UDC 614.878.086
DOI: 10.21668/health.risk/2024.4.05.eng

Read online

Research article

ALGORITHM FOR PREDICTING WATER QUALITY INDICATORS IN WATER BODIES USING A NEURAL NETWORK

M.A. Shiryaeva[1,2], O.O. Sinitsyna[1], M.V. Pushkareva[1], V.V. Turbinsky[1]

[1]F.F. Erisman Federal Scientific Center of Hygiene, 2 Semashko St., Mytishchi, Moscow region, 141014, Russian Federation
[2]Russian State Agrarian University – Moscow Timiryazev Agricultural Academy, 49 Timiryazeva St., Moscow, 127550, Russian Federation

Clean and safe drinking water is a fundamental necessity for human health and well-being and a critical component in sustainable ecosystem development. In recent decades, water quality issues have become even more urgent due to population growth, industrial expansion and climate change.

A series of works by foreign researchers report results obtained by applying neural networks. There are studies confirming results of water quality prediction generated by neural networks to be quite valid.

In this research, we used Google Earth Pro, Microsoft Excel, water flow sensor based on Arduino UNO board with author's modification (tail feathering and built-in plugin for calculation of flow velocity), Python, Tensorflows keras2.2.0, Scikit-learn, Pandas libraries for machine learning and development of neural network architecture. In this study, two ANNs were combined to build a hybrid neural network model for predicting water quality indicators.

Neural network models offer unique opportunities to improve water resources management at various levels, ranging from local to global one. A key advantage of such models is a possibility to adapt them to specific conditions and requirements, which provides more accurate prediction and timely decision making under uncertainty. The relevance of the work is determined by application of neural networks for water quality prediction. This can improve systems for early warning about pollution, help optimize operational processes at water treatment plants and develop effective water management strategies.

In this research, an innovative hybrid neural network model has been developed for predicting water quality indicators. It is based on integrating deep convolutional neural network and bidirectional recurrent neural network, which consists of three functional parts.

Keywords: *neural network, Tensorflows keras2.2.0, water bodies, drinking water, risk factor, negative impact, water pollution, determination coefficient, optimization algorithm.*

Assessment of water resources plays an exceptional role in contemporary life, even more so, given growing human-induced burdens on water ecosystems and climate change manifestations [1, 2, 3]. Rivers, lakes, and reservoirs are main sources of centralized drinking water supply to population; a key component in agricultural irrigation; a major source of water resources for various industries. They also occupy a substantial place in recreational infrastructure [4, 5].

Over the last decades, overall quality of water in water bodies has deteriorated considerably due to intensifying human-induced impacts and growing pollution [6]. This calls for developing and implementing innovative ap-

Margarita A. Shiryaeva – Junior Researcher of Water Hygiene Department (e-mail: Shiryaeva.MA@fncg.ru; tel.: +7 (903) 161-14-04; ORCID: https://orcid.org/0000-0001-8019-1203).

Oxana O. Sinitsyna – Corresponding Member of the Russian Academy of Sciences, Doctor of Medical Sciences, Professor, Deputy Director for Research, Director of the Institute of Complex Hygiene Issues (e-mail: sinitsyna.oo@fncg.ru; tel.: +7 (926) 447-08-74; ORCID: http://orcid.org/0000-0002-0241-0690).

Maria V. Pushkareva – Doctor of Medical Sciences, Professor, Chief Researcher of Water Hygiene Department (e-mail: pushkareva.mv@fncg.ru; tel.: +7 (912) 980-92-74; ORCID: https://orcid.org/0000-0002-5932-6350).

Viktor V. Turbinsky – Doctor of Medical Sciences, Head of Water Hygiene Department (e-mail: turbinskii.vv@fncg.ru; tel.: +7 (920) 666-72-73; ORCID: https://orcid.org/0000-0001-7668-9324).

proaches to monitoring and prediction of water conditions; they should surpass conventional methods in precision, reliability and promptness in getting required results [7].

Use of machine learning seems a most promising trend within this context. It should employ neural networks for simulating and predicting dynamics of water quality determinants. Such models are capable of considering the most complex non-linear interrelations between multiple influencing factors and are able to self-learn. This makes them a highly effective instrument for solving relevant tasks [8, 9].

The aim of this study was to develop and employ an innovative algorithm for predicting quality indicators of water bodies using a neural network. The aim was achieved by step-by-step accomplishment of several learning and perfecting stages. Comprehensive study of the existing methods for water quality assessment was the primary task. Detailed analysis of research publications revealed strong and weak points of various approaches and assessed their effectiveness in variable natural conditions. Due to it, we succeeded in identifying the most promising trends for developing this new innovative algorithm. The next stage involved designing our own neural network and matched initial parameters and training it using actual data. Chemical composition of water was considered in the process. Maximum precision in prediction was successfully achieved by calibrating the algorithm. Next, effectiveness of the developed neural model was estimated. Comparative tests were performed, in which prediction results were compared with actual data and indicators determined by using conventional approaches. This allowed us to establish what advantages this new algorithm had and where its use would be the most advisable. The final stage involved comparing the new approach with conventional procedures, which made it possible to identify advantages and drawbacks of the suggested model. In conclusion, we developed recommendations on using the algorithm in practice and established future prospects for further improvement. This study makes a substantial contribution to solving a strategic task, which is to predict quality of water resources.

Materials and methods. Studies aimed at analyzing the sanitary state and water quality indicators in dynamics were conducted at a section of the Oka River bed that was geographically a part of Ryazan City agglomeration. It provided a representative example of interaction between a large water flow and a highly urbanized territory.

Complex and consistent monitoring of the analyzed water flow is necessary for making effective predictions of Oka River water quality. This allows identifying negative trends in due time and taking relevant measures aimed at preventing sanitary-epidemiological troubles. Optimal location of control points for data collection plays the key role in supporting highly effective and representative monitoring. This location should consider both peculiar hydrological conditions of a given river and spatial distribution of potential pollution sources [10]. Figure 1 provides a draft scheme for locating module meteorological stations integrated with water auto-sampling systems. The red circles represent location of measuring devices; the green sector covers the observation territory of 400 km^2; the blue sector shows a zone where observations of two neighboring stations overlap, which provides necessary data excessiveness for obtaining more reliable results.

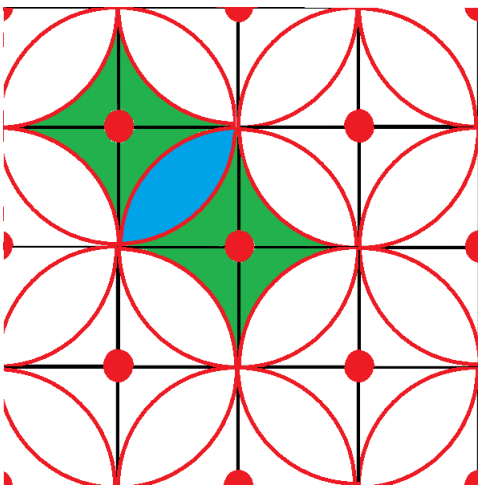

Figure 1. Scheme to show location of module meteorological stations

The suggested scheme for locating control-measuring points was developed in order to optimize the economic component together with preserving the maximum possible effectiveness of the systemic monitoring. This is especially important given strict budgetary limitations imposed on funding allocated for environmental protection programs. Use of mathematical modeling and optimization techniques, which considers hydrological peculiarities of a water body, specific distribution of human-induced pollution sources and economic limitations, makes it possible to reliably determine necessary and sufficient quantity of control-measuring points to effectively cover the whole water intake area of the analyzed section of the Oka River bed. This quantity ensures systemic monitoring and allows predicting sanitary conditions of the river considering multifactorial dynamics of external exposures [11].

Distances between module stations are:

$$l_1 = 21.6\sqrt{2} = 30.5 \text{ km};$$

and

$$l_2 = 21.6 \cdot 2 = 43.2 \text{ km}.$$

Next, we can define the square of overlapped observations by two stations:

$$F = 2 \cdot \left(\frac{21.6^2 \cdot \Pi}{4} - \frac{21.6^2}{2} \right) =$$

$$= \frac{21.6^2 \cdot \Pi}{2} - \frac{2 \cdot 21.6^2}{2} =$$

$$= \frac{21.6^2 (\Pi - 2)}{2} \approx 266 \text{ km}^2.$$

The following theoretical deductions can be made based on analyzing the conceptual model shown in Figure (1) that describes overlap zones covered by operations of hydrometric stations. Each two neighboring stations have intersection where boundaries of their measuring zones cross; three stations have two intersections, and so on in conformity with the established regularity.

By using this tendency, it seems possible to predict the necessary quantity of monitoring posts to provide full coverage of the analyzed water intake basin with control measurements. To quantify the required number, we suggest building a mathematical model by formulating an equation[1], the initial parameter (n) in which is represented by the number of stations that cover an area of 400 km^2.

Then the number of operational zone intersections between the stations in an area of 266 km^2 can be given as (n – 1). Solving this equation makes it possible to determine the minimal necessary number of hydrometric posts to create a regular observation network and ensure qualitative hydrometeorological monitoring on the whole water intake territory as well as to design an optimal configuration of the observation network.

The next equation is then derived:

$$400n + 266(n-1) = 245{,}000 \text{ km}^2,$$

$$n \approx 368 \text{ units}.$$

Considering the results derived from mathematical modeling of the minimum necessary number of hydrometric stations, we can conclude that it is advisable to make an observation network consisting of 368 universal module hydrological posts. This will insure spatial coverage of the whole Oka River water area in order to ensure qualitative monitoring of its climatic, hydrochemical and microbiological characteristics.

Investigations and primary data collection relied on using up-to-date software and innovative equipment. This included Google Earth Pro, geographical software for spatial analysis and data visualization; Microsoft Excel for statistical and preliminary analysis of the obtained results; a water-surface unmanned drone designed by the authors and equipped with water autosampler; Garmin Striker Cast GPS, a high-precision sonar device for depth measurements and creating bathymetrical maps of the analyzed river section; as well as

[1] Evgrafov A.V. Metrologiya, standartizatsiya i sertifikatsiya [Metrology, standardization and certification]: manual. Moscow, RGAU-MSHA Publ., 2015, 83 p. (in Russian).

an innovative water-flow sensor based on Arduino UNO microcontroller board with authors' modifications that included an optimized tail unit for stabilizing a position in a flow and an integrated software plugin for calculating flow velocity based on measured flow parameters.

A specialized software plugin was designed to achieve higher measuring precision and expand functional capabilities of the water flow sensor based on Arduino UNO microcontroller board. It was based on a mathematically derived formula for transforming data on water flow into flow velocity values considering geometrical parameters of the sensor, in particular, inlet and outlet diameters equal to 11.9 mm. This ensured the optimal ratio between the device sensitivity and its resistance to clogging with suspended particles.

Accordingly, the following formulas were introduced into the transformation plugin to determine the flow velocity (m/sec) from water flow (l/sec):

$$V = \frac{4W}{\pi \cdot D^2 \cdot 1000}, \quad (1)$$

$$V = \frac{4W}{\pi \cdot 0,0119^2 \cdot 1000},$$

where $\pi = 3.14$, W means baseline data of the water flow sensor (l/sec), D is the inlet and outlet diameters of the sensor (mm).

The complex model for machine learning, which was developed within this study, will primarily allow achieving more effective prediction of quality of surface waters as a strategically important water supply source, first of all, by estimating its conformity to safe standards. It will also help develop scientifically grounded recommendations for industrial enterprises, agricultural complexes and other potential pollution sources how to minimize negative impacts on a water body and reduce volumes of pollutants discharged into it [12, 13].

Data on water chemical composition were obtained by laboratory tests.

The algorithm for complex prediction of water quality, which is suggested in this study, includes the following successive stages.

Stage 1: Data cleansing. Prior to direct prediction of water quality, the iForest method is employed to detect anomalies in a data array on water quality $X_{n\times m}$ (where n is the number of water quality indicators and m is the number of data groups; within this study, n and m are constants: n = 9, m = 1360); these detected anomalies are replaced with null values. Later the Lagrange interpolation is used to fill in the null values since the method ensures data integrity and continuity [14, 15].

Stage 2: Data expansion. At the first stage, a predicted goal is removed from the data array $X_{n\times m}$ and, as a result, a new data array $X_{n\times(m-1)}$ is created. Bearing in mind, that data on water quality are collected with a 4-hour interval, the data windowing method is employed for averaging with a window size equal to 6 in order to create a set of moving averages $Z_{n\times(m-1)}$. This minimizes influence exerted by accidental factors of variations in data on water quality and allows more precise tracking of daily changes in water quality indicators. At the second stage, the principal component analysis (PCA) is used to decrease $X_{n\times(m-1)}$ dimensionality and to preserve two principal components $P_{2\times m}$. To prevent the model from overtraining, $Z_{n\times(m-1)}$, $P_{2\times m}$ and data on water quality $X_{n\times(m-1)}$ are simultaneously introduced into the model input without any target parameters, whereas a target prediction is formed at the model output.

Stage 3: Model training. The available data array on water quality is divided into a training dataset and a test dataset in the ratio 8:2. Within this study, the training dataset included 1100 dataset that covered the period from June 25, 2021 to February 16, 2022, whereas the test dataset included 272 datasets collected over the period from February 17, 2022 to April 01, 2022. The data windowing technique is employed, taking into account a long-term relationship between data on water quality and time factors [16, 17], to divide the training set into fixed training windows with a step of the i length in time sequence. After that, data of the first j training windows are used to predict a j+1 training window. In each new cycle, the oldest window is excluded from

the analysis and the next new window is included into it and the process continues until the last training window is reached. Such an approach, which involves excluding stale data, ensures the model training considering future trends. At the final stage, in accordance with the test dataset for each station, the trained model is used to predict key water quality indicators including total nitrogen and phosphor levels as well as permanganate oxidability.

Within the accomplished complex research, a comprehensive assessment was performed to establish effectiveness of the suggested hybrid neural network model for predicting water quality indicators, which included comparative analysis with reference methods employed in the sphere. To obtain quantitative characteristics of prediction precision, several conventional metrics were applied including mean absolute error (*MAE*), which describes average difference between predicted and actual values; mean absolute percentage error (*MAPE*), which estimates a relative value of prediction error; root-mean-square error (*RMSE*), which considers squared differences and gives greater weight to big errors; as well as the determination coefficient (R^2), which characterizes a share of dependable variable dispersion explained by the model [18].

At the initial stage in the study, the isolation Forest method (iForest) was applied as an effective algorithm for detecting anomalies in multidimensional data arrays. With its use, spikes in initial dada on water quality at the analyzed stations were identified and quantified. They were equal to approximately 1.1, 1.7 and 3.2 % of the total data volume respectively. All detected spikes, which could influence the model precision considerably, were removed thoroughly; after that, the remaining null values amounted to approximately 3.9, 4.5 and 5 % for the stations 1–3 respectively. This required using certain methods for data recovery; in particular, the Lagrange interpolation was employed to recover the continuous function per the discrete set of points.

To assess whether it was acceptable to use the developed model in practice, conventional prediction models ARIMA and SMA were tested within this study.

ARIMA (Autoregressive Integrated Moving Average) and SMA (Simple Moving Average) models are very popular for predicting time series, water quality included. ARIMA models consider autocorrelation and autoregression in data, which allows them to capture dynamics of changes in water quality over time. SMA is eligible for predicting water quality with more stable time series and fewer spikes.

Results and discussion. The accomplished complex study included assessment of qualitative characteristics of water resources in the Oka River, which was considered a surface water supply source. The assessment was based on analyzing average long-term values of 52 control indicators including organoleptic, microbiological and chemical ones over a long period between 2014 and 2022. The research results showed that water taken at the Sokolovskii water intake had a considerably lower average long-term ammonia level equal to 0.48 mg/l, which was considerably lower than the same indicators established at the Okskii and Borkovskii water intakes, 1.6 and 2.1 times respectively ($p < 0.05$). It should be noted that ammonia concentrations above its maximum permissible level (MPL) were detected in practically each fifth sample taken at the Borkovskii water intake gates, whereas the same indicator established for the Okskii water intake was 2.8 times lower and amounted to 7.5 %. It is interesting that ammonia ions in levels higher than MPL were not found in any one-time water sample taken at the control point of the Sokolovskii water intake. Statistical analysis did not establish any significant differences in average long-term chemical oxygen demand (COD) or biochemical oxygen demand (BOD_5) in water of the analyzed water intakes. The proportion of one-time samples, in which the analyzed indicators did not conform to safety requirements, varied within 22.7–32.5 and 61.8–75.0 % respectively. Our study also showed that average levels of total coliform bacteria (TCB) amounted to 813.3 and 818.9 CFU/100 ml in water from the

Okskii and Borkovskii water intakes respectively; this was 1.5 times higher against the levels established at the control gates of the Sokolovskii water intake ($p < 0.05$). Some tests results are shown in the graphs (Figure 2).

In this study, an innovative hybrid neural network model was developed for predicting water quality indicators. It was based on integrating a deep convolutional neural network and bidirectional recurrent neural network, which consisted of three functional parts. At the initial stage, the model is employed to identify and extract potential non-linear interrelations between time series data about water quality in the Oka River in order to create effective low-dimensional attributes. Next, a vector of water quality indicators is built based on the extracted attributes. This vector is used as an input signal for the network of the deep convolutional neural network. When being trained, the network regulates weights and shifts constantly considering dependence of short-term, long-term and contextual attributes of a data time series for further optimization of data on water quality in order to achieve higher precision in attribute expression. At the final stage, the layer of complete connection becomes involved at the upper part of the model. It serves as an output layer for generating predicted values of water quality indicators.

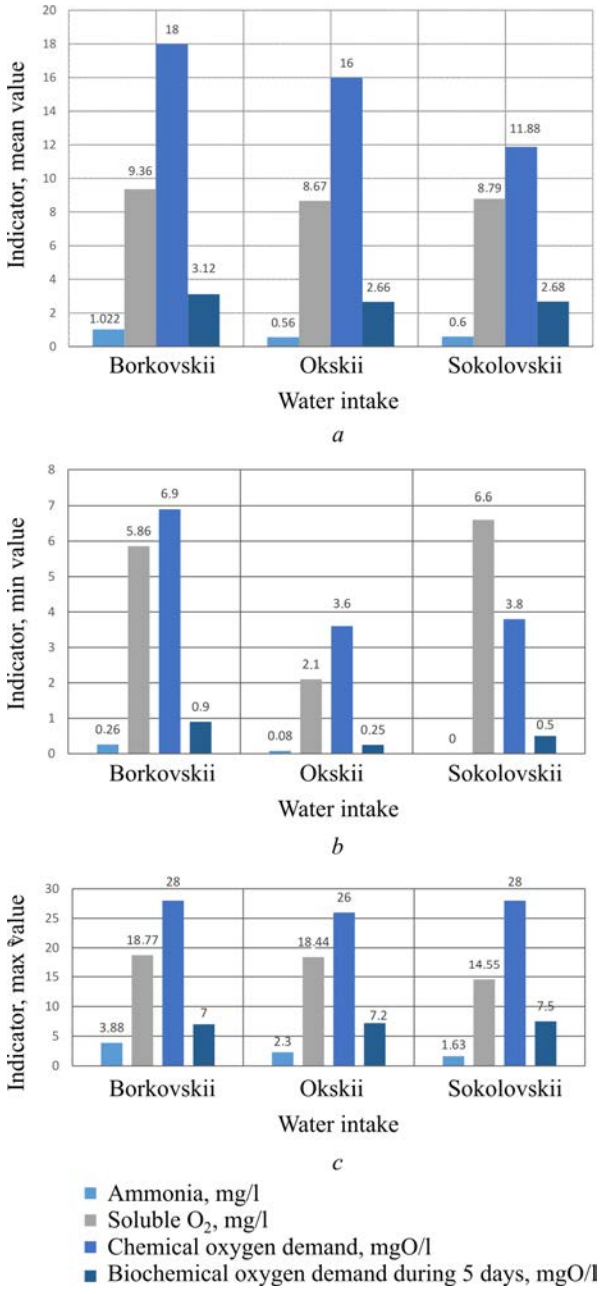

Figure 2. Some water quality indicators for the Oka River established at three analyzed water intakes for the period 2014–2022, where: *a* shows the mean values; *b*, the minimum values; *c*, the maximum values

The developed hybrid neural network prediction model was realized as software using highly-productive library for deep learning Tensorflows keras, version 2.2.0, which provides a wide range of instruments employed to build and train neural networks. The model training went on for 50 epochs using 120 time intervals, which allowed achieving an optimal balance between prediction precision and computational expenses. The Adam method was employed as an optimization algorithm for correcting the model weights and shifts. It combines advantages provided by adaptive gradient descend methods and the method of moments. When the model convergence was reached indicating that the loss function was minimized, final weight coefficients were obtained, which were later used to predict water quality at the analyzed water intake stations (Sokolovskii, Okskii and Borkovskii). The model architecture and parameters were selected thoroughly and presented in the following way: the number of hidden layers was two, which enabled the model to effectively detect complex non-linear relationships in data; the conjugate gradient method was selected as an optimization algorithm due to its known capability to rapidly converge to an optimal solution; minimal relative change in the training

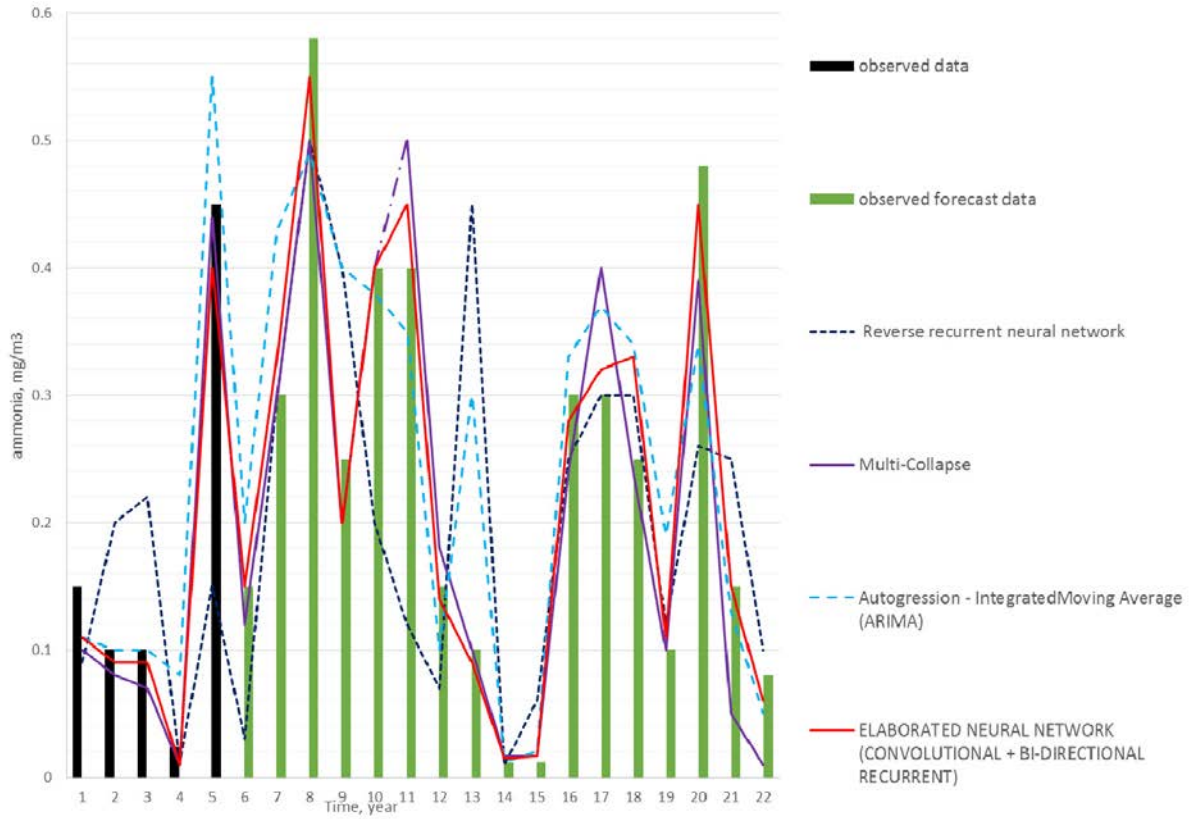

Figure 3. Predictive results obtained for the ammonia level at the Okskii water intake based on neural network in comparison with existing neural network models and the classical moving average model. Baseline data (period of 1–5 years – elapsed time period)

error coefficient was fixed at 0.001, which ensured the relevant balance between the model precision and prevented of overtraining.

This allows using this neural network model to fill in gaps in data by calculating missing concentrations of certain chemicals. Some results obtained by using neural network modeling that cover a 25-year observation period (n = 25) are shown as an example in Figure 3 for the ammonia level as a water quality indicator at the Okskii water intake. The obtained regularities can be used for predicting future dynamics of the analyzed indicators.

To determine whether it was acceptable to use the developed neural network model, it was compared with the classical moving average model (Figure 3, observed forecast data). This method was selected due to being quite a common technique for analyzing time series such as nitrate levels in water [19, 20]. It helps identify trends by reducing influence of accidental fluctuations and noise in data sets.

The moving average (MA) method involves calculating an average value of several previous data points for each time moment. The formula for moving average calculation is given as:

$$SMA_t = \frac{X_{t-n+1} + X_{t-n+2} + X_t}{n}, \qquad (2)$$

where SMA_t is the moving average value at the time moment t;

X are observed values (for example, nitrate levels);

n is the number of periods (interval) for smoothing.

The coefficient α determines a weight, which is given to the latest observation: higher α values lead to more rapid response to changes in data.

To analyze levels of pollutants, both methods were employed for data smoothing

per time series and identification of long-term trends:

1. Moving average makes data on pollutant levels more stable and helps visualize trends in changes over time.

2. Exponential smoothing ensures more rapid response to changes in pollutant levels, which is especially useful in case data are susceptible to drastic variations.

In addition to SMA_t (Figure 3, observed forecast data), we applied the conventional prediction using ARIMA (Autoregressive Integrated Moving Average or a statistical analysis model, which uses time series data to predict future values in a series).

In this study, we suggest considering an alternative approach to water quality prediction. It relies on using neural networks for analyzing large historical data arrays and this method is fundamentally different from conventional mechanistic models, which are widely used for the purpose. Mechanistic models for water quality prediction that include such well-known systems as QUAL, WASP, MIKE, SWAT, BASINS and some others are based on detailed description of a structure of an analyzed water system and on considering multiple limitations associated with a set of physical, biological and chemical processes in water environment. This determines their complexity and requires substantial volumes of input data to create and then solve a system of equations that describe changes in water quality in dynamics over time and space [21–23].

Despite being common and recognized by experts society, mechanistic models tend to be very complex in their essence and require large scopes of input data including multiple modeling parameters, conditions of water sources and pollution discharges as well as other specific characteristics of water environment. A process of building such models is extremely labor-consuming and also it is very difficult to identify optimal model parameters; this imposes substantial limitations on their applicability for a wide range of water bodies, especially if data on their hydrological and sanitary conditions are insufficient and not detailed [24, 25].

The suggested neural model, which is based on up-to-date deep architectures, has turned out to be very effective in solving the task. Underlying non-linear multi-layer mechanisms for data analysis make it possible to reveal complex interrelationships between water quality indicators and external factors thereby creating significant predictions. The accomplished investigations confirmed high prediction reliability due to the model being capable of analyzing and predicting non-linear processes in uncertain conditions quite effectively.

In addition, the model is universal and eligible for variable water bodies including rivers, lakes and reservoirs. This substantially expands the sphere where the model could potentially be used for water quality monitoring and management of water resources. An obvious advantage the model has in comparison with traditional numeric algorithm is determined by its higher prediction precision and computational effectiveness. This model provides new opportunities for developing promising approaches to water quality monitoring and management of water resources.

Prediction of nitrate levels using time series models such as ARIMA (Autoregressive Integrated Moving Average) is a powerful method for data analysis, which considers trends, seasonality, and autocorrelation. With this procedure, we are going to consider several steps necessary for the ARIMA model implementation and present relevant formulas. Over the last few decades, investigations with their focus on time series predictions have been mostly concentrating on two approaches [26]. One of them is based on mathematical statistics, for example, on autoregression models or integrated moving average (ARIMA). These models often have advantages in a situation when a data array in a time series is small since they require relatively smaller data volumes for assessment of model parameters [21].

Historical data on nitrate levels at water intake gates have been collected. The data are given as a time series, where each element corresponds to a nitrate level at a certain time moment. ARIMA is denoted as ARIMA (p, d, q), where p is the autoregression order, d is the differentiation level, q is the moving average order.

The graphs ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) were applied to establish p and q values. ACF shows time series autocorrelation at different lags. If ACF declines rapidly, then q can be low. PACF shows partial autocorrelation and can help establish p.

Next, the ARIMA model was built [27]:

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots + \Phi_p Y_{t-p} + \\ + \Theta_1 e_{t-1} + \Theta_2 e_{t-2} + \ldots + \Theta_q e_{t-q} + e_t, \quad (3)$$

where Φ are the AR model coefficients, Θ are the MA model coefficients, e_t is the error.

The model parameters were estimated by using the maximum likelihood method. The model quality was tested by using the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion).

Next, prediction was accomplished using the assessed ARIMA model:

$$\hat{Y}_{t+h} = \hat{Y}_t + \sum_{i=1}^{p} \Phi_i \hat{Y}_{t-i} - \sum_{j=1}^{q} \Theta_j e_{t-j}, \quad (4)$$

where h is the time horizon.

This procedure is a structured approach to predicting pollutant levels in river water intakes using the ARIMA model [27, 28]. Water quality management can be improved considerably due to precise predictions, which allow making more grounded decisions on protection of ecosystems and human health.

To estimate advantages and drawbacks of the suggested prediction model and other etalon neural networks, such as LSTM and a reverse recurrent one, we compared mean average error (*MAE*), which shows average difference between predicted and actual values; mean average percentage error (*MAPE*); root-mean-square error (*RMSE*), which consider the square of differences; the determination coefficient (R^2) (Table). The ARIMA model was selected as the etalon method. The developed model yielded the highest results in comparison with the etalon ARIMA model and the reverse recurrent neural network model. The multi-convolutional model turned out to be the strongest competitor; its root-mean-square error value was 0.0557 whereas it was 0.0248 lower in the developed model (that is, the average *RMSE* = 0.0309).

To visualize the comparison results, a graph was built to show the results for ammonia levels at the Okskii water intake obtained by using the conventional ARIMA model and the developed neural network (Figure 4).

The developed neural network underestimated the analyzed indicator for 2021 in training on input data over 2018–2022 (the 4th year is the corresponding indicator in the graph). The developed neural network provided underestimated ammonia levels in comparison with the etalon ARIMA model. However, precision of the developed neural network was higher than that of the ARIMA model ($RMSE_{ARIMA}$ is higher than $RMSE_{NN}$ by 1.17).

The determination coefficient and root-mean-square error beamed by using the developed neural network model: comparison with the parameters already established for the Okskii water intake

Quality indicator	Parameters of statistical analysis model	Autoregression – Integrated Moving Average (ARIMA)	Neural network		
			Reverse recurrent neural network	Multi-convolutional LSTM	Developed neural network (convolutional + bidirectional recurrent)
BOD, mgO_2/l	R^2	0.9408	0.9920	0.9996	0.9996
	RMSE	1.2030	0.5360	0.0566	0.0299
Total nitrogen, mg/l	R^2	0.8760	0.9933	0.9999	0.9996
	RMSE	1.0000	0.5400	0.0542	0.0315
Ammonia	R^2	0.9400	0.9945	0.9999	0.9999
	RMSE	0.9850	0.5466	0.0520	0.0310
O_2, mg/l	R^2	0.9308	0.9900	0.9996	0.9999
	RMSE	1.0000	0.5280	0.0600	0.0312

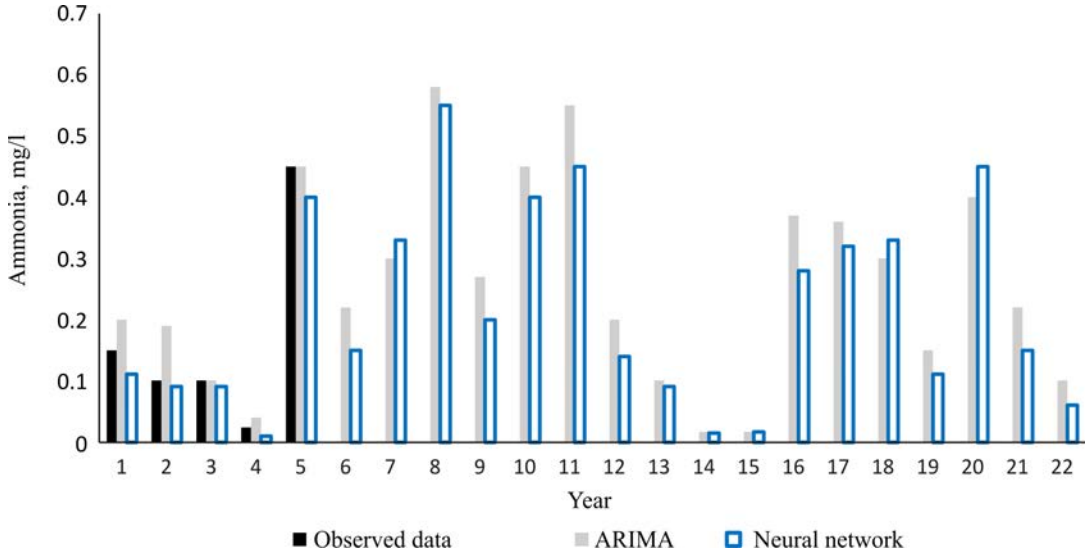

Figure 4. Comparison of mean ammonia levels per observation data (period of 1–5 years equal to 2018–2022) obtained by using classical ARIMA prediction and developed neural network for the Okskii water intake

Conclusion. The laboratory test results showed that water taken at the Sokolovskii water intake had a considerably lower average long-term ammonia level equal to 0.48 mg/l. This was substantially lower than the same levels established at the Okskii and Borkovskii water intakes, which were 1.6 and 2.1 times higher respectively ($p < 0.05$). We did not establish any considerable differences in average long-term levels of chemical oxygen demand (COD) and biochemical oxygen demand (BOD_5) in water taken at the analyzed water intakes. The proportions of one-time samples with these indicators deviating from the valid safe standards varied within the range between 22.7 and 32.5 % for COD and between 61.8 and 75.0 % for BOD_5.

To provide high reliability, most advanced approaches to data processing were applied at the preliminary stage of the accomplished study including the isolating forest algorithm and the Lagrange interpolation. This made it possible not only to effectively raise the integrity of the data array but also to minimize potential impacts exerted by inaccuracies and anomalies on the subsequent modeling. In addition to preliminary data processing, the moving average method and the principal component analysis were applied; they allowed optimizing water quality indicators and preventing the model overtraining, which was a crucial factor for providing high precision of prediction calculations in the long-term outlook. Therefore, we developed an innovative hybrid neural network model for predicting water quality indicators based on integration of a deep convolutional neural network and bidirectional recurrent neural network, which consisted of three functional parts.

The results obtained by experimental approbation of the developed model clearly demonstrate high stability and generalizing capability of the suggested approach, which is manifested through lower prediction inaccuracy in comparison with the conventional methods. This also provides new prospects for using this concept in prediction of one-dimensional time series of various objects within natural sciences and technical analysis in comparison with such models as Autoregression Integrated Moving Average (ARIMA) or reverse recurrent neural networks. The multi-convolutional model turned out to be the strongest competitor; its root-mean-square error value was 0.0557. At the same time, it was 0.0248 lower in the developed model, which means that the average $RMSE = 0.0309$.

The accomplished study with its focus on using recurrent neural networks to predict pollution of the Oka River, a key water body in Central Russia, allows making a well-grounded conclusion that precise and timely prediction of changes in water quality is quite

possible. This provides new opportunities for implementing effective environmental-protection activities and providing sustainable development of the region in the long-term outlook. Use of the developed model to predict dynamics of the Oka River pollution for the next two decades provides a unique opportunity to reveal potential environmental hazards and to take revenant actions to prevent them. This is significant step towards more effective preservation of natural resources, provision of sanitary-epidemiological safety and improvement of life quality of the regional population.

Funding. The research was not granted any sponsor support.

Competing interests. The authors declare no competing interests.

References

1. Liao Z., Wang X., Zhang Y., Qing H., Li C., Liu Q., Cai J., Wei C. An integrated simulation framework for NDVI pattern variations with dual society-nature drives: A case study in Baiyangdian Wetland, North China. *Ecological Indicators*, 2024, vol. 158, pp. 111584. DOI: 10.1016/j.ecolind.2024.111584

2. Karpenko N.P., Glazunova I.V., Shiryaeva M.A. Analysis of geo ecological problems and assessment of the availability of drinking water in the Klinsky district of the Moscow region. *Prirodoobustroistvo*, 2023, no. 5, pp. 88–94. DOI: 10.26897/1997-6011-2023-5-88-94 (in Russian).

3. Shivam K., Tzou J.-C., Wu S.-C. Multi-step short-term wind speed prediction using a residual dilated causal convolutional network with nonlinear attention. *Energies*, 2020, vol. 13, no. 7, pp. 1772. DOI: 10.3390/en13071772

4. Wu G.-D., Lo S.-L. Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. *Engineering Applications of Artificial Intelligence*, 2008, vol. 21, no. 8, pp. 1189–1195. DOI: 10.1016/j.engappai.2008.03.015

5. Ho J.Y., Afan H.A., El-Shafie A.H., Koting S.B., Mohd N.S., Jaafar W.Z.B., Hin L.S., Malek M.A. [et al.]. Towards a time and cost effective approach to water quality index class prediction. *Journal of Hydrology*, 2019, vol. 575, pp. 148–165. DOI: 10.1016/j.jhydrol.2019.05.016

6. Juwana I., Muttil N., Perera B.J.C. Uncertainty and sensitivity analysis of West Java Water Sustainability Index – A case study on Citarum catchment in Indonesia. *Ecological indicators*, 2016, vol. 61, pp. 170–178. DOI: 10.1016/j.ecolind.2015.08.034

7. Rosenthal O.M., Fedotov V.Kh. Identification of water polluting enterprises based on neural network analysis. *Prirodoobustroistvo*, 2023, no. 1, pp. 62–68. DOI: 10.2689711997-6011-2023-1-62-68 (in Russian).

8. Shamsutdinova T.M. Application of Neural Network Modeling in Problems of Predicting the Level of River Floods. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnye tekhnologii*, 2023, vol. 21, no. 2, pp. 39–50. DOI: 10.25205/1818-7900-2023-21-239-50 (in Russian).

9. Shitikov V.K., Zinchenko T.D., Golovatiyuk L.V. Methods of neural networks for estimation of superficial waters quality by usage of hydrobiological exponents. *Izvestiya Samarskogo nauchnogo tsentra Rossiiskoi akademii nauk*, 2002, vol. 4, no. 2, pp. 280–289 (in Russian).

10. Ratkovich L.D., Markin V.N., Glazunova I.V. Voprosy ratsional'nogo ispol'zovaniya vodnykh resursov i proektnogo obosnovaniya vodokhozyaistvennykh sistem: monografiya [Issues of rational use of water resources and design substantiation of water management systems: a monograph]. Moscow, K.A. Timiryazev Russian State Agrarian University – K.A. Timiryazev MSHA, 2013, 256 p. (in Russian).

11. Karpenko N.P., Lomakin I.M., Drozdov V.S. Management issues of geoenvironmental risks in the assessment of groundwater quality in urban areas. *Prirodoobustroistvo*, 2019, no. 5, pp. 106–111. DOI: 10.34677/1997-6011/2019-5-106-111 (in Russian).

12. Litvinova A.A., Dement'yev A.A., Lyapkalo A.A., Korshunova E.P. Comparative Characteristics of Quality Parameters of Waters of the Oka River in Places of Water Intake of Utility and Drinking Water System in Ryazan. *Rossiiskii mediko-biologicheskii vestnik imeni akademika I.P. Pavlova*, 2022, vol. 30, no. 4, pp. 481–488. DOI: 10.17816/PAVL0VJ89568 (in Russian).

13. Zholdakova Z.I., Sinitsyna O.O., Turbinsky V.V. About adjustment of requirements to zones of sanitary protection of sources of the centralized economic and drinking water supply of the population. *Gigiena i sanitariya*, 2021, vol. 100, no. 11, pp. 1192–1197. DOI: 10.47470/0016-9900-2021-100-11-1192-1197 (in Russian).

14. Karpenko N.P., Shiryaeva M.A. Three-dimensional modeling as a system for displaying total chemical soil pollution. *Prirodoobustroistvo*, 2021, no. 1, pp. 6–14. DOI: 10.26897/1997-6011-2021-1-6-14 (in Russian).

15. Lagutina N.V., Novikov A.V., Sumarukova O.V., Naumenko N.O. Assessment of the water quality of the Rybinsk reservoir as a result of the water level lowering. *Prirodoobustroistvo*, 2019, no. 2, pp. 122–125. DOI: 10.34677/1997-6011/2019-2-122-126 (in Russian).

16. Naumenko N.O. Vvedenie ratsional'nogo normirovaniya na ob"emy sbrosov zagryaznya-yushchikh veshchestv v vodnye ob"ekty s tsel'yu podderzhaniya ustoichivosti ekosistemy [Introduction of rational rationing of the volume of pollutant discharges into water bodies in order to maintain sustainability of an ecosystem]. *Sovremennye problemy i perspektivy razvitiya rybokhozyaistvennogo kompleksa: materialy VII nauchno-prakticheskoi konferentsii molodykh uchenykh s mezhdunarodnym uchastiem*. Moscow, Russian Federal Research Institute of Fisheries and Oceanography Publ., 2019, pp. 344–346 (in Russian).

17. Liu H., Zhang F., Tan Y., Huang L., Li Y., Huang G., Luo S., Zeng A. Multi-scale quaternion CNN and BiGRU with cross self-attention feature fusion for fault diagnosis of bearing. *Meas. Sci. Technol.*, 2024, vol. 35, no. 8, pp. 086138. DOI: 10.1088/1361-6501/ad4c8e

18. Jiang Y., Li C., Sun L., Guo D., Zhang Y., Wang W. A deep learning algorithm for multi-source data fusion to predict water quality of urban sewer networks. *Journal of Cleaner Production*, 2021, vol. 318, pp. 128533. DOI: 10.1016/j.jclepro.2021.128533

19. Veerendra G.T.N., Kumaravel B., Kodanda Rama Rao P., Dey S., Phani Manoj A.V. Forecasting models for surface water quality using predictive analytics. *Environment, Development and Sustainability*, 2024, vol. 26, no. 6, pp. 15931–15951. DOI: 10.1007/s10668-023-03280-3

20. Chen X., Jiang Z., Cheng H., Zheng H., Cai D., Feng Y. A novel global average temperature prediction model – based on GM-ARIMA combination model. *Earth Science Informatics*, 2023, vol. 17, no. 1, pp. 853–866. DOI: 10.1007/s12145-023-01179-1

21. Jiao G., Chen S., Wang F., Wang Z., Wang F., Li H., Zhang F., Cai J., Jin J. Water quality evaluation and prediction based on a combined model. *Appl. Sci.*, 2023, vol. 13, no. 3, pp. 1286. DOI: 10.3390/app13031286

22. da Silva A.C., das Graças Braga da Silva F., de Mello Valério V.E., Lima Silva A.T.Y., Marques S.M., Tosta dos Reis J.A. Application of data prediction models in a real water supply network: comparison between arima and artificial neural networks. *Revista Brasileira de Recursos Hídricos*, 2024, vol. 29, pp. e12. DOI: 10.1590/2318-0331.292420230057

23. Deng T., Chau K.-W., Duan H.-F. Machine learning based marine water quality prediction for coastal hydro-environment management. *J. Environ. Manage.*, 2021, vol. 284, pp. 112051. DOI: 10.1016/j.jenvman.2021.112051

24. Lu X., Dong Y., Liu Q., Zhu H., Xu X., Liu J., Wang Y. Simulation on TN and TP distribution of sediment in Liaohe estuary national wetland park using mike21-coupling model. *Water*, 2023, vol. 15, no. 15, pp. 2727. DOI: 10.3390/w15152727

25. Kim J., Seo D., Jang M., Kim J. Augmentation of limited input data using an artificial neural network method to improve the accuracy of water quality modeling in a large lake. *Journal of Hydrology*, 2021, vol. 602, no. 4, pp. 126817. DOI: 10.1016/j.jhydrol.2021.126817

26. Wongburi P., Park J.K. Prediction of Wastewater Treatment Plant Effluent Water Quality Using Recurrent Neural Network (RNN) Models. *Water*, 2023, vol. 15, no. 19, pp. 3325. DOI: 10.3390/w15193325

27. Jaya N.A., Arsyad M., Palloan P. Estimation of Groundwater River Availability in Leang Lonrong Cave Using ARIMA Model and Econophysics Valuation Approach. *Advances in Social Humanities Research*, 2024, vol. 2, no. 5, pp. 737–754. DOI: 10.46799/adv.v2i5.240

28. Tiyasha, Tung T.M., Yaseen Z.M. Deep learning for prediction of water quality index classification: tropical catchment environmental assessment. *Natural Resources Research*, 2021, vol. 30, no. 6, pp. 4235–4254. DOI: 10.1007/s11053-021-09922-5

Shiryaeva M.A., Sinitsyna O.O., Pushkareva M.V., Turbinsky V.V. Algorithm for predicting water quality indicators in water bodies using a neural network. Health Risk Analysis, 2024, no. 4, pp. 50–62. DOI: 10.21668/health.risk/2024.4.05.eng

Received: 22.07.2024
Approved: 29.11.2024
Accepted for publication: 19.12.2024