# Modeling water level using downstream river water level observations and machine learning methods

*Mikhail* Sarafanov[1,2], *Eduard Kazakov*[1], and *Yulia Borisova*[1,3]

[1]State Hydrological Institute, 23 2-nd line, Vasilievsky Island,199004, St. Petersburg, Russia
[2]ITMO University, 49 Kronversky prospekt, 197101, St. Petersburg, Russia
[3]Saint Petersburg State University, 7/9 Universitetskaya nab, 199034, St. Petersburg, Russia

**Abstract.** The article presents the results of the development of a model for calculating levels at one gauging station using the levels at another. To link the levels at two gauging stations, the data on levels, temperature and precipitation were used. The use of machine learning methods to solve the problem of predicting water levels made it possible to achieve an accuracy of about 6 cm. At the same time, traditional statistical models (linear regression, polynomial regression) have 14-16 cm error.

## 1 Introduction

One of the most common hydrological tasks is the assessment of relationship between water levels on upstream and downstream river sections. It arises if there is a need, having received the value of the water level on one gauging station, to quickly obtain estimated values of the water levels on the other. Sometimes, if the observations in one of the stations are stopped for any reason, there is a need to have at least calculated values. Traditionally, such calculations are based on hydraulic models [1, 2] (if reliable data on the configuration of water surface slopes are available) or statistical models (if an archive of data on level observations at both sections is available) [3]. Some authors discuss using time series predictions [4], which makes it possible to obtain estimates of the water level during spring floods.

In practice, statistical models are more common due to their simplicity, as well as the undemanding variety of source information (for example, information about biases is often missing). Usually regression dependencies (linear, power or polynomial) of one level from another are used. The accuracy of such models is not always satisfactory, since they do not take into account the seasonal variability and sensitivity of the relationship to meteorological conditions.

We tried to develop a model using machine learning methods for linking water levels at the gauging station of river Tihvinka – Goreluha (predictor) and river Tihvinka – station Tihvin (response). Station Goreluha is located 20 km downstream of Tihvin. At the gauging station Goreluha an automatic level gauge is functioning, data from which can be obtained in real time. Moreover, for this station, the approximating relation between the level and flow rate of water based on long-term observations is well known. At the gauging station

Tihvin there is no operational automatic level gauge, telegram data is delayed, and there is no reliable information about the relationship between the level and the water flow. This makes it impossible to use hydrological forecasting models. In such a situation, it is necessary to provide the most accurate mechanism for calculating levels between sections.

## 2 Materials and methods

The study used data provided by the North-West Department of Hydrometeorology and Environmental Monitoring (Russia). To implement the algorithm for converting the values of the water level at one gauging station to the water levels at another, we used the average daily data on levels (cm), air temperature (degrees Celsius) and the amount of precipitation (mm) for the period from January 1, 2013 to January 1, 2018. Programming language Python was used as a tool for processing data and building models. Machine learning models were implemented using the library scikit-learn [5]. The mean absolute error, mean median error, and root mean square error were used as the metrics of quality for models.

Simple linear regression and a polynomial of degree 5 were used as the "basic" models for calculating. The predictor is the water level on Goreluha, the response is the water level in Tihvin. The regression curves and the separation of data into test and training samples are shown at Fig. 1. 1434 objects were included in the training sample, and 359 were included in the test sample.

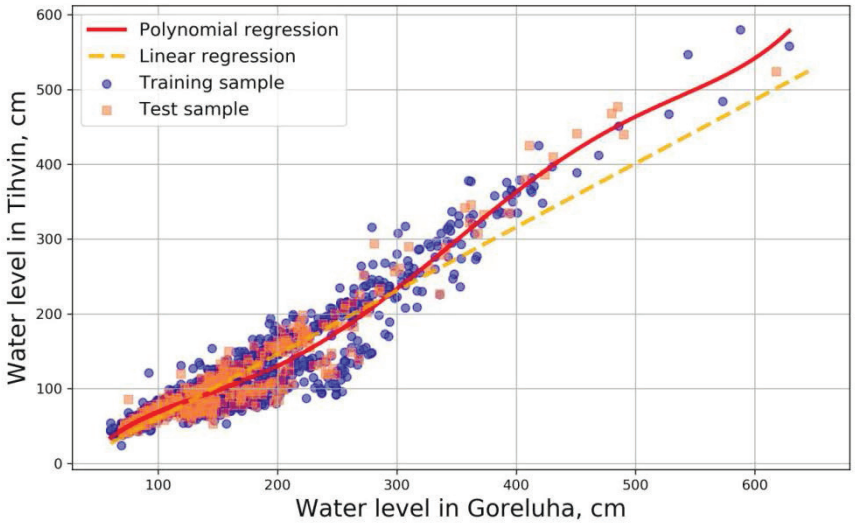


**Fig. 1.** Regression dependences of the water level in Tihvin on the water level in Goreluha.

To improve the predictive ability of the linear regression model, an additional predictor was added – a month. This categorical feature helped to take into account the seasonal variability of water levels (Fig. 2).

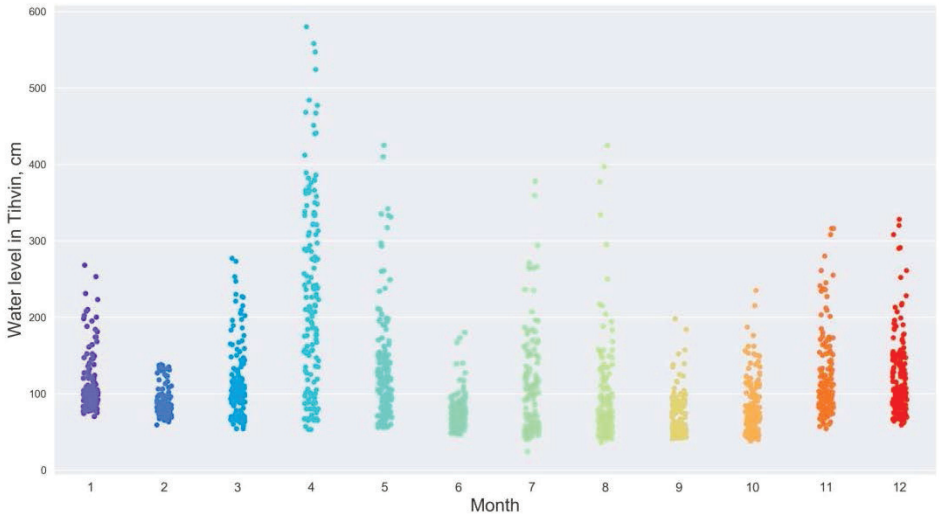


**Fig. 2.** Data on water levels at a gauging station in Tihvin grouped by months.

The graph shows that the minimum values are observed in June and July (summer low water) and February (winter low water). In April and May, the values of the level are maximum, since at this time, high water are usually observed.

In order to improve the model, in addition to the month, the average temperature for 10 days preceding the observed level, the amplitude of air temperature for 10 days preceding the observed level, the amount of precipitation and dispersion for 10 days preceding the observed level were used as additional predictors. The last 4 features were calculated from data from the nearest weather station and included in the training and test sample using Python.

After constructing all the indicated models, a categorical attribute was included as an additional categorical predictor, which took values equal to 1 if the difference in water level at Goreluha and the water level in Tihvin was greater than zero over the previous day, and values equal to 0 if there was a difference in the previous day the water level on Goreluha and the water level on Tihvin was less than zero. This allowed us to improve the quality of the model and avoid some of the large errors (more than 80 cm).

The approach was justified by the following observation in the data: the water level in Goreluha in the vast majority of cases is higher than the water level in Tihvin. This is due to the fact that Goreluha is located downstream. However, in some cases, the level in Tihvin was higher, which can be explained by the spatial heterogeneity of precipitation. To account for such short-term changes, it was decided to include a new predictor in the model.

## 3 Results and discussion

As a result of the experiments, more than 7 different models were tested. The selection of hyperparameters for machine learning models was carried out by a grid search. The results of the comparison of models are shown in Table.

**Table.** Metrics for level prediction algorithms in a test sample.

| Model | Set of predictors included in the model | Mean absolute error, cm | Mean median error, cm | Root Mean Square Error, cm |
|---|---|---|---|---|
| Linear regression | Goreluha water level | 16.7 | 11.0 | 23.6 |
| Polynomial regression | Goreluha water level | 14.4 | 10.8 | 19.4 |
| Linear regression | Goreluha water level, month | 12.0 | 6.9 | 17.8 |
| Linear regression | Goreluha water level, month, level difference code on the previous observation, precipitation amount and variance for 10 days, average temperature and amplitude of air temperature for 10 days | 10.7 | 8.0 | 14.8 |
| K-nearest neighbors | | 8.9 | 4.6 | 14.6 |
| Random forest | | 7.6 | 4.6 | 11.6 |
| Support Vector Machine | | 6.7 | 4.2 | 10.2 |

As can be seen from the table, the support vector machine method turned out to be the most accurate, which ultimately had a mean absolute error of 6.7 cm. From the graph of the kernel density estimation of residuals (Fig. 3) for the 4 most accurate models, it can be seen that the residuals of the support vector machine method are normally distributed.

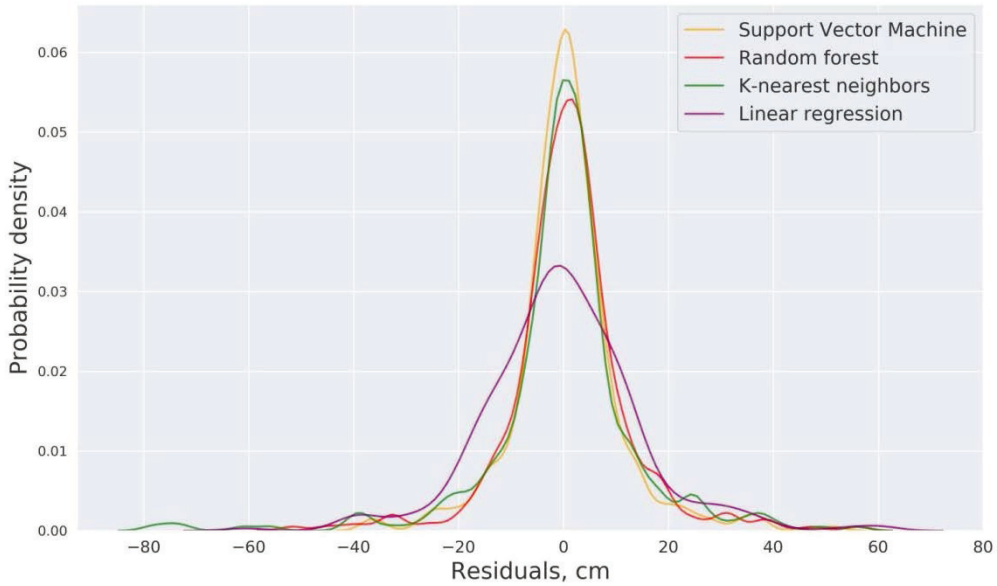


**Fig. 3.** Kernel density estimation of residuals for different models.

Thus, according to the results on the test sample, the quality of the model can be considered satisfactory.

## 4 Future research

In addition to using the presented models, the stacking technology was used [6]. As the first level models, 3 models of the support vector machine with different hyperparameters were used, as well as two ensemble algorithms-random forests with different values of the maximum depth of trees. The meta-algorithm was the K-nearest neighbors.

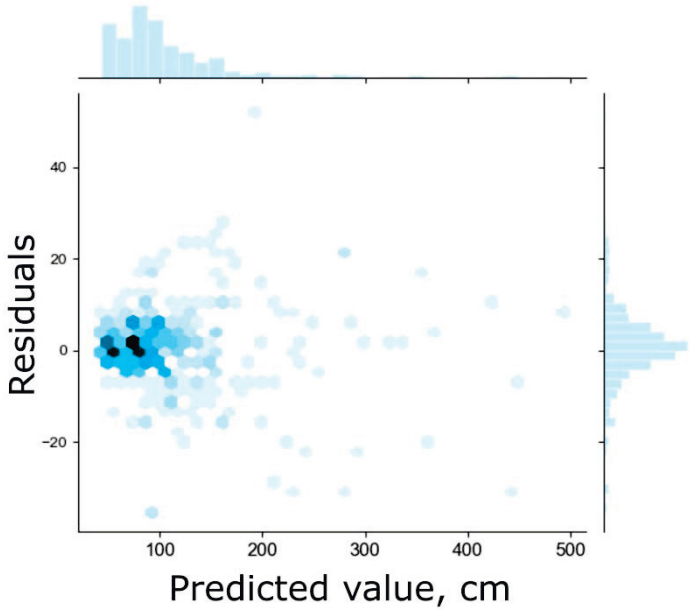


**Fig. 4.** Residuals plot for ensemble of models using stacking.

As a result of stacking, the accuracy was improved: 6.1 cm – mean absolute error, 3.7 cm – mean median error, 9.8 cm – root mean square error. We plan to continue our research in the direction of combining algorithms to improve the quality of forecasts.

## 5 Conclusion

Various ensemble algorithms (for example, random forest) show greater efficiency in regression problems in a number of studies [7, 8]. However, to solve our problem, the random forest algorithm was not suitable. First of all, because in addition to linking the levels within the available data, it was important for us to be able to use the model even for those level values that the model had not previously seen in the training set. A random forest cannot adequately perform the extrapolation procedure.

As a result of the work, a model for forecasting water levels at the Tihvin gauging station was implemented and verified. The mean absolute error was 6.7 cm (support vector machine) and 6.1 cm (using stacking technique), which far exceeds the quality of prediction using linear regression (16.7 cm) or polynomial regression (14.4 cm). The disadvantages of the proposed dependence approximation approach using machine learning methods include the need to attract additional data, such as daily average temperatures and precipitation from a weather station.

## References

1. A. Romanov, V. Ilyinich, Environmental management, **5**, 66-70. (2012) (in Russian)
2. N. Velikanov, V. Naumov, S. Koryagin, Technical and technological problems of service, **3(41)**, 32-35 (2017) (in Russian)
3. D. Buyanov, R. Fedotov, P. Tkachenko, Scientific and educational problems of civil protection, **2**, 112-118 (2015) (in Russian)
4. G. Struchkova, V. Timofeeva, T. Kapitonova, D. Nogovitsyn, K. Kusatov, Bulletin of the Samara Scientific Center of the Russian Academy of Sciences, **18(2),** 213-216 (2016) (in Russian)
5. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, Journal of Machine Learning Research, **12**, 2825-2830 (2011)
6. M. Graczyk, T. Lasota, B. Trawiński, K. Trawiński, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5991, 340-350 (2010)
7. U. Grömping, American Statistician, **63(4)**, 308-319 (2009)
8. T. Cootes, M. Ionita, C. Lindner, P. Sauer, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **7578** LNCS(PART 7), 278-291 (2012)