Early Detection and Prediction of Some Threats in Complex Distributed Systems Based on Data Mining

Artur Gizatullin[1], Andrey Ivantsov[1], Alexander Pavlov[1], Sergey Pavlov[1] and Olga Khristodulo[1]

[1] *Ufa State Aviation Technical University, K. Marx Street, 12, Ufa, 450008, Russia*

Abstract

A method for predicting threats in complex distributed systems is proposed, based on the intelligent analysis of large data arrays on the results of monitoring changes in water level in water bodies and air temperature at the measurement point, which makes it possible to increase the efficiency of planning and implementing measures to fend off such and similar threats. The method is based on general approaches and mathematical models previously used by the authors to develop adaptive algorithms for controlling gas turbine engines, which is especially relevant in the context of the increasingly widespread introduction of automatic means for monitoring the state of complex distributed systems and the exponential growth in the number of data used to support decision-making. The choice of the future value of the water level at the measurement point is carried out based on the results of processing the data accumulated for all previous flood periods on the compliance of the water level and its changes per day with the values of air temperature and its changes for the same day. The results of an experimental assessment of the accuracy of predicting the water level in the water bodies of the Republic of Bashkortostan in the flood period of 2021 are presented, which confirm the applicability of the proposed forecasting method to support decision-making to fend off threats in complex distributed systems from a sharp rise in water.

Keywords

Forecasting, threats, complex distributed systems, data mining, spring flood, decision support

1. Introduction

The increasing use of new, highly automated means of monitoring the development of biophysical processes in complex distributed systems (CDS), in which the possibility of quickly obtaining and processing a large number of parameters characterizing their state is realized, it becomes possible to use (with appropriate processing and adaptation) well-proven models and management methods (including tasks of monitoring, predicting the state and fending off threats) by technical systems.

The authors of this article have extensive experience in developing and using statistical methods for processing a large number of measured parameters for controlling such complex technical systems as gas turbine engines (GTE) [1, 12-14]. The analysis of the development of some threats in the CDS, which include a large number of objects that are different in nature and significantly remote from each other, showed the possibility of using these data analysis methods to formally describe the dependence of the parameters determining the state of the CDS on the most significant factors and subsequent threat forecasting based on the revealed dependence.

GraphiCon 2021: 31st International Conference on Computer Graphics and Vision, September 27-30, 2021, Nizhny Novgorod, Russia
EMAIL: gizartur@yandex.ru (Artur Gizatullin); andreyiv0508@gmail.com (Andrey Ivantsov); asp.gis@gmail.com (Alexander Pavlov); psvgis@mail.ru (Sergey Pavlov); o-hristodulo@mail.ru (Olga Khristodulo)
ORCID: 0000-0003-1951-0484 (Artur Gizatullin); 0000-0003-1129-9506 (Andrey Ivantsov); 0000-0001-9206-9870 (Alexander Pavlov); 0000−0001−9672−7623 (Sergey Pavlov); 0000−0002−3987−6582 (Olga Khristodulo)

As an example, let's consider one of the most common types of threats to the security of the population and territory in the Republic of Bashkortostan – the spring flood (which is more often called a flood), which consists in flooding and underflooding of individual territories and objects located on them due to the rise of water in water bodies due to snow melting. The size of the flooded territories (their boundaries, area and depth) depends on the water level in water bodies, measured at stationary posts of Roshydromet, as well as at increasingly widely used automatic monitoring stations owned by local authorities. The height of the water rise at each of these observation posts (at a specific point in the territory) depends on many natural and man-made factors. The main natural factors include: water reserves in the soil, the depth of freezing of the soil, water reserves in the snow, the area and depth of snow cover, air temperature and other meteorological parameters, natural ice jams and forest blockages. Technogenic factors include: planned or emergency water discharges from hydraulic structures located upstream above the measurement point, construction of engineering structures on the water bodies themselves (bridges, water crossings of pipelines, etc.) or near them (dams, artificial reservoirs, embankments, etc.), ice congestion (resulting from human activity). Some of these factors have a long-term impact on the possibility of water rising, and the other part is short-term.

2. A method for predicting threats in complex distributed systems based on the intelligent analysis of large data arrays (on the example of the problem of predicting changes in the water level in water bodies)

Earlier in their works the authors of this article proposed one of the possible approaches to predicting future changes in the water level, based on changing artificial neural networks for intelligent analysis of the measured values of only the water level. In this paper, we consider the problem of operational assessment of changes in the water level (forecast) for one day ahead, under the influence of the most significant factors at stationary observation posts. At each of these posts, the water level is measured daily *h*, and at each of these posts, the water level is measured daily and the task of the operational forecast is to measure at a specific time t_i for each post, determine the future value (for the next moment in time t_{i+1}) the water level, which we will denote *hp*. It should be noted here that the predicted (future) value of the water level hp_{i+1} differs from the value actually measured in a day h_{i+1}, therefore, a special designation is introduced for it

$$hp_{i+1} \neq h_{i+1} . \quad (1)$$

In this paper, it is proposed to determine the future value of the water level based on the analysis of its changes in similar conditions in the past, while it is assumed that the main factor influencing a sharp rise in the water level (namely, it poses a threat to the objects of the CDS) is a sharp warming, that is, a large (sharp) change in air temperature per day. In other words, the change in the water level at a particular measurement point most significantly depends on how much the air temperature has changed at this point. Given that the modern system of meteorological observations and weather forecasting gives a fairly accurate forecast of air temperature changes for 1-3 days ahead, to identify the dependence of the water level on changes in air temperature and then use the identified dependence to predict the water level, these predicted values can be used as an actual change in air temperature.

It is proposed to select the future value of the water level hp_{i+1} based on the results of processing the data accumulated for all previous flood periods on the compliance of the water level and its changes per day with the values of air temperature and its changes for the same day. The analyzed data are measured at equidistant time points t_i air temperature values T_i and the water level h_i [2]. Since the forecast consists in determining the future value, that is, the value of the change in the water level is calculated $hp_i = h_i + \Delta h_i$ depending on the temperature change $TP_i = T_i + \Delta T_i$ then, for the implementation of the proposed method of forecasting additional, changes in the water level Δh_i and temperature ΔT_i are calculated

$$\Delta h_i = h_{i+1} - h_i,$$
$$\Delta T_i = T_{i+1} - T_i. \quad (2)$$

The action of various natural and man-made factors, examples of which were given above), differently leads to a change in the water level at the measurement point in accordance with changes

in temperature at the same point during the same day, and for forecasting it is necessary to determine the statistical dependence Δh_i, from the corresponding values h_i, T_i, ΔT_i in the form of some function

$$\Delta h = f(h, T, \Delta T), \tag{3}$$

and the use of this dependence in the future to determine (calculate) future *hp* values.

Let's represent all the values of the parameters $h, T,\ \Delta h\ and\ \Delta T$ measured at each individual observation post for the previous time period as a set

$$W = \{W_i\}_{i=\overline{1,p}}, \tag{4}$$

where is each element of the set

$$W_i = \{h_i, T_i, \Delta h_i, \Delta T_i\}_{i=\overline{1,p}} \tag{5}$$

it represents the measured values of the parameters at the i-th moment of time t_i, p is the total number of observations.

The ranges of possible changes in each of these parameters are divided into a fixed number of segments $h^0, h^1, \dots, h^{M_1}$;

$$\begin{gathered} T^0, T^1, \dots, T^{M_2}; \\ \Delta h^0, \Delta h^1, \dots, \Delta h^{M_3}; \\ \Delta T^0, \Delta T^1, \dots, \Delta T^{M_4}, \end{gathered} \tag{6}$$

where M^1, M^2, M^3, M^4 is the number of segments of the partition of the possible values of the corresponding parameter.

Based on the analysis of data from long-term observations of the flood situation (that is, sets (4) and (6)), a new set is constructed

$$N = \{N_{jklm}\}_{j=\overline{1,M_1},k=\overline{1,M_2},l=\overline{1,M_3},m=\overline{1,M_4}}, \tag{7}$$

each element of which shows the number of elements of the set (4) satisfying the following conditions:

$$\begin{gathered} h^{j-1} < h_i \le h^j; \\ T^{k-1} < T_i \le T^k; \\ \Delta h^{l-1} < \Delta h_i \le \Delta h^l; \\ \Delta T^{m-1} < \Delta T_i \le \Delta T^m, \end{gathered} \tag{8}$$

when running through the index *i* of all possible values, $i = \overline{1,p}$ [3-6].

In other words, the number N_{jklm} represents the rate of change of water level Δh that fall in the interval $[\Delta h^{l-1}, \Delta h^l]$ that occurred while the values of the parameters $h, T\ and\ \Delta T$, respectively trapped in segments $[h^{j-1}, \Delta h^j]$, $[T^{k-1}, T^k]$, $[\Delta T^{m-1}, \Delta T^m]$.

At the prediction stage, the current values of the parameters $h_i\ and\ T_i$ are measured at each specific moment of time t_i. As already noted above, modern methods of forecasting air temperature allow predicting changes in air temperature with acceptable accuracy, so at this point in time, the future predicted value is known

$$\Delta Tp_i = TP_{i+1} - T_i, \tag{9}$$

which we will consider actual, that is, we suppose

$$\Delta T_i = \Delta Tp_i. \tag{10}$$

Next, the numbers of the segments of the partition (6) are determined, in which the current values $h_i, T_i, \Delta T_i$ fall, that is, the current values of the indices j_T, k_T, m_T are determined, for which

$$\begin{gathered} j_T: h_i \in \left[h^{j_T-1}, h^{j_T}\right], \\ k_T: T_i \in [T^{k_T-1}, T^{k_T}], \\ m_T: \Delta T_i \in [\Delta T^{m_T-1}, \Delta T^{m_T}]. \end{gathered} \tag{11}$$

A new set *N1⊂N* is formed from the elements of the set *N*

$$N1 = \{N1_l\}_{l=\overline{1,M^3}}, \tag{12}$$

$$N1_l = N_{j_T k_T l m_T}, l = \overline{1, M^3} \tag{13}$$

which is the set of frequencies of occurrence of Δh at the values of the other three parameters satisfying the relations (11). As the predicted value of the water level change, it is proposed to choose the middle of the segment of the partition from (6) by Δh for which the frequency of occurrence of such a value Δh is the greatest, that is, the segment satisfying the condition is selected as the current value of the l_T index

$$l_T: N1_{l_T} = \max N1_l, l = \overline{1, M^3}, \quad (14)$$

as the predicted value of the water level change, the following is selected

$$\Delta hp_i = \frac{1}{2}\left(\Delta h^{l_T-1}, \Delta h^{l_T}\right), \quad (15)$$

and the predicted value of the water level at the next time t_{i+1} is determined by a simple ratio

$$hp_{i+1} = h_i + \Delta hp_i. \quad (16)$$

Upon the occurrence time of the next control water levels are measured actual values of T_{i+1}, h_{i+1} are computed and actual values $\Delta T_i \; and \; \Delta h_i$ that allows you to adjust the value of one element of the set *N*, the corresponding segments of the split ranges of parameters (8), which hit the actual value of the item $(h_i\ , T_i, \Delta h_i, \Delta T_i)$, by increasing its value by one. This means that the frequency of occurrence of the four values $(h_{j_T}, T_{k_T}, \Delta h_{l_T}, \Delta T_{m_T})$ increases by one.

$$\begin{gathered} h_i \in \left[h^{j_T-1}, h^{j_T}\right]; \\ T_i \in [T^{k_T-1}, T^{k_T}]; \\ \Delta h_i \in [\Delta h^{l_T-1}, \Delta h^{l_T}]; \\ \Delta T_i \in [\Delta T^{m_T-1}, \Delta T^{m_T}]; \end{gathered} \quad (17)$$

$$N_{j_T k_T l_T m_T} = N_{j_T k_T l_T m_T} + 1 \quad (18)$$

and for each subsequent prediction, a set *N* with updated values of its elements is used, that is, in the process of conducting a flood situation, the process of training the forecasting model continues.

3. Experimental verification of the applicability of the proposed forecasting method for planning and conducting measures to counter threats

The accuracy of the forecast using this forecasting method was studied during the flood in the Republic of Bashkortostan in 2021. For each stationary observation post of the Federal Hydrometeorological Service (there are 41 of them in the Republic), on the basis of archival data on the observation of water level and air temperature values (about 12 thousand values of each parameter in total) and calculated values of daily changes in these parameters (also about 12 thousand values of each of them), a set N was built according to the above algorithm. The number of segments of splitting the possible values of each of the parameters was assumed to be equal 10: $M^1 = M^2 = M^3 = M^4 = 10$. On each i - th day of a flood situation, the beginning of which is characterized by a significant rise in the water level, based on the measured values $h_i \; and \; T_i$ and value forecast ΔTp_i, carried out the forecast values Δhp_i and calculation hp_{i+1} by the ratio (16). [7-8].

These values of $\Delta hp_i \; and \; hp_{i+1}$ for each of the observation posts were transmitted to the Ministry of Emergency Situations, where they were used for planning and carrying out measures to fend off the flood threat to the population and territory (including all infrastructure and industrial facilities) [9-10].

On the next *i+1*-th day, when the actual value of $h_{i+1}$was obtained, the weighted average quadratic error of the forecast was calculated

$$E_{i+1} = \frac{(hp_{i+1}-h_{i+1})^2}{h_{i+1}}. \quad (19)$$

At the end of the flood, the average forecast error for the entire flood was calculated for each observation post

$$E = \frac{1}{p}\sum_{i=1}^{p} E_i, \quad (20)$$

where p is the number of days of monitoring the flood situation. An example of the correspondence of the forecast and actual values of the water level at one of the observation posts is shown in Figure 1. For this example, the average forecast error was *E=0.031*, which is in good agreement with other forecasting methods and is applicable to support decision-making on planning and conducting measures to fend off threats from a flood situation.

As noted at the beginning of this article (in the introduction), the main motive for the use of methods of intelligent processing of large data arrays, which have shown their high efficiency in managing complex technical systems (for example, aviation gas turbine engines), was the increasingly widespread use of highly automated monitoring tools for the development of processes dangerous to the population and territories in the CDS. All this has a direct bearing on the control of the development

of the flood situation. To date, the main source of information for the early detection and parrying of flood threats is the stationary posts of the hydrometeorological service, which measure the necessary parameters once a day, and often by non – automated methods with low accuracy. That is, a relatively small amount of not quite accurate data is used to support decision-making, and in these conditions, the use of the proposed approaches is limited by the small amount and low accuracy of the available data.

This year, the pilot operation of automatic flood control stations was carried out, and next year it is planned to put them into commercial operation with the gradual decommissioning of "manual" monitoring tools. At the same time, continuous dynamic control of all necessary parameters is carried out and the possibility of their continuous use arises, similar to how it has been happening for a long time for technical objects and systems. In this case, the amount of data used will increase by several orders of magnitude (from one value per day to 24-240 or more), which will lead to even greater efficiency and demand for the methods proposed in this article.

A significant increase in the number of data available for analysis and processing will improve the accuracy of the forecast by increasing the numbers $M^j, j = \overline{1,4}$, since this will reduce the length of the segment determined by the ratio (14) for calculating the predicted value from the ratios (15) and (16). At the same time, the time required for the actual calculation of the forecast will practically not change even with a significant increase in the amount of analyzed data, since the forecast itself is still calculated by the ratios (14-16). The computational load will increase only at the stage of training the prediction model, that is, calculating the elements of sets N and N1, due to checking a large number of inequalities (8). In other words, the time for training the model will increase (in the above experiment it was about 10 minutes), but there will still be less time for implementing measures to parry the predicted threats, that is, it will still be quite acceptable [11].

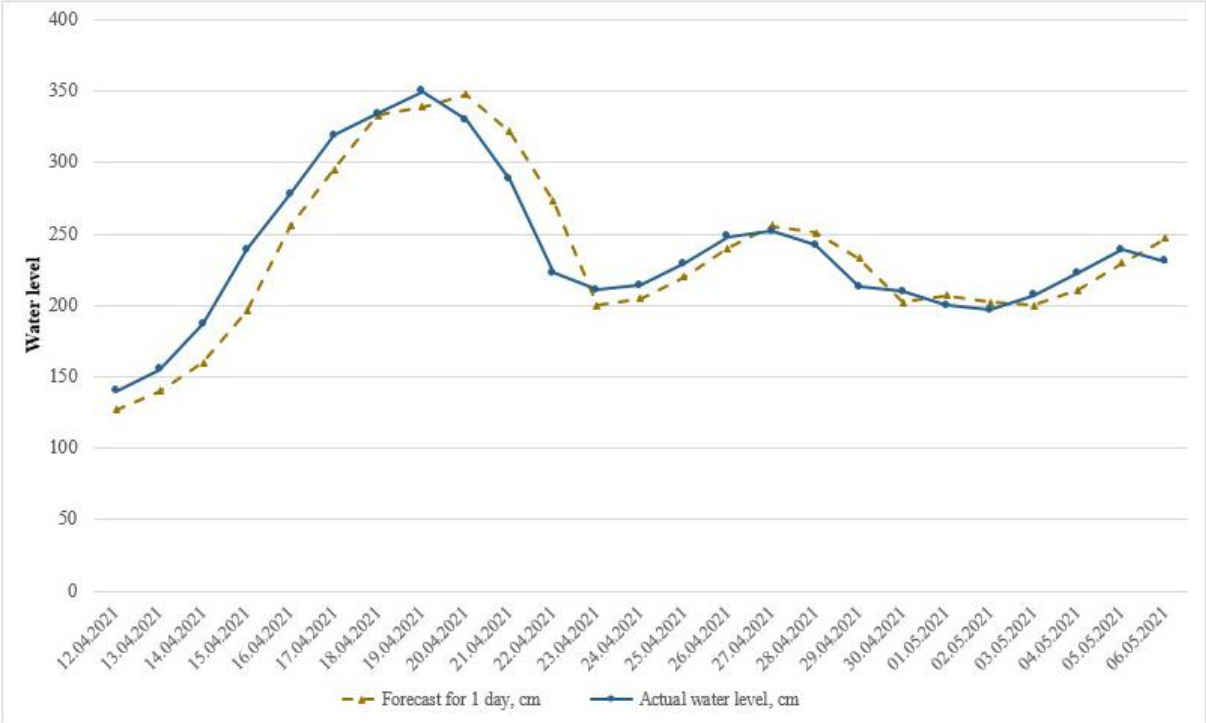

Figure 1: Changes in the forecast and actual values of the water level at the hydrological post No. 76275, Belaya River, Arsky Kamen

4. Conclusion

The proposed method of intellectual analysis of data on the results of monitoring and forecasting the development of a flood threat in complex distributed systems, which is a development of the

previously proposed adaptive method of controlling gas turbine engines, showed a fairly high accuracy of predicting the water level in reservoirs. The use of this method by the relevant authorities for the early detection and prediction of flooding of territories and economic and vital objects located on them will allow for more effective planning and implementation of measures to fend off this type of threats, and similar ones. The timeliness of the development and implementation of such methods of intelligent analysis of large arrays of measurement information is confirmed by the increasingly widespread use of automatic means for monitoring the state of the CDS and their individual objects and subsystems, which leads to an exponential increase in the number of data used to support decision-making.

5. Acknowledgements

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

6. References

[1] G. G. Kulikov, S. V. Pavlov, V. V. Stepanov, Design of adaptive automatic control systems for gas turbine engines as objects with aftereffect, Issues of designing information and cybernetic systems, Ufa, UAI (1987), pp. 53-61. In Russian.

[2] E.V. Palchevsky, O.I. Khristodulo, S.V. Pavlov, A.M. Kalimgulov, Intelligent data analysis for forecasting threats in complex distributed systems, CEUR Workshop Proceedings. Vol. 2744, 2020, pp. 285-296. DOI: 10.51130/graphicon-2020-2-3-79

[3] J. Noymanee, T. Theeramunkong, Flood Forecasting with Machine Learning Technique on Hydrological Modeling, Procedia Computer Science 156 (2019) 377–386. DOI: 10.1016/j.procs.2019.08.214

[4] 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, Transactions of the Hydrometeorological Research Center of the Russian Federation 356 (2015) 1–247.

[5] S. Han, P. Coulibaly, Bayesian flood forecasting methods: a review, Journal of Hydrology 551 (2017) 340–351. DOI: 10.1016/j.jhydrol.2017.06.004

[6] Y.V. Grebnev, A.V. Spring, Flood monitoring and forecasting in the Krasnoyarsk Territory using neural network algorithms, Siberian Fire and Rescue Bulletin 3 (2018) 13–16.

[7] A. Bukvic, J. Harrald, Rural versus urban perspective on coastal flooding: The insights from the U.S. Mid-Atlantic communities, Climate Risk Management 23 (2019) 7–18. DOI: 10.1016/j.crm.2018.10.004

[8] Y. Zhou, S. Guo, F. Chang, Explore an evolutionary recurrent ANFIS for modelling multi-step-ahead flood forecasts, Journal of Hydrology 570 (2019) 343–355. DOI: 10.1016/j.jhydrol.2018.12.040

[9] 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 (2016) 128–141.

[10] V.G. Mokhov, V.I. Tsimbol, Electrical energy consumption prediction of the federal district of Russia on the based of the reccurent neural network, Journal of computational and engineering mathematics 5(2) (2018) 3–15. DOI: 10.14529/jcem180201

[11] 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 11(1) (2008) 18–27.

[12] G. G. Kulikov, V. V. Antonov, M. A. Shilina, Mathematical and the Software for Creation and Implementation of Subject-Oriented Management Information Systems from Conditions of Identifiability and Traceability, Bulletin of the South Ural State University 16 (2016) 143-15. DOI: 10.14529/ctcr160316

[13] G. G. Kulikov, V. A. Trushin, A. I. Abdulnagimov, Monitoring of the parameters of the thermally stressed state of the turbine blades of an aviation gas turbine engine and assessment of their residual resource, Bulletin of the Ufa State Aviation Technical University 21 (2017) 100-104.

[14] G.I. Pogorelov, G.G. Kulikov, A.I. Abdulnagimov, B.I. Badamshin, Application Of Neural Network Technology And High-Performance Computing For Identification And Real-Time Hardware-In-The Loop Simulation Of Gas Turbine Engines, Proceedings of the 3rd International Conference on Dynamics and Vibroacoustics of Machines, DVM 2016 (2017), pp. 402-408. DOI: 10.1016/j.proeng.2017.02.338