Advances in Intelligent Systems Research, volume 174
Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)

A Randomized Algorithm for Restoring Missing Data in the Time series of Lake Areas Using Information on Climatic Parameters

Yuri S. Popkov*
Federal Research Center "Informatics and Control" RAS
Moscow, Russia
popkov.yuri@gmail.com

Evgeny S. Sokol
Ugra Research Institute of Information Technologies
Khanty-Mansiysk, Russia
eugen137@gmail.com

Andrey V. Melnikov
Ugra Research Institute of Information Technologie,
Khanty-Mansiysk, Russia
melnikovav@uriit.ru

Vladimir Y. Polishchuk
Institute of Monitoring of Climatic and Ecological Systems
Tomsk, Russia,
liquid_metal@mail.ru

Yury M. Polishchuk
Ugra Research Institute of Information Technologies
Khanty-Mansiysk, Russia
yupolishchuk@gmail.com

***Abstract*—In the tasks of predicting the volumes of methane emissions from thermokarst lakes in the Arctic territories, as one of the causes of modern global warming, it is necessary to use, along with climatic characteristics, data on the dynamics of lake areas, which are usually obtained using satellite imagery. Due to the large number of cloudy days in the northern territories, it is possible to obtain only a small number of cloudless images, which leads to significant omissions in the time series of lake areas. To restore the missing values of the area of the lakes, it is proposed to use a new approach to the restoration of missing values based on the methods and algorithms of entropy-randomized machine learning. The work is supposed to restore the missing values in the experimental data on the areas of thermokarst lakes using time series of average annual temperature and annual precipitation. As experimental data on the thermokarst lakes areas and climatic parameters (temperature and amount of precipitation), we used the results of studies conducted in the Arctic zone of Western Siberia from 1973 to 2007. Studies were conducted in nine test sites selected in different permafrost zones (continuous, discontinuous and insular). Data on the average annual temperature and annual precipitation for each test site were obtained by reanalysis. The developed algorithm for recovering missing values within the framework of this approach is implemented using the MATLAB R2019a tools. The missing values are calculated for the selected nine test sites. To illustrate, the time series of the values of the area of lakes, temperature and precipitation in one of the test sites are shown. An analysis of the omissions recovery errors was carried out, which showed that the developed algorithm allows us to restore the missing values of the lake areas from the data on changes in temperature and precipitation with practically acceptable accuracy.**

Keywords—randomization, machine learning, entropy criteria, thermokarst lakes, climatic parameters, modeling, restoration of missing data

I. INTRODUCTION

The randomized approach is of particular importance for solving the problems of predicting the dynamics of the accumulation of greenhouse gases in the thermokarst lakes of the Arctic zone in connection with their influence on global climatic changes. Solving these may be the basis for the development and functioning of adaptation systems [1] to changing environmental conditions at various control lazy levels.

With the coming global warming in the coming decades, the processes of thawing of frozen rocks will accelerate [2], leading to an additional release of methane as a vital product of microorganisms that process thawed organic matter. This is capable of making an additional tangible contribution to climate warming, which raises global public concern. Awareness of this was the reason for the adoption of the Paris Climate Agreement (2016), which envisages the development in different countries of measures capable of preventing an increase in the average annual temperature of the Earth by more than 1.5 ° C until 2050. The development of such measures at the regional level for the Arctic regions is impossible without the formation of reasonable forecasts of methane and carbon dioxide emissions, which can be based on knowledge of the spatio-temporal dynamics of lake fields [3,4] in the regions. The tasks of predicting the dynamics of accumulation of greenhouse gases in thermokarst lakes for the next decades require the use of data on the time series of lake areas and climatic parameters (temperature, precipitation).

Due to the high degree of bogging of the Arctic territories, data on lake areas can only be obtained using satellite imagery [4,5]. Due to the large number of cloudy days in the northern territories, it is possible to obtain cloudless images only in some years. As a result of this, the obtained time series of lake areas have a significant number of missing values. Issues of restoring gaps in time series of the lake area data are currently underdeveloped. The use of enthropy-randomized methods, which have shown, according to [6-8], is highly effective in solving problems of the global economy, demography, etc., is considered the most promising approach to restoring passes in our conditions. However, the methodological issues of restoring missing values in the time series of lake areas within the framework of entropy-randomized approach have not been developed, which was the purpose of this work.

II. DATA

In Fig. 1 shows the layout of test sites (TS) for conducting studies aimed at obtaining data on the time series of areas of

thermokarst lakes, average annual temperature, and annual precipitation.

Landsat medium-resolution satellite images (30 m) obtained from 1973 to 2007 were used to collect the data on lake areas. Methodological issues for conducting these studies are described in [4]. Data on climatic characteristics were obtained using data reanalysis procedures [3].

For example in Tab. 1 shows data on the average area of thermokarst lakes $\tilde{S}_m$, average annual temperature $\tilde{T}$, annual precipitation $\tilde{R}$ at TS - 1 from 1998 to 2007 years.

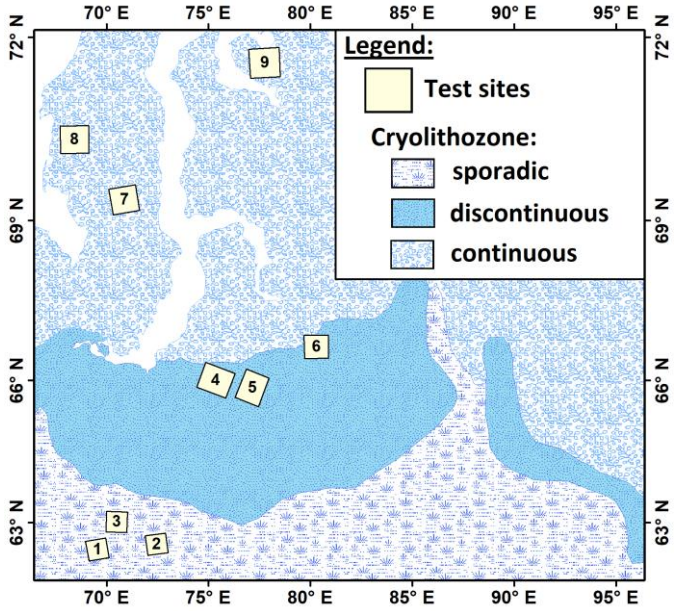

Fig. 1. The layout of the test sites in the study area of Western Siberia

As can be seen from the Tab. 1 (column 2), the data on the areas of lakes have a large number of missing values, which is characteristic of all the studied test sites.

TABLE I. DATA ON LAKE AREA, TEMPERATURE AND PRECIPITATION AT TS-1

Data / Years	$\tilde{S}_m$, ha	$\tilde{T}$, °C	$\tilde{R}$, mm
1	2	3	4
1998	-	-3,50	544,80
1999	-	-2,50	563,15
2000	-	-1,00	561,30
2001	62,36	-2,00	430,20
2002	-	-2,06	520,15
2003	62,31	-0,48	429,20
2004	-	-1,80	338,40
2005	-	0,44	343,90
2006	-	-2,78	222,00
2007	66,96	0,14	325,40

Consider the issues of preparing data for calculations. Data on the area $\tilde{S}$, temperature $\tilde{T}$, and precipitation $\tilde{R}$ are collected for each of the nine test sites studied. For calculations, we transform the data to the standard (normalized) form using the following formulas:

$$S = \frac{\tilde{S}-\tilde{S}_{min}}{\tilde{S}_{max}-\tilde{S}_{min}},$$

$$T = \frac{\tilde{T}-\tilde{T}_{min}}{\tilde{T}_{max}-\tilde{T}_{min}} \quad (1)$$

$$R = \frac{\tilde{R}-\tilde{R}_{min}}{\tilde{R}_{max}-\tilde{R}_{min}}.$$

where $\tilde{S}$, $\tilde{T}$, $\tilde{R}$ – average area, average annual temperature and the amount of annual precipitation in natural units (ha, ° C, mm, respectively), S, T, R - normalized values of area, temperature and precipitation (displayed on the interval [0,1]), the lower index is of all indicators means the minimum and maximum value of the sample.

III. MODEL AND ALGORITHMS

The technology of entropy-randomized forecasting is implemented in below following sequence of steps [6]. First, a predictive randomized parametric model (RPM) is formed, its defining parameters are synthesized and the necessary information support is agreed with it. RPM transforms an array of real input data

$$X = [x^1, \dots, x^s], \quad \xi(\varphi) \in R^n$$

into a model output characterized by a matrix

$$Z = [z(1),\dots,z(s)], \; z(j) \in R^m.$$

In the general case, this transformation is assumed to be dynamic, i.e. the model output observed at time j depends on the input observed on a certain historical interval $j - \varrho, \dots, j$, i.e. from the matrix $X_\varrho(j) = [x^{j-\varrho}, \dots, x^j]$. The mathematical expression of this connection is the vector functional $\hat{\Omega}(X_\varrho(j)|\alpha, P(\alpha))$ with random parameters α of the interval type

$$\alpha \in \mathcal{A} = [\alpha^-, \alpha^+]. \quad (2)$$

The probabilistic properties of the parameters are characterized by a probability distribution density (PDD) P(α), which is assumed to be continuously differentiable. The RPM output at the jth moment in time is an ensemble $\hat{Z}(j \mid P(\alpha))$ of random vectors

$$\hat{z}(j|\alpha) = \hat{\Omega}(X_\varrho(j)|\alpha, P(\alpha)), \qquad j = \overline{1,s}. \quad (3)$$

To simulate the influence of measuring errors, random noise $\xi \in R^m$ of an interval type is introduced at the output of the object:

$$\xi^j \in \Xi_j = [\xi^j_-, \xi^j_+], \quad j = \overline{1,s} \quad (4)$$

with continuously differentiable PDD functions $Q_j(\xi^j)$, $j = \overline{1,s}$, according to which the ensemble $\mathcal{F}\left(j \middle| Q_j(\xi^j)\right)$ is generated for each moment of measuring the output of the object. A set of random vectors (measurement noise) for the entire measurement interval is described by a matrix

$$K = \left[\xi^{(j)},\ j = \overline{1,s}\right], \quad (5)$$

which is characterized by the joint function PDD Q(K). If the noise in the measurements is statistically independent, then

$$Q(K) = \prod_{j=1}^{s} Q_j\left(\xi^{(j)}\right) \quad (6)$$

The observed RPM output can be represented as:

$$v\left(j|\alpha,\ \xi^{(j)}\right) = \hat{\Omega}(X_0(j)|\alpha, P(\alpha)),\ \ j = \overline{1,s}. \quad (7)$$

Random vectors (6) form ensembles, the mathematical expectation of which has the form:

$$\mathcal{M}\{v(j|\alpha,\xi^{(j)})\} = \int_{\mathcal{A}} \hat{z}(j|\alpha)P(\alpha)\,d\alpha + \int_{\Xi_j} Q_j\left(\xi^{(j)}\right)\xi^{(j)}\,d\xi^{(j)} =$$

$$= \mathcal{W}[P(\alpha), Q_j\left(\xi^{(j)}\right)],\qquad j = \overline{1,s}. \quad (8)$$

The second stage of the technology of randomized forecasting [6-9] is associated with the training of RPM. It is implemented using the following algorithm [10]:

$$[P^*(\boldsymbol{\alpha}), Q^*(K)] = arg\ max\ \mathcal{H}[P^*(\boldsymbol{\alpha}), Q^*(K)] \quad (9)$$

on the set of normalized functions $P^*(\boldsymbol{\alpha}), Q^*(K)$ for which the conditions of empirical balances are satisfied:

$$\mathcal{W}[P(\alpha), Q_j\left(\xi^{(j)}\right)] = y^{(j)},\ \ j = \overline{1,s} \quad (10)$$

where $y^{(j)} \in R^m$ is the vector of real measurements of the object's output.

Problem (9-10) belongs to the class of functional entropy-linear equations of the Lyapunov type [10], which have an analytical solution obtained using Lagrange factors $\Theta = [\theta^j, j = \overline{1,s}]$ (vectors $\theta^j \in R^m$):

$$P^*(\alpha) = \frac{\exp\left(-\sum_{j=1}^{s}\langle\theta^j,\ \hat{z}(j|\alpha)\rangle\right)}{\mathcal{P}(\Theta)},$$

$$Q_j^*\left(\xi^{(j)}\right) = \frac{\exp\left(-\langle\theta^j,\ \xi^{(j)}\rangle\right)}{\mathcal{Q}_j(\theta^j)},\ \ j = \overline{1,s};$$

$$Q_j^*\left(\xi^{(j)}\right) = \frac{exp\left(-\langle\theta^j,\ \xi^{(j)}\rangle\right)}{\mathcal{Q}_j(\theta^j)},\ \ j = \overline{1,s};$$

$$Q(K) = \prod_{j=1}^{s} Q_j^*\left(\xi^{(j)}\right). \quad (11)$$

The denominators of these expressions are normalization constants

$$\mathcal{P}(\Theta) = \int_{\mathcal{A}} \exp\left(-\sum_{j=1}^{s}\langle\theta^j,\ \hat{z}(j|\alpha)\rangle\right)\,d\alpha,$$

$$\mathcal{Q}_j(\theta^j) = \int_{\Xi_j} \exp(-\langle\theta^j,\ \xi^{(j)}\rangle)\,d\xi^{(j)},\ \ j = \overline{1,s} \quad (12)$$

The optimal PDD and normalization constants are parameterized by the Lagrange multipliers, which are determined by the solution of the following balance equations:

$$\frac{\mathcal{E}(\Theta)}{\mathcal{P}(\Theta)} + \frac{\mathcal{T}_j(\theta^j)}{\mathcal{Q}_j(\theta^j)} = y^{(j)},\quad j = \overline{1,s} \quad (13)$$

where

$$\mathcal{E}(\Theta) = \int_{\mathcal{A}} \hat{z}(j|\alpha)\exp\left(-\sum_{j=1}^{s}\langle\theta^j,\hat{z}(j|\alpha)\rangle\right)\,d\alpha, \quad (14)$$

$$\mathcal{T}_j(\theta^j) = \int_{\Xi_j}\xi^{(j)}\exp(-\langle\theta^j,\ \xi^{(j)}\rangle)\,d\xi^{(j)},\ \ j = \overline{1,s}.$$

In our case, the array of measured data on the lakes area has a large number of missing values. To restore the missing data, we use the principle of entropy randomization by area, using the available data on temperature and precipitation. It is known that temperature and precipitation affect the area of lakes, and to a first approximation this effect can be described by a linear dependence with noise in the form:

$$S[n] = \alpha T[n] + \beta R[n] + \xi[n]. \quad (15)$$

Coefficients α, β - random, interval:

$$\alpha \in \mathcal{A} = [\alpha^-, \alpha^+],\quad \beta \in \mathcal{B} = [\beta^-, \beta^+]. \quad (16)$$

Denote the PDD parameters $P(\alpha), F(\beta)$.

Noise is also standardized and interval:

$$\xi[n] \in \Xi_j = [\xi^-, \xi^+]. \quad (17)$$

Denote the PDD of noise $Q_n(\xi[n])$.

Further, using the algorithm of randomized machine learning, we obtain:

$$\mathcal{H} = -\int_{\mathcal{A}} P(\alpha)\ln P(\alpha)\,d\alpha - \int_{\mathcal{B}} F(\beta)\ln F(\beta)\,d\beta -$$

$$-\sum_{m=1}^{k}\int_{\Xi_m} Q_m(\xi[m])\ln Q_m(\xi[m])d\xi[m] \Rightarrow \max \quad (18)$$

under condition of normalization:

$$\int_{\mathcal{A}} P(\alpha)\,d\alpha = 1,$$

$$\int_{\mathcal{B}} F(\beta)\,d\beta = 1,$$

$$\int_{\Xi_m} Q_m(\xi[m])d\xi[m] = 1,\quad m = \overline{1,k} \quad (19)$$

and empirical balances:

$$\int_{\mathcal{A}} P(\alpha)\alpha T[m]\,d\alpha + \int_{\mathcal{B}} F(\beta)\beta R[m]\,d\beta +$$

$$+\int_{\Xi_m} Q_m(\xi[m])\xi[m]d\xi[m] = S[m],\quad m = \overline{1,k}. \quad (20)$$

The solution to problem (18) has the form:

$$P^*(\alpha,\theta) = \frac{\exp(-\alpha\, l_r(\theta))}{\mathcal{P}(\theta)},$$

$$F^*(\beta,\theta) = \frac{\exp(-\beta\, h_r(\theta))}{\mathcal{F}(\theta)}, \quad (21)$$

$$Q^*_m(\xi[m],\theta) = \frac{\exp(-\theta_m,\xi[m])}{\mathcal{Q}_j(\theta_m)},\ m = \overline{1,k}$$

where $\theta = \{\theta_1,\dots,\theta_k\}$ - Lagrange multipliers;
- normalization coefficients:

$$\mathcal{P}(\Theta) = \int_{\mathcal{A}} \exp(-\alpha l_r(\theta))\, d\alpha\ );$$

$$\mathcal{F}(\theta) = \int_{\mathcal{B}} \exp(-\beta h_r(\theta))\, d\beta \qquad (22)$$

$$\mathcal{Q}_j(\theta_m) = \int_{\mathcal{B}} \exp(-\theta_m \xi[m])\, d\xi[m],\quad m = \overline{1,k}$$

$$l_r(\theta) = \sum_{m=1}^{k} \theta_m T[m],\quad h_r(\theta) = \sum_{m=1}^{k} \theta_m R[m].$$

To determine the values of the Lagrange multipliers, it is necessary to solve the following system of equations:

$$L(\theta)\underset{1,k}{T(m)} + K(\theta)R(m) + G_m(\theta) = S(m), m = \qquad (23)$$

$$L(\theta) = \frac{\exp(-\alpha^- l_r(\theta))(\alpha^- l_r(\theta)+1)}{\exp(-\alpha^- l_r(\theta))-\exp(-\alpha^+ l_r(\theta))} +$$

$$+\frac{\exp(\alpha^+ l_r(\theta))(-\alpha^+ l_r(\theta)+1)}{\exp(\alpha^+ l_r(\theta))(-\alpha^+ l_r(\theta)+1)}$$

$$K(\theta) = \frac{\exp(-\beta^- h_r(\theta))(\beta^- h_r(\theta)+1)+}{\exp(-\beta^- h_r(\theta))-\exp(-\beta^+ h_r(\theta))} +$$

$$+\frac{\exp(\beta^+ h_r(\theta))(-\beta^+ h_r(\theta)+1)}{\exp(-\beta^- h_r(\theta))-\exp(-\beta^+ h_r(\theta))} \qquad (24)$$

$$G_m(\theta) = \frac{\exp(-\xi^-[m]\theta_m)(\xi^-[m]\theta_m+1)}{\exp(-\xi^-[m]\theta_m)-\exp(-\xi^+[m]\theta_m)} +$$

$$+\frac{\exp(-\xi^+[m]\theta_m)(\xi^+[m]\theta_m+1)}{\exp(-\xi^-[m]\theta_m)-\exp(-\xi^+[m]\theta_m)}.$$

Using the model to calculate all the missing data and sampling the PRV (24), we can construct an ensemble of trajectories S [n]. We calculate the average trajectory and fill in the missing data.

After that, we transform the data from normalized to natural values (ha). To do this, we perform the action inverse to (1):

$$\tilde{S}_{res} = S * (\tilde{S}_{max} - \tilde{S}_{min}) + \tilde{S}_{min}. \qquad (25)$$

IV. RESULTS

In accordance with the above algorithm, data was restored on all test sites. To illustrate, Fig. 2 shows the results of recovery data presented in the form of graphs of the time course of the restored values of the area of lakes in the test sites 3, 5 and 8. The hollow dots show the real (measured) values.

The restoration error was calculated as an estimate of the average deviation of the restored values from the measured data.

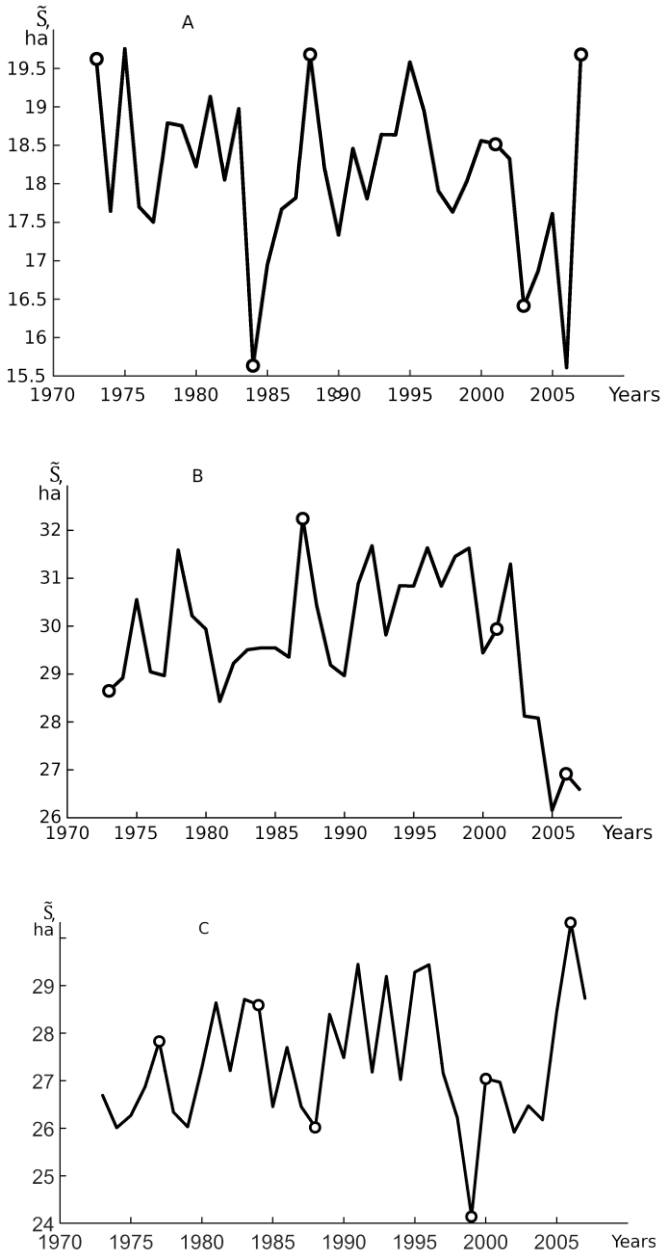

Fig. 2. Time series of values of the reconstructed data on the average area of lakes in three test sites 3, 5, and 8 in different permafrost zones: sporadic (A), discontinous (B), and continuous (C), respectively

For each test site, we calculate the average deviation of the restored area data from the measured values using the following formula:

$$\Delta_s = \frac{1}{n}\sum_{i=1}^{n} \frac{|\tilde{S}_{res_i} - \tilde{S}_{m_i}|}{\tilde{S}_{m_i}}, \qquad (26)$$

where $\tilde{S}_{res_i}$ is the value of the average area obtained as a result of measurements, $\tilde{S}_{m_i}$ are the restored values of the average area, n is the number of measured points on a each TS.

The data obtained are given in Tab. II.

TABLE II. RESTORATION ERRORS

TS	1	2	3	4	5	6	7	8	9	mean
Δ_s	0,04	0,15	0,08	0,04	0,04	0,06	0,04	0,07	0,04	0,05

In addition, the mean deviation of the model area values from the measured values in all test sites was calculated using the formula:

$$\Delta = \frac{1}{k}\sum_{i=1}^{k}\frac{|\tilde{S}_{res_i}-\tilde{S}_{m_i}|}{\tilde{S}_{m_i}}, \quad (27)$$

where k is the total number of all measured values of the average area for all test sites.

In our case, the following value was obtained: $\Delta = 0.06$, while the standard deviation calculated by the standard formula is 0.09. Consequently, the error in the recovery of gaps according to the developed algorithm does not exceed 9% in the mean square, which can be considered as a practically acceptable result.

V. CONCLUSION

The article discusses the methodological issues of a new approach to the restoration of missing values in experimental data on the areas of thermokarst lakes using time series of average annual temperature and annual precipitation. As experimental data on the areas of thermokarst lakes and climatic parameters (temperature and amount of precipitation), we used the results of studies conducted in the Arctic zone of Western Siberia from 1973 to 2007. Nine test sites were selected for research, three in each of the permafrost zones. Data on the area of lakes are the results of remote measurements from satellite images. Due to the large number of cloudy days in the north of Western Siberia during the indicated period, a small number of values of the average area of lakes in those years when there were cloudless images. Therefore, to restore the missed values of the time series of lake areas, the most prospective approach was based on entropy-randomized modeling.

An algorithm for recovering missing values within the framework of this approach is developed, implemented using the MATLAB R2019a tools. For illustration, time series of the area of lakes, temperature and precipitation in one of the test areas are shown. The analysis of the errors in the restoration of the passes showed that the developed algorithm allows us to restore the missing values of the lake areas from the data on changes in temperature and precipitation with practically acceptable accuracy.

ACKNOWLEDGMENT

The reported study was funded by the RFBR according to the research project No. 19-07-00282.

REFERENCES

[1] Y.Z. Tsypkin, Fundamentals of Learning Systems Theory, Moscow: Nauka, 1970. 398 p.

[2] Y.M. Polishchuk, A.N. Bogdanov, N.A. Bryksina, I.N. Muratov, V.Y. Polishchuk, M.A. Kupriyanov, O.A. Baysalyamova, and V.P. Dneprovskaya, "Experience and results of remote research of cryolithozone lakes in Western Siberia from satellite images of various resolutions over a 50-year period," Current problems of remote sensing of the Earth from the Space, vol. 14, no. 6, pp. 42–55, 2017.

[3] V.Y.Polishchuk, and Y.M.Polishchuk, Geo-simulation modeling of fields of thermokarst lakes in permafrost zones: monograph, Khanty-Mansiysk: Ugra State University Press, 2013, 129 p.

[4] V.Y. Polishchuk, and Y.M. Polishchuk, "Modeling of thermokarst lake dynamics in West-Siberian permafrost," in Permafrost: Distribution, Composition and Impacts on Infrastructure and Ecosystem, O.S. Pokrovsky, Ed. New York: Nova Science Publishers, 2014, vol. 6, pp. 205– 234.

[5] Y.M. Polishchuk, A.N. Bogdanov, and I.N. Muratov, "Methodological issues of constructing generalized histograms of the distribution of lake areas in the permafrost zone based on satellite images of medium and high resolution," Current problems of remote sensing of the Earth from Space, vol. 13, no. 6, pp. 224–232, 2016.

[6] Y.S. Popkov, A.Y. Popkov, and Y.A. Dubnov, Randomized machine learning under limited data, Moscow: URSS, 2019, 310 p.

[7] K.V. Vorontsov, Mathematical teaching methods on precedents: Lecture course, Moscow: MIPT, 2013, 245 p.

[8] M. Vidyasagar, "Statistical Learning Theory and Randomized Algorithms for Control," IEEE Control System Magazine, vol. l, no. 17, pp. 69-88, 1998.

[9] O.N. Granichin, and B.T. Polyak, Randomized estimation and optimization algorithms with almost arbitrary interference, Moscow: Nauka, 2002, 291 p.

[10] Y.S. Popkov, and A.Y. Popkov, "New method of entropy-robust estimation for ramdomized models under limited data," Enthropy, vol. 16, pp. 675-698, 2014.