

Intelligent Data Analysis for Forecasting Threats in Complex Distributed Systems*

Evgeny Palchevsky^[0000-0001-9033-5741], Olga Khristodulo^[0000-0002-3987-6582],
Sergey Pavlov^[0000-0001-9672-7623] and Artur Kalimgulov^[0000-0002-9969-0712]

Ufa State Aviation Technical University, Ufa 450008, Russia
teelxp@inbox.ru, o-hristodulo@mail.ru, psvgis@mail.ru,
kalimgulovartur@gmail.com

Abstract. A threat prediction method based on the intellectual analysis of historical data in complex distributed systems (CDS) is proposed. The relevance of the chosen research topic in terms of considering the flood as a physical process of raising the water level, which is measured at stationary and automatic hydrological posts, is substantiated. Based on this, a mathematical formulation of the problem is formulated, within the framework of which an artificial neural network based on the freely distributed TensorFlow software library is implemented. The analysis of the effectiveness of the implemented artificial neural network was carried out, according to which the average deviation of the predicted water level when forecasting for one day at a stationary hydrological post was 3.323%. For further research on forecasting water levels, an algorithm is proposed for evaluating historical data at automatic posts, which will allow using these data to predict water levels according to the proposed method and at automatic posts. Thus, the neural network allows predicting the flood situation with acceptable accuracy, which allows special services to take measures to counter this threat.

Keywords: Complex Distributed Systems, Threat Prediction, Data Mining, Neural Networks, Flood Situation, Water Level Prediction

1 Introduction

The ever-more rapid development of digital technologies allows us to take a fresh look at the interaction of various components of the real world: physical, biological, social, etc. Moreover, digital technologies themselves become the most important component of this world, and these technologies allow people to influence the interaction of all components among themselves and especially to people. This complex interconnected set of objects of various complexity and physical nature can be con-

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

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

sidered as a complex distributed system (CDS), which is characterized by a significant remoteness of the component components from each other and a quick change in their characteristics over time.

The components of such CDS themselves are complex distributed systems and have (or potentially can have), including negative impact on each other, that is, they constitute or pose a threat to each other. For example, such a biological (natural) component of the CDS as a river network when a natural (spring or rain flood) or artificial (dam destruction) high water rise occurs on it can have a negative impact on such technical components of the CDS as power lines, piping systems, etc., or on the social components of the CDS – places of residence and recreation of people. Today, one of the urgent tasks is the development of various, as a rule, highly computerized, means and methods of countering various threats based on forecasting the development of processes that form the basis of these threats.

This article discusses the task of countering one of the types of threats - flood, based on the physical process of raising the water level in water bodies, and capable of causing significant material damage to individual components of the CDS in the territory of the subject of the Russian Federation (for example, the Republic of Bashkortostan). To counter this threat, complex technical systems with an increasingly significant digital component are created and intensively developed, which allow monitoring and predicting water levels in water bodies. These systems are based on the integration of modern technical means of obtaining the information necessary for monitoring (measuring the level and temperature of water and air, water flow rate, etc.) and modern information technologies for processing this information for forecasting (analysis of large, including poorly structured data; distributed databases data; the Internet of things (or the Internet of everything); artificial intelligence, etc.) and can be classified as cyberphysical systems.

Traditionally, the water level is measured at stationary hydrological observation posts (gauging stations) of the regional department of hydrometeorology and environmental monitoring (Bashhydromet) and planning of measures to counter the flood threat is carried out on the basis of the forecast of the water level according to these data (for these posts). Recently, water bodies have additionally begun to install automatic water level monitoring posts equipped with video cameras for early detection and recording of a sharp rise in the water level, which is dangerous for the components of the CDS. It would be logical to use the values of water levels obtained from automatic posts as additional data for forecasting water levels in water bodies, however, there are no historical data needed for forecasting.

There are various approaches to solving the problem of predicting water levels in water bodies, based on the analysis of hydrogeological and hydrodynamic parameters of the state of water bodies [1–3], a number of works by domestic and foreign authors are devoted to various aspects of the development of methods in this area [4–6]. However, in these works, insufficient attention was paid to the problem of predicting water levels based on the intellectual analysis of retrospective data, including from automatic posts using artificial neural networks (ANNs).

2 The use of artificial intelligence technologies to predict the levels of water rise during the spring flood

One of the main parameters of the possible negative impact of the flood situation on various components of the emergency control system is h – the water level in water bodies, measured daily at n stationary hydrological posts by Bashhydromet employees. We introduce the notation: h_{ji}^k is the value of the water level measured at the k -th hydrological post on the i -th date of the j -th year. Here $k = \overline{1, n}$, where n is the number of hydrological posts involved in the calculations, j is the number of the year, i is the specific measurement date. At the stage of making a short-term forecast of water levels, the forecasting task is to calculate the value of the level of water rise on the next $i+1$ day, that is, h_{ji+1}^k , for any $k = \overline{1, n}$ on a specific current i -day of measurement.

To solve this problem, it is proposed to use the results of previous measurements of the level of water rise h_{ji}^k at all stationary hydrological posts located in the territory under consideration in the same climatic and hydrological conditions for all previous years.

In each specific territory (including the Republic of Bashkortostan), various water bodies can be in different conditions that affect the nature of the development of the flood situation. For example, in one part of the territory the river flows in the mountains, and in another part – along the plain. Also, on the same territory, different rivers can belong to different basins with different hydrological and climatic characteristics. Therefore, to predict the water level at a particular gauging station, it is logical to use data on water levels at those hydrological posts that are in the same conditions with it and the number of hydrological posts participating in the forecasting can be less than n . Determining the uniformity of conditions in which hydrological posts are located is an independent scientific task, and in this article, for simplicity of presentation of the forecasting method, but without violating the generality of reasoning, we assume that data from n hydrological posts are used.

A similar remark must be made with respect to the possible values of the index i denoting the date of measurement. In fact, the value of the date itself for forecasting is not significant, since in different years at different posts the rise in the water level and the return of the water level to the normal value occur on different days. The serial number of the measurement since the start of the flood is significant. The number of measurements, which corresponds to the duration of the flood situation, for each fixed pair of values of the indices k and j is also different, therefore we introduce its notation l_{kj} . In the framework of the comments made, we denote the entire set of previously measured water level values as

$$H1 = \{h_{ji}^k\}, k = \overline{1, n}; j = \overline{1, m}; i = \overline{1, l_{kj}}, \quad (1)$$

where m is the number of years of observations. But in the current $(m+1)$ th year, by the time the flood starts and during the flood, measurements are also taken and can be

used to predict the values of the water level on each specific day on the next $(i_0 + 1)$ th day, therefore into consideration another set

$$H2 = \{h_{m+1,i}^k\}, k = \overline{1, n}; i = \overline{1, t_0}. \quad (2)$$

In this case, of course, it is important to choose the day the flood starts, from which the forecasting starts (that is, to which day $i=1$ corresponds). This article is not considered in this article, since it affects the specific values of l_{kj} , the number and list of elements of the sets $H1$ and $H2$, which in turn affects the time for calculating forecast values, but does not affect the method and forecasting algorithm itself.

It is also necessary to introduce an additional notation for the predicted value of the water level – hp , since in the future, to evaluate the effectiveness of forecasting methods, it will be necessary to use the notation h and hp together.

Based on the introduced notation, predicting the water level in the current $m+1$ year at a fixed hydrological station k_0 , $1 \leq k_0 \leq n$ at some fixed point in time t_0 will be described one day in advance by a certain function of the set of previously performed measurements:

$$hp_{m+1,i_0+1}^{k_0} = f(H1, H2). \quad (3)$$

There are many approaches and methods for constructing this function, using all or part of the $H1$ and $H2$ data, including various statistical [7–9], hydrological [10] and intelligent methods [11–13].

Currently, the so-called artificial intelligence methods and, in particular, the methods of constructing artificial neural networks (ANNs) are widely used to solve various problems [14, 15]. In this paper, it is proposed to predict the future value of the water level using an artificial neural network with training without a teacher based on the integration of the back propagation method of error and the Rosenblatt method, which are carried out in two stages (Fig. 1).

At the first stage, which is carried out before the development of the flood situation, the parameters are selected and ANNs are trained based on existing values from the set $H1$. The result of the training are the values of the weight coefficients of the synapses, which are subsequently used for forecasting. At the second stage, the forecast values of hp_{m+1,i_0}^k are calculated daily for all n observation posts ($k = \overline{1, n}$) using a trained ANN.

Since ANNs only process data that varies in the range $[0,1]$, it is necessary to convert (normalize or normalize) all measured (archived and current) water levels according to the most common ratio:

$$\vec{h}_{ji}^k = \frac{h_{ji}^k - h_{min}}{h_{max} - h_{min}}, \quad (4)$$

where $h_{min} = \min_{k,j,i}(h_{ji}^k)$ and $h_{max} = \max_{k,j,i}(h_{ji}^k)$, for all $k = \overline{1, n}; j = \overline{1, m}; i = \overline{1, l_{kj}}$.

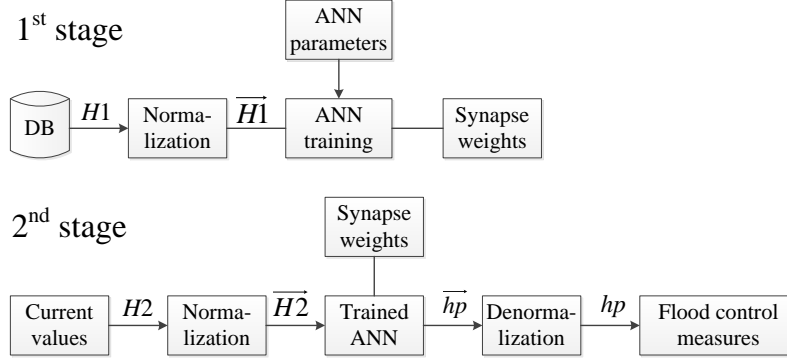


Fig. 1. Scheme of using a neural network for prediction.

Subsequently, normalized values of \vec{h}_{ji}^k are fed to the input of the ANN both at the training stage and at the forecasting stage, as a result of which the result (predicted value) is also normalized. Therefore, before applying the predicted values to counter the threat of flood, they are denormalized (reduced to the usual values of the water level) in a ratio that is the opposite (4):

$$h_{ji}^k = \vec{h}_{ji}^k \cdot (h_{max} - h_{min}) + h_{min}. \quad (5)$$

3 Analysis of the effectiveness of the proposed approach for predicting water levels at stationary hydrological posts

To predict water levels, the authors developed the software module “Forecaster” [16], based on the use of the freely distributed library of machine learning programs “TensorFlow” [17, 18]. The analysis of the effectiveness of applying the proposed approach for predicting water levels was carried out on retrospective (for the last 20 years) data at stationary gauging stations for the period from January 1, 2000 to May 22, 2019. The total amount of data is 22,341, of which 66% (data long-term observations of water levels h_{ji}^k from 01.01.2000 to 31.12.2014) are fed to the input sample of an artificial neural network to create an image of an array of input data for direct analysis for the purpose of further training, and the remaining 34% (01.01.2015–22.05.2019) – for training.

Every day, since the beginning of monitoring the development of the flood situation (in 2020 it was April 18, that is, $i=1$), using the Forecaster program, the forecast for the next day was made for 3 hydrological posts (that is, $n=3$; this value is n taken to reduce the time of the experiment), that is, it was determined

$$hp_{m+1,i+1}^k = \text{Forecaster}(H1, H2). \quad (6)$$

The next day, the actual value of the water levels at the same stations $h_{m+1,i+1}^k$ was measured and the predicted value was compared with the actual value based on the value of the relative difference, which most often characterizes the forecast accuracy:

$$E_i^k = \frac{(hp_{m+1,i+1}^k - h_{m+1,i+1}^k)^2}{h_{m+1,i+1}^k}, \quad (7)$$

for all hydrological posts $k = \overrightarrow{1, n}$. Carrying out forecasting daily until May 17 (that is, $l_{k,m+1} = 30$), we obtain 30 forecasts for each post and actual values of water levels for each of n posts, which makes it possible to determine the average forecast accuracy for each k -th post:

$$E^k = \frac{1}{l_{k,m+1}} \sum_{i=1}^{l_{k,m+1}} E_i^k, \quad (8)$$

and for all posts for the entire forecasting period in 2020:

$$E = \frac{1}{n} \sum_{k=1}^n E^k. \quad (9)$$

In fig. 2 shows the actually measured and predicted water levels in 2020 at one of the observation posts. The calculation of the accuracy of the forecast showed the value of average accuracy for each post $E1 = 2.900\%$; $E2 = 3.511\%$; $E3 = 3.560\%$, and the average forecast accuracy for the entire flood period of 2020 is $E = 3.323$, which corresponds to the forecasting accuracy by other methods [2].

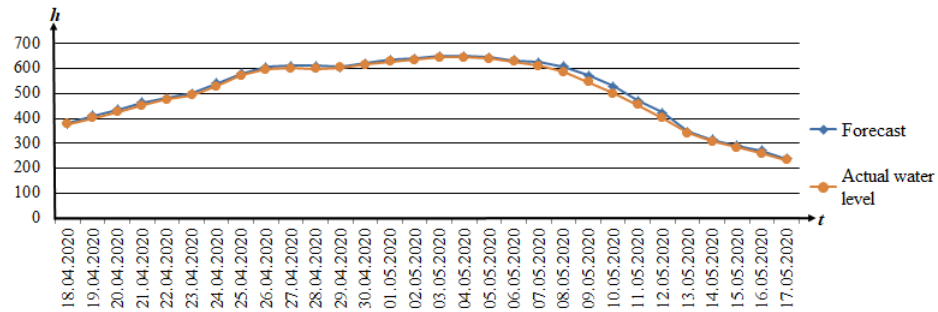


Fig. 2. An example of the results of forecasting levels of water rise at a stationary gauging station 76289 (Ufa, Belaya river).

Thus, the use of artificial intelligence technologies in the form of a recurrent neural network gives a sufficiently accurate result in the framework of forecasting water levels at stationary hydrological posts, which allows the relevant services to quickly respond to this threat and take the necessary measures to counter it.

4 Using data from automatic posts to predict the level of water rise

In connection with the emergence of new technical capabilities for automatic measurement of the state of the CDS (water level, temperature, wind speed and direction, etc.), automatic stations for measuring and recording water level in water bodies have recently been increasingly used [19]. As a rule, they have departmental affiliation different from Roshydromet divisions: they are part of the structure of local or regional executive authorities that deal with the prevention and counteraction of threats. The main purpose of these stations is the early detection of a threat and informing the governing bodies and the population about it, at the same time, the data from these posts can be used to predict a possible threat according to the technique proposed in the previous paragraphs of this article. To use this technique, there are no archival measurements at automatic posts (after all, they simply did not exist before), therefore, it is proposed to introduce an estimate of the water level at the locations of automatic hydrological posts in the past (since, despite the fact that there were no automatic posts, the water level in this point was at some point and the regularity of its (level) change is the same as at stationary hydrological posts) based on archival values at neighboring stationary gauging stations. This offer can be used only for those automatic posts that are located between stationary hydrological posts (one upstream and one lower), and the water level value is interpolated for them. It is necessary to take into account the fact that at stationary gauging stations one value of the water level is measured discretely every day at a fixed point in time (usually at 10 a.m. local time), and at automatic posts the water level is measured continuously. For the correct application of relations (1)–(6) (that is, for comparability of measurement results at automatic and stationary hydrological stations), the average of continuously measured values over 10 minutes is selected as the water level value at the k -th hydrological station for a specific date (from 955 to 1005) on this date.

Figure 3 shows the location of stationary hydrological posts (34 objects) and automatic posts (38 objects) at the water bodies of the Republic of Bashkortostan [20], an analysis of their relative position (Fig. 3) showed that 7 automatic posts (during the flood in 2020) are located between stationary gauging stations (table 1).

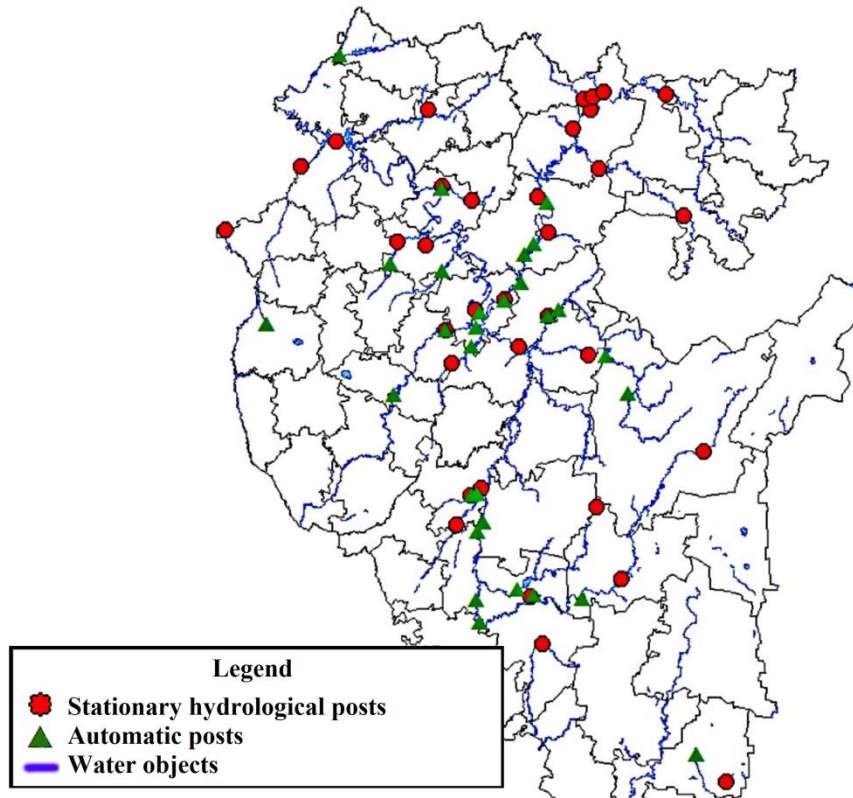


Fig. 3. Placement of stationary and automatic gauging stations on water bodies of the Republic of Bashkortostan.

Table 1. Automatic posts located between stationary gauging stations at water bodies of the Republic of Bashkortostan

#	Downstream hydrological post	Automatic post title	Upper hydrological post
1	Krasnaya Gorka village (Ufa river)	Red key	Pavlovskaya hydroelectric station, n. pool (r. Ufa)
2	Krasnaya Gorka village (Ufa river)	Yaman Port	Pavlovskaya hydroelectric station, n. pool (r. Ufa)
3	Andreevka village (Belaya river)	Birsk	Birsk city (Belaya river)
4	Lyakhovo village (Urshak river)	Bulgakovo	Okhlebinino village (Belaya river)
5	Sterlitamak city (Belaya river)	Sterlitamak (st. B. Khmel'nitsky)	Novofedorovskoe village (Ashkadar river)
6	Sterlitamak city (Belaya river)	Sterlitamak (Vodolazhenko St.)	the village of New Otradovka (p. Sterlya)
7	Meteli village (Ay river)	Bolsheustikinsky	Lakly village (Ay river)

Denote the total number of automatic posts located between two stationary posts by na . In our case, today $na=7$, but over time, automatic posts can be closed, moved to another place or liquidated, so na , generally speaking, is a variable. Since water level measurements at automatic posts (real and interpolated) were not included in the definition of the set $H1$ by relation (1), we denote ha_{ji}^k as the interpolated estimated value of the water level at the location of the k -th automatic post, $k = \overline{1, na}$, in the i -th day of the j -th year, where $j = \overline{1, m}$, $i = \overline{1, l_k}$, as before. Actual measurements at stationary posts adjacent to the selected k th automatic post are included in the set $H1$, but the order of the numbers of these posts in the set $H1$ and in table 1 do not coincide, therefore, to automate the calculations, we introduce additional notation. We denote by KD the set of numbers of stationary gauging stations located downstream, and by KU the set of numbers of stationary gauging stations located upstream of the corresponding automatic post so that

$$\begin{aligned} KD &= (kd_1, kd_2, \dots, kd_{na}), \\ KU &= (ku_1, ku_2, \dots, ku_{na}), \end{aligned} \quad (10)$$

and the automatic post with number k is located between two stationary posts with numbers kd_k and ku_k .

In the calculations (interpolation) of the ha_{ji}^k value, the distance between the automatic and neighboring stationary gauging stations is involved, therefore we introduce the variable x , which denotes the distance (distance) of the corresponding post from the river mouth, and, generally speaking, the water level at any point of the river can be considered as a function from this distance:

$$h = h(x). \quad (11)$$

For each k -th stationary post, this distance is a fixed number x_k , $k = \overline{1, n}$; accordingly, for automatic posts this distance is denoted by xa_k , $k = \overline{1, na}$. By virtue of the notation introduced, the location of some automatic post between two stationary hydrological posts is described by the relation

$$\begin{aligned} x_{kd_k} &< xa_k < x_{ku_k}, \\ h_{ji}^k &= h_{ji}^k(x_k), k = \overline{1, n}, \\ ha_{ji}^k &= h_{ji}^k(xa_k), k = \overline{1, na}. \end{aligned} \quad (12)$$

In the general case, this dependence has the same nature, defined by relation (11), but different notations are introduced to explain the interpolation algorithm for stationary and automatic posts.

Assuming that the change in the water level in the river between two points (the locations of gauging stations) occurs linearly (which is quite acceptable for small distances), the interpolated value of ha_{ji}^k is on a straight line connecting the water level points at neighboring stationary posts and is calculated (in accordance with known rules Euclidean geometry, see Fig. 4) by the relation

$$ha_{ji}^k = h_{ji}^{kd_k} + \left(h_{ji}^{ku_k} - h_{ji}^{kd_k} \right) \cdot \frac{xa_k - x_{kd_k}}{x_{ku_k} - x_{kd_k}}. \quad (13)$$

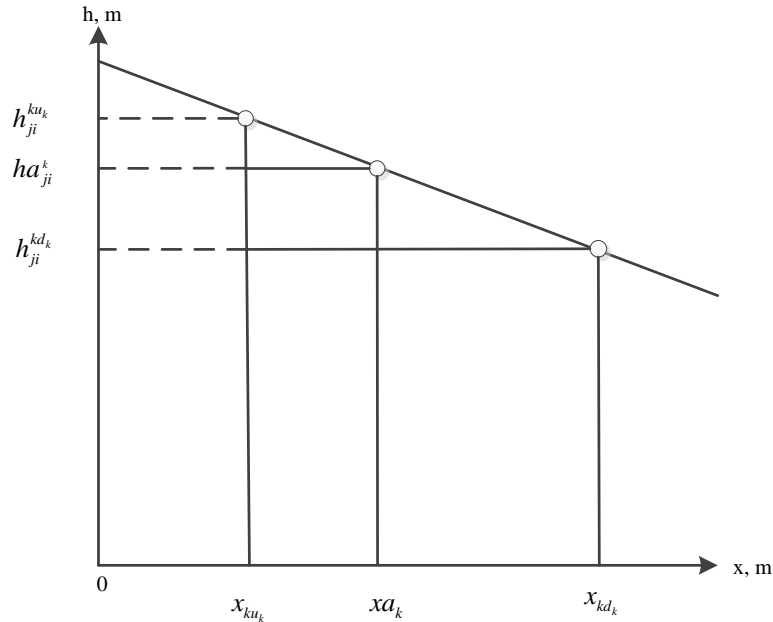


Fig. 4. Diagram of changes in water level at various (neighboring) gauging stations for a fixed date depending on their distance from the river mouth.

Fig. 5 shows an example of interpolation of the values of the level of water rise for the automatic hydrological post “Red Key”, based on data from neighboring stationary hydrological posts – “p. Krasnaya Gorka (Ufa River)” and “Pavlovskaya Hydroelectric Power Station, n. pool (Ufa River)”. As an example, we use the historical data of water levels at these hydrological posts from 04.05.2009; water level at the post “s. Krasnaya Gorka (p. Ufa)” was 197 cm, at the post “Pavlovskaya hydroelectric power station, n. pool (p. Ufa)” – 413cm. As a result of the interpolation for the “Red Key” automatic post, the water level values were obtained – 223 cm, which we will consider as historical data for this automatic post on 04.05.2009. The water level values thus obtained as a result of interpolation for each automatic post and for all the remaining dates will be considered historical data for this automatic post. The amount of these data is the same as the number of observations at neighboring stationary gauging stations. Now, the array of the estimated data can be used to predict water levels according to the method proposed in paragraph 2 using the ANN for all automatic posts included in table 1.

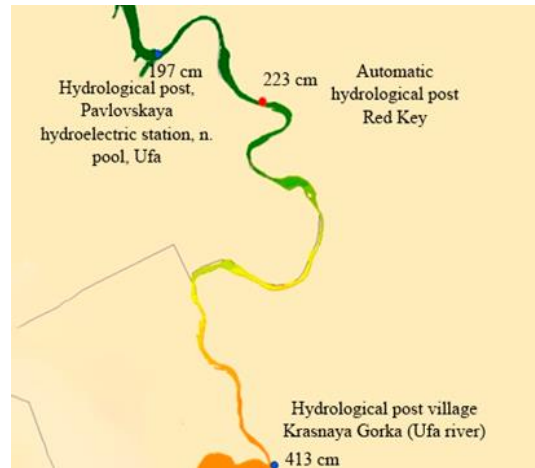


Fig. 5. The result of the interpolation of data on the levels of water rise for automatic post.

5 Conclusion

Solving complex socially and economically significant tasks for managing the CDS, which undoubtedly includes parrying such types of threats as floods, requires the integration of modern breakthrough technologies, such as data mining, distributed databases, the Internet of things and artificial intelligence, which are inextricable parts of a single process for obtaining and analyzing heterogeneous data on the state of the CDS and its individual components. The article proposes an approach that allows for the joint analysis and processing of data on water levels from automatic and stationary gauging stations to predict flood threats using artificial neural networks. The analysis of the results of applying the proposed approach for predicting the water level in the water bodies of the Republic of Bashkortostan during the 2020 flood showed its effectiveness, as it gives a fairly accurate forecast, and thereby allows the relevant services to quickly respond to this threat and take the necessary measures to counter it.

6 Acknowledgments

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

References

1. O.E.J. Wing, C.C. Sampson, P.D. Bates: A flood inundation forecast of Hurricane Harvey using a continental-scale 2D hydrodynamic model. *Journal of Hydrology* X 4, 1–17 (2019).

12 E. Palchevsky, O. Khristodulo, S. Pavlov, et al.

2. S.V. Borsch, Yu.A. Simonov, A.V. Khristoforov: Flood forecasting system and early warning of floods on the rivers of the Black Sea coast of the Caucasus and the Kuban basin. Hydrometeorological Research Center of the Russian Federation, Russia (2014).
3. M.Ya. Zdereva, V.F. Bogdanova, N.A. Khluchina: Evaluation of the possibility of using model precipitation forecasts for forecasting rain floods in the mountain rivers of Altai. Transactions of the Hydrometeorological Research Center of the Russian Federation 359, 128–141 (2016).
4. Y.V. Grebnev, A.V. Spring: Flood monitoring and forecasting in the Krasnoyarsk Territory using neural network algorithms. Siberian Fire and Rescue Bulletin 3(10), 13–16 (2018).
5. J. Noymanee, T. Theeramunkong: Flood Forecasting with Machine Learning Technique on Hydrological Modeling. Procedia Computer Science 156, 377–386 (2019).
6. Y. Zhou, S. Guo, F. Chang: Explore an evolutionary recurrent ANFIS for modelling multi-step-ahead flood forecasts. Journal of Hydrology 570, 343–355 (2019).
7. S. Han, P. Coulibaly: Bayesian flood forecasting methods: a review. Journal of Hydrology 551, 340–351 (2017).
8. S.V. Pavlov, R.R. Sharafutdinov, O. I. Khristodulo: Development of a geoinformation model of a river network taking into account cartographic, hydrological and morphometric information to determine the boundaries of flood zones when the water level in water bodies changes. Bulletin of USATU 1(11), 18–27 (2008).
9. A. Bukvic, J. Harrald: Rural versus urban perspective on coastal flooding: The insights from the U.S. Mid-Atlantic communities. Climate Risk Management 23, 7–18 (2019).
10. S.V. Pavlov, V.G. Krymsky, O.I. Khristodulo: Information support for assessing and managing water resources based on geoinformation technologies, Russia (2010).
11. V.G. Mokhov, V.I. Tsimbol: Electrical energy consumption prediction of the federal district of Russia on the basis of the recurrent neural network. Journal of computational and engineering mathematics 2(5), 3–15 (2018).
12. F. Policelli, A. Hubbard, H. Chul Jung, B. Zaitchik, C. Ichoku: A predictive model for Lake Chad total surface water area using remotely sensed and modeled hydrological and meteorological parameters and multivariate regression analysis. Journal of Hydrology 568, 1071–1080 (2019).
13. A. Moghar, M. Hamiche: Stock Market Prediction Using LSTM Recurrent Neural Network. Procedia Computer Science 170, 1168–1173 (2020).
14. L. Yang, Y. Huang, M. Xia, H. Li: Neural-network based boost pressure prediction for two-stage turbocharging system of diesel engine. IFAC-PapersOnLine 52, 178–184 (2019).
15. F. Yang, H. Zhang, X. Hou, Y. Tian, Y. Xu: Experimental study and artificial neural network based prediction of a free piston expander-linear generator for small scale organic Rankine cycle. Energy, vol. 175, 2019, pp. 630–644.
16. E.V. Palchevsky: Forecaster. Certificate of state registration of computer programs No. 2020611923 Russian Federation, 12.02.2020.
17. M. Olazaran: A Sociological History of the Neural Network Controversy. Advanced in Computers 37, 335–425 (1993).
18. TensorFlow, <https://www.tensorflow.org>. Last accessed 9 Jun 2020.
19. I.S. Belyuchenko: Regional monitoring – the scientific basis of nature conservation. Ecological bulletin of the North Caucasus 1(2), 25–40 (2006).
20. Flood System 2.0, <http://portal.introgis.com/CoGIS/FloodRB>. Last accessed 5 Jun 2020.