

Possibilities of application of logistic regression in hydrological forecasts (on the example of the mountain river Samur)

Ekaterina V. Gaidukova^{1*}, Vardui G. Margaryan², and Igor O. Vinokurov¹

¹ Russian State Hydrometeorological University, 79, Voronezhskaya Str., St. Petersburg, 192007, Russian Federation

² Yerevan State University, 1, Alek Manukyan, Yerevan Str., 375025, Armenia

Abstract. The possibilities of constructing regression equations for predicting the runoff of mountain and semi-mountain rivers are considered. A predictive equation for the Samur river catchment has been obtained, which connects water discharge and predictors by casual relationships: water levels, air temperature, air humidity, atmospheric pressure, precipitation, dew point temperature, wind direction, and cloudiness. Logistic regressions are obtained, which allow using categorical variables as independent variables. The result of a logistic regression forecast is the probability of the occurrence or non-occurrence of the event of the predicted value. The positive and negative aspects of this approach for mountain rivers are revealed, which consist of the interpretation of the predicted probability of the event. Actions are proposed that allow for obtaining more reliable forecasts.

1 Introduction

A feature of logistic regression is the possibility of using it to study the logical relationships between categorical variables. This type of regression is not common in the practice of hydrological forecasts, since hydrological characteristics are mostly quantitative. But when assessing the relationship between hydrological and meteorological quantities, which can be categorical, the establishment of a logistic regression acquires practical relevance.

The effective use of regression dependencies for forecasting the hydrological characteristics of mountain rivers is noted by some researchers. So, for example, for some mountain rivers of Uzbekistan, multifactorial dependencies between river flow during the growing season and precipitation are calculated [1]. To predict the water level of the Mzymta River (Krasnodar Territory), methods based on regression analysis and the use of neural network technologies are proposed that give approximately equal results [2] and based on the theory of Markov processes with discrete time [3]. The use of multiple regression with two predictors – water discharge for the previous period and precipitation, led to an improvement in forecasts for the Narym River (a river in the East Kazakhstan region of Kazakhstan, the right tributary of the Irtysh) [4], and for the Amyl River (a mountain river in the Krasnoyarsk Territory) a comparative Analysis of methods for forecasting maximum water levels showed that one-factor dependencies have higher determination coefficients than the multiple regression model [5].

The mountain rivers of the Caucasus have significant rainfall throughout the warm hydrological year, and solid

precipitation exceeds 40–50% of the total. Using the Terek River as an example, the work [6] presents a dependence for a short-term forecast of water discharges, and for the Gusurchay, Velvelichai, and Zagemchay rivers, [7] obtained multiple linear regression equations for predicting runoff based on precipitation.

The purpose of the study was to test the method of logistic regression to the watershed of the river Samur and evaluate the possibilities and effectiveness of this method for predicting water discharges.

2 Materials and Methods

2.1 Initial data

Identification of the possibilities of using logistic regression for hydrological forecasts was carried out in the catchment area of the Samur River (fig. 1). Discharges and water levels at station c. Usukhchay for 2013, 2014, and 2015 [8]. Meteorological data for Usukhchay station included quantitative variables – air temperature, air humidity, atmospheric pressure, precipitation, and dew point temperature; categorical variables are wind direction, and cloudiness [9]. On fig. 2 shows examples of chronological graphs of changes in meteorological quantities.

The Samur river is located in southern Dagestan, along the part of the channel the border of the Russian Federation with Azerbaijan passes. The length of the river is 213 kilometers, the total fall is 2910 meters, the average slope is 17.7 ‰, the catchment area is 4990 km², and the average height is 1970 meters. The main Caucasian ridge is the boundary of the river basin. Samur in the southwest, in the northeast of the border –

* Corresponding author: oderiut@mail.ru

the northern spurs of the Side Range. In the lower reaches of the river, the boundaries of the basin are not expressed.

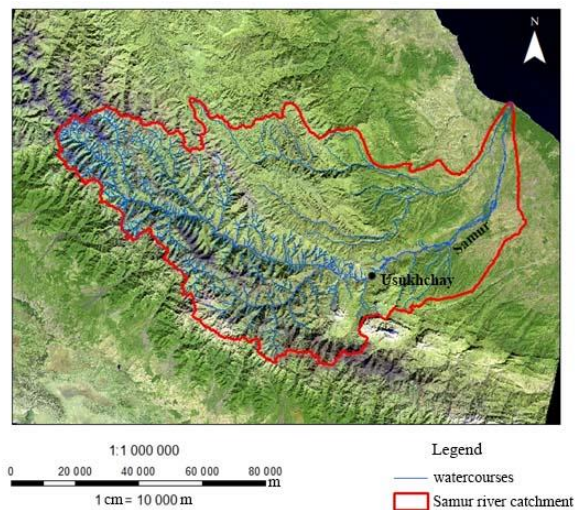


Fig. 1. Map of the Samur River catchment.

In the annual series of water discharges, the periods of spring flood and rain flood were singled out as the most demanded when issuing hydrological forecasts. Since river Samur is a mountain river with a peculiar formation of river flow [10], it was difficult to identify the indicated periods in the available initial data: in 2013 and 2015, the spring flood smoothly turns into a rain flood, in 2014 there is a low water period the period between spring flood and rain flood (see Fig. 3).

2.2 Methods

Regression models are a way to study causal relationships, i.e. assessment of the interaction of variables. When constructing a logistic regression, dependent variables (predictants) must be categorical, while independent variables can be both categorical and quantitative [11]. At the same time, the dichotomized dependent variable evaluates not the probability of occurrence, but the logarithm of the ratio of the probability of occurrence and non-occurrence:

$$\ln(P_{occur} / P_{non-occur}) - \text{logit}.$$

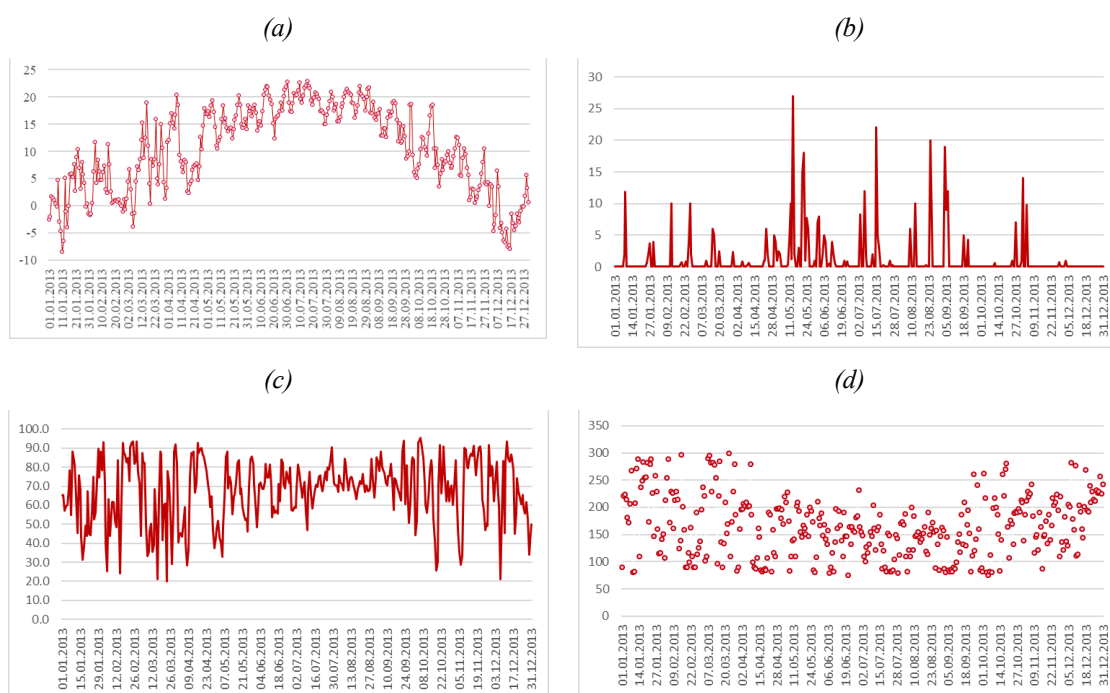


Fig. 2. An example of changes in meteorological characteristics (air temperature (a), precipitation (b), air humidity (c), wind direction (d)) for 2013.

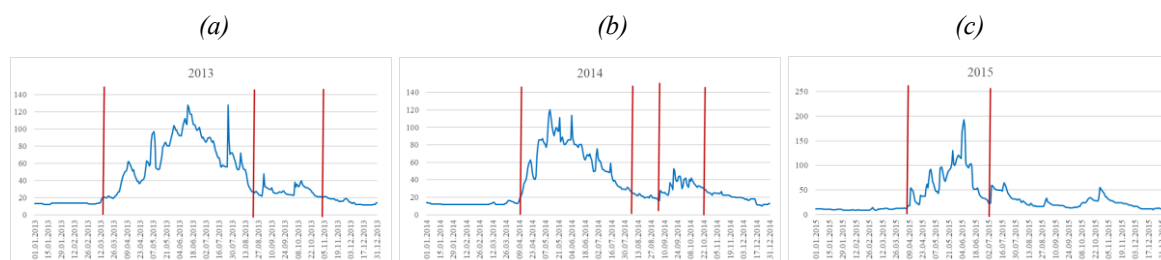


Fig. 3. Identification of periods of spring floods and autumn floods in the years under consideration: (a) – 2013, (b) – 2014, (c) – 2015.

The probability of the event P is encoded by the value 1 (100 % probability of occurrence), respectively, the probability of non-occurrence is $1-P$, $P/(1-P)$ is the chance of occurrence or non-occurrence.

Logit is calculated by the expression:

$$\ln \left[\frac{P}{1-P} \right] = B_0 + \sum_{i=1}^k B_i x_i = \quad (1)$$

$$= B_0 + B_1 x_1 + B_2 x_2 + \dots + B_k x_k$$

where x_i ($i = \overline{1, k}$) – independent variables; B_i – coefficients of multiple linear regression (show how much the logit will change on average when the independent variable changes).

The probability of an event occurrence is calculated using the following expressions:

$$\frac{P}{1-P} = \ell^{B_0+B_1x_1+B_2x_2+\dots+B_kx_k}, \quad (2)$$

$$P = \frac{\ell^{B_0+B_1x_1+B_2x_2+\dots+B_kx_k}}{1 + \ell^{B_0+B_1x_1+B_2x_2+\dots+B_kx_k}}. \quad (3)$$

Qualitative assessment – the resulting dichotomized dependent variable can characterize possible events: $P > 0.5$ – occurrence; $P < 0.5$ – non-occurrence.

3 Results and Discussion

In the first stage, the relationships between all hydrometeorological variables were evaluated using the calculated correlation matrix. The greatest relationship was naturally found between discharges and water levels, as well as between water discharges and air temperatures and dew points.

For categorical variables, the Spearman rank correlation coefficient was calculated, which belongs to the method of correlation analysis and reflects the ratios of variables sorted by increasing values [12]. The value of the Spearman correlation coefficient lies in the range of +1 and -1, characterizing the direction of the relationship between the features measured in the rank scale.

Spearman correlation coefficients were calculated for the number of clouds, wind direction, and water flow. The correlation between water flow and cloud amount reaches 0.77, and between water flow and wind direction 0.56.

When constructing regression equations, the lead time of the forecast was taken into account, which was taken equal to 1 day; with a longer lead time, the connection between the predictor and predictors is lost.

On fig. Figure 4 shows an example of hydrographs built based on actual and calculated data using the regression equation, with and without water levels (it is not advisable to exclude the water level from consideration).

As an example, Figure 5 shows the results of forecasting for the spring flood of 2015.

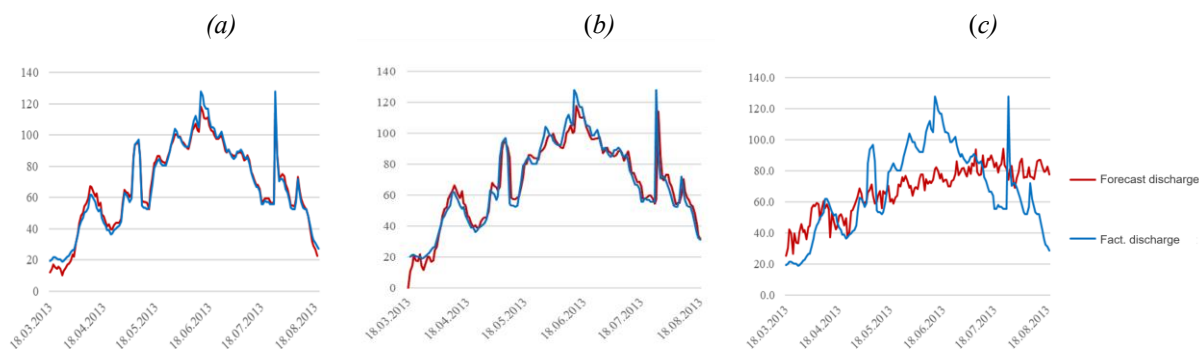


Fig. 4. An example of actual and calculated by the regression equation hydrographs on the river Samur for the period of the spring flood of 2013: *a* – calculation without time shift; *b* – calculation for a lag of one day; *c* – calculation according to the regression equation without taking into account water levels.

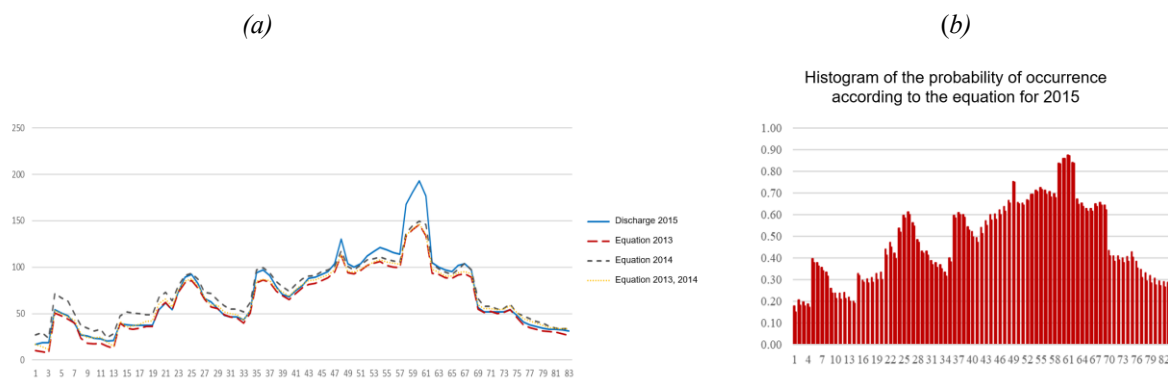


Fig. 5. An example of the result of forecasting is by the regression equation (*a*) and by logistic regression (*b*) on the river. Samur during the spring flood of 2015.

The following results were obtained:

- regression equations were built for 2013 and 2014, and the forecast was given for 2015, and for spring flood the technique is effective since the forms of flood hydrographs are similar for all three years;
- hydrographs of rain floods are not similar, and the technique showed unsatisfactory results.

Expanding the initial database for constructing regression equations to include various forms of hydrographs, or vice versa, using only an analog year with the corresponding hydrograph form, will increase the efficiency of the approach under consideration.

The use of logistic regression made it possible to estimate the probability of occurrence of each predicted water discharge. But according to the calculated expression (3), a regularity is obvious – for large expenses, a high probability of occurrence is inherent, which can be seen in Fig. 5b.

4 Conclusion

The use of regression models for prognostic purposes is the simplest and most physically reasonable approach to predicting the characteristics of natural processes. Regression equations allow you to study the causal relationships that characterize the interaction of variables, in which some are causes and others are consequences. In practice, models that use several independent variables are most often used, and multiple regressions are built.

During the approbation of the method of logistic regression to the watershed of the river Samur and assessing the capabilities and effectiveness of this method for predicting water flow, the following was revealed:

- the method of logistic regression for hydrological forecasts is effective if there is an analog year for the forecast year in the initial data;
- the approach to assessing the probability of non-occurrence is not reliable in forecasting extreme water flows;
- the rank correlation method has shown its effectiveness for categorical hydrometeorological variables.

It is planned to develop an approach that would automatically select years for compiling a regression model [13] and it is possible to use artificial neural networks [14].

Acknowledgments

The study was financially supported by the National Research Committee of the Republic of Armenia and the Russian Foundation for Basic Research (RF) within the framework of the joint scientific study “Short-term probabilistic forecast of river flow during the spring flood” No. 20RF-039 and No. 20-55-05006/20, respectively.

References

1. Z.F. Khakimova, N.R. Sobirova, Issues of long-term forecasts of mountain river runoff for the growing season. In: *Use of water resources in the context of climate change* (Ufa, Bashkir State Agrar. Univ., 2022)
2. E.A. Semenchin, N.G. Titov, M.M. Kuzyakina, K.A. Lebedev, Comparative analysis of methods of mathematical modeling of the water level in the river mountain type (for example, the river Mzymta). *Kuban State Univ.* **12-5**, 952–957 (2014)
3. N.G. Titov, M.V. Kuzyakina, K.A. Lebedev, Applying the Markov equation to predicting the water level in a river with a steep dip. *Sci. almanac* **9(11)**, 1126–1129 (2015). DOI: 10.17117/na.2015.09.1126
4. Zh.Zh. Karamoldoev, O.Yu. Kalashnikova, Forecast of water inflow into the Toktogul reservoir for the growing season. *Bishkek: Bull. of BSU* **3(23)**, 146–150 (2012)
5. V.P. Galakhov, O.V. Lovtskaya, S.Yu. Samoilova, E.V. Mardasova, Comparative analysis of methods for forecasting maximum levels and volumes of flood runoff of a mountain river. *Bull. of the Tomsk Polyt. Univ. Georesource engin.* **2**, 193–203 (2022). DOI: 10.18799/24131830/2022/2/3438
6. V.M. Mukhin, Methodological bases of physical and statistical types of short-term forecasts of mountain river runoff. *Proc. of the Hydrometeorol. Res. Center of the Rus. Fed.* **349**, 5–46 (2013)
7. R.G. Verdiev, Computation and prediction of the flood runoff of the eastern Caucasus rivers. *Rus. Meteor. and Hydrol.* **34**, 46–50 (2009)
8. *Automated information system for state monitoring of water bodies*. Retrieved from: <https://gmvo.skniivh.ru/> Accessed: 20-April-2022
9. *All-Russian Research Institute of Hydrometeorological Information, World Data Center*. Retrieved from: <http://meteo.ru/> Accessed: 21-April-2022
10. E.V. Gaidukova, V.G. Margaryan, I.O. Vinokurov, A.Yu. Romashchenko, Short-term forecasting of water consumption on the river. *Samur. Int. Res. J.* **6-3(108)**, 17–23 (2021). DOI: 10.23670/IRJ.2021.108.6.064
11. O.V. Tereshchenko, N.V. Kurilovich, E.I. Knyazeva, *Multivariate statistical data analysis in social sciences* (Minsk, BGU, 2012)
12. M.A. Kharchenko, *Correlation analysis* (Voronezh, VSU Publishing House, 2010)
13. E.V. Gaidukova, A.E. Kachalova, K.A. Litvinova, Accounting for the periodicity of water content in forecasting river runoff on the example of the rivers of the North-West region. *N. word in scien.: dev. pros.* **4-1(10)**, 72–77 (2016)
14. A.E. Sumachev, N.V. Myakisheva, V.G. Margaryan, A.E. Misakyan, Long-term forecasting of water levels in Lake Ilmen using probabilistic approaches. *Nat. and tech. scien.* **6(157)**, 96–102 (2021)