2009) Forecasting weekly evapotranspiration with ARIMA and artificial neural network models. *J Irrig Drain Eng* 135:323–334. [https://doi.org/10.1061/\(ASCE\)IR.1943-4774.0000008](https://doi.org/10.1061/(ASCE)IR.1943-4774.0000008)
- Maier HR, Jain A, Dandy GC, Sudheer KP (2010) Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environ Model Softw* 25:891–909. <https://doi.org/10.1016/j.envsoft.2010.02.003>
- Monteiro M, Costa M (2018) A Time Series Model Comparison for Monitoring and Forecasting Water Quality Variables. *Hydrology* 5:37. <https://doi.org/10.3390/hydrology5030037>
- Mount NJ, Maier HR, Toth E, Elshorbagy A, Solomatine D, Chang FJ, Abrahart RJ (2016) Data-driven modelling approaches for socio-hydrology: Opportunities and challenges within the Panta Rhei

- Science Plan. *Hydrol Sci J* 61:1192–1208. <https://doi.org/10.1080/02626667.2016.1159683>
- Ömer Faruk D (2010) A hybrid neural network and ARIMA model for water quality time series prediction. *Eng Appl Artif Intell* 23:586–594. <https://doi.org/10.1016/J.ENGAPPAL.2009.09.015>
- Rafael A, Parmezan S, Souza VMA, Batista GEAPA (2019) Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Inf Sci* 484:302–337. <https://doi.org/10.1016/j.ins.2019.01.076>
- Raman H, Sunilkumar N (1995) Multivariate modelling of water resources time series using artificial neural networks. *Hydrol Sci J* 40:145–163. <https://doi.org/10.1080/02626669509491401>
- Salmani MH, Salmani Jajaei E (2016) Forecasting models for flow and total dissolved solids in Karoun river-Iran. *J Hydrol* 535:148–159. <https://doi.org/10.1016/J.JHYDROL.2016.01.085>
- Schmidhuber J (2015) Deep learning in neural networks: An overview. *Neural Netw* 61:85–117. <https://doi.org/10.1016/J.NEUNET.2014.09.003>
- Sheikhy Narany T, Aris AZ, Sefie A, Keesstra S (2017) Detecting and predicting the impact of land use changes on groundwater quality, a case study in Northern Kelantan, Malaysia. *Sci Total Environ* 599–600:844–853. <https://doi.org/10.1016/J.SCITOTENV.2017.04.171>
- Shi B, Wang P, Jiang J, Liu R (2018) Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies. *Sci Total Environ* 610–611:1390–1399. <https://doi.org/10.1016/j.scitotenv.2017.08.232>
- Tiyasha, Tung TM, Yaseen ZM (2020) A survey on river water quality modelling using artificial intelligence models: 2000–2020. *J Hydrol* 585:124670. <https://doi.org/10.1016/j.jhydrol.2020.124670>
- Valipour M, Banihabib ME, Behbahani SMR (2013) Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *J Hydrol* 476:433–441. <https://doi.org/10.1016/J.JHYDROL.2012.11.017>
- Wu W, May RJ, Maier HR, Dandy GC (2013) A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks. *Water Resour Res* 49:7598–7614. <https://doi.org/10.1002/2012WR012713>
- Wu W, Dandy GC, Maier HR (2014) Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. *Environ Model Softw* 54:108–127. <https://doi.org/10.1016/j.envsoft.2013.12.016>
- Zhang X, Liang F, Yu B, Zong Z (2011) Explicitly integrating parameter, input, and structure uncertainties into Bayesian Neural Networks for probabilistic hydrologic forecasting. *J Hydrol* 409:696–709. <https://doi.org/10.1016/j.jhydrol.2011.09.002>
- Zhou J, Wang Y, Xiao F, Wang Y, Sun L (2018) Water Quality Prediction Method Based on IGRA and LSTM. *Water* 10:1148. <https://doi.org/10.3390/w10091148>

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