Threat prediction in complex distributed systems using artificial neural network technology

E.V. Palchevsky1, O.I. Khristodulo1, S.V. Pavlov1

teelxp@inbox.ru | o-khristodulo@mail.ru | psygis@mail.ru

1Ufa State Aviation Technical University, Ufa, Russia

In the context of this article, a method for detecting threats based on their forecasting and development in complex distributed systems is proposed. Initially, the relevance of the research topic is substantiated from the point of view of the prospective use of various methods in the framework of threat management and their forecasting in complex distributed systems. Based on the analysis of these methods, a proprietary forecasting method based on the second generation recurrent neural network (RNN) was proposed. The mathematical formulation of the problem is presented, as well as the structure of this neural network and its mathematical model of self-learning, which allows achieving more accurate (with less error) results in the framework of threat prediction (in this case, the level of water rise at gauging stations) in complex distributed systems. An analysis was also made of the effectiveness of the existing and proposed forecasting methods, which showed the stability of the neural network in relation to other forecasting methods: the error of the neural network is 3-20% of actual (real) water levels; the least squares method reaches up to 34.5%, the numerical method in a generalized form - up to 36%; linear regression model – up to 47.5%. Thus, the neural network allows a fairly stable forecast of the flood situation over several days, which allows special services to carry out flood control measures.

Keywords: water level forecasting, flood situation, neural networks, neural network for forecasting.

1. Introduction

Currently, complex distributed systems include various components in the form of physical, biological and digital systems [1]. In the framework of this article, complex distributed systems are understood to mean technical objects, for example, potentially dangerous objects, oil pipelines, since technical objects located in a certain area, as defined in the literature (Reimers N.F. Nature Management, 1990; Mikhailov N.I. Physical-geographical zoning, 1985, etc.) - belong to the class of complex unique geotechnical objects and have all the features of complex systems. For such systems, as a rule, there is a danger in the form of external threats that contribute to causing substantial material damage. For example, in the Republic of Bashkortostan, such a threat is spring flood, which threatens complex distributed systems due to possible flooding and flooding. To counter such threats, it is proposed to proactively predict the level of water rise in order to plan further activities by specialized services to prevent the negative impact of the flood on complex distributed systems falling into its distribution zone.

Many scientists (both domestic and foreign) are involved in this problem in a wide variety of scientific fields. To solve this problem, many approaches and methods are used, such as numerical methods, regression models, etc. [2, 4-23]. But due to the lack of works containing a description of the method of early detection of threats based on their early forecasting, it becomes relevant to use neural network approaches and technologies to solve this problem.

Thus, it seems important and necessary, using a recurrent neural network, to develop a method for early detection of threats on the basis of predicting the level of water rise in the flood period to counter them in complex distributed systems. This will give the necessary specialized services some time to carry out flood control measures.

2. Existing solutions

Currently, the literature describes many methods for predicting flood situations (including water levels), the most used of which, with the exception of hydrological ones, are as follows:

- least squares method;
- numerical methods;
- general regression models.

The main objective of these approaches is to use mathematical methods and models capable of producing fairly accurate short-term forecasting of water levels.

Least square method

It is a way to solve various mathematical problems and is based on minimizing the sum of the squared deviations between the original and calculated values. The main working formula for forecasting:

\[ Y_{t+1} = (x + b) \cdot a, \]  \hspace{1cm} (1)

where \( Y_{t+1} \) is the predicted indicator, \( t+1 \) is the period for which the forecast is made, \( a \) and \( b \) are the coefficients of forecast indicators and period, \( x \) is the symbol of time. Calculation of coefficient \( a \):

\[ a = \frac{\sum_{i=1}^{n}((Y_{i+1}-x) \cdot n_{1}) - \sum_{i=1}^{n}x \cdot \sum_{i=1}^{n}Y_{i}}{\sum_{i=1}^{n}x^{2}-(\sum_{i=1}^{n}x)^{2}}, \] \hspace{1cm} (2)

where \( Y_{i} \) is the actual value of the time series, \( n_{1} \) is the number of levels in the time series. The coefficient \( b \) is calculated by the following formula:

\[ b = \frac{\sum_{i=1}^{n}Y_{i}}{n_{1}} - a \frac{\sum_{i=1}^{n}x}{n_{1}}. \] \hspace{1cm} (3)

And for smoothing the time series by the least squares method, in order to obtain and reflect the patterns of the forecast, it is necessary to correctly determine the type of curve and the time analytical dependence:

\[ M = \frac{\sum_{i=1}^{n}Y_{i}^{2}}{(n_{1}-p-1) \cdot a}, \] \hspace{1cm} (4)

where \( Y_{p} \) are the calculated values of the time series, \( p \) is the number of parameters of the described trend.
Numerical methods

They represent a way to solve a mathematical problem in numerical form. In the case of forecasting, information on previously obtained data is used. Accordingly, for this, the formulas (in generalized form) of the forecast (5) and correction (6) are used:

\[ y_{i+1} = y_{i-3} + (h_1 \cdot (2y_i' + y_{i-1} + y_{i-2}) + O(h_1^3)), \quad (5) \]

where \( O(h_1^3) \) the calculated error in the forecast, \( h_1 \) is the iteration step, \( i+1 \) is the forecast period.

\[ y_{i+1} = y_{i-1} + (h_1 \cdot (y_{i+1} + 4y_i' + y_{i-1}) + O(h_1^3)). \quad (6) \]

General regression models

These models are used in many problems of data analysis and forecasting. One of the most common regression models is multivariate. A general view of this model is presented in the following formula:

\[ Y_p = f^2 \cdot \sum_{i=1}^{n} x(x_{1i}, x_{2i}, x_{3i}, \ldots x_{ni}) = f(x), \quad (7) \]

where \( Y_p \) is the predicted indicator, \( x_1, x_2, x_3, \ldots x_n \) are the factors affecting the forecast of water levels. In this case, the initial information is provided in the form of time series, and the following functions can be used to calculate the forecast: linear (8), power (9), exponential (10), exponential (11), hyperbolic (12). In formulas 8-12: \( a \) and \( b \) are the coefficients of forecast indicators.

\[ Y_p = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + \ldots + b_nx_n. \quad (8) \]

\[ Y_p = a_0 \cdot (b_1x_1) + a_1 \cdot (b_2x_2) + a_2 \cdot (b_3x_3) + \ldots + a_n \cdot (b_nx_n). \quad (10) \]

In practice, the linear (8) function of the multivariate regression model is often used because of the simplicity of constructing the multiple regression equation.

The results of predicting water levels at gauging stations and a comparison of these methods with the proposed solution of the authors are presented in section 4.

3. Development of a method for predicting water levels based on a second generation recurrent neural network

One of the main parameters of the possible impact of the flood situation in a certain territory (for example, the Republic of Bashkortostan) is \( H \) — the level of water rise in water bodies, measured daily at n-posts by employees of the regional department of hydrometeorology and environmental monitoring. We introduce the following notation: \( H_{ji}^k \) is the water level measured at the \( k \)-th post on the \( i \)-th day of the \( j \)-th year. Here \( i = 1, n \), \( j = 1, \ldots, n \), \( n \) is the number of measuring posts involved in the calculations, \( j \) is the number of the year, \( i \) is the specific measurement date.

The task of forecasting is to calculate the water level value for the next \( i+1 \) day on a specific current \( i \)-th day of measurement, i.e. \( H_{ji+1}^k \), or after 2 days on \( i+2 \) day, i.e. \( H_{ji+2}^k \) or after \( l \) days on the \( i+l \)-th day, i.e. \( H_{ji+l}^k \) for any \( k \).

To solve this problem, it is proposed to use the results of previous measurements of the water level \( H_{ji}^k \) at all control posts located in the considered territory (in our case, the Republic of Bashkortostan) for all previous years. The proposed forecasting method is based on the construction of a recurrent neural network, the structure and algorithm of work (with training stages) of which are presented in fig. 1 and fig. 2: Initially, data (gauging stations codes, dates, water levels) enter the input layer of the neural network for further processing in the intermediate link (layers) of the RNN for the purpose of training. Teaching without a teacher is based on the integration of the methods of back propagation of error and Rosenblatt for a more accurate forecast of water level at hydrological posts. At the output of the neural network, we obtain the predicted values of the water levels for a given period of days.
\[ C(i) = (\dot{C} \cdot (i)) \cdot (\dot{C} \cdot (i + 1)) \cdots (\dot{C} \cdot (i + N - 1)), \]
\[ ((C, (i - P)) \cdots (C) \cdot (i + (P + 1)) \cdots (C, (i + 1)))). \]  \tag{14}

The neuron itself has a sigmoidal activation function with a range of values \((0, 1)\), described by the equation:
\[ \psi = \frac{1}{1 + e^{-\psi}} \]  \tag{15}

We denote \( U_i \) as the sum of the signals of each \( i \)-th neuron of all \( j \)-th intermediate layers, and \( g \) is the sum of the signals of each \( i \)-th neuron in the output layer, then
\[ U_i = \sum_{j=0}^{N} w_{ij} \cdot C_j, \]  \tag{16}
where \( w_{ij} \) is \( f(U_i) \). Respectively,
\[ g = \sum_{i=0}^{i} w_i \cdot C(i), \]  \tag{17}
where \( C(i) = f(g) \). Thus, the output is the result in the form of an output data signal \((I_i)\). A mathematical model of training a neural network for predicting water levels is presented in subsection 3.1.

### Neural network training for predicting water levels

The main idea of the training method with the back propagation of the error is to spread the error signals from the outputs of the RNN to its inputs. Partially, this method of training a recurrent neural network was used in predicting and countering cyber threats [3]. The initial task is to minimize the error function:
\[ E = \frac{1}{2} \sum_{j=1}^{p} (y_j - d_j)^2, \]  \tag{18}
where \( y_j \) is the obtained value of the \( j \)-th output of the RNN, a \( d_j \) is the reference value of the \( j \)-th output of the neural network. Accordingly, the minimization of \( E \) is determined by the gradient descent method. At the first stage, there is an automatic adjustment of the weight coefficients of the synapses:
\[ \Delta W_{ij} = -m \cdot \frac{df}{dw_{ij}}, \]  \tag{19}
where \( w_{ij} \) is the synapse weight, \(-m(0 < m < 1)\) is the neural network learning rate, expressed by a coefficient. Next, it is necessary to disclose (8) for more accurate training of the neural network. In this way,
\[ \frac{df}{dw_{ij}} = \frac{dy_j}{dy_i} \cdot \frac{dy_i}{ds_i} \cdot \frac{ds_i}{dw_{ij}}, \]  \tag{20}
where \( s_i \) is the sum of the input signals of each neuron RNN, and \( y_i \) is the output of the \( i \)-th neuron. Respectively, \( \frac{dy_j}{dy_i} = \sum_k \frac{dy_j}{dy_k} \cdot \frac{dy_k}{ds_k} = \sum_k \frac{dy_j}{dy_k} \cdot \frac{dy_k}{ds_k} \cdot \frac{ds_k}{uw_{ij}}, \)  \tag{21}
where \( k \) is the number of neurons in the \( n+1 \) layer. Also for disclosure (19) we introduce a new variable:
\[ \delta_{j}^{n+1} = \frac{dy_j}{dy_i} \cdot \frac{dy_i}{ds_i} \]  \tag{22}

After entering the variable, we obtain recursive formulas for the output (23) and input (24) layers of the RNN:
\[ \delta_{i}^{n} = \frac{dy_i}{ds_i} \cdot (y_i^{n} - d_i), \]  \tag{23}
\[ \delta_{j}^{n} = \frac{dy_j}{dy_i} \cdot \sum_k \delta_{k}^{n+1} \cdot w_{ij}^{n+1}. \]  \tag{24}

Thus, the disclosed formula (25) for automatically adjusting the weight coefficients of the synapses is as follows:
\[ \Delta W_{ij} = -m \cdot \delta_{j}^{n+1} \cdot y_i^{n-1}. \]  \tag{25}

In the case of the Rosenblatt method, everything is different, since it was originally intended for the training of a single-layer perceptron. Thus, the rule for training a recurrent neural network is as follows:
\[ q_{ij}(t + 1) = q_{ij}(t) + a_1 \cdot x_i \cdot d_j \]  \tag{26}

Next, it is necessary to integrate the training rules (26), the error minimization functions (18) and the weight adjustment of the synapse coefficients (25) in order to increase the accuracy of training and, as a result, the predicted values:
\[ H^k_{n+1} = \left( q_{ij}(t+1) \cdot \delta_{j}^{n+1} \cdot (y_i^{n+1} - d_i)^2 \right) \frac{H^k_n}{S_{ij}} \]  \tag{27}

Thus, based on the Rosenblatt methods and the back propagation of the error, formula (27) was obtained, which allows one to increase the accuracy of forecasting water levels at hydrological posts (the results are presented in Section 4).

### 4. Analysis of the effectiveness of the proposed method for predicting water levels

To analyze the effectiveness of the proposed method for predicting water levels, we used long-term data on measuring water levels at gauging stations provided by the Bashkir Administration for Hydrometeorology and Environmental Monitoring (Bashhydromet) from 01.01.2000 to 15.05.2019 years in the form of: code of a hydrological post, date and water level at the gauging station.

During the experiment, many iterations were carried out to calculate the predicted levels of water rise for the entire period of long-term observations – from 01.01.2001 to 31.12.2019. The total array of data used in the experiment is 13 800 (hydrological station code, date and water level), of which 66%, these are data from long-term observations from 01.01.2000 to 31.12.2013, fed to the input neural networks, and the remaining 34% (01.01.2014 – 31.03.2019) – for training. From 34% of the data, the last 10 days are taken to predict water levels for the next 5 days, and the rest of the information is used to analyze and improve the accuracy of the forecast as part of the training.

As an example, the article presents the results of an experiment predicting the levels of water rise at three hydrological posts in Ufa, Shakhia district of Ufa and the village of Okhlebinino, Iginsky district of the Republic of Bashkortostan during the spring flood of 2019. The experiment was carried out in two stages. The first stage was a comparison of real and predicted levels of water rise at hydrological posts at different periods of the flood – when there is a rise in water levels, a peak period and a decline. The results are shown in tables 1-3, where the actual water level is actually the measured value at the hydrological station, and the predicted water level is the water level value obtained using the neural network.

The predicted values of water levels by the neural network at gauging stations at the beginning of the flood period (increase (rise) in water level) are presented in Table 1.

<table>
<thead>
<tr>
<th>The date</th>
<th>Hydropost number</th>
<th>Real water level, cm.</th>
<th>Predicted water level, cm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>01.04.19</td>
<td>76289</td>
<td>-11.00</td>
<td>0.00</td>
</tr>
<tr>
<td>02.04.19</td>
<td></td>
<td>8.00</td>
<td>20.00</td>
</tr>
</tbody>
</table>
The predicted values of the water levels by the neural network at gauging stations during the flood peak are presented in Table 2.

**Table 2.** Predicted values for five days of water levels at gauging stations during the flood peak

<table>
<thead>
<tr>
<th>The date</th>
<th>Hydropost number</th>
<th>Real water level, cm.</th>
<th>Predicted water level, cm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>03.04.19</td>
<td>(Ufa, Belaya River)</td>
<td>39.00</td>
<td>34.00</td>
</tr>
<tr>
<td>04.04.19</td>
<td></td>
<td>63.00</td>
<td>50.00</td>
</tr>
<tr>
<td>05.04.19</td>
<td></td>
<td>74.00</td>
<td>86.00</td>
</tr>
<tr>
<td>01.04.19</td>
<td></td>
<td>124.00</td>
<td>150.00</td>
</tr>
<tr>
<td>02.04.19</td>
<td></td>
<td>131.00</td>
<td>147.00</td>
</tr>
<tr>
<td>03.04.19</td>
<td>3000014 (microdistrict Shaksha, Ufa)</td>
<td>134.00</td>
<td>160.00</td>
</tr>
<tr>
<td>04.04.19</td>
<td></td>
<td>134.00</td>
<td>161.00</td>
</tr>
<tr>
<td>05.04.19</td>
<td></td>
<td>160.00</td>
<td>180.00</td>
</tr>
<tr>
<td>01.04.19</td>
<td>76288 (Iglinsky district, the village of Okhlebinino)</td>
<td>291.00</td>
<td>300.00</td>
</tr>
<tr>
<td>02.04.19</td>
<td></td>
<td>270.00</td>
<td>290.00</td>
</tr>
<tr>
<td>03.04.19</td>
<td></td>
<td>281.00</td>
<td>305.00</td>
</tr>
<tr>
<td>04.04.19</td>
<td></td>
<td>296.00</td>
<td>312.00</td>
</tr>
<tr>
<td>05.04.19</td>
<td></td>
<td>323.00</td>
<td>330.00</td>
</tr>
</tbody>
</table>

The predicted values of the water levels by the neural network at gauging stations during the flood period are presented in Table 3.

**Table 3.** Predicted values for five days of water levels at gauging stations during the flood period

<table>
<thead>
<tr>
<th>The date</th>
<th>Hydropost number</th>
<th>Real water level, cm.</th>
<th>Predicted water level, cm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>29.04.19</td>
<td>76289 (Ufa, Belaya River)</td>
<td>151.00</td>
<td>190.00</td>
</tr>
<tr>
<td>03.05.19</td>
<td></td>
<td>140.00</td>
<td>180.00</td>
</tr>
<tr>
<td>01.05.19</td>
<td></td>
<td>166.00</td>
<td>199.00</td>
</tr>
<tr>
<td>02.05.19</td>
<td></td>
<td>188.00</td>
<td>205.00</td>
</tr>
<tr>
<td>03.05.19</td>
<td></td>
<td>182.00</td>
<td>200.00</td>
</tr>
<tr>
<td>01.04.19</td>
<td>3000014 (microdistrict Shaksha, Ufa)</td>
<td>313.00</td>
<td>300.00</td>
</tr>
<tr>
<td>02.04.19</td>
<td></td>
<td>204.00</td>
<td>200.00</td>
</tr>
<tr>
<td>03.04.19</td>
<td></td>
<td>275.00</td>
<td>260.00</td>
</tr>
<tr>
<td>02.05.19</td>
<td></td>
<td>313.00</td>
<td>300.00</td>
</tr>
<tr>
<td>03.05.19</td>
<td></td>
<td>303.00</td>
<td>295.00</td>
</tr>
<tr>
<td>01.04.19</td>
<td>76288 (Iglinsky district, the village of Okhlebinino)</td>
<td>365.00</td>
<td>390.00</td>
</tr>
<tr>
<td>02.04.19</td>
<td></td>
<td>387.00</td>
<td>400.00</td>
</tr>
<tr>
<td>03.05.19</td>
<td></td>
<td>405.00</td>
<td>420.00</td>
</tr>
<tr>
<td>02.05.19</td>
<td></td>
<td>364.00</td>
<td>380.00</td>
</tr>
<tr>
<td>03.05.19</td>
<td></td>
<td>329.00</td>
<td>350.00</td>
</tr>
</tbody>
</table>

Thus, the error between the predicted and actual values of the water levels, according to the results of the experiment, is 3-20% (tables 1-3).

The second stage of the experiment consisted in a comparative analysis of the data of the levels of water rise obtained using the described forecasting method based on the recurrent neural network and the known, most described in the literature sources, forecasting methods (least squares method, numerical methods and regression models).

As an example, the article presents the results of comparing the actual levels of water rise at the Ufa gauging station during the flood peak according to Bashhydromet and the forecast levels of water rise obtained on the basis of calculations by known methods (table 4).

An important difference of the proposed method for predicting water levels using a recurrent neural network in comparison with other known methods is the speed of obtaining the forecast and its correctness (more accurate) when forecasting ahead of time (for 5 days). The remaining methods considered in the experiment are more accurate only with short-term forecasting (1-2 days), the analysis results are shown in fig. 3.

![Fig. 3. An example of the results of forecasting water levels at hydrological posts](image-url)
According to the data in Table 4 and Fig. 3, it can be seen that in the first two days the forecast of water rise levels by known methods is quite accurate: the difference between the predicted values from the real ones varies from 1 to 3%. However, in the following days, the gap between the predicted and real values increases: for example, when calculating by the least squares method, the error reaches 34.5%, by numerical method – 36%, and for a linear regression model this indicator is 47.5%. It is worth noting that the implemented recurrent neural network does not give an accurate result in the first two days (in contrast to the known forecasting methods considered), but in the following days shows a more stable result.

<table>
<thead>
<tr>
<th>The date</th>
<th>Hydropost number</th>
<th>Real water level, cm.</th>
<th>Predicted values (least squares method), cm.</th>
<th>Predicted values (numerical method), cm.</th>
<th>Predicted values (general regression model), cm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.04.19</td>
<td>76289 (Ufa, Belaya River)</td>
<td>350.00</td>
<td>390.00</td>
<td>360.00</td>
<td>353.00</td>
</tr>
<tr>
<td>16.04.19</td>
<td>Ufa, Belaya River</td>
<td>368.00</td>
<td>410.00</td>
<td>370.00</td>
<td>378.00</td>
</tr>
<tr>
<td>17.04.19</td>
<td>(Ufa, Belaya River)</td>
<td>366.00</td>
<td>470.00</td>
<td>420.00</td>
<td>460.00</td>
</tr>
<tr>
<td>18.04.19</td>
<td></td>
<td>346.00</td>
<td>490.00</td>
<td>540.00</td>
<td>540.00</td>
</tr>
<tr>
<td>19.04.19</td>
<td></td>
<td>315.00</td>
<td>480.00</td>
<td>280.00</td>
<td>600.00</td>
</tr>
</tbody>
</table>

5. Conclusions

A method is proposed for early detection of threats (for example, the Republic of Bashkortostan) for parrying them in complex distributed systems. The proposed forecasting method is based on the construction of a recurrent neural network, the structure and operation algorithm of which is described in the article. The results of the analysis of the effectiveness of the proposed method for predicting water levels showed an error of predicted values from 3 to 20%. A comparative analysis of the data of the levels of water rise obtained using the described forecasting method based on the recurrent neural network and the well-known, most described in the literature sources, forecasting methods (least squares method, numerical methods and regression models), which revealed errors in other forecasting methods up to 47.5%. Thus, the use of artificial neural network technology has shown more stable results in forecasting threats, using the example of spring flood, which will allow the special services to give the necessary time for flood control measures to prepare for the protection of complex distributed (including technical objects) systems.

Acknowledgments

The reported study was funded by RFBR, project number 20-08-00301.

References


About the authors

Evgeny V. Palchevsky, PhD student, Department of Geographic Information Systems, Ufa State Aviation Technical University. E-mail: teelxp@inbox.ru.

Olga I. Khristodulo, Doctor of Technical Sciences, Professor, Department of Geoinformation Systems, Ufa State Aviation Technical University. E-mail: o-khristodulo@mail.ru.

Sergey V. Pavlov, Doctor of Technical Sciences, Professor, Department of Geoinformation Systems, Ufa State Aviation Technical University. E-mail: psvgis@mail.ru.