Journal of Hydroinformatics

Journal of Hydroinformatics Vol 25 No 1, 140 doi: 10.2166/hydro.2023.127

Real-time water quality detection based on fluctuation feature analysis with the LSTM model

Lixiang Wang†, Huilin Dong†, Yuqi Cao, Dibo Hou* and Guangxin Zhang

State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
*Corresponding author. E-mail: houdb@zju.edu.cn
†These authors contributed equally.

ABSTRACT

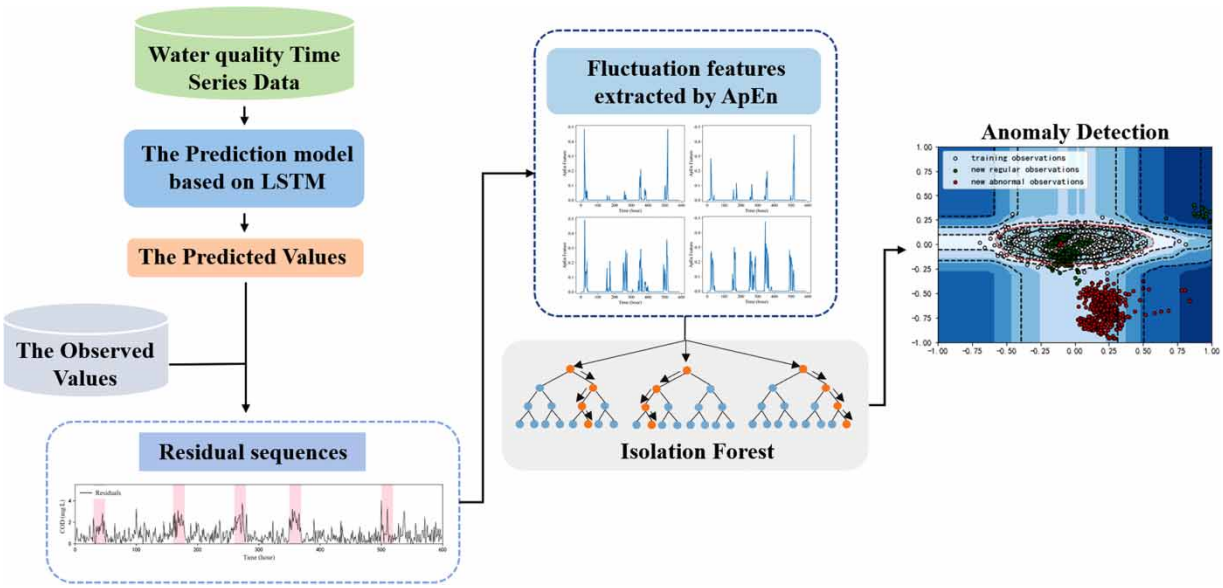
Signal analysis and anomaly detection for water pollution early warning systems are important and necessary. In view of the nonlinear and volatile characteristics of water quality time series, this paper proposes a new method for water anomaly detection based on fluctuation feature analysis. The method has two steps. First, the water quality time series data are used to calculate the residuals between the observed value and the predicted value with the long short-term memory (LSTM) network. Second, the dynamic features are extracted by sliding time window and described by the Approximate Entropy (ApEn) which are input to the anomaly detection model with Isolation Forest. Compared with traditional anomaly detection methods, the results obtained by the proposed method show better performance in distinguishing water quality anomalies. The proposed method can be applied to real-time water quality anomaly detection and early warning.

Key words: anomaly detection, feature extraction, LSTM networks, water time series prediction

HIGHLIGHTS

- A prediction model based on LSTM networks is constructed to predict six water quality indicators.
- Dynamic features of water time series are extracted by the Approximate Entropy (ApEn).
- Combining with the high-dimensional ApEn characteristics, the Isolation Forest method is applied to identify anomalies of water quality.
- This research has the potential for the improvement of water quality early warning system.

GRAPHICAL ABSTRACT



INTRODUCTION

Water is an essential resource for living organisms and human society. However, with the accelerated development of industrialization and urbanization, river pollution incidents have increasingly occurred in the last decades, damaging the ecological environments and endangering the safety of people's lives and property (Zhu *et al.* 2018; Liu *et al.* 2020). Thus, anomaly detection of water quality is crucial to protect the ecological environment and human beings.

Conventional water quality indicators, such as total phosphorus (TP), total nitrogen (TN) and chemical oxygen demand (COD), are the most direct tools for water quality monitoring, which are mainly collected chronologically in the form of time series at a fixed time interval (Shi *et al.* 2018; Tinelli & Juran 2019; Jiang *et al.* 2020). Using these conventional water quality indicators, many water quality anomaly detection approaches have been proposed (Byer & Carlson 2005; Koch & McKenna 2010; Arad *et al.* 2012; Olsen *et al.* 2012; Perelman *et al.* 2012; Wechmongkhonkon *et al.* 2012; Zhang *et al.* 2014; Azhar *et al.* 2015), including threshold methods (Byer & Carlson 2005; Zhang *et al.* 2014), statistical analysis models (Koch & McKenna 2010; Perelman *et al.* 2012) and artificial intelligence methods (Arad *et al.* 2012; Olsen *et al.* 2012; Wechmongkhonkon *et al.* 2012; Azhar *et al.* 2015), such as clustering classification (Azhar *et al.* 2015), time series analysis and artificial neural network (ANN) algorithm (Wechmongkhonkon *et al.* 2012). However, most threshold methods and statistical analysis methods with a single water quality indicator are difficult to describe the fluctuation of water quality. Thus, it is necessary to propose an ideal method, which can capture the fluctuation characteristics of water quality and can be applied to real-time water quality anomaly detection.

With the development of water quality monitoring technologies, more water quality indicators can be obtained and used in water quality prediction and anomaly detection. Now the relevant research teams introduce different methods (Bouamar & Ladjal 2007; Durdu 2010; Bao & Meng 2015; Wang *et al.* 2016; Wang *et al.* 2019; Baek *et al.* 2020; Liu *et al.* 2020) to improve the accuracy of the water quality prediction model and anomaly detection algorithm, such as data assimilation, machine learning and ANNs. Durdu (2010) proposed a hybrid model combining a neural network and autoregressive moving average (ARMA) sequence for water quality prediction, which achieved better performance than a single model. Bouamar & Ladjal (2007) utilized ANNs and support vector machines (SVMs) to classify water quality data into normal and anomalous groups. According to strong water quality fluctuation, Bao & Meng (2015) proposed a wavelet-based method to extract the features of the water quality index under different scales, and realized feature recognition using energy spectrum analysis, to implement water quality anomaly detection. Baek *et al.* (2020) used long short-term memory (LSTM) networks accurately simulate the water quality including TP, TN and total organic carbon (TOC). Liu *et al.* (2020) used Bayesian autoregressive (BAR) model for water quality variation prediction and Isolation Forest (IF) algorithm for water quality anomaly detection. However, these methods less consider the trend of changes and the dependence properties in long-term time series, and rarely explore the correlation among the multidimensional features, resulting in low accuracy of long-term prediction of water quality (Ahmed *et al.* 2019).

Therefore, it is a challenge to predict the changes in water quality indicators and extract the fluctuation features of water quality accurately over a long period.

In this paper, we have proposed an unsupervised method for real-time water quality anomaly detection, combing six kinds of water quality indicators and using the Approximate Entropy (ApEn) to reduce the impact of normal water quality fluctuations. This method had two steps. At first, the missing data were filled by the exponential smoothing method. A water quality prediction model with LSTM networks was constructed to predict the water quality indicators, in order to mine the relationship between these indicators and the dynamic changes of water quality data in time series. Second, the ApEn was calculated to extract the fluctuation characteristics of water quality, which represented the water quality prediction residuals. The extracted features would be the input to the IF algorithm to identify water quality anomalies. The flowchart of this proposed method is shown in Figure 1. This work can be applied to anomaly detection and early warning in the process of online monitoring of urban river water quality.

MATERIALS AND METHODS

Dataset

Out of practical needs and the analysis of common pollution sources in urban rivers, according to water quality standards such as environmental quality standards for surface water (Zhang *et al.* 2020), sanitary standards for drinking water (Zhang *et al.* 2022) and technical specifications for surface water and sewage monitoring (Qi *et al.* 2006), we had chosen

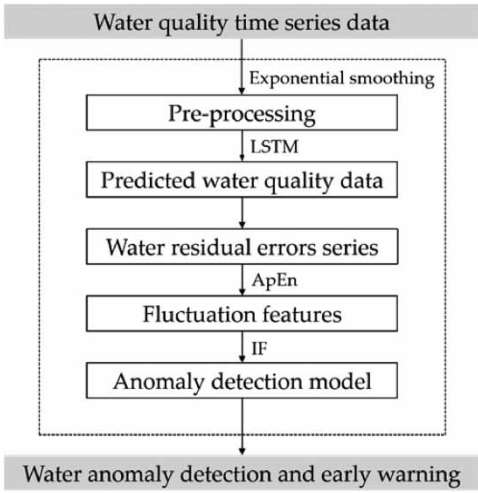


Figure 1 | Flowchart of the water quality anomaly detection method.

six conventional water quality indicators, including dissolved oxygen (DO), NH_3-N, electrical conductivity (Cond), pH, COD and turbidity.

In this paper, the normal water quality time series data were measured each hour from the water quality monitoring stations deployed in an urban river in south China. The obtained dataset comprised 3,000 continuous data points in total. Each data point contained the measurements of six types of conventional water quality indicators, including DO, NH_3-N, Cond, pH, COD and turbidity.

In the water quality prediction stage, all normal water quality time series data were used to construct the prediction models. Also, 85% of the normal water quality time series data were used for training and 15% were used for testing.

In the water anomaly detection stage, only the last 600 data points were used as the observed values. In order to verify the performance of the detection algorithm, the Gaussian inverted U-shaped anomalies were superimposed artificially on the normal observed values to simulate different strengths of water pollution events. The time periods of adding anomaly events were (30, 60), (150, 180), (250, 270), (350, 370) and (500, 520), respectively.

Water quality prediction model based on LSTM networks

When a water pollution incident occurred, the related water quality indicators would fluctuate and differ from the normal level. To accurately capture the fluctuation and change of water quality indicators, we used the LSTM network for water quality prediction in this paper, combining six water quality indicators which included DO, NH_3-N, pH, electrical conductivity, COD and turbidity.

The monitoring data of water quality indicators were arranged in chronological order, and the sliding window structure could be used as the input of the LSTM model for multivariate time series prediction.

Assuming that the time series $\boldsymbol{Y} = (y_1, \cdots, y_{T-1}, y_T) \in \boldsymbol{R}^T$ of any water quality indicators was the prediction target, the time series matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T) = (\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^M)^T$, using the observation data of the multiple water quality parameters at the historical time T as the relevant characteristic variables, could be expanded as follows:

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^M \\ \vdots & \vdots & \vdots & \vdots \\ x_{T-1}^1 & x_{T-1}^2 & \vdots & x_{T-1}^M \\ x_T^1 & x_T^2 & \cdots & x_T^M \end{bmatrix} \in R^{T \times M} \tag{1}$$

where M referred to the dimension of the water quality parameters, $\boldsymbol{x}^m = (x_1^m, \cdots, x_{T-1}^m, x_T^m)$ was the time series of the mth dimensional water quality parameter at the historical time T, then x_i^m represented the observed value of the mth variable at time i.

In the LSTM model, we used the sliding time window of size $\boldsymbol{d}$ and the sequence $\boldsymbol{x}(t) = \boldsymbol{x}_{t-d+1}, \ldots, \boldsymbol{x}_{t-1}, \boldsymbol{x}_t$ as the input, then the predicted value $\boldsymbol{y}_{t+n}$ of any water quality indicators could be predicted by the following iterative process in Equation (2).

$$\begin{aligned} y_t = [x_t^m] &= F(\boldsymbol{x}_{t-d} \cdots, \boldsymbol{x}_{t-2}, \boldsymbol{x}_{t-1}) \\ y_{t+1} = [x_{t+1}^m] &= F(\boldsymbol{x}_{t-d+1} \cdots, \boldsymbol{x}_{t-1}, \boldsymbol{x}_t) \\ &\vdots \\ y_{t+n} = [x_{t+n}^m] &= F(\boldsymbol{x}_{t-d+n} \cdots, \boldsymbol{x}_{t+n-2}, \boldsymbol{x}_{t+n-1}) \end{aligned} \tag{2}$$

When the prediction model based on LSTM was trained, the observed water quality time series data could be input into the prediction model to predict the current water quality indicators, and then the deviation between the observed value and the predicted value at the current time was calculated to obtain a residual vector group.

Feature extraction method with ApEn

ApEn, combining the advantages of good anti-interference ability of information entropy in signal processing, was suitable for extracting the statistical complexity characteristics of irregular nonlinear signals, such as water quality time series data (Huang *et al.* 2017; Ma *et al.* 2019). Because the entropy amplitude was greatly different between the normal and abnormal water quality, the ApEn was used as a feature to detect water quality anomalies in this paper. The ApEn of water quality time series data could be calculated by the following steps:

- The reconstruction subsequence X_i was defined in Equation (3).

$$X_i = [u(i), u(i+1), \cdots, u(i+m-1)], i = 1 \sim N - m + 1 \tag{3}$$

where $u(i)$ was the one-dimensional time series, N was the length of time series, m was the size of the sliding time window.

- The distance d between the vector X_i and X_j was calculated in Equation (4).

$$d(X_i, X_j) = \max_{k=0\sim m-1}^{i,j=1\sim N-m+1} (|u(i+k) - u(j+k)|) \tag{4}$$

- Given a threshold r, counted the number of $d < r$ and the ratio of it to the total number of vectors was noted in Equation (5).

$$C_i^m(r) = \frac{num[d(X_i, X_j) < r]}{N - m + 1} \tag{5}$$

- Took the logarithm of $C_i^m(r)$, and the average $\Phi^m(r)$ of it was expressed in Equation (6).

$$\Phi^m(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \ln C_i^m(r) \tag{6}$$

- For sliding time window size $m + 1$, repeated the steps above to get $\Phi^{m+1}(r)$.
- The ApEn of the sequences of length N was estimated in Equation (7).

$$ApEn(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r) \tag{7}$$

Based on empirical values, took $m = 2$ and $r = 0.25\mathrm{SD}(u)$ in this paper, where $\mathrm{SD}(u)$ represents the standard deviation of original water quality time series.

According to the calculation principle of the ApEn value, after selecting different subsequences from the original data sequence by using the sliding window, we could calculate the ApEn of each subsequence and analyze the abnormal volatility of the data by the fluctuation in the ApEn values.

IF algorithm for anomaly detection

The construction process of IF (Liu *et al.* 2008) could be divided into the following steps:

- Constructed each iTree based on sampling without replacement to improve the diversity among iTrees.
- Set the depth of iTrees and the stop condition when construct an iTree from samples.
- Calculated the weighted path length of multiple iTrees for real-time anomaly detection. When there were abnormal fluctuations in multidimensional water quality characteristics, the weighted path length would be shorter.
- Used ensemble learning to carry out convergence calculation of the fused results of multiple iTrees. Details of calculation could be found in Equations (8) and (9).

$$c(n) = 2H(n-1) - 2\frac{n-1}{n} \tag{8}$$

$$s(x, n) = 2^{\left(-\frac{E[h(x)]}{c(n)}\right)} \tag{9}$$

where n was the number of given samples, $H(n-1)$ was a harmonic number, $c(n)$ was the average path length of failed search in binary tree, $h(x)$ was the standardized path length of any sample, $s(x, n)$ was the anomaly score.

- The LSTM network could be used for high-precision and dynamical prediction of the time series of water quality. Then, the IF algorithm was proposed to detect the abnormal water quality by using ApEn to select anomalous fluctuation characteristics.

The combination of the prediction-based and isolation-based method could achieve water quality anomaly detection and early warning.

RESULTS AND DISCUSSION

Water quality prediction

Take the prediction results of COD and DO for examples, we had compared and analyzed the performance of the LSTM model, back propagation (BP) neural network model, recurrent neural network (RNN) model and gate recurrent unit (GRU) model.

In BP prediction model, the number of hidden nodes was set to 50, the optimizer was set as 'adam' and the MSE was used as the loss function. In the RNN prediction model, the unit was set to 64, the optimizer was set as Root Mean Square Propagation (RMSprop) and the MSE was used as the loss function. In GRU prediction model, the unit was set to 50, the optimizer was set as 'adam' and the MSE was used as the loss function.

The accuracy and effectiveness of different prediction models were evaluated by root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and R^2, as shown in Figure 2 and Table 1.

The red columns in Figure 3 represent the time periods during which the water quality anomaly events occur. As we could see in Figure 3, the prediction model based on LSTM had a good tracking effect on the water quality background signal, with the residual sequence retaining the water quality anomalies. At the same time, it could reduce the impact of non-steady

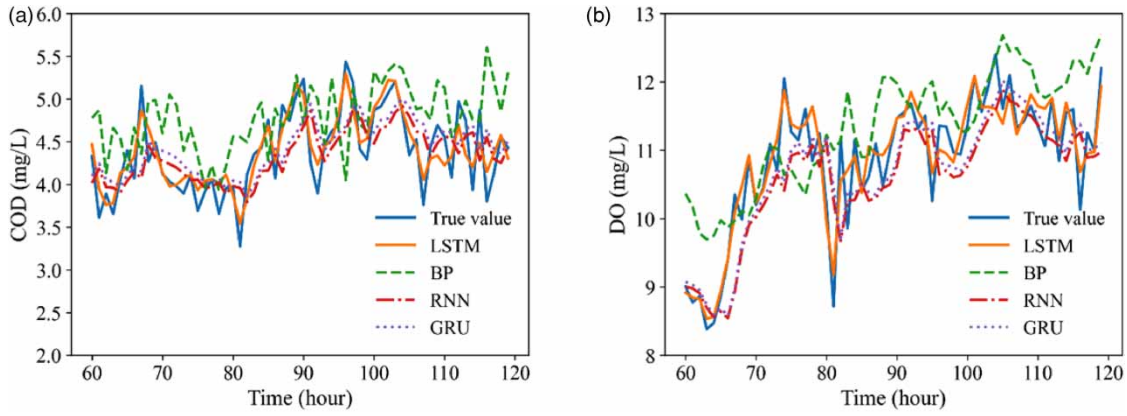


Figure 2 | Prediction results on different models. (a) Predicted results of COD on a LSTM model, BP model, RNN model and GRU model and (b) predicted results of DO on a LSTM model, BP model, RNN model and GRU model.

Table 1 | Prediction accuracy on different models

Model	RMSE	MAE	MAPE (%)	R^2
BP	0.8197	0.6727	10.7912	0.7373
RNN	0.5880	0.4271	6.4838	0.9510
GRU	0.5713	0.4163	6.2606	0.9513
LSTM	**0.2842**	**0.2169**	**3.1484**	**0.9878**

Bold entries emphasize that LSTM model performs better than others.

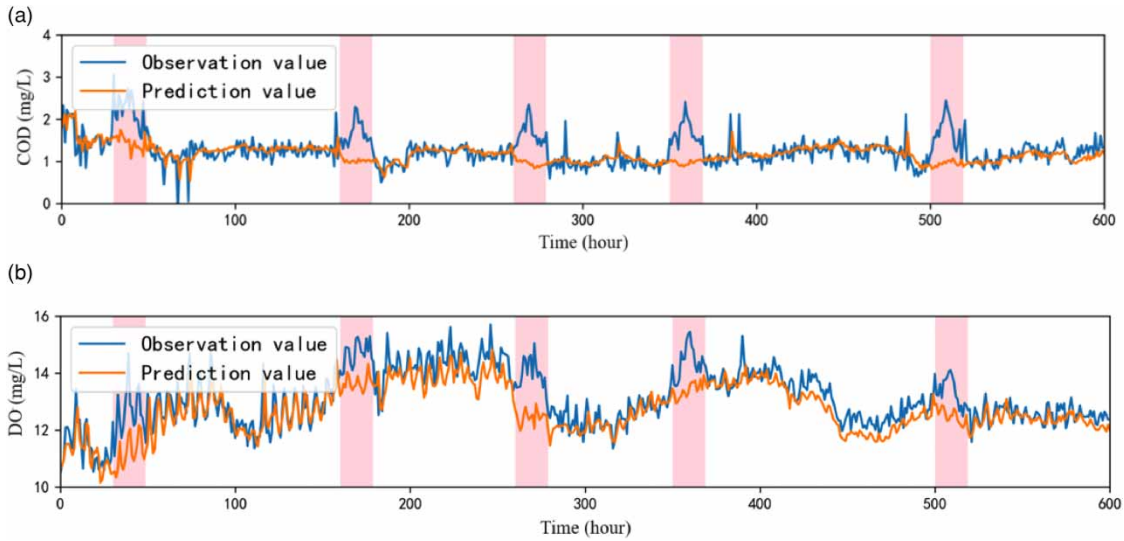


Figure 3 | Results of predicting COD and DO on the LSTM model under the anomaly event strength of 1.5. (a) Predicted results of COD on the LSTM model and (b) predicted results of DO on the LSTM model. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/hydro.2023.127.

fluctuations in water quality and the significant difference between the predicted values and the observed values indicated the moment at which an abnormal water quality incident might occur.

Anomaly detection

In this paper, the performance of anomaly detection results based on different inputs of IF algorithm was evaluated by the precision, recall, F1-score, the receiver operating characteristic (ROC) curve and its area under curve (AUC) value. As shown in Figure 4, the ApEn had separated abnormal water quality from the normal to the maximum extent and there was a common feature in the ApEn detection results under four different event strengths. That was, after adding the anomalies, the ApEn value of the water residual when $T > 30$ was significantly larger than that when $T < 30$, which indicated that water quality fluctuation could be divided into two different evolutionary stages with $T = 30$ as the boundary. A larger value of ApEn meant a more complex sequence. Therefore, it could be judged that during the time period from 30 to 50, water pollution events were highly likely to occur as water quality fluctuates abnormally.

However, the change in the statistical value increased slowly, when water pollution events occurred. So, it was difficult to distinguish abnormal water quality from normal water quality only by the ApEn values, especially in the early stage. Therefore, we used IF for anomaly detection after extracting the fluctuation characteristics of water quality with ApEn. It could be seen in Figure 5 that the isolated forest constructed the boundary between the normal features and the abnormal features. Intuitively, the data could be divided into two categories: within and outside the abnormal decision boundary. The detection effect was relatively stable in normal conditions where the normal water quality samples were all clustered within the decision-making boundary. However, when an anomaly occurred, the anomaly samples were presented outside the decision-making boundary in a global outlier manner.

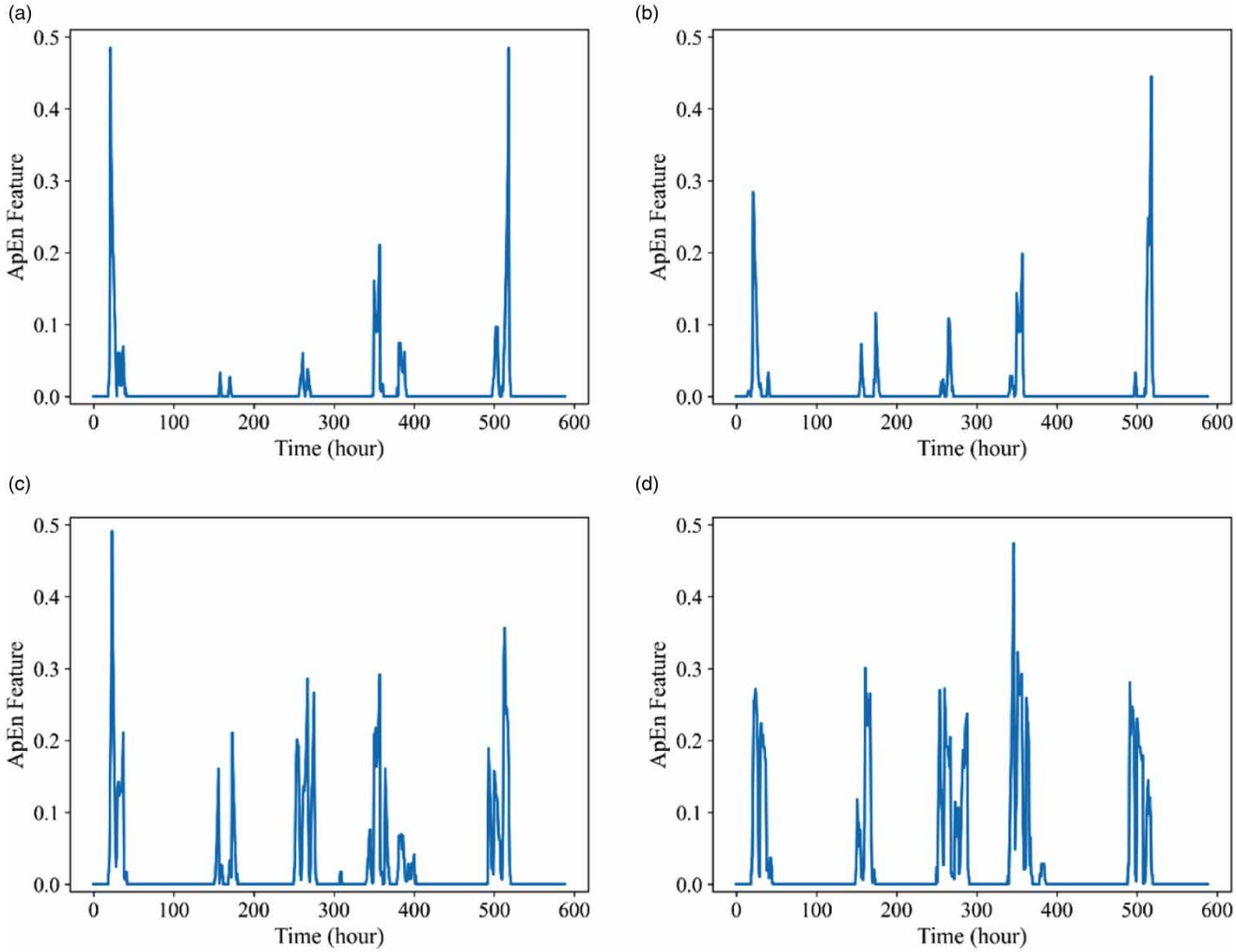


Figure 4 | ApEn features of abnormal water quality time series data under different anomaly event strengths. (a) ApEn features of COD under anomaly event strength of 0.8; (b) ApEn features of COD under anomaly event strength of 1.0; (c) ApEn features of COD under anomaly event strength of 1.5 and (d) ApEn features of COD under anomaly event strength of 2.0.

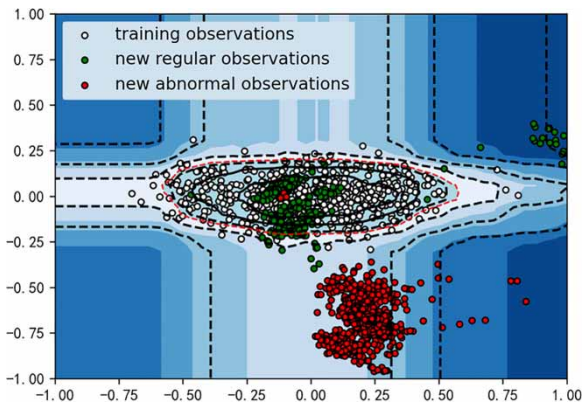


Figure 5 | Anomaly detection result under event strength of 2.0, using the ApEn feature as the input of the IF algorithm.

To compare the performances of the anomaly detection method under different anomaly event strengths, we used the original abnormal water quality data, the residuals obtained by the LSTM model and the ApEn feature as the input of IF. The results are shown in Figure 6 and Table 2. We could learn from the results that when there were abnormal fluctuations in the water quality time series, the proposed method performed better than other methods under different event strengths. It was because that the residual method could give the statistical average of the signal's entropy but usually did not consider the localized characteristics or small fluctuations of the water quality signal, while the ApEn feature could reflect the frequency of the signal change and the mutation time accurately. Therefore, the IF algorithm, combining high-dimensional ApEn characteristics of multiple water quality principal components, could identify anomalies of water quality in an early stage.

CONCLUSIONS

This study presented a method for water anomaly detection based on the entropy features and water prediction model, which was effective in early warning of water anomaly. The results and analysis demonstrate the following:

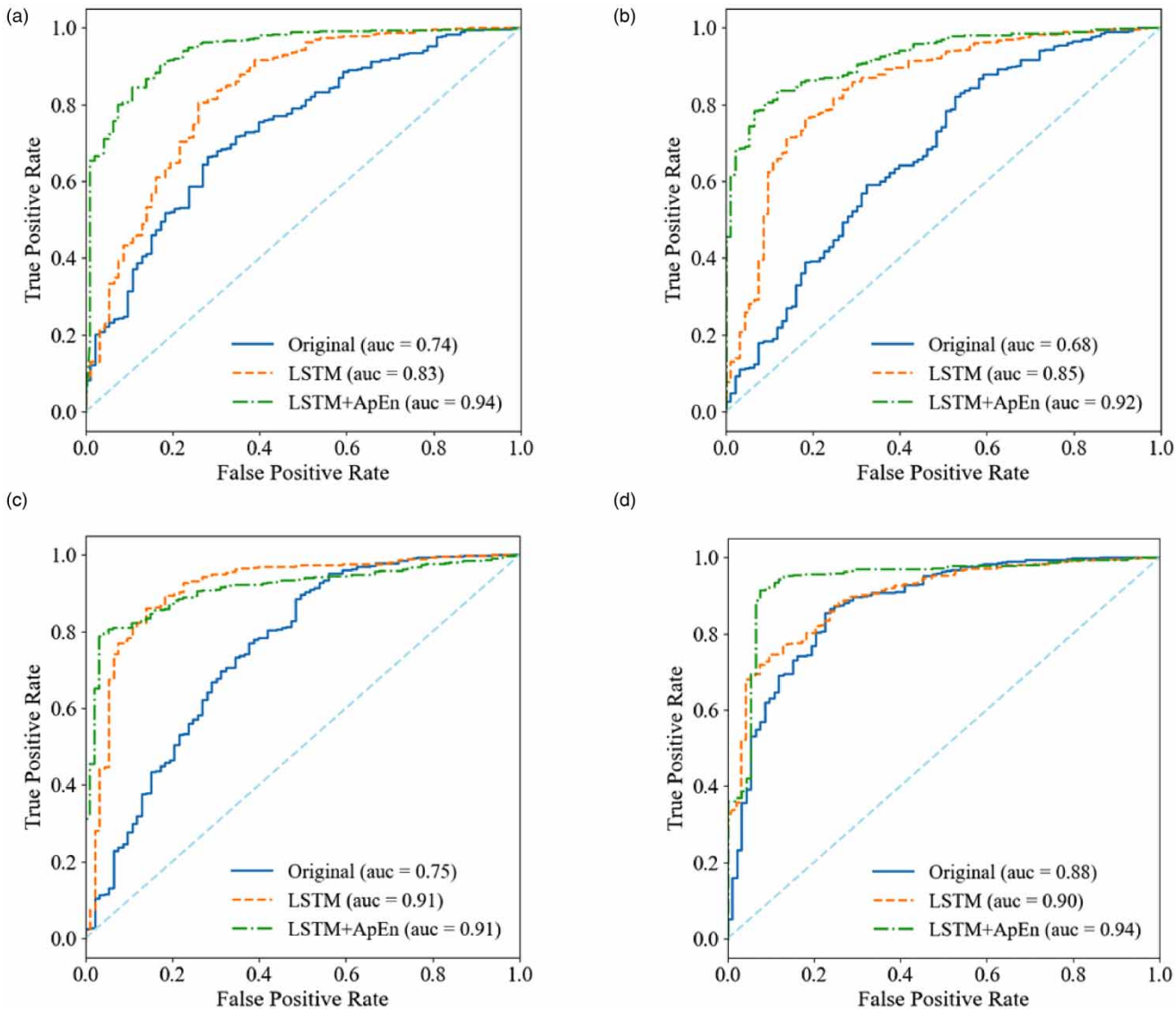


Figure 6 | Comparsion of ROC curves under different anomaly event strengths, using abnormal water quality time series data, residuals with LSTM model and ApEn features of COD as the input of IF algorithm, respectively. (a) ROC curve under anomaly event strength of 0.8; (b) ROC curve under anomaly event strength of 1.0; (c) ROC curve under anomaly event strength of 1.5 and (d) ROC curve under anomaly event strength of 2.0.

Table 2 | Anomaly detection results of different feature extraction methods

Event strength	Evaluation index	Original	LSTM	LSTM + ApEn
0.8	Precision	0.66	0.76	0.78
	Recall	0.63	0.77	0.87
	F1-score	0.64	0.76	0.81
1	Precision	0.67	0.77	0.90
	Recall	0.64	0.77	0.81
	F1-score	0.65	0.77	0.83
1.5	Precision	0.71	0.80	0.90
	Recall	0.68	0.83	0.80
	F1-score	0.69	0.81	0.82
2	Precision	0.83	0.90	0.87
	Recall	0.76	0.82	0.91
	F1-score	0.79	0.84	0.89

- The water prediction model based on LSTM revealed that trends of water baseline. Compared with the prediction model based on RNN, the proposed prediction method could obtain higher prediction accuracy with less time.
- The ApEn feature selection method played an important role in the detection performance. The residual sequences of each principal component were combined and the ApEn was used to select features of water prediction residuals. The ApEn had different trends and distributions when the water system was under pollution or chaos.
- The developed IF algorithm, using multiple water time series, proved to be effective in detecting water quality anomalies.

Abnormal water quality events often cause correlation changes of multiple water quality indicators at the same time (Mao *et al.* 2017). Different water quality parameters have different sensitivity to different pollutants. Kroll (2006) has concluded that residual chlorine and TOC are sensitive parameters to the pollutants such as sodium citrate, sodium cyanide, nicotine. Relevant researches have proved that using multiple water quality indicators (Vugrin *et al.* 2009; Liu *et al.* 2014) can reduce false alarm caused by non-water pollution events such as operating conditions.

In the process of water quality prediction, due to the complexity of prediction problems and the limitation of data availability, this proposed method only considers the historical data of conventional water quality indicators that directly reflect the impact of water quality changes. However, the river water body is more complex and is often affected by external factors such as geography and meteorology, the water quality prediction model proposed in this paper does not consider the hydrological information (Herath *et al.* 2020; Jiang *et al.* 2022), resulting in lack of interpretability and physical consistency (Herath *et al.* 2021). How to take into account more background knowledge and enhance the interpretability of the prediction model are the content of my follow-up research.

FUNDING

This work was funded by the Key Technology Research and Development Program of Zhejiang Province (2021C03177 and 2022C03078) and National Natural Science Foundation of China (U21A20519 and 61803333).

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossian, M. S., Ehteram, M. & Elshafie, A. 2019 Machine learning methods for better water quality prediction. *J. Hydrol.* **578**, 124084.

Arad, J., Perelman, L. & Ostfeld, A. 2012 A coupled decision trees bayesian approach for water distribution systems event detection. In *Proc of the 2012 WEWR*, 20–24 May, Albuquerque, New Mexico.

Azhar, S. C., Aris, A. Z. & Yusoff, M. K. 2015 Classification of river water quality using multivariate analysis. *Procedia Environ. Sci.* **30**, 79–84.

Baek, S., Pyo, J. & Chun, J. A. 2020 Prediction of water level and water quality using a CNN-LSTM combined deep learning approach. *Water* **12** (12), 3399.

Bao, Y. & Meng, W. 2015 Research on water quality multiscale feature extraction and anomaly detection method based on wavelet packet energy spectrum. In *Proc of the 26th CPCC*, 31 July, Nanchang, China.

Bouamar, M. & Ladjal, M. 2007 Evaluation of the performances of ANN and SVM techniques used in water quality classification. In *Proc of the 14th ICECS*, 11–14 December, Marrakech, Morocco.

Byer, D. & Carlson, K. H. 2005 Real-time detection of intentional chemical contamination in the distribution system. *J. Am. Water Works Assoc.* **97** (7), 130–133.

Durdu, F. A. 2010 Hybrid neural network and ARIMA model for water quality time series prediction. *Eng. Appl. Artif. Intell.* **23** (4), 586–594.

Herath, H., Chadalawada, J. & Babovic, V. 2020 Hydrologically informed machine learning for rainfall-runoff modeling: a genetic programming-based toolkit for automatic model induction. *Water Resour. Res.* **56** (4), e2019WR026933.

Herath, H., Chadalawada, J. & Babovic, V. 2021 Hydrologically informed machine learning for rainfall-runoff modelling: towards distributed modelling. *Hydrol. Earth Syst. Sci.* **25** (8), 4373–4401.

Huang, P. J., Wang, K. & Hou, D. 2017 In situ detection of water quality contamination events based on signal complexity analysis using online ultraviolet-visible spectral sensor. *Appl. Opt.* **56** (22), 6317–6323.

Jiang, J., Zheng, Y., Pang, T., Wang, B. & Tian, Y. 2020 A comprehensive study on spectral analysis and anomaly detection of river water quality dynamics with high time resolution measurements. *J. Hydrol.* **589**, 125175.

Jiang, S., Zheng, Y., Wang, C. & Babovic, V. 2022 Uncovering flooding mechanisms across the contiguous United States through interpretive deep learning on representative catchments. *Water Resour. Res.* **58** (1), e2021WR030185.

Koch, M. W. & McKenna, S. A. 2010 Distributed sensor fusion in water quality event detection. *J. Water Res. Plann. Manage.* **137** (1), 10–19.

Kroll, D. 2006 *Securing our Water Supply: Protecting a vulnerable resource*. PennWell Books, Tulsa, Oklahoma.

Liu, F., Ting, K. & Zhou, Z. 2008 Isolation forest. In *Proc of the 8th ICDM*, 15–19 December, Pisa, Italy.

Liu, S., Che, H. & Smith, K. 2014 Contamination event detection using multiple types of conventional water quality sensors in source water. *Environ. Sci. Processes Impacts* **16** (8), 2028–2038.

Liu, J., Wang, P. & Jiang, D. 2020 An integrated data-driven framework for surface water quality anomaly detection and early warning. *J. Clean. Prod.* **251**, 119–145.

Ma, W., Kang, Y. & Song, S. 2019 Analysis of streamflow complexity based on entropies in the Weihe River Basin, China. *Entropy* **22** (1), 38.

Mao, Y., Qi, H. & Jie, Q. 2017 M-TAEDA: Detection algorithm for abnormal events of multivariate water quality parameter time series data. *Comput. Appl.* **37** (01), 138–144.

Olsen, R. L., Chappell, R. W. & Loftis, J. C. 2012 Water quality sample collection, data treatment and results presentation for principal components analysis–literature review and Illinois river watershed case study. *Water Res.* **46** (9), 3110–3122.

Perelman, L., Arad, J., Housh, M. & Ostfeld, A. 2012 Event detection in water distribution systems from multivariate water quality time series. *Environ. Sci. Technol.* **46** (15), 8212–8219.

Qi, W., Lian, J. & Sun, Z. 2006 Technical specifications for surface water and sewage monitoring. *China Environ. Monit.* **1**, 54–57.

Shi, B., Peng, W., 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**, 1390–1399.

Tinelli, S. & Juran, I. 2019 Artificial intelligence-based monitoring system of water quality parameters for early detection of non-specific bio-contamination in water distribution systems. *Water Sup.* **19** (6), 1785–1792.

Vugrin, E., McKenna, S. & Hart, D. 2009 Trajectory clustering approach for reducing water quality event false alarms. In *Proc of the 2009 WEWRC*, 17–21 May, Missouri, America.

Wang, X., Zhang, J. & Babovic, V. 2016 Improving real-time forecasting of water quality indicators with combination of process-based models and data assimilation technique. *Ecol. Indic.* **66**, 428–439.

Wang, X., Zhang, J. & Babovic, V. 2019 A comprehensive integrated catchment-scale monitoring and modelling approach for facilitating management of water quality. *Environ. Modell. Software* **120**, 104489.

Wechmongkhonkon, S., Poomtong, N. & Areerachakul, S. 2012 Application of artificial neural network to classification surface water quality. *World Acad. Sci. Eng. Technol.* **6** (9), 574–578.

Zhang, Q., Qi, G. & Yan, S. 2014 Construction and application of water quality warning system in drinking water source. *Environ. Sci. Manage.* **39** (2), 123–125.

Zhang, Y., Lin, J. & Wang, H. 2020 Study on environmental quality standards for surface water in China. *Environ. Sci. Res.* **33** (11), 2523–2528.

Zhang, Y., Li, C. & Zhang, J. 2022 Analysis of sanitary standard for drinking water. *Water. Sup. Technol.* **16** (5), 38–43.

Zhu, W., Shi, B., Jiang, J. & Wang, P. 2018 Dynamic early warning method based on abnormal detection of water quality time series. *Environ. Sci. Technol.* **41** (12), 131–137.

First received 24 August 2022; accepted in revised form 5 January 2023. Available online 12 January 2023