Hydrology and meteorology

A TWO-STAGE HYBRID MODEL FOR HYDROLOGICAL TIME SERIES FORECASTING

Volchak A., Sidak S., Parfomuk S.
Brest state technical University
Str. Moscow f. 267, Brest, Republic of Belarus
E-mail: harchik-sveta@mail.ru

Annotation. A model based on the combined use of the empirical mode decomposition (EMD) method and the autoregressive integrated moving average (ARIMA) model is proposed to obtain forecast estimates of the maximum river runoff. The analysis and comparison of the results of modelling the maximum water runoff of the Dnieper River at the Rechitsa station allow concluding that the use of the hybrid EMD-ARIMA model is preferable to the classical ARIMA model. The decomposition approach of forecasting maximum water runoff series of the spring flood allows to take into account all its local features, internal structure, as well as abnormal discharges.

Keywords: empirical mode decomposition, Hilbert-Huang transform, hybrid model, EMD, ARIMA, river runoff forecast estimates.

INTRODUCTION

Obtaining forecast estimates of the maximum flow according to long-term observations is one of the urgent problems of hydrology. The availability of reliable information about the future values of the maximum river flow is one of the fundamental factors for effective planning, management and stable operation of the water resources system.

In the conditions of a changing climate and increasing anthropogenic pressure on water resources, the forecast assessment of maximum water consumption acts as an important economic factor. This factor allows business entities to choose an objective development strategy, take timely protective measures to prevent or minimize damage from adverse and dangerous hydrometeorological phenomena. The scientific problem of forecasting the maximum water runoff is of obvious importance from the point of view of design, construction and operation of hydraulic structures, implementation of measures to prevent the negative impact of water.

During the XX-beginning of the XXI century, methods for calculating and forecasting hydrological characteristics based on the study of the patterns of long-term flow fluctuations under the condition of stationary climate in the past and future were developed. However, at present, the validity and correctness of the application of these methods is being questioned in connection with climate changes caused by the processes of climate warming. The ongoing climate changes have already led to changes in the maximum flow of rivers. According to a number of studies, there was a significant decrease in the maximum water runoff on all major rivers of Belarus, as well as the repeatability of the maximum water runoff of low availability significantly decreased [1]. It is assumed that the main reason for the change in the maximum runoff in the territory of Belarus was the climate mitigation in winter and an increase in the frequency of winter thaws, as a result of which part of the spring runoff passes into the minimum winter runoff [2].

Increasing the requirements for the economic efficiency and safety of the water systems operation leads to the need to improve the existing methods of long-term forecasting of the maximum river flow, increasing its accuracy and timeliness.

Today a large number of mathematical models and methods of analysis and forecasting of hydrological time series have been developed [3]. The most commonly methods use: 1) correlation and regression analysis, classification and regression trees; 2) predictive extrapolation; 3) spectral analysis, wavelet analysis; 4) Markov chains; 5) pattern recognition theory; 6) neural networks, genetic algorithms.

Nowadays, hydrology often uses methods based on the construction of statistical distributions of maximum runoff rates based on the available series of observations of runoff and further extrapolation of these distributions to the region of small probabilities to assess the risk of spring floods. The use of this approach implies the fulfilment of the condition of uniformity and stationarity of the runoff observation series.

One of the most common classical statistical forecasting models is the ARIMA model proposed by Box and Jenkins [4]. This model is effectively applied in such industries as hydrology, economics, environment and politics. However, hydrological processes are complex, reflecting the interaction of a large number of climate-forming factors, each of which can be described by different models. The mathematical model of the hydrological process, constructed as a result of the application of traditional methods, as a single and indivisible one, is practically not feasible and is of little use for use in forecasting problems.

Recently, many works devoted to hybrid decomposition models and forecasting methods obtained by combining two or more methods in order to obtain the best characteristics of a combined hydrological model and the possibility of using them for forecasting non-stationary hydrological series [5]. The main goal of the decomposition approach to forecasting is to divide the initial time series into a set of series with a simpler structure, considered independently of each other.

The purpose of this article is to obtain forecast estimates of the maximum water runoff of the spring flood of the rivers in Belarus using a two-stage hybrid model based on the joint use of the methods of empirical mode decomposition and Box-Jenkins.

INITIAL DATA AND METHODS

The object of the study is the first largest and watery river of Belarus, the Dnieper. The Dnieper River flows through the territory of three countries: Russia, Ukraine and Belarus. The total length of the river is 2,145 km, almost 700 km of which is located in the territory of Belarus. The river originates on the Valdai Upland in Russia and flows into the Black Sea. The main right tributaries in the territory of Belarus are the Drut and the Berezina Rivers, the left is the Sozh River.

As a rule, the flood passes in one wave in the Dnieper River basin on the Dnieper, Sozh, and Berezina Rivers. During the spring flood, flooding of the floodplain is typical for most rivers of the Dnieper basin. The longest spring flooding in the Dnieper River basin was observed in 1956, 1958, 1962, 1970 and 1979. The last significant flood was in 1999.

The Dnieper River basin has an important natural and socio-economic significance due to the fact that socially significant natural resources (for example, water, land and forest resources) are concentrated in the territory of the basin, and it is also a valuable resource base for industrial enterprises, land users, water users, government structures, state control and regulatory bodies. The Dnieper River basin region in the territory of Belarus is developed both in industrial and agricultural terms, and therefore, the impact of surface and underground water resources on social development and the main sectors of the economy are significant.

All this justifies the relevance of studying the current changes in the maximum runoff of the Dnieper Basin Rivers, both in connection with its economic significance and from the influence of the changing climate on the characteristics of river flow.

The study uses data from hydrological observations of the maximum water runoff of the spring flood of the Dnieper River during the period of instrumental observations of the State Institution "Republican Center for Hydrometeorology, Control of Radioactive Contamination and Environmental Monitoring" of the Ministry of Natural Resources and Environmental Protection of the Republic of Belarus. The forecast estimates of the change in the maximum runoff of the Dnieper River due to the influence of climatic factors were obtained for the Rechitsa section. The gaps in the data series were restored using the computer software complex "Hydrolog" [6]. The study period was 141 years (1877-2017).

A hybrid EMD-ARIMA model proposed for predicting the maximum river runoff is based on the joint use of the empirical mode decomposition (EMD) method and the autoregressive integrated moving average

(ARIMA) model. EMD is used to decompose the initial non-stationary series into a series of mode functions (intrinsic mode functions, IMF) and the remainder, each of which can be applied to the ARIMA methodology (Fig. 1).

The main stages of obtaining forecast estimates using the EMD-ARIMA hybrid model are:

1) Sequential operations for extracting mode functions from the original time series, starting with high-frequency ones, and the remainder;
2) Development of a suitable ARIMA model for each IMF function and remainder;
3) Performing general calculations to predict the initial time series based on the forecast of each sub-series;
4) Compare the performance of the EMD-ARIMA hybrid model with the standard ARIMA model.

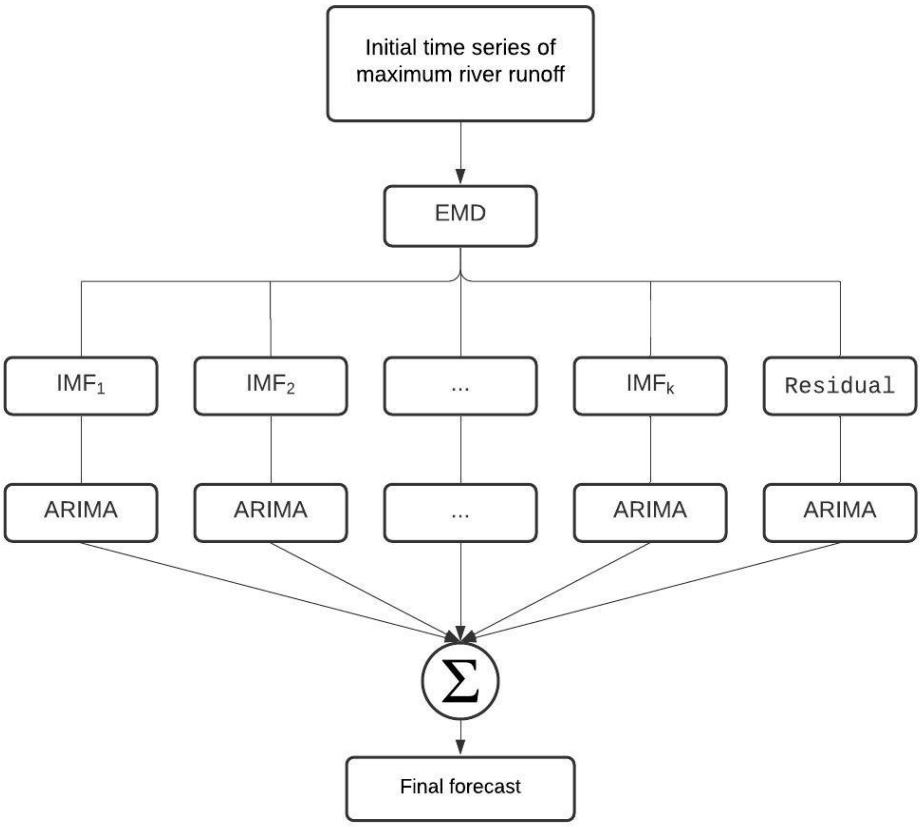


Fig. 1. Block diagram of the EMD-ARIMA hybrid model

EMD Method

Since many works have been devoted to the description of the ARIMA model [7], below will focus in more detail on the description of the EMD method, which is the basis of the hybrid model. This method is a relatively new form of time series decomposition. The peculiarity of the method is that the time series may not be linear or stationary. A necessary condition for the correct representation of nonlinear and non-stationary processes is the possibility of forming an adaptive basis that is functionally dependent on the content of the data itself.

The EMD method is the most important component of the Huang-Hilbert Transform (HHT), which is widely used in various fields of science and technology, along with the Fourier transform and wavelet analysis [8].

The difference between the EMD method and the wavelet analysis is that the EMD process decomposes into a series of mode functions that are not specified analytically and are determined exclusively by the analyzed data sequence itself, and the basic transformation functions are formed adaptively, directly from the input data. In the wavelet transform, as well as in the Fourier transform, the decomposition is performed in a fixed basis of functions. This basis must be pre-defined, that is, a specific wavelet function used in the transformation process must be selected.

The main idea of the EMD method is the assumption that the process under study is an additive combination of various internal oscillations, each of which is a mode function with extremes and zero values [9].

Each IMF function must meet two criteria:

1) The number of functions extreme and the number of zero intersections must be equal or differ by one;
2) At any point of the functions the average value of the envelopes interpolating local maxima and minima must be zero.

Let Q (t) is a multi-year series of maximum water flow rates. Then the main idea of the EMD method is to decompose the time series Q(t) into IMF functions and the remainder r(t). As a result of this decomposition, the series Q (t) can be represented as:

$$Q(t) = \sum_{i=1}^{k} IMF_i(t) + r_k(t), \quad (1)$$

where $r_k(t)$ – the residual component of the time series decomposition, $IMF_i(t)$ – the i-th internal mode function.

The algorithm of EMD process consists of the following steps:

1) Selection of a *Q(t)* series in the sifting process; putting into consideration three variables *i*, *j* and *k*, where *i* – number calculated IMF, *j* – number of its approach, *k* – number of IMF (initial values of variables: *i*=1, *j*=1, *k*=0);
2) Introducing of an additional series $S_i(t) = Q(t)$;
3) Identification of all local extrema in the time series *Q(t)*;
4) Forming the upper *p(t)* and lower *q(t)* envelopes by connecting all local maxima and local minima, respectively;
5) Definition of the average function $\phi_j(t) = \frac{p(t)+q(t)}{2}$;
6) Calculation of the first filtering component $\psi_j(t) = Q(t) - \phi_j(t$. If $\psi_j(t)$is an IMF function, we proceed to the next step. Otherwise, we use$\psi_j(t)$ as the values of the series *Q(t)* and increase the value of the variable j by 1 (*j* = *j*+1). For the updated series *Q(t)* repeat steps 3-5.
7) This stage consists of two steps. First – persistent $\psi_j(t)$ obtained in the previous step as $IMF_i(t)$, i.e. $IMF_i(t) = \psi_j(t)$, the value of the variable *k* is incremented by one. The second –determine the remainder $r_i(t)$:

$$r_i(t) = S_i(t) - IMF_i(t). \quad (2)$$

8) In accordance with the characteristics of the function $r_i(t)$ obtained at stage 7, a decision is made to stop the calculations. Calculations are stopped in the following cases:
   - $r_i(t)$ is a constant or constant function from which no more IMF functions can be extracted;
   - the remainder of $r_i(t)$ over the entire study interval becomes insignificant in its values compared to the initial series.
9) If the criteria for stopping the EMD process are not met, we use $r_i(t)$ as the values of the series *Q(t)*, we increase the value of the variable *i* by 1 (*i* = *i*+1), we assign the value 1 to the variable j. Next, we proceed to step 2 and continue the EMD process.

The most important stage in the implementation of the EMD method is the process of constructing the upper and lower envelopes of local extreme. Traditionally, cubic splines and B-splines are used for these purposes. However, when these types of splines are interpolated, such a phenomenon as the edge effect may occur (Fig. 2). The reason for the end distortions of the mode functions is the unpredictability of the approximation by the extremes of the upper and lower envelopes at the end sections of the modes.

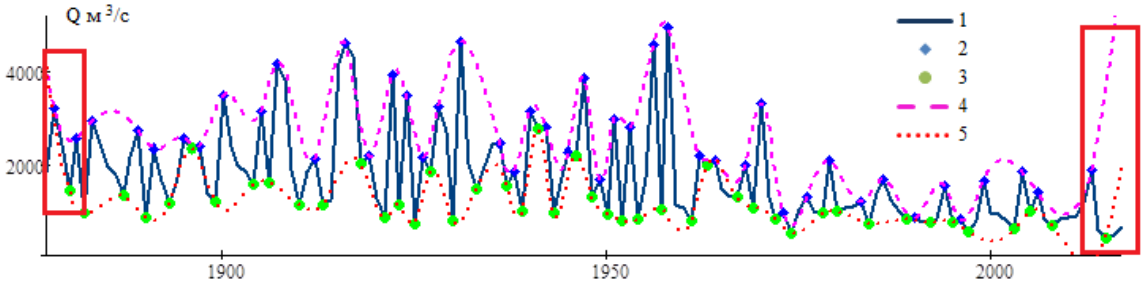


Fig. 2. Edge effect during cubic spline interpolation
1 – observed data, 2 – local maxima, 3 – local minima, 4 – upper envelope, 5 – lower envelope

Due to the fact that the process of constructing mode functions is a sequential subtraction of the current calculated mode from the previous input data of the maximum water runoff, errors in the approximation of envelopes at the end points lead to distortions of the calculated modes and recursive accumulation of errors in calculating mode functions. This significantly affects the results of EMD decomposition.

To weaken the edge effects at the ends of the series, we propose the use of a source-like approximation in the construction of envelope extreme [10]. The application of this approach to reduce the influence of end effects on the EMD process allows us to standardize the process of reducing the edge distortions of mode functions.

ARIMA Model

In general, the ARIMA (*p*, *d*, *q*) model for a non-stationary series y_t is expressed by the formula

$$\Delta^d y_t = c + \sum_{i=1}^{p} \alpha_i \Delta^d y_{t-i} + \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t , \quad (3)$$

where *p* – the autoregressive parameter, *d* – the order of the operations of taking the differences, *q* – the order of the moving average, *c* – constant, α_i, b_j – coefficients of the model, $\Delta^d y_t$ – the difference operator *d*-th order ($\Delta y_t = y_t - y_{t-1}$ – difference of the first order), ε_t – "white noise".

The forecasting using ARIMA models is based on the Box-Jenkins methodology, which contains three stages:

1) Identification of the model (determination of parameters *d*, *p*, *q*);
2) Evaluation and verification of the adequacy of the model;
3) Forecasting.

At the first stage of the model it is necessary to analyze the series for stationarity and select an ARIMA model for further evaluation. At the second stage the parameters of ARIMA models are evaluated by the maximum likelihood method and the adequacy of the obtained ARIMA models is checked. Several criteria are used for their comparison: the estimates of the model coefficients should be statistically significant; the remnants of the model should have the properties of white noise. If several ARIMA models turn out to be adequate, it is necessary to choose the model with the smallest number of parameters and the best statistical characteristics of the quality of the model fit. In this paper, the Akaike information criterion (AIC) is used for this purpose:

$$AIC(p,q) = ln\,\sigma^2 + \frac{2k}{N}, \quad (4)$$
$$\sigma^2 = \frac{RSS}{N-p-q},\ k = p + q,$$

where p, q – the parameters of the ARIMA model, N – the number of observations, RSS – the residual sum of squares.

Model Effectiveness Evaluation

The values of the standard error (RMSE), the average absolute error(MAE), the average absolute error in percent (MAPE), the correlation coefficient (R) for the training and test samples were obtained to evaluate the effectiveness of the EMD-ARIMA hybrid model using the formulas (5)-(8).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q^o{}_i - Q^m{}_i)^2}\,, \tag{5}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Q^o{}_i - Q^m{}_i\right|\,, \tag{6}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Q^o{}_i - Q^m{}_i}{Q^o{}_i}\right| \cdot 100\%\,, \tag{7}$$

$$R = \frac{\sum_{i=1}^{N}(Q^m{}_i - \overline{Q^m})(Q^o{}_i - \overline{Q^o})}{\sqrt{\sum_{i=1}^{N}(Q^m{}_i - \overline{Q^m})^2}\sqrt{\sum_{i=1}^{N}(Q^o{}_i - \overline{Q^o})^2}}, \tag{8}$$

where N – the sample volume, $Q^o{}_i$ – observed data, $Q^m{}_i$ – simulated data, $\overline{Q^o}$ – the average value of the observed data, $\overline{Q^m}$ – the average value of the simulated data.

RESULTS AND DISCUSSION

Fig. 3 shows a chronological graph of long-term fluctuations in the maximum runoff of the Dnieper River at the Rechitsa station for the period from 1877 to 2017. In this study, the data of maximum water flow rates from 1877 to 1997 were used to train both ARIMA and EMD-ARIMA models. Then the resulting model is applied to the test data (the period from 1998 to 2017).

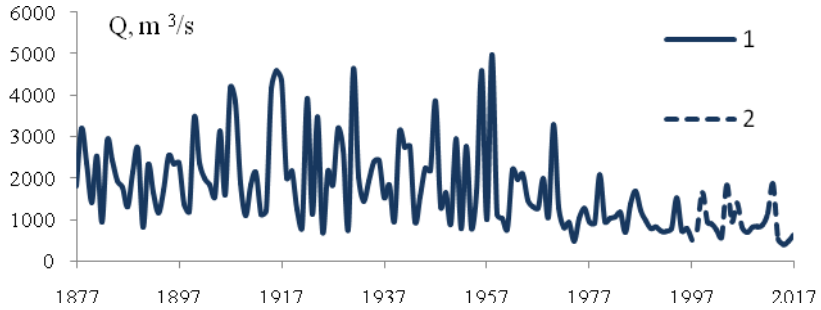


Fig. 3. Long-term changes in the maximum runoff of the Dnieper River at the Rechitsa station for the period 1877-2017
1 – training sample, 2 – test sample

Fig. 4 shows the first stage of the EMD process-the extraction of local extremes of a number of maximum river runoff.

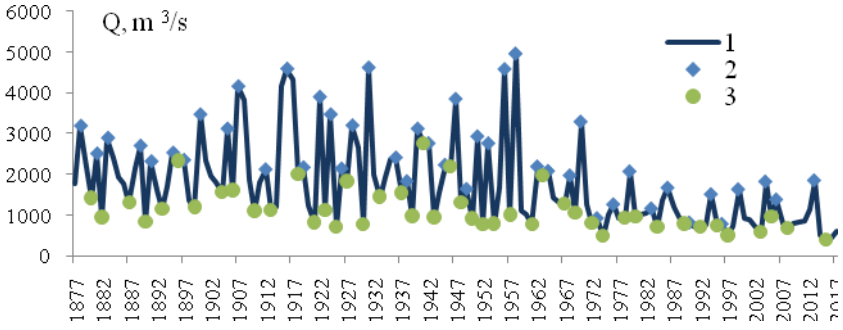


Fig. 4. Extraction of local extremes in the original data series
1 – observed data, 2 – local maxima, 3 – local minima

Fig. 5 shows an example of building IMF1. In this example, the extracted function does not satisfy the IMF conditions (the number of extremes and zero intersections differs by more than one) and, therefore, the sifting procedure must be performed in the future.

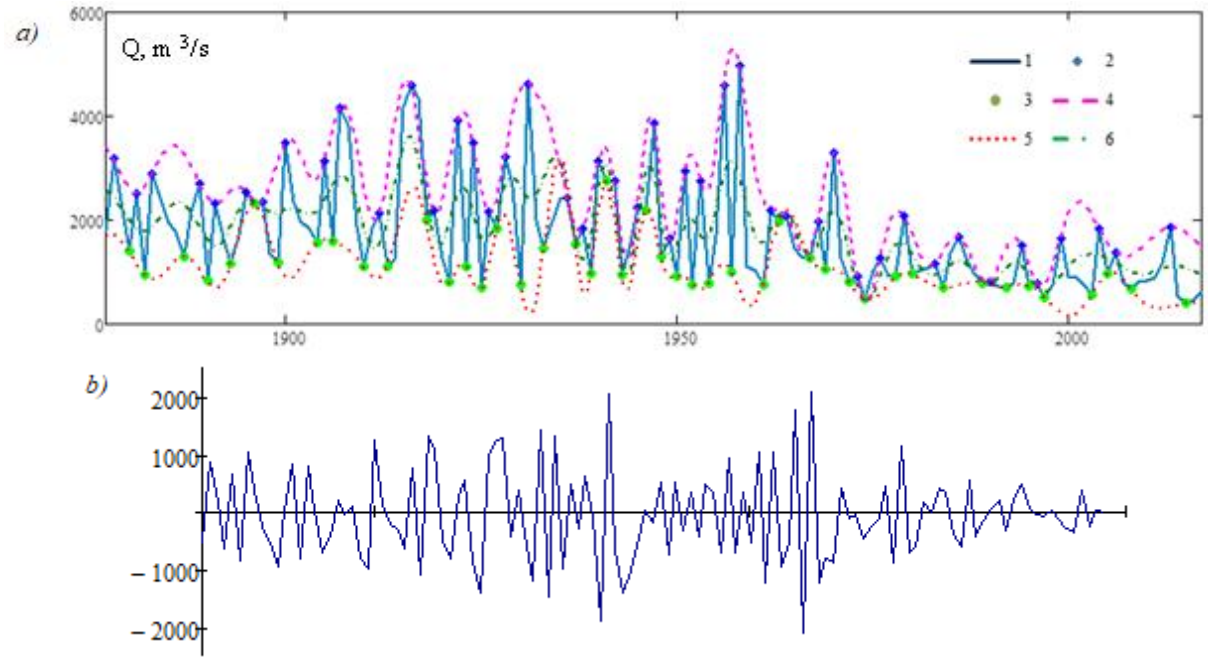
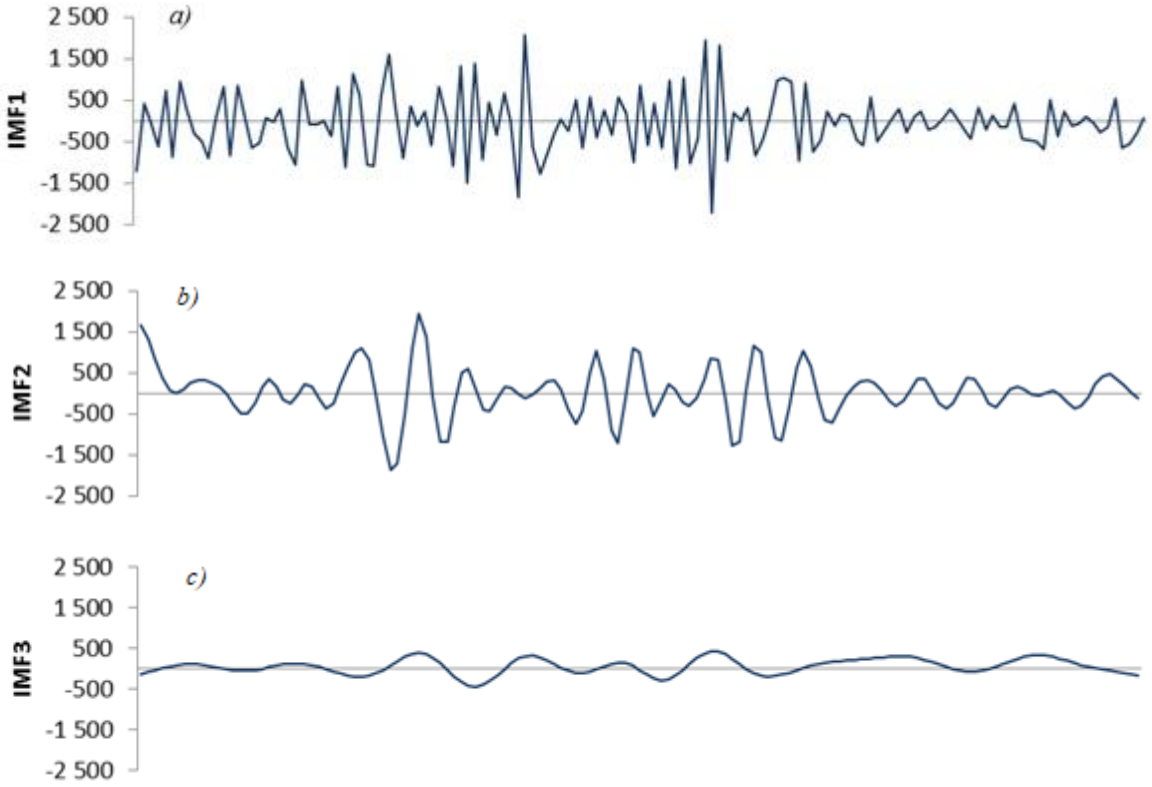


Fig. 5. Construction of the first approximation to IMF1:
***a*) 1 – observed data, 2 – local maxima, 3 – local minima, 4 – upper envelope, 5 – lower envelope, 6 – function of the average values of the envelopes;**
***b*) the first component of screening**

The result of the process of empirical mode decomposition of a series of maximum river runoff is shown in Fig. 6.



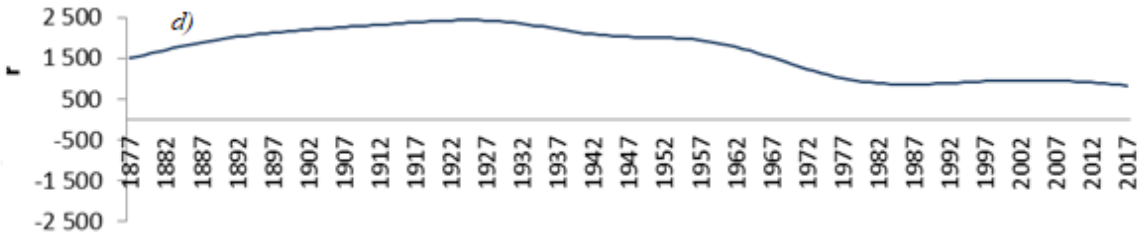


Fig. 6. Decomposition of the series of maximum river runoff into the functions of IMF and the remainder r
a – IMF1, b – IMF2, c – IMF3, d – the remainder r

For a reasonable conclusion about the stationarity of the obtained IMFs and the remainder *r*, their autocorrelation (ACF) and partial autocorrelation functions (ACF) are analyzed. The analysis showed that the components of IMF and *r* are neither pure AR nor MA models, but ARMA nor ARIMA models. To predict each IMF and the remainder of the entire portfolio of models, those that have minimum AIC values are selected (Table 1).

Table 1

Results of building ARIMA models

Model	IMF	ARIMA	AIC
ARIMA	ARIMA(7,1,0)		13.24
Hybrid model EMD-ARIMA	IMF1	ARIMA(3,0,5)	13.15
	IMF2	ARIMA(7,0,4)	11.20
	IMF3	ARIMA(1,4,1)	3.15
	Remainder	ARIMA(4,4,2)	-0.71

The parameters for evaluating the effectiveness of the ARIMA and EMD-ARIMA models for training and test samples are presented in Table 2. As can be seen from Table 2 and Fig. 7, the model data obtained using the hybrid EMD-ARIMA model are in better agreement with the observed data compared to the classical ARIMA model. The EMD-ARIMA method gives smaller error values both on the training data set and on the test sample.

Table 2

Values of the performance indicators of forecasting models

	Training sample				Testing sample			
Model	**RMSE, м^3/с**	**MAE, м^3/с**	**MAPE, %**	**R**	**RMSE, м^3/с**	**MAE, м^3/с**	**MAPE, %**	**R**
ARIMA	684,46	511,70	30	0,76	373,15	273,85	30	0,66
EMD-ARIMA	630,90	467,79	29	0,80	278,45	217,31	24	0,76

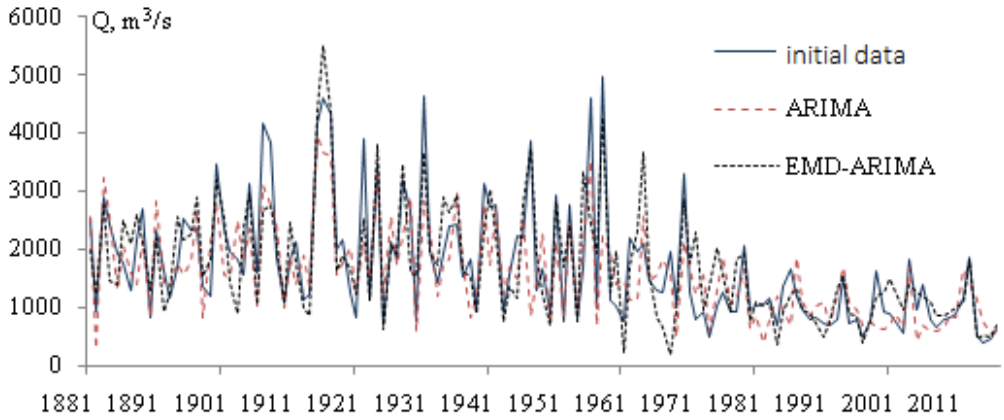


Fig. 7. Comparison of the actual and modelled values of the maximum runoff of the Dnieper River at the Rechitsa station

Fig. 8 shows the forecast estimates of the maximum runoff of the Dnieper River at the Rechitsa station for the period 2018-2027, obtained using the best of the two models considered in the work − the hybrid model EMD-ARIMA.

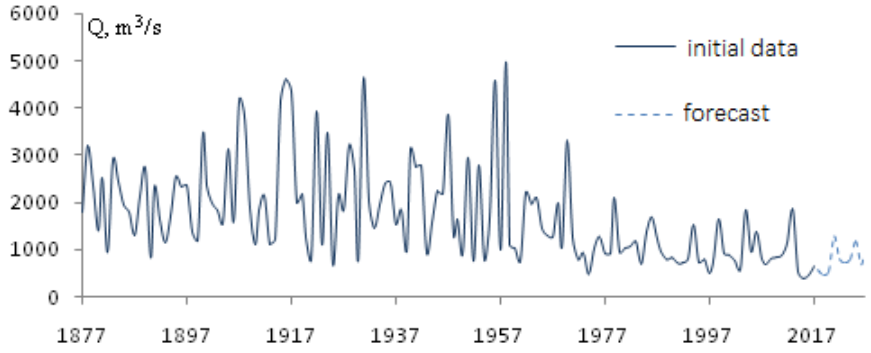


Fig. 8. Forecast estimates of the maximum runoff of the Dnieper River at the Rechitsa station for the period 2018-2027

CONCLUSION

A hybrid EMD-ARIMA model is proposed to predict a non-stationary time series of maximum river flow rates. The main idea of using the proposed model is to use the EMD method to decompose a number of maximum river runoff into individual IMFs and the remainder and to select suitable ARIMA models for the obtained series of IMFs. The final forecast is obtained by combining the results of forecasts by different ARIMA models of each series. The method is relatively simple in computational terms and does not require the fulfilment of the condition of stationarity of hydrological series.

The analysis of efficiency indicators for modelling the maximum water runoff allows concluding that the proposed hybrid model EMD-ARIMA has an advantage over the classical ARIMA model. . The results obtained in the work showed that the proposed hybrid model is able to predict the values of the maximum river runoff with high accuracy.

The work was carried out with the support of the BRFFI (grant no. X20M064)

REFERENCES

1. Volchek, A.A. Assessment of modern changes in the maximum flow of rivers of Belarus / A.A. Volchek, An. A. Volchek, S. V. Sidak / / Geography. - Minsk, 2020. – № 4(167). – P. 26-33.
2. Loginov, V. F. Spring floods on the rivers of Belarus: spatial and temporal fluctuations and forecast / V. F. Loginov, A. A. Volchek, An. A. Volchek - Minsk : Belorusskaya nauka, 2014. - 244 p.
3. Georgievsky, Yu. M. Shanochkin, S. V. Hydrological forecasts. S.-Pb.: RGGMU, 2007.
4. G.E.P. Box and G.M. Jenkins, Time Series Analysis Forecasting and Control, Holden-Day, San Francisco, 1970.
5. Wang W, Van Gelder P, Vrijling JK, Ma J (2006) Forecasting daily stream flow using hybrid ANN models. J Hydrol 324: 383–399
6. Volchek, A.A. Automation of hydrological calculations / A. A. Volchek // Water management construction and environmental protection: proceedings of the International Scientific and Practical Conference on problems of water management, industrial and civil Construction and economic and social transformations in market relations. / Brest. polytechnic University. Institute.- Biberach-Brest-Nottingham, 1998. - pp. 55-59.
7. Sahoo GB, Schladow SG, Reuter JE (2009) Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models. J Hydrol 378: 325–342.
8. Norden E. Huang, Samuel S.P. Shen. The Hilbert-Huang transform and its applications // World Scientific Publishing Co. Pte. Ltd. – 2005. – 311 c.
9. Peel MC, G.G.S. Pegram, and T.A. McMahon. 2007: Empirical Mode Decomposition: Improvement and application. In International Congress on Modeling and Simulation, edited by Oxley, L. and D. Kulasiri. Modelling and Simulation Society of Australia and New Zealand, December 2007, 2996-3002.
10. Dolgal, A.S. Application of empirical mode decomposition in the processing of geophysical data /Dolgal, A.S., Khristenko, L.A. // Izv. Tomsk Polytechnic University. un-ta. Resource engineering. - 2017. - Vol. 328. - No. 1. - pp. 100-108.