remote sensing

MDPI

Article

Estimation of the Water Level in the Ili River from Sentinel-2 Optical Data Using Ensemble Machine Learning

Ravil I. Mukhamediev [1,2], Alexey Terekhov [2], Gulshat Sagatdinova [2], Yedilkhan Amirgaliyev [2,*], Viktors Gopejenko [3,4], Nurlan Abayev [2,5], Yan Kuchin [1,2], Yelena Popova [6] and Adilkhan Symagulov [1,2]

1 Institute of Automation and Information Technology, Satbayev University (KazNRTU), 22 Satbayev Street, Almaty 050013, Kazakhstan; r.mukhamediev@satbayev.university (R.I.M.); ykuchin@mail.ru (Y.K.); a.symagulov@satbayev.university (A.S.)
2 Institute of Information and Computational Technologies, Pushkin Str., 125, Almaty 050010, Kazakhstan; aterekhov1@yandex.ru (A.T.); alchemdzz@gmail.com (G.S.); abayev_n@meteo.kz (N.A.)
3 International Radio Astronomy Centre, Ventspils University of Applied Sciences, LV-3601 Ventspils, Latvia; viktors.gopejenko@isma.lv
4 Department of Natural Science and Computer Technologies, ISMA University of Applied Sciences, LV-1019 Riga, Latvia
5 RSE Kazhydromet, 11/1 Mangilik El avenue, Astana 010000, Kazakhstan
6 Transport and Telecommunication Institute, LV-1019 Riga, Latvia; popova.j@tsi.lv
* Correspondence: amir_ed@mail.ru

Abstract: Monitoring of the water level and river discharge is an important task, necessary both for assessment of water supply in the current season and for forecasting water consumption and possible prevention of catastrophic events. A network of ground hydrometric stations is used to measure the water level and consumption in rivers. Rivers located in sparsely populated areas in developing countries of Central Asia have a very limited hydrometric network. In addition to the sparse network of stations, in some cases remote probing data (virtual hydrometric stations) are used, which can improve the reliability of water level and discharge estimates, especially for large mountain rivers with large volumes of suspended sediment load and significant channel instability. The aim of this study is to develop a machine learning model for remote monitoring of water levels in the large transboundary (Kazakhstan-People's Republic of China) Ili River. The optical data from the Sentinel-2 satellite are used as input data. The in situ (ground-based) data collected at the Ili-Dobyn gauging station are used as target values. Application of feature engineering and ensemble machine learning techniques has achieved good accuracy of water level estimation (Nash–Sutcliffe model efficiency coefficient (NSE) >0.8). The coefficient of determination of the model results obtained using cross-validation of random permutations is $NSE = 0.89$. The method demonstrates good stability under different variations of input data and ranges of water levels ($NSE > 0.8$). The average absolute error of the method ranges from 0.12 to 0.18 meters against the background of the maximum river water level spread of more than 4 meters. The obtained result is the best current result of water level prediction in the Ili River using the remote probing data and can be recommended for practical use for increasing the reliability of water level estimation and reverse engineering of data in the process of river discharge monitoring.

Keywords: remote sensing data; Sentinel-2; machine learning; ensemble machine learning; water level estimation; the Ili River

Citation: Mukhamediev, R.I.; Terekhov, A.; Sagatdinova, G.; Amirgaliyev, Y.; Gopejenko, V.; Abayev, N.; Kuchin, Y.; Popova, Y.; Symagulov, A. Estimation of the Water Level in the Ili River from Sentinel-2 Optical Data Using Ensemble Machine Learning. *Remote Sens.* **2023**, *15*, 5544. https://doi.org/10.3390/rs15235544

Academic Editor: Raffaele Albano

Received: 12 September 2023
Revised: 14 November 2023
Accepted: 24 November 2023
Published: 28 November 2023

1. Introduction

Contemporary production and life-support systems are consumers of large quantities of water. To meet these needs in Kazakhstan, located mainly in the arid zone, hundreds of hydraulic structures and complexes, including dams, weirs, canals, reservoirs, etc., are designed, constructed and operated. These facilities, as well as other constructions of

considerable size, are a source of substantial danger, which increases in case of improper design, operation and insufficient control of their current condition. Moreover, hydrotechnical facilities are operated under conditions of significant natural anomalies, which can be aggravated by anthropogenic impacts. These factors lead to serious failures, including those of catastrophic nature, significant damage and loss of human life. For example, in Kazakhstan, catastrophic floods associated with dam failures, including those in neighboring countries, resulted in the deaths of several dozen people and material damage of several tens of millions of dollars [1]. There is a need for more detailed monitoring of river discharge, water conditions, dams, etc., to prevent catastrophic events and to improve the quality of operation of equipment and structures [2]. At the same time, the increase in manual monitoring leads to significant costs. Nevertheless, there are examples of application of satellite data for assessment of water quality [3–5], drainage volumes [6,7], forecasting of possible flood damage [8,9], assessment of sediment load [10] and landslide dams [11] on the basis of statistical methods. There are also examples of applying machine learning methods for water quality assessment [12,13] on the basis of remote probing of the earth's surface, detection of hydraulic structures failures [14], etc. In this regard, it seems reasonable to evaluate the possibility of using machine learning to determine the level and associated water flow in rivers based on satellite images.

It is particularly interesting to develop appropriate methods for relatively small (from the point of view of satellite monitoring) meandering rivers flowing through sparsely populated areas with a sparse network of hydrometric monitoring stations and characterized by significant variability of water consumption. The transboundary (PRC-Kazakhstan) Ili River, which is about 1400 km long and has a basin area of 140,000 km^2 and flows into Lake Balkhash, is one such river. Depending on the weather conditions, the water level in the river (Ili-Dobyn hydrometric station) can vary by more than 4 m, which is accompanied by variations in water consumption from about 80 to 2300 m^3/s.

In this study, we aim to describe a method for determining the river level using an approach that has been referred to in the literature as a space hydraulic station [15] or virtual station [16]. The meaning of this concept is the application of data of remote probing of the earth's surface for solving the tasks that are traditionally performed by the hydrometric stations installed on the river. One of these tasks is to determine the water level in the river, which is closely related to the task of estimating the water discharge. It should be noted that during the warm season, which is important for water level estimation, cloud cover in the south of Kazakhstan is insignificant. This provides favorable conditions for the use of satellite data of optical range. As noted above, as the object of the study we have chosen the transboundary Ili River (PRC-Kazakhstan), which is the largest river in the Xinjiang Uygur Autonomous Region (XUAR) of China and plays an important role in water supply in the south-east of Kazakhstan. The river is the main tributary of the large Balkhash Lake, with a water mirror area of about 16 thousand km^2.

Our contribution to the current state of the research field is as follows:

- We obtained the state-of-the-art results in the problem of determining the water level in the Ili River using the optical remote sensing data and machine learning methods.
- We compared the machine learning algorithms and found that ensemble machine learning methods (Random Forest, eXtreme Gradient Boosting (XGBoost) and LightGBM) demonstrated the best and most robust water level estimation results.
- A set of input variables and corresponding feature engineering techniques is identified, which allows significant improvement of the original result and a good value of the Nash–Sutcliffe model efficiency coefficient.
- For the proposed model, the input parameters that ensure its stability depending on the volume and quality of input data were identified.

The paper consists of the following sections:

- Section 2 briefly describes the study area.
- The current state of the research area is discussed in Section 3.
- In Section 4, we describe the proposed method.

- Section 5 describes the results.
- Section 6 is devoted to discussion of the results.
- Finally, we refer to the limitations of the proposed approach and formulate the objectives of future research.

2. Study Area

The Ili River, flowing in the territory of Kazakhstan along the dry plain terrain (Figure 1), was chosen as the study area in the present research.

Figure 1. The Ili River basin. The study area is marked with a rectangle.

It should be noted that modeling of discharge and water level of the Ili River arouses natural interest of researchers due to its high importance in the hydrological system of southern Kazakhstan. Much attention is paid to the Ili biosystem, the effects of climate change and the state of glaciers, water use, etc. [17–24]. For example, in one study [25] the productivity (net primary productivity-NPP) of the Ili River is considered. The proposed model of the spatial temporal distribution of NPP showed high efficiency with a coefficient of determination equal to 0.65. The author of [26] modeled the relationship between the state of wetland biomass in the delta of the Ili River and changes in the water level in Lake Balkhash. The accuracy of the achieved forecast amounted to 76%. In another study [27], the task of hydrological monitoring of the Ili River is considered on the basis of assessment of the water mirror of the Kapshagai reservoir on the Tekes River, which is the main tributary of the Ili River in the upper part of the basin in the territory of China. Due to the meager network of hydrometric stations, the task of remote monitoring of the water level of the Ili River is very urgent. Further, we consider the methods of solving a similar problem described in the literature.

3. Related Works

As indicated above, the task of calculating the river discharge is of considerable interest not only from the point of view of estimating the water reserves, but also for predicting the

excessive discharge that may cause catastrophic events [28]. A common practice of river runoff volume estimation is the use of hydrometric models estimating the discharge level on the basis of data from hydrometric stations. Such models are based on both the level and width of the river beds [29].

In addition to hydrometric models, machine learning (ML) methods have been successfully applied to solving such problems. For instance, in one study [30] the artificial neural networks (ANN) were used to forecast the water level in the Bedup River. The accuracy of the forecast was 83.5%. ANNs were also used in [31] to anticipate the water discharge and water level in the Ramganga River catchment of the Ganga Basin (India). In the studies [32,33], a multilinear regression and long short-term memory (LSTM) model was applied to predict water levels in the Guam River and the Han River, South Korea. In the latter case, it was possible to forecast the water level one hour ahead in the tidal section of the river with a fairly high accuracy (RMSE = 0.08 m). However, the accuracy of the forecast dropped significantly when forecasting for a twenty-four-hour period (RMSE = 0.28 m). A Gaussian process regression (GPR) model was applied to forecast the daily levels of the Durian Tunggal River in tropical peninsular Malaysia [28]. Models based on classical machine learning algorithms were used in the study [34] to simulate the water level in the tidal zone of the rivers of this peninsula.

To forecast the river discharge of the Hunza River, Pakistan D. Hussain et al. [35] performed a comparison of multilayer perceptron (MLP), support vector regression (SVR) and random forest (RF) algorithms. These models showed the following results: R^2 = 0.910, 0.831 and 0.993, respectively. An in situ dataset of historical river flow data for the period from 1962 to 2008 was used to train the models. In [36], RF showed the results closest to the results of the GR2M rainfall-runoff model, which justifies the possibility of applying ML in areas for which there are no physical characteristics of the basin and hydrological information. Thanh, H.V. et al. [37] reconstructed the average daily discharge in the Mekong River megadelta, Vietnam. They used RF, Gaussian process regression (GPR), support vector regression (SVR), decision tree (DT), least squares support vector machine (LSSVM), and multivariate adaptive regression spline (MARS) models. RF and MARS showed the best results (MAE = 517, 722, 200 m^3/s, for year-round, flood and dry months, respectively).

As can be seen, in all of the above cases, ground data collected over a long (several years) period of time were used to train and test ML models.

Due to the widespread reduction in hydrometric stations, remote sensing data are becoming increasingly popular for assessing the river discharge regimes [38]. In general, the use of satellite data increases the accuracy of calculation [39], reconstructing [40] and forecasting the volume of water discharge with a sufficiently high value of the Nash–Sutcliffe model coefficient of efficiency (NSE), such as in [41], where the value of NSE = 0.8 was achieved.

To solve the latter task, both optical and microwave data are used, including satellite altimetry [42,43], which is practically applicable on large water surfaces due to low spatial resolution [44]. It is noted that microwave radiometric data have a spatial resolution of about 25 km [45]. In other words, the use of such data is possible only for the largest rivers [46], lakes or reservoirs with an accuracy in the range of 0.2 to 1.05 meters [47]. The additional disadvantage of satellites with radar altimetry is their relatively low periodicity from 10 (Jason-2,3) to 35 days (SARAL/AltiKa). The satellite products based on Synthetic Aperture Radar (SAR) are considered as the most viable methods for observing the extent and level of flooding [48]. A number of studies have considered the use of SAR for estimating the lake levels with decimeter accuracy [49]. It was even reported that if the CryoSat-2 satellite trajectory mapping to the Earth surface is perpendicular to the channel or river bed, such accuracy is ensured even for river beds only a few meters wide [50].

Recently, the use of hybrid approaches, combining both hybrid machine learning models and a combination of different data sources, has been gaining popularity for estimating the water level and discharge volumes. For example, in [51], a method using several data sources (satellite-derived, climate mode indices and ground-based meteorological obser-

vations) and hybrid deep learning models for forecasting the water level of the Murray River (Australia) was proposed. Due to this, a high forecasting result was obtained (mean error—0.020 meters, accuracy 98%). The authors used data from 19 hydrometric stations. A hybrid approach of a slightly different kind is also described in the study [16]. The authors combined the hydrometric model GR6J with the ML model (LSTM) to simulate discharge in the vicinity of the virtual stations on the Yangtze River. The model is calibrated using Gravity Recovery and Climate Experiment (GRACE) data. The method is recommended for flood monitoring in the areas where no ground hydrometric stations are available.

One study [52] deals with calibration and validation of the suspended sediment and discharge models for the Tisza and the Maros rivers (Hungary) based on Sentinel-2 data. The developed models are to be used to estimate sediment discharge at engaged periods, since such measurements at the hydrometric stations are made once a month. The RF and combined models showed the suspended sediment concentration (SSC) results where R2 = 0.87, 0.82 for the specified rivers, respectively.

The problem of determining the water level in the Ili River was first considered in the study [53], where an empirical step dependence between the test fragment of the river bed and the data of the hydrometric station "164 km" was formed. The results of the analysis showed a rather high Pearson correlation coefficient (0.9). Recent work [54] used shoals on the river bed as indicators of water level and provides NSE value equal to 0.74, but with limitations in the range of water levels (not more than 280 centimeters).

The results of the literature analysis are summarized in Table 1.

Table 1. Machine learning methods and remote sensing technology in the water monitoring tasks.

Task	Study Area	Machine Learning Methods	Remote Sensing Data	Result	Ref.
1	Ramganga River catchment of the Ganga Basin, India	ANN	-	Ac = 83.5%	[30,31]
1	Guam River and the Han River, South Korea	multilinear regression and LSTM	-	RMSE = 0.08 m	[32,33]
1	Catchment located in the east coast of tropical peninsular Malaysia	SVR	-	NSE = 0.986	[28,34]
1	Murray River, Australia	CNN, LSTM, BiLSTM	MODIS	Ac = 98%	[51]
2	Lakes or reservoirs	-	satellite altimetry	accuracy in the range of 0.2 to 1.05 meters	[42–44]
2	Lakes or reservoirs	-	satellite SAR	decimeter accuracy	[49]
2	Ili River	-	Sentinel-2	*NSE* = 0.74	[54]
3	Hunza River, Pakistan	MLP, SVR, RF		*NSE* = 0.993	[35]
3	Mekong River megadelta, Vietnam	RF, GPR, SVR, DT, LSSVM, MARS	-	MAE = 200 m^3/s for dry month	[37]
4	Brahmani River basin, India	ANN, RF, SVR	Aqua-MODIS, Landsat	*NSE* > 0.85	[38]
4	Midstream Yangtze River basin	LSTM, RF	-	*NSE* = 0.69	[16]
5	Tisza and the Maros rivers, Hungary	RF and combined model	Sentinel-2	*NSE* = 0.87	[52]

Note. Task 1—Water level prediction problem; 2—water level estimation; 3—forecast of the river discharge; 4—river discharge estimation; 5—sediment discharge estimation.

For medium-sized and small meandering rivers flowing through sparsely populated or inaccessible territories and having, as a consequence, a very limited network of gauging stations, the solution of the problem of determining water levels and water discharge volumes from satellite data remains a difficult task. Application of the paradigm of space gauging stations, virtual stations or virtual gauging stations is one of the ways to solve this problem. At the same time, when using optical range data, it is important to select the distinctive locations on the riverbed that determine the water level with high accuracy [15,55].

4. Method

In the flat part of Kazakhstan, the Ili River bed has a width of 100 meters or more, which allows using Sentinel-2 satellite data with a resolution of ten meters to monitor the riverbed filling with water. In its turn, filling of the bed makes it possible to assess the water level in the river and the subsequent calculation of the volume of the river discharge.

In the present work, we have extended the scope of the approach proposed in studies [53,54]. The basic idea of the method is quite natural and consists in using the area of the river bed filled with water within some section as a water level indicator. Additionally, the assumption is made that the use of river banks for a given river can improve the accuracy of the estimation within some water level boundaries. It is assumed that under the conditions of relatively low water, the shoals are in a stable state and that their overwater area varies sufficiently with water level. In this case, the area of the shoals can be estimated using space images. It is quite reasonable to expect that within a certain water level boundary, the change in the area of the shoals is linearly related to the water level in the river [54]. However, at high water levels, shoals may disappear and the riverbed sections may change. Moreover, the previous studies did not investigate the stability of the method under significant changes in water level and variations in model input parameters. In this paper we cover this gap and show that the use of ensemble machine learning together with satellite spectral canal data gives a good result under significant water level variations in the river and, in addition, gives a slightly worse but still good result without the use of river shoals, the area of which varies significantly from year to year. We use some preprocessing and feature engineering techniques to achieve the best and the most stable result of water level forecasting.

The proposed method includes the following steps (Figure 2):

1. Formation of Sentinel-2 satellite imagery dataset.
2. Preprocessing and features engineering.
3. Training and tuning of machine learning models.
4. Evaluation of results using a specified set of quality metrics.

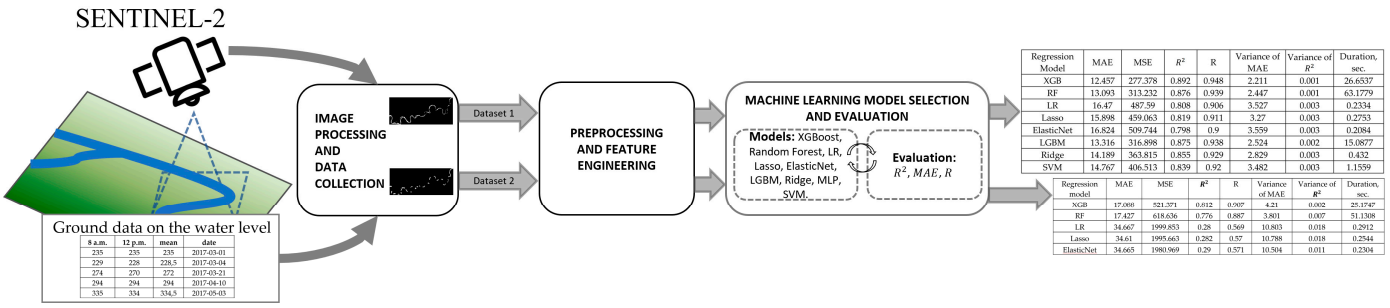

Figure 2. The structure of the proposed method.

Figure 3 shows the investigated part of the Ili River's bed from the China–Kazakhstan border to the Ili-Dobyn gauging station, which is approximately 40 km long. The river, flowing through gravelly sandy sediments, is characterized by a variable riverbed morphology, where shoals and bed sections are formed, transformed and disappear.

To estimate such a variable object, it is quite natural to apply not only linear methods of water level reconstruction based on water surface area, but also a wider range of machine learning methods.

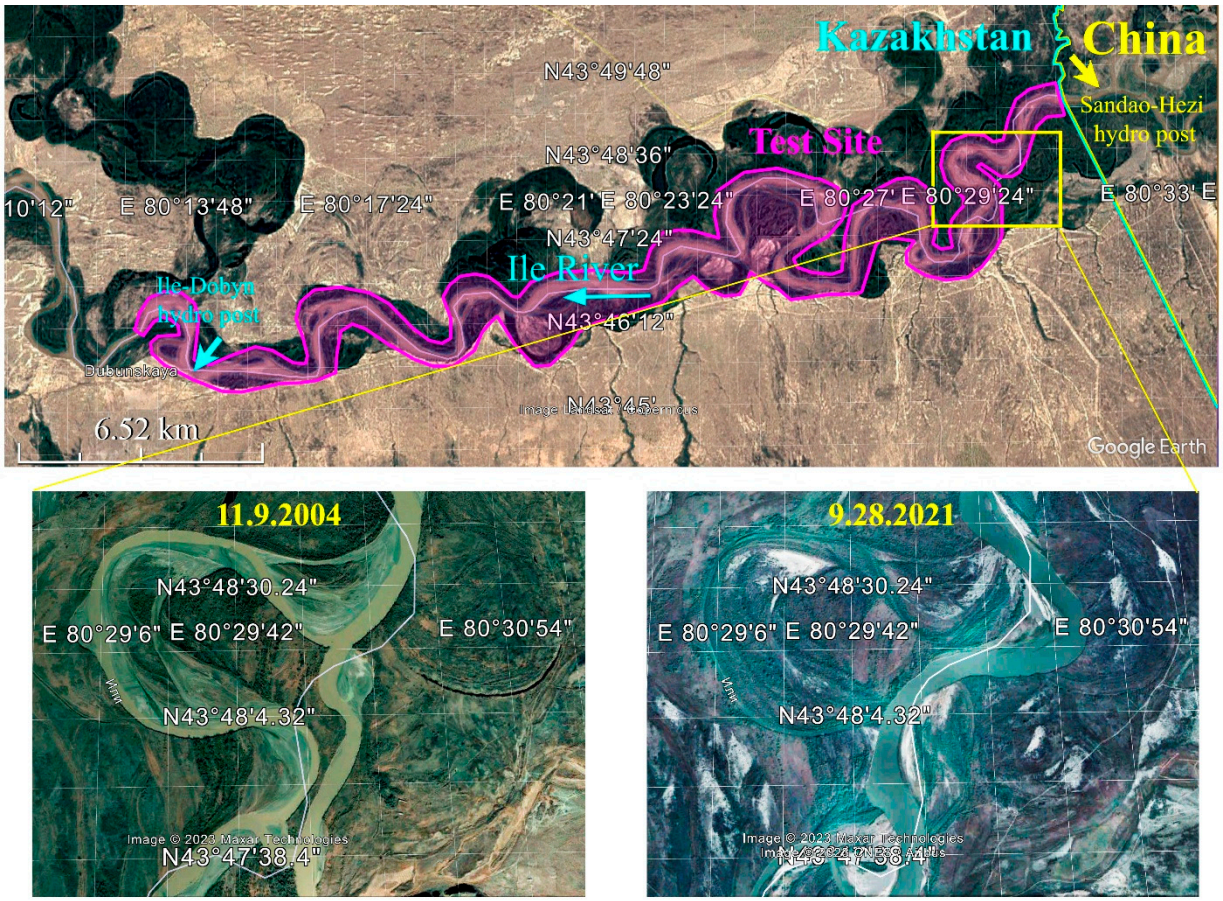

Figure 3. Investigated part of the Ili River bed (**top**). Illustration of the riverbed variability (**bottom**).

4.1. Generation of the Dataset

In the supervised learning tasks, it is important to select a target variable or target column of data, which is matched with one or several input columns (input variables) or features. If features and target variables are defined, the further research scheme is quite trivial: the machine learning algorithm looks for a relationship between the inputs in the target variable. As a target column in our task, we used ground data on the water level in the Ili River obtained at the hydrometric station "Ili-Dobyn" (43°45′31.15″N; 80°13′53.04″E) of the "Kazhydromet" system.

The hydrometric station measures the water level as the distance between the water line (in a measuring well connected to the river) and the "zero" mark of the station. This value characterizes the river's water content (water flow) at the time of measurement, at 8 a.m. and 8 p.m. (manual measurement, year 2022) or at 8 a.m., 12 p.m., 4 p.m. and 8 p.m. (automatic measurements, years 2016–2021). This water flow is considered representative for the transit part of the river bed where there is no additional water inflow. Satellite data (spectral characteristics of the channel) are obtained as an instantaneous picture at the moment of passage of Sentinel-2 (a solar synchronous satellite with a local time of passage, approximately at 11.50). The average water speed in the river varies between 2–4 m/sec, depending on the water flow, in other words, water passes through the analyzed fragment of the channel in approximately 3–6 hours. The gauging station is located in the lower part of the test section of the channel (see Figure 3); therefore, the satellite estimates the amount of water that passes through the test section from approximately 9 a.m. to 12 p.m. (high water) or from 6 a.m. to 12 p.m. (low water). In such conditions, it seems correct to compare the satellite data obtained at 11.50 with the average water level obtained at 8 a.m. and 12 p.m. Therefore, the target column was formed as an average of the gauging station readings at 8 a.m. and 12 p.m.

4.1.1. Method of Data Preparation

The input data were generated in a such way that the water surface area was used as a main input parameter. In the first case, the input parameter was the water surface area measured at the specially selected locations along the riverbed (Dataset-1). These selected zones are shown in Figure 4.

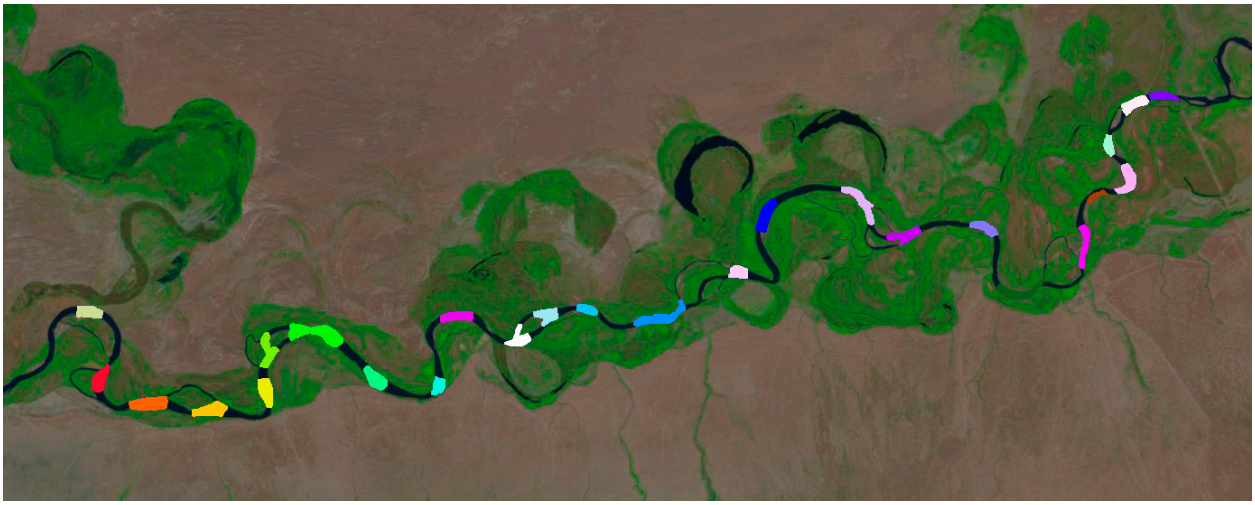

Figure 4. Distinctive sections of the Ili River's bed (marked in color), within which the total water mirror area was determined. Water flow is directed from right to left.

The assumption was made that at average seasonal water levels the morphology of the river bed is stable and that the river bed shoulders with average inundation rates between 45 and 85% can be used as water level markers. Such areas are highlighted in Figure 4 with colors. The water surface on space images was distinguished using Modified Normalized Difference Water Index (MNDWI1) [56,57]. The third and eleventh channels of satellite images were used to calculate the index:

$$\text{MNDWI1} = (\text{B3} - \text{B11})/(\text{B3} + \text{B11}) \quad (1)$$

where: B3—third (559 nm); B11—eleventh (1610 nm) channels of Sentinel-2.

To separate land from water, we used the threshold value MNDWI1 = +0.25 (Figure 5).

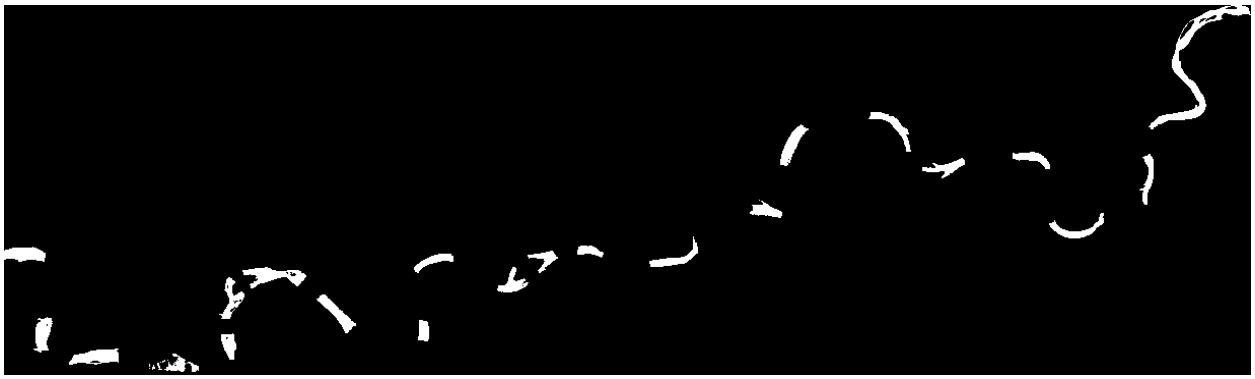

Figure 5. Water surface area of the river bed section extracted with MNDWI1.

In the course of preliminary experiments, different MNDWI1 threshold values were tested on a separate test site based on 2019 data. It was found that the best correlation between the number of "water" pixels in the shoals mask and the water level in the river, according to the Ili-Dobyn gauging station, is given by the threshold MNDWI1 = +0.25.

In the second case, for comparative experiments and assessing the influence of expert marking on the result of water level prediction, a "simplified" dataset was developed in which the entire river channel was used as an initial one without identifying the "sensitive" zones.

Once the water surface of the riverbed was extracted (Figure 6), its area was calculated and used as one of the model input parameter (Dataset 2).

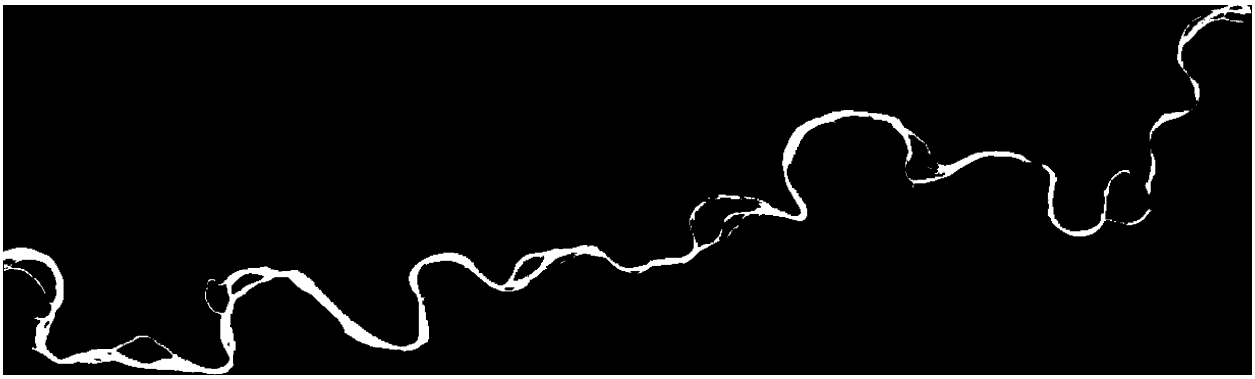

Figure 6. Riverbed delineation with application of MNDWI1.

In addition to calculating the total number of water pixels, averaged values of spectral indices of space images from B1 to B12 were additionally calculated for the selected riverbed sections. Taking into account the fact that vegetation can significantly affect the channel outline, NDVI vegetation index was additionally calculated.

$$NDVI = \frac{B5 \ - B4}{B4 \ + B5}, \tag{2}$$

where B5—Band 5, Near-Infrared (0.85–0.88 nm), B4—Band 4, Red (0.64–0.67 nm).

4.1.2. Datasets

As a result of the actions described above, two datasets (Dataset-1 and Dataset-2) were generated, each of which is based on two initial sets of data (Tables 2 and 3).

Table 2. Water levels and area of selected river bed sections (in pixels).

Mean	Date	pixelCount	pixelCount_Clo
235	1 March 2017	35,928	24,580.00
228.5	4 March 2017	12,125	6105.00
272	21 March 2017	31,851	41,113.00
294	10 April 2017	33,249	41,113.00
334.5	3 May 2017	34,572	0.00
...	...	...	...
273	30 November 2021	35,983	0.00

Table 2 contains the following indicators:

- Mean—average water level obtained from 8 a.m. and 12 p.m. measurements (target value).
- Date—date of measurements.
- pixelCount—number of pixels in the river mask.
- pixelCount_Clo—number of pixels in the image distorted by cloudiness.

A total of 276 low-cloud images (with a 10 m resolution) from the Sentinel-2 image archive for the period from 2017 to 2021 were selected.

Table 3. Sentinel-2 satellite spectral data averaged over the area of the selected water areas.

Date	NDVI	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
6 March 2017	−0.133	0.1591	0.1368	0.1267	0.1327	0.1340	0.1113	0.1129	0.1027	0.0406	0.0026	0.0743	0.0538
29 March 2017	0.0188	0.6219	0.6118	0.5793	0.6180	0.6302	0.6395	0.6554	0.6413	0.4458	0.1065	0.4236	0.3761
5 April 2017	0.0077	0.1860	0.1634	0.1524	0.1649	0.1721	0.1707	0.1787	0.1674	0.0768	0.0312	0.1130	0.0937
28 April 2017	−0.021	0.1931	0.1759	0.1694	0.1765	0.1801	0.1787	0.1894	0.1738	0.0753	0.0151	0.1373	0.1090
...	...	...	...	...	...	...	...	...	...	...	...	...	...
30 November 2021	−0.306	0.167	0.14	0.119	0.104	0.097	0.069	0.066	0.06	0.033	0.002	0.034	0.024

Note. B1, B2, . . ., B12—Band 1, Band 2, . . ., Band 12 of satellite Sentinel-2 [58] (see Appendix A Table A1).

Dataset-1 is obtained by using expert-identified zones on the riverbed (Figure 5). The pixelCount value is the sum of the water surface areas of these zones. The parameters described in Table 2 are the average values of the spectral ranges obtained over the area of these zones.

Dataset-2 uses, instead of selected zones, the entire riverbed in the 40-kilometer section under consideration (Figure 6). The pixelCount value is the water surface area of this riverbed section. As in the previous case, the average values of spectral ranges are calculated for this area.

In the process of computational experiments, some the data were considered as data with anomalous values. Firstly, these are data for the year 2017, when abnormally high volumes of river flow were observed. Secondly, a large disproportion between the value of pixelCount and mean (a large value of pixelCount and a small value of mean and vice versa) may be caused by an error in the estimation of pixelCount. Data rows with such values were excluded in some experiments. This process is described below in Section 5.1.

4.2. Machine Learning Models

Deep learning methods are highly effective in solving many practical problems. However, such methods either require large amounts of labeled data for training or the availability of a pre-trained model that can be tuned using the transfer learning technique [59]. In the current case, neither of the above features are not available. Therefore, we decided to compare several types of models of the conventional architecture, including ensemble machine learning models (gradient boosting [60] and bagging technique), support vector machines and classical regression algorithms. Machine learning models are summarized in Table 4, which slightly extends the version of the table presented in the paper [61].

Table 4. Machine learning models.

Regression Model	Abbreviation	About Method	References
Linear regression	LR	Method is based on linear approach	[62]
Lasso regression	Lasso	Based on the use of a regularization mechanism that not only helps in reducing over-fitting but it can help in feature selection	[63]
Ridge regression	Ridge	The regularization mechanism is used to prevent over-fitting	[64,65]
Elastic net	ElasticNet	Hybrid of ridge regression and lasso regularization	[66]
XGBoost	XGB	Ensemble learning method based on the gradient boosted trees algorithm	[67]
LightGBM	LGBM	Ensemble learning method based on the gradient boosted trees algorithm	[68–70]
Random forest	RF	Ensemble learning method based on bagging technique	[71]
Support vector machines	SVM	Method is based on the kernel technique	[72]
Artificial neural network or multilayer perceptron	ANN or MLP	Feed forward neural network	[73,74]

Linear regression models generally minimize the cost function in the form:

$$J(\theta) = \min \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n}\theta_j^2\right],$$

where m is the set of training examples; $x^{(i)}$ is values of parameters or properties (features) for the i-*th* object; $y^{(i)}$ is the actual value of the target variable for the i-*th* example; λ is the regularization factor; h_θ is the hypothesis function in the form $h_\theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_k$, k is the number of input parameters and $\theta_0, \theta_1, \theta_2, ..\theta_n \in \mathbf{\Theta}$ are the regression parameters.

Lasso and ridge regressions differ depending on the setting of the regression parameter.

The XGB, LGBM models implement the technique of boosting using an ensemble of algorithms of the same type (ensemble learning method based on the gradient boosted trees algorithm). The ensemble algorithms are selected so that each next algorithm is trained taking into account the error gradient of the previous ones. In other words, the next algorithm (b) is tuned so that the target value is the antigradient of the error function of the previous algorithm: $-L'(y^{(i)}, h_\theta\left(x^{(i)}\right))_{i=1}^{m}$, meaning that when training algorithm b, instead of the traditional pairs $\left(x^{(i)}, y^{(i)}\right)$ we use the pairs $(x^{(i)}, -L'\left(y^{(i)}, h_\theta\left(x^{(i)}\right)\right)$, where $h_\theta(x)$ is the previous algorithm's hypothesis function.

The RF model implements the bootstrap aggregation technique. The idea is that for each random subsample of the training dataset, a separate decision tree is built based on only part of the features. Then, voting is performed between the trees to form the final result.

The well-known SVM algorithm implements a technique based on changing the parameters of the function (core) that determines the distance between objects of different classes. The general expression of the SVM cost function is as follows:

$$J(\Theta) = C\sum_{i=1}^{m} y^{(i)} S_1(\mathbf{\Theta}^T, f_k{}^{(i)}\left(x^{(i)}\right)) + \left(1 - y^{(i)}\right) S_0\left(\mathbf{\Theta}^T, f_k{}^{(i)}\left(x^{(i)}\right)\right) + \frac{1}{2}\sum_{j=1}^{m}\theta_j^2,$$

where S_1 and S_0 are functions, which are usually piecewise linear functions and f_k is a core function that determines the significance of the objects of the training set in the feature space. A very popular Gaussian function $f_k{}^{(i)}\left(x^{(i)}\right) = exp(\frac{\left|x - x^{(i)}\right|^2}{2\delta^2})$, which for any x allows estimating its proximity to the "marker" object $x^{(i)}$ and thus forming boundaries between classes by setting the value of δ, C being the regularization parameter ($C = 1/\lambda$).

The following metrics are widely used to assess the quality of the regression models [75–77]: Mean Squared Error (MSE), Mean Absolute Error (MAE) and correlation coefficient (R). To evaluate the hydrological models, the Nash–Sutcliffe model efficiency (NSE) coefficient [78] is widely used, the formula for calculating of which coincides with the formula for calculating the coefficient of determination (R^2) (Table 5).

According to [79], a forecast model is considered to be good ($NSE \geq 0.80$), satisfactory ($0.36 \leq NSE < 0.80$) or not satisfactory ($NSE < 0.36$). However, it is recommended in [80] to compare the model with alternative models to evaluate the model. The results of such comparisons are summarized in the next section.

Since the amount of data is relatively small, the method of cross-validation of random permutations (ShuffleSplit) was employed for the model estimation, as it was used in [76]. To achieve a statistically significant result, splitting the data into test and training set is performed 100 times with averaging of the obtained result.

In other words, during the computational experiments, the dataset was randomly divided 100 times into training (90%) and test (10%) sets. Each time, machine learning models were trained and the results were evaluated. The resulting machine learning model score is the average of these evaluations.

Table 5. Evaluation metrics of regression models.

Evaluation Index	Equation
Mean Absolute Error	$MAE = \frac{\sum_{i=1}^{n} \left\| y^{(i)} - h^{(i)} \right\|}{n}$ where n is sample size;
Mean Squared Error	$MSE = \frac{\sum_{i=1}^{n} \left(y^{(i)} - h^{(i)} \right)^2}{n}$
Nash–Sutcliffe model efficiency (or determination coefficient)	$NSE = 1 - \frac{\sum_{i=1}^{n} \left(y^{(i)} - h^{(i)} \right)^2}{\sum_{i=1}^{n} \left(y^{(i)} - \overline{y} \right)^2}$, where $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$.
Linear correlation coefficient (or Pearson correlation coefficient)	$R(y,h) = \frac{\sum_{i=1}^{n} \left(h^{(i)} - \overline{h} \right)(y_i - \overline{y})}{\sum_{i=1}^{n} (y_i - \overline{y})^2 \sum_{i=1}^{n} \left(h^{(i)} - \overline{h} \right)^2}$, $\overline{h} = \frac{1}{n} \sum_{i=1}^{n} h^{(i)}$

Computational experiments were performed using a specially developed program in Python with the use of such libraries as numpy, sklearn, matplotlib, pandas, statistics, xgboost, pickle, time, shap, which provides reading, preprocessing of initial data, formation of dataframes, application of machine learning models, output of results and evaluation of the impact of input parameters. The computational experiments were performed on a computer with an Intel(R) Core(TM) i7-10750H processor, equipped with 32 GB of RAM and discrete video card Nvidia GeForce GTX 1650 Ti.

The purpose of the computational experiments is not only to select the most accurate algorithm, but also to evaluate its robustness under different combinations of input parameters and dataset content.

5. Results

5.1. Preprocessing and Features Engineering

The preliminary experiments performed using Dataset-1 showed that using pixelCount as input data and Mean as a target column gives a relatively low result (NSE = 0.46). To improve it, the initial values were pre-processed and additional input values were generated. First of all, "pixelCount_Clo", "month", "year", season (1—spring, 2—summer, 3—autumn), gradient of water mirror area change, day number in a year and the value of the area in the previous measurement were used as input values in addition to pixelCount. Secondly, we removed anomalous values, which include those in which pixelCount is high and mean water level is low and vice versa. We believe that such errors were made in the pixel count calculation process. Moreover, the area-averaged spectral ranges of the Sentinel-2 satellite were added (Table 2). These measures made it possible to increase the coefficient of determination by two times. The process of computational experiments was carried out in such a way that the new input variables were sequentially added and measures were taken to clean the input data from the anomalous values:

- Columns "pixelCount_Clo", "month", "year" were added (NSE = 0.79).
- The values for the year 2017 were removed, since many anomalous values (NSE = 0.81) were found. The records were cleared of anomalous values in the following way. First, the average pixelCount values were calculated. Then, there were removed those rows in which the pixelCount value is high (by $\gamma\sigma$ sigma is greater than the mean) and the level is low (by $\gamma\sigma$ sigma less) and vice versa, where σ is the variance of the values, γ is the empirical coefficient controlling the allowable spread of the data (the best value of NSE is obtained at γ = 0.5).
- Gradient of total river surface area values (sign of the difference between the current pixelCount and the previous one), season (1—spring, 2—summer, 3—autumn) (NSE = 0.87) were added.
- pixelCount value for previous date was added (NSE = 0.881).
- Area-averaged values of spectral ranges was added (NSE = 0.892).

The process of tuning parameters of machine learning models, the obtained optimal parameters of the models, and the process of obtaining the final result are given in Appendix B.

The results obtained by using Dataset-1 are summarized in Table 6.

Table 6. Quality metrics for machine learning models.

Regression Model	MAE	MSE	*NSE*	R	Variance of MAE	Variance of *NSE*	Duration, sec.
XGB	12.457	277.378	0.892	0.948	2.211	0.001	26.6537
RF	13.093	313.232	0.876	0.939	2.447	0.001	63.1779
LR	16.47	487.59	0.808	0.906	3.527	0.003	0.2334
Lasso	15.898	459.063	0.819	0.911	3.27	0.003	0.2753
ElasticNet	16.824	509.744	0.798	0.9	3.559	0.003	0.2084
LGBM	13.316	316.898	0.875	0.938	2.524	0.002	15.0877
Ridge	14.189	363.815	0.855	0.929	2.829	0.003	0.432
SVM	14.767	406.513	0.839	0.92	3.482	0.003	1.1559

Table 6 shows the mean values of MAE, MSE, NSE and R, as well as the variance of the estimates. It can be seen that the best results are demonstrated by boosting (XGB, LGBM) and bootstrap aggregation (RF) models. The mean error of water level estimation does not exceed 12.46 cm. Good results are also shown by the models based on linear regression. The results presented in Tables 5–9 obtained by using cross-validation of random permutations methods as mentioned in Section 4.

5.2. Robustness Analysis

In practice, incoming data may not be cleared of anomalous values, or anomalous values may not be due to physical reasons (disappearance or appearance of new shoals at high and low water levels, non-linear nature of changes in the area of the river bed, etc.).

Additional experiments were performed to check the stability of the algorithms (Appendix C):

(a) Using the full dataset for the period from 2017 to 2021
   1. Without removal of anomalous values (277 records)
   2. Removal of anomalous values at γ = 1.0 (270 records)
   3. Removal of anomalous values at γ = 0.5 (270 records)

(b) Using a reduced dataset for the period 2018 to 2021
   4. No removal of anomalous values (244 records)
   5. Removal of anomalous values at parameter γ = 1.0 (244 records)
   6. Removal of anomalous values at parameter γ = 0.5 (232 records)
   7. Additionally, all records with water levels greater than 280 cm were deleted. (164 records).

The experimental results are grouped in Figure 7, where the corresponding experiments from 1 to 7 are color-coded.

It is evident that the most robust results for all variants of data processing are demonstrated by the XGB and LGBM boosting algorithms. The smallest mean error (10.25 cm) is provided by the RF algorithm, but with a significantly reduced set of input values (164).

In the second experiment, the "simplified" dataset (Dataset-2) was used. A simplified dataset was generated based on the water surface of the entire 40-kilometer section of the river bed (Figure 6). Despite such a "simplified" approach, the XGB model results are in the "good" category ($R^2 \geq 0.80$) (Table 7). All statistical results of computational experiments mentioned in Tables 7–9 are presented in Supplementary Materials.

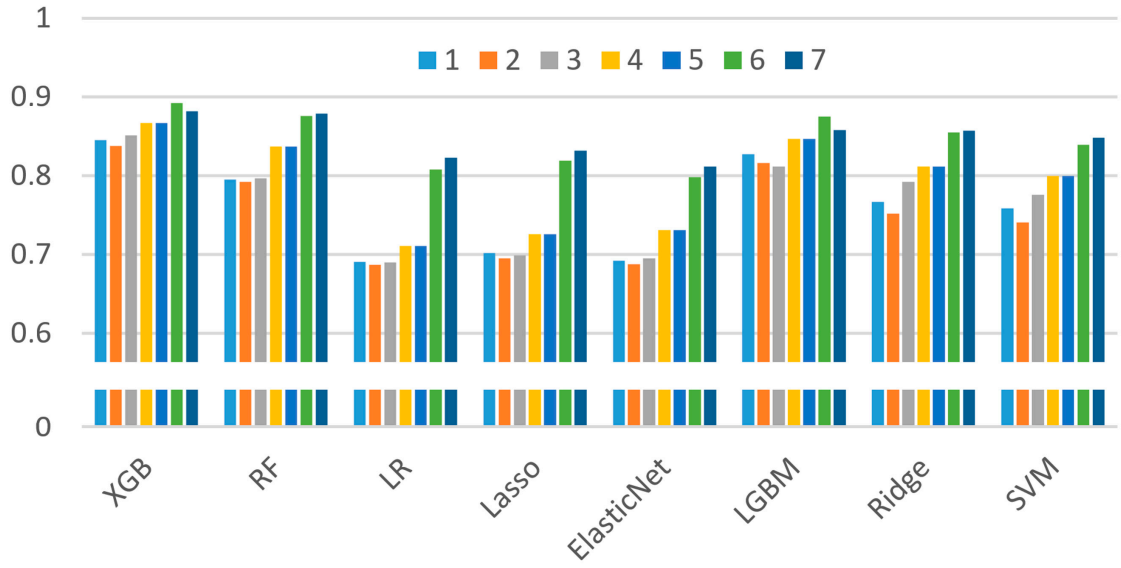

Figure 7. Comparative results of machine learning algorithms (R^2) under different data preprocessing parameters.

Table 7. Quality metrics of the "simplified" model.

Regression Model	MAE	*NSE*	R
XGB	17.088	0.812	0.907
RF	17.427	0.776	0.887
LR	34.667	0.28	0.569
Lasso	34.61	0.282	0.57
ElasticNet	34.665	0.29	0.571
LGBM	18.858	0.764	0.879
Ridge	28.191	0.493	0.719
SVM	28.941	0.415	0.687
MLP	31.728	0.393	0.697

The simulation results obtained by using the simplified approach can be slightly improved by performing the data cleaning with parameter γ = 1.0 (30 out of 277 records are marked as erroneous) (Table 8).

Table 8. Quality metrics of the "simplified" model after cleaning the input data from the anomalous values.

Regression Model	MAE	*NSE*	R
XGB	14.039	0.827	0.914
RF	14.256	0.808	0.904
LR	30.52	0.223	0.529
Lasso	30.43	0.231	0.534
ElasticNet	30.598	0.241	0.532
LGBM	15.499	0.788	0.891
Ridge	23.385	0.516	0.735
SVM	24.374	0.429	0.701
MLP	19.075	0.642	0.831

6. Discussion

The results obtained are illustrated in Figures 8 and 9. Figure 8 shows a scatterplot of the water level measured and predicted by XGB for Dataset-1. The diagonal (red) line in the figure shows the optimum line, where the prediction value coincides with the actual value (y).

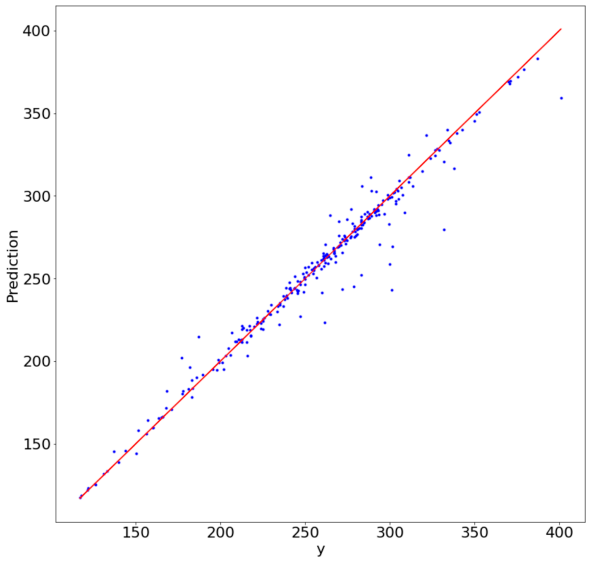

Figure 8. Scatterplots of actual and predicted water level (blue dots). The red straight line is the optimal line, where the prediction value coincides with the actual value.

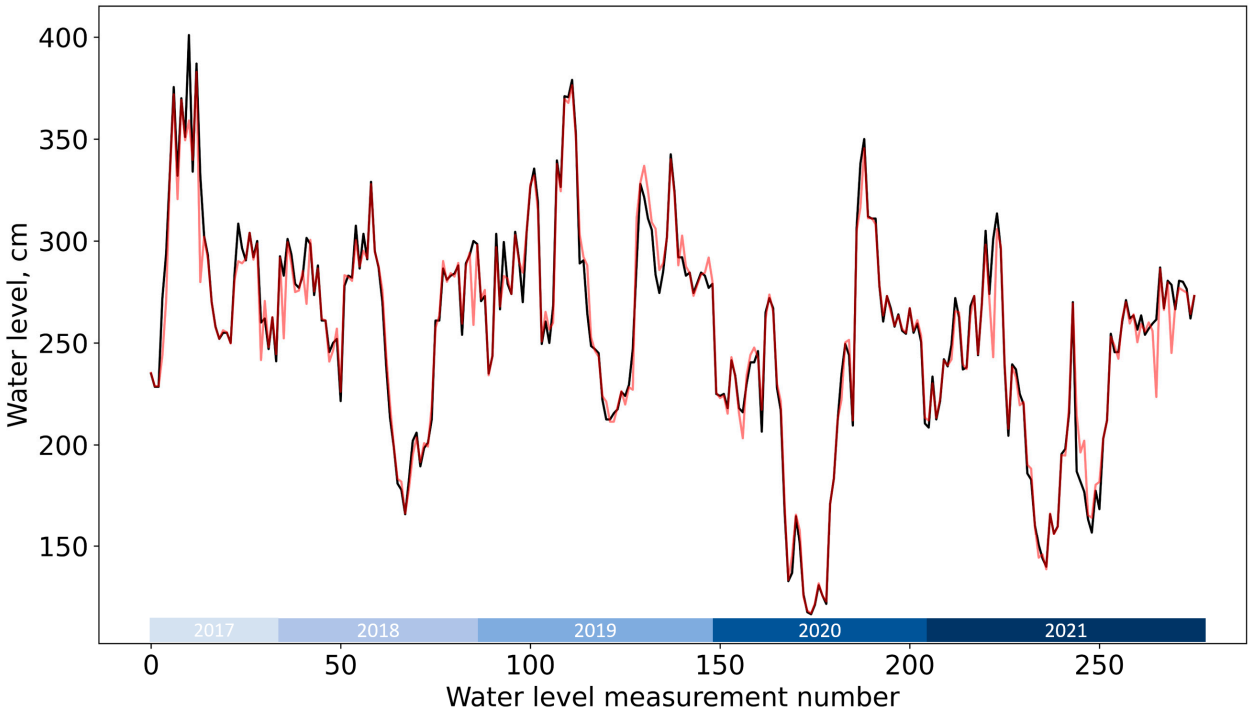

Figure 9. Comparison predicted (red line) and actual (black) water level.

Figure 9 shows the dynamics of changes in the water level measured at the gauging station (black line) and the predicted values (red line). Calculations were performed using the XGB model without removing extreme water level values.

Overall, the ensemble machine learning models demonstrated good stability results, providing an NSE value close to 0.8 using datasets 1 and 2. At the same time, the linear

models show the results that significantly depend on the preliminary marking of the river bed and data preprocessing. If the riverbed is marked by an expert (Dataset-1) and the set of input data is significantly reduced (extreme water level values are removed), linear models show a good result ($NSE > 0.8$). Linear models show unsatisfactory results (NSE = 0.36) (with the exception of ridge regression) if the markup is not performed (Dataset-2) and extreme data are not removed. The markup of the riverbed may be required quite often due to annual changes in the morphology of the Ili River channel. Therefore, the obtained results allow us to recommend the ensemble machine learning models for further use since they are more accurate and less dependent on expert marking of the river bed.

Part of the process of analyzing the output of machine learning models is to rank the input parameters based on their contribution to the model output. This allows us to identify the most influential parameters, the accuracy of which is of the most importance.

The estimation of the impact of the model input parameters was performed using the well-known agnostic model SHAP [81], which is a part of the group of explainable machine learning algorithms [82]. In contrast to the widely used Gini index, SHAP allows estimating the direction of influence and can work in the case of multicollinearity (Appendix D).

Figure 10 summarizes the results obtained. Input parameters are ranked by their influence on the modeling result.

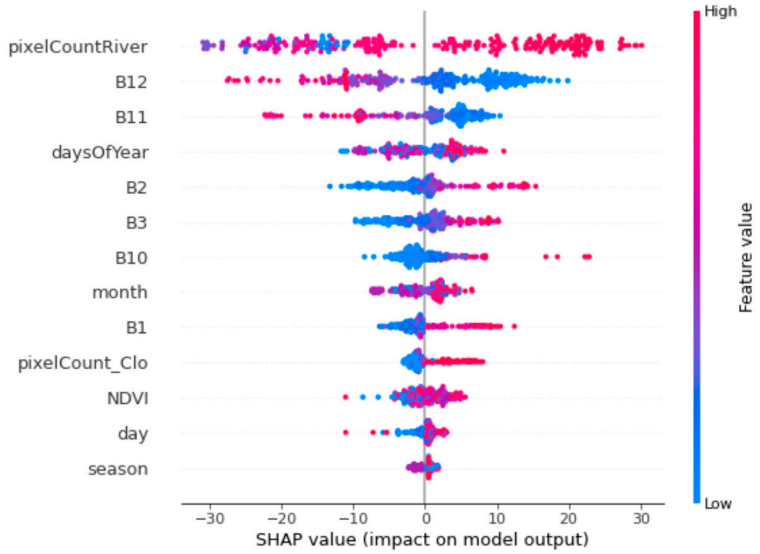

Figure 10. Level of influence of input parameters of the "simplified" model.

The most influential input parameter is PixelCountRiver (water surface area). The least influential input parameter is season. The value of each parameter affects the model outputs to a greater or lesser extent. For example, high values of PixelCountRiver (red) and low values of B12 (blue and light blue) are associated with a high value of the target variable (mean—average water level). High water level is more specific for the second half of the year (high value of daysOfYear), etc. We can see, that in some cases the high value of feature means a high water level (daysOfYear, B2, B3, B1), but in other cases it is the other way around (B12, B11). It should be noted that input variables affect the result in the aggregate, so simple removal of some insignificant variables can lead to a sharp deterioration of the result. The quality of the obtained level estimates is close to that demonstrated by the satellite altimetry and methods based on SAR data with a marginal accuracy of about 1 decimeter. The developed method is sufficiently robust to variations of input parameters and significantly exceeds the results described in publications [53,54].

The developed method was used to predict the discharge of the Ili River. For this purpose, the same riverbed section was used, but the value of daily river discharge was used as the target value of the model. The obtained results are summarized in Table 9.

Table 9. Quality metrics of machine learning models predicting the Ili River discharge.

Regression Model	MAE	*NSE*	R
XGB	37.219	**0.766**	0.88
RF	38.169	**0.756**	0.876
LR	53.537	0.528	0.746
Lasso	54.367	0.515	0.738
ElasticNet	62.273	0.431	0.688
LGBM	40.808	**0.72**	0.854
Ridge	47.68	0.586	0.779
SVM	49.489	0.585	0.779

Once again, the ensemble machine learning (XGB, RF, LGBM) demonstrates the best results (highlighted in bold).

Once again, the ensemble machine learning (XGB, RF, LGBM) demonstrates the best results (highlighted in bold). The results are close to good in accordance with [79] (R^2 close to 0.8).

The average error is about 37 cubic meters per second. In the available dataset, the average discharge value is 378 cubic meters per second and the maximum discharge value is 904 cubic meters per second.

However, in general, the quality indicators of the model are lower than for water level forecasting. It can be assumed that forecasting the discharge requires a longer bed length or a threshold function MNDWI1 as a function of suspended material content in the water. The large amount of waterborne suspended sediment changes the typical MNDWI1 values for the water mirror.

Despite the good results, the proposed method has certain drawbacks. The following main limitations of the method can be identified:

- Limited spatial resolution, making the method suitable only for relatively large rivers.
- Dependence on the state of the atmosphere. The method does not allow determining the water level in case of significant cloud cover.
- Limited applicability of the model for other regions with different riverbed morphology.
- High accuracy of the method depends on expert marking of the riverbeds.

7. Conclusions

Changes in the volume of river discharge not only have a significant impact on economic activity but can lead to catastrophic events. For this reason, estimation of the water level in rivers and calculation of discharge volumes are important tasks that are widely performed at hydrometric stations. In the case of a sparse network of hydrometric stations or their absence in inaccessible or sparsely populated areas, it is possible to use the remote sensing data as a temporary or relatively permanent measure of water level control. The use of such virtual gauging stations reduces costs, increases the reliability of level measurements and allows implementation of the reverse analysis of data, which is important in reconstructing events resulting in emergency situations. For mountain rivers with meandering beds, this method of water level estimation is additionally justified by the fact that the installation of permanent hydrometric stations on the constantly changing channel is difficult, and the accuracy of measurements is not guaranteed. In this paper, a relatively simple method of water level estimation based on the use of machine learning is proposed; its advantages are satisfactory accuracy (NSE > 0.80) and robustness in a wide range of values of input parameters. The obtained results are the best ones for today in the task of predicting the water level in the Ili River using the remote sensing data.

At the same time, the implementation of the virtual stations requires calibration of the remote sensing data using the information from ground hydrometric stations, which is limited by the frequency of satellite flights, spatial resolution of images, temporal syn-

chronization of satellite flights with measurements at the station and weather. Despite the limitations, the proposed method of the river water level estimation using the ensemble machine learning provides acceptable accuracy for the practice; it is quite stable and can be used as a duplicate in case of a limited number of hydrological gauging stations. In the future it is planned:

- To assess the applicability of the methods for other large rivers of South Kazakhstan.
- To evaluate the possibility of using SAR data to improve the accuracy of estimating the width of the river bed.
- To apply the methods of image processing using deep learning models, for example, convolutional networks, which will require a significant increase in the set of input data.
- To investigate the relationship between the length of the virtual gauging station and the accuracy of the forecast.
- More precise tuning of parameters and hyperparameters of machine learning models using, for example, evolutionary programming.

Supplementary Materials: The following supporting information can be downloaded at: https://www.dropbox.com/sh/01vasbyvom5ckz3/AADvTeJhyeOTL3HFZ6D1cxDYa?dl=0, (accessed on 23 November 2023).

Author Contributions: Conceptualization, A.T. and R.I.M.; methodology R.I.M. and A.T.; software, R.I.M., G.S., A.S. and Y.K.; validation, Y.P., V.G. and A.S..; formal analysis, R.I.M.; investigation, R.I.M., Y.K.; resources, N.A. and V.G.; data curation, N.A. and G.S.; writing—original draft preparation, R.I.M. and A.T.; writing—review and editing, Y.P. and A.S.; visualization, R.I.M., A.T. and G.S.; supervision, R.I.M.; project administration, Y.A. and A.T.; funding acquisition, Y.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan Grant No. BR18574144 "Development of a data mining system for monitoring dams and other engineering structures under the conditions of man-made and natural impacts" and Grant No. BR21881908 "Complex of urban ecological support (CUES)".

Data Availability Statement: The data presented in this study and program codes are openly available in https://www.dropbox.com/sh/x3ejdbi69k5amgb/AADk3khzCjTl6Rv-FLjeAz55a?dl=0, (accessed on 23 November 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ANN	Artificial Neural Network
BiLSTM	Bidirectional LSTM
CNN	Convolution Neural Network
DT	decision tree
GPR	Gaussian process regression
LSSVM	least squares support vector machine
LSTM	long short-term memory
MAE	Mean Absolute Error
MARS	multivariate adaptive regression spline
ML	Machine Learning
MLP	multilayer perceptron
MSE	Mean Squared Error
NSE (R^2)	Nash–Sutcliffe model coefficient
RBF	Radial Basis Function Neural Network
RF	Random Forest
SHAP	Shapley Additive exPlanations
SVM	Support Vector Machine
SVR	Support Vector Regression
XGB	eXtreme Gradient Boosting

Appendix A

Table A1. Spectral ranges Sentinel-2.

Sentinel-2 Bands	Sentinel-2A		Sentinel-2B		
	Central Wavelength (nm)	Bandwidth (nm)	Central Wavelength (nm)	Bandwidth (nm)	Spatial Resolution (m)
Band 1—Coastal aerosol	442.7	21	442.2	21	60
Band 2—Blue	492.4	66	492.1	66	10
Band 3—Green	559.8	36	559.0	36	10
Band 4—Red	664.6	31	664.9	31	10
Band 5—Vegetation red edge	704.1	15	703.8	16	20
Band 6—Vegetation red edge	740.5	15	739.1	15	20
Band 7—Vegetation red edge	782.8	20	779.7	20	20
Band 8—NIR	832.8	106	832.9	106	10
Band 8A—Narrow NIR	864.7	21	864.0	22	20
Band 9—Water vapor	945.1	20	943.2	21	60
Band 10—SWIR—Cirrus	1373.5	31	1376.9	30	60
Band 11—SWIR	1613.7	91	1610.4	94	20
Band 12—SWIR	2202.4	175	2185.7	185	20

Appendix B. Tuning Parameters and Hyperparameters of Machine Learning Models

Several steps were preceded obtaining the final result characterizing the quality of the machine learning models.

(1) At the first step, we searched for the optimal combination of input parameters of the models. For this we used the MLextend library [83,84].

(2) In the second step, hyperparameters were configured using the GrigSearch() method. The custom algorithms were trained by splitting the dataset once into training and testing. Table A2 lists the tunable hyperparameters of the regression models and their best combinations found using GrigSearch().

Table A2. Hyperparameters of regression models and their best combinations found using GrigSearch().

Regression Model	Model Parameters	Best Params
RF	'max_depth': [32, 16, 8, 4, 2], 'n_estimators': [50, 100, 400, 1000], 'max_features': [4, 7, 14]	max_depth = 16, max_features = 4, n_estimators = 100
SVR	'kernel': ['linear','rbf'], 'C': [0.015, 0.03, 0.05, 0.025, 0.03, 1, 100, 1000, 2000, 3000], 'gamma': [0.01, 0.08, 0.1, 0.15, 0.2, 0.25, 1]	C = 1000, Gamma = 0.2, Kernel = 'rbf'
LGBM	'learning_rate': [0.01, 0.1, 0.25, 0.6, 0.7], 'max_depth': [32, 16, 8, 4, 2], 'n_estimators': [1000, 400, 50, 100], 'min_child_samples': [2, 10, 20, 50], 'min_child_weight': [0.0001, 0.001, 0.01, 0.1, 1,2]	learning_rate = 0.01, max_depth = 2, min_child_weight = 2, min_child_samples = 2, n_estimators = 1000
XGB	'gamma': [0, 0.1, 0.2, 0.8, 3.2, 12.8, 25.6, 102.4, 200], 'learning_rate': [0.01, 0.1, 0.25, 0.6, 0.7], 'max_depth': [32, 16, 8, 4, 2] 'n_estimators': [50, 100, 400, 1000] 'colapse_bytree': [0.1, 0.2, 0.4], 'min_child_weight': [2, 4, 6]	Gamma = 0.0, learning_rate = 0.25, max_depth = 8 n_estimators = 400 colapse_bytree = 0.1 min_child_weight = 2

(3) Then all algorithms were performed with a 50-fold split into training and test. Algorithms with poor performance were excluded from their totality, the execution time

of which was much (several times) above average (some regression models based on SVM).

(4) At the last stage, the algorithms were again trained and tested with a 200-fold split into training and test datasets, using the method of cross-validation of random permutations (ShuffleSplit). The results obtained are shown in Tables 5–9.

For the LR, Lasso, Ridge, ElasticNet, SVM, MLP models, the input data was normalized using MinMaxScaler().

Appendix C. Results of Computational Experiments at Different Processing of Initial Data

Table A3. Experiment 1. The full dataset for the period from 2017 to 2021 is used (277 records).

Regression Model	MAE	MSE	NSE	R	Variance of MAE	Variance of R^2	Duration
XGB	15.024	425.157	0.845	0.924	3.881	0.002	28.9665
RF	16.884	561.712	0.795	0.897	3.609	0.004	73.8551
LR	21.171	846.091	0.691	0.842	5.241	0.007	0.2573
Lasso	20.934	816.489	0.702	0.847	5.194	0.006	0.2942
ElasticNet	22.077	846.867	0.692	0.842	5.312	0.005	0.2165
LGBM	16.258	473.927	0.827	0.913	3.435	0.003	17.3745
Ridge	17.877	631.565	0.767	0.885	5.53	0.015	0.4069
SVM	18.049	657.847	0.759	0.879	6.382	0.012	1.1061

Table A4. Experiment 2. The anomalous values at $\gamma = 1.0$ are removed from dataset (270 records are used).

Regression Model	MAE	MSE	NSE	R	Variance of MAE	Variance of R^2	Duration
XGB	15.068	432.232	0.838	0.92	3.301	0.002	26.9609
RF	16.46	540.478	0.792	0.894	3.876	0.004	73.702
LR	20.741	816.747	0.687	0.84	6.301	0.007	0.2575
Lasso	20.493	793.558	0.695	0.843	6.016	0.007	0.2713
ElasticNet	21.513	813.917	0.688	0.838	5.291	0.005	0.2077
LGBM	16.173	478.723	0.816	0.907	4.111	0.003	16.5896
Ridge	17.654	639.832	0.752	0.879	5.602	0.018	0.4608
SVM	17.887	670.769	0.741	0.871	5.114	0.016	1.1679

Table A5. Experiment 3. The anomalous values at $\gamma = 0.5$ are removed from dataset (270 records) (270 records are used).

Regression Model	MAE	MSE	NSE	R	Variance of MAE	Variance of R^2	Duration
XGB	14.659	401.71	0.851	0.926	4.214	0.002	26.0439
RF	16.24	514.742	0.797	0.899	3.697	0.005	70.7347
LR	20.177	786.327	0.69	0.845	4.884	0.011	0.2214
Lasso	19.696	765.638	0.699	0.847	4.352	0.009	0.2673
ElasticNet	20.734	776.286	0.695	0.843	3.917	0.007	0.2154
LGBM	16.097	477.845	0.812	0.905	3.957	0.004	16.6644
Ridge	16.502	526.122	0.792	0.897	3.918	0.006	0.4448
SVM	17.013	569.296	0.776	0.888	4.308	0.006	1.1549

Table A6. Experiment 4. The reduced dataset for the period 2018 to 2021 is used (244 records).

Regression Model	MAE	MSE	NSE	R	Variance of MAE	Variance of R^2	Duration
XGB	13.485	334.454	0.867	0.935	3.172	0.001	27.979
RF	14.91	413.625	0.837	0.919	4.017	0.003	68.1341
LR	19.406	726.094	0.711	0.857	6.697	0.012	0.2553
Lasso	18.865	688.109	0.726	0.863	6.663	0.011	0.356
ElasticNet	19.008	678.672	0.731	0.863	6.495	0.008	0.2832
LGBM	14.915	388.547	0.847	0.924	3.043	0.002	16.4032
Ridge	15.794	473.48	0.812	0.909	4.2	0.005	0.4099
SVM	16.517	509.542	0.8	0.9	5.605	0.004	1.0911

Table A7. Experiment 5. The anomalous values at parameter γ = 1.0 were removed (244 records are used).

Regression Model	MAE	MSE	NSE	R	Variance of MAE	Variance of R^2	Duration
XGB	13.485	334.454	0.867	0.935	3.172	0.001	26.9818
RF	14.908	413.881	0.837	0.919	4.145	0.003	66.2989
LR	19.406	726.094	0.711	0.857	6.697	0.012	0.2304
Lasso	18.865	688.109	0.726	0.863	6.663	0.011	0.3212
ElasticNet	19.008	678.672	0.731	0.863	6.495	0.008	0.1995
LGBM	14.915	388.547	0.847	0.924	3.043	0.002	16.1009
Ridge	15.794	473.48	0.812	0.909	4.2	0.005	0.363
SVM	16.517	509.542	0.8	0.9	5.605	0.004	1.126

Table A8. Experiment 6. The anomalous values at parameter γ = 0.5 were removed (232 records are used).

Regression Model	MAE	MSE	NSE	R	Variance of MAE	Variance of R^2	Duration
XGB	12.457	277.378	0.892	0.948	2.211	0.001	26.6537
RF	13.093	313.232	0.876	0.939	2.447	0.001	63.1779
LR	16.47	487.59	0.808	0.906	3.527	0.003	0.2334
Lasso	15.898	459.063	0.819	0.911	3.27	0.003	0.2753
ElasticNet	16.824	509.744	0.798	0.9	3.559	0.003	0.2084
LGBM	13.316	316.898	0.875	0.938	2.524	0.002	15.0877
Ridge	14.189	363.815	0.855	0.929	2.829	0.003	0.432
SVM	14.767	406.513	0.839	0.92	3.482	0.003	1.1559

Table A9. Experiment 7. The reduced dataset for the period 2018 to 2021 are used. Additionally, all records with water levels greater than 280 cm were deleted. (164 records are used).

Regression Model	MAE	MSE	R^2	R	Variance of MAE	Variance of R^2	Duration
XGB	10.267	192.933	0.882	0.943	3.021	0.003	24.6675
RF	10.253	196.141	0.879	0.942	2.84	0.004	54.0638
LR	12.432	288.842	0.823	0.916	3.572	0.007	0.2274
Lasso	12.663	274.545	0.832	0.92	2.764	0.005	0.2245
ElasticNet	13.607	308.322	0.812	0.909	3.294	0.005	0.1985
LGBM	11.423	232.186	0.858	0.931	3.157	0.004	12.1241
Ridge	11.269	232.433	0.857	0.932	3.045	0.003	0.3597
SVM	11.616	245.88	0.848	0.929	3.064	0.004	0.8737

Appendix D. The Difference between SHAP Value and Gini Impurity Index

In the group of machine learning methods using decision trees, the importance of a feature can be computed as the total reduction in the criterion brought by that feature (Gini importance). However, Gini importance does not allow us to specify in which direction this property is affected when it is increased or decreased. Moreover, its main use is to properly partition a decision tree into subtrees. At the same time, SHAP (Shapley Additive exPlanations) allows estimating the direction of influence and can work in the case of a significant dependence between the input parameters. It uses a game theory approach to determine the feature importance in the machine learning models. Its essence is as follows. We must assign an importance value to each property that reflects the impact on model prediction when that property is enabled. To calculate this effect, the model f(S∪{i}) is trained with this property, and the other model—f(S)—is trained with the property excluded. The predictions of these two models are then compared on the current input signal f(S∪{i} (xS∪{i}))—fS (xS), where xS represents the values of the input properties in the set S. Since the effect of eliminating a feature depends on other features in the model, the specified difference is calculated for all possible subsets of S⊆n\{i}. The weighted average of all possible differences is then calculated:

$$\varphi_i = \sum_{S \subseteq \{1,2,\ldots,n\} \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} (f(S \cup \{i\}) - f(S)),$$

These quantities are called SHAP values. Having calculated these values for all parameters, we can then compare them with each other to identify the most significant ones.

References

1. Terekhov, A.; Abaev, N.; Lagutin, E. Satellite monitoring of the Sardobinsky reservoir in the Syrdarya River Basin (Uzbekistan) before and after the dam breach on May 1, 2020. *Mod. Probl. Earth Remote Sens. Space* **2020**, *17*, 255–260.
2. In Kazakhstan, 268 Dams Were Recognized as Dangerous. Available online: https://vesti.kz/society/v-kazahstane-268-plotin-priznali-opasnyimi-44002 (accessed on 4 September 2023).
3. Wang, X.; Yang, W. Water quality monitoring and evaluation using remote sensing techniques in China: A systematic review. *Ecosyst. Health Sustain.* **2019**, *5*, 47–56. [CrossRef]
4. Kapalanga, T.S.; Hoko, Z.; Gumindoga, W.; Chikwiramakomo, L. Remote-sensing-based algorithms for water quality monitoring in Olushandja Dam, north-central Namibia. *Water Supply* **2021**, *21*, 1878–1894. [CrossRef]
5. Yang, H.; Kong, J.; Hu, H.; Du, Y.; Gao, M.; Chen, F. A review of remote sensing for water quality retrieval: Progress and challenges. *Remote Sens.* **2022**, *14*, 1770. [CrossRef]
6. Tarpanelli, A.; Brocca, L.; Lacava, T.; Melone, F.; Moramarco, T.; Faruolo, M.; Pergola, N.; Tramutoli, V. Toward the estimation of river discharge variations using MODIS data in engaged basins. *Remote Sens. Environ.* **2013**, *136*, 47–55. [CrossRef]
7. Riggs, R.M.; Allen, G.H.; David, C.H.; Lin, P.; Pan, M.; Yang, X.; Gleason, C. RODEO: An algorithm and Google Earth Engine application for river discharge retrieval from Landsat. *Environ. Model. Softw.* **2022**, *148*, 105254. [CrossRef]

8. Psomiadis, E.; Tomanis, L.; Kavvadias, A.; Soulis, K.X.; Charizopoulos, N.; Michas, S. Potential dam breach analysis and flood wave risk assessment using HEC-RAS and remote sensing data: A multicriteria approach. *Water* **2021**, *13*, 364. [CrossRef]
9. Bhattacharya, B.; Mazzoleni, M.; Ugay, R. Flood inundation mapping of the sparsely gauged large-scale Brahmaputra Basin using remote sensing products. *Remote Sens.* **2019**, *11*, 501. [CrossRef]
10. Zeng, Y.; Meng, X.; Zhang, Y.; Dai, W.; Fang, N.; Shi, Z. Estimation of the volume of sediment deposited behind check dams based on UAV remote sensing. *J. Hydrol.* **2022**, *612*, 128143. [CrossRef]
11. Zou, W.; Zhou, Y.; Wang, S.; Wang, F.; Wang, L.; Zhao, Q.; Liu, W.; Zhu, J.; Xiong, Y.; Wang, Z. Using a single remote-sensing image to calculate the height of a landslide dam and the maximum volume of a lake. *Nat. Hazards Earth Syst. Sci.* **2022**, *22*, 2081–2097. [CrossRef]
12. Silveira Kupssinskü, L.; Thomassim Guimarães, T.; Menezes de Souza, E.; Zanotta, D.; Roberto Veronez, M.; Gonzaga Jr, L.; Mauad, F.F. A method for chlorophyll-a and suspended solids prediction through remote sensing and machine learning. *Sensors* **2020**, *20*, 2125. [CrossRef] [PubMed]
13. Xiao, Y.; Guo, Y.; Yin, G.; Zhang, X.; Shi, Y.; Hao, F.; Fu, Y. UAV multispectral image-based urban river water quality monitoring using stacked ensemble machine learning algorithms—A case study of the Zhanghe river, China. *Remote Sens.* **2022**, *14*, 3272. [CrossRef]
14. Feng, C.; Zhang, H.; Wang, S.; Li, Y.; Wang, H.; Yan, F. Structural damage detection using deep convolutional neural network and transfer learning. *KSCE J. Civ. Eng.* **2019**, *23*, 4493–4502. [CrossRef]
15. Mukhamedjanov, I.; Konstantinova, A.; Lupyan, E.; Umirzakov, G. Assessment of capabilities of satellite monitoring of the river discharge dynamics on the example of analyzing the Amudarya river condition. *Mod. Probl. Remote Sens. Earth Space* **2022**, *1*, 87.
16. Xiong, J.; Guo, S.; Yin, J. Discharge estimation using integrated satellite data and hybrid model in the midstream Yangtze River. *Remote Sens.* **2021**, *13*, 2272. [CrossRef]
17. Imentai, A.; Thevs, N.; Schmidt, S.; Nurtazin, S.; Salmurzauli, R. Vegetation, fauna, and biodiversity of the Ili Delta and southern Lake Balkhash—A review. *J. Great Lakes Res.* **2015**, *41*, 688–696. [CrossRef]
18. Talipova, E.; Shrestha, S.; Alimkulov, S.; Nyssanbayeva, A.; Tursunova, A.; Isakan, G. Influence of climate change and anthropogenic factors on the Ili River basin streamflow, Kazakhstan. *Arab. J. Geosci.* **2021**, *14*, 1756. [CrossRef]
19. Kogutenko, L.; Severskiy, I.; Shahgedanova, M.; Lin, B. Change in the Extent of Glaciers and Glacier Runoff in the Chinese Sector of the Ile River Basin between 1962 and 2012. *Water* **2019**, *11*, 1668. [CrossRef]
20. Duskayev, K.; Myrzakhmetov, A.; Zhanabayeva, Z.; Klein, I. Features of the sediment runoff regime downstream the Ile river. *J. Ecol. Eng.* **2020**, *21*, 117–125. [CrossRef]
21. Thevs, N.; Nurtazin, S.; Beckmann, V.; Salmyrzauli, R.; Khalil, A. Water consumption of agriculture and natural ecosystems along the Ili River in China and Kazakhstan. *Water* **2017**, *9*, 207. [CrossRef]
22. Pueppke, S.G.; Zhang, Q.; Nurtazin, S.T. Irrigation in the Ili River basin of Central Asia: From ditches to dams and diversion. *Water* **2018**, *10*, 1650. [CrossRef]
23. Pueppke, S.G.; Nurtazin, S.T.; Graham, N.A.; Qi, J. Central Asia's Ili River ecosystem as a wicked problem: Unraveling complex interrelationships at the interface of water, energy, and food. *Water* **2018**, *10*, 541. [CrossRef]
24. Li, Y.; Song, Y.; Fitzsimmons, K.E.; Chen, X.; Wang, Q.; Sun, H.; Zhang, Z. New evidence for the provenance and formation of loess deposits in the Ili River Basin, Arid Central Asia. *Aeolian Res.* **2018**, *35*, 1–8. [CrossRef]
25. Jiao, W.; Chen, Y.; Li, W.; Zhu, C.; Li, Z. Estimation of net primary productivity and its driving factors in the Ili River Valley, China. *J. Arid Land* **2018**, *10*, 781–793. [CrossRef]
26. Propastin, P.A. Simple model for monitoring Balkhash Lake water levels and Ili River discharges: Application of remote sensing. *Lakes Reserv. Res. Manag.* **2008**, *13*, 77–81. [CrossRef]
27. Terekhov, A.; Pak, I.; Dolgikh, S. LANDSAT 5, 7, 8 and DEM data in the task of monitoring the hydrological regime of the Kapshagai reservoir on the Tekes River (Chinese part of the Ile River Basin). *Mod. Probl. Remote Sens. Earth Space* **2015**, *12*, 174–182.
28. Ahmed, A.N.; Yafouz, A.; Birima, A.H.; Kisi, O.; Huang, Y.F.; Sherif, M.; Sefelnasr, A.; El-Shafie, A. Water level prediction using various machine learning algorithms: A case study of Durian Tunggal river, Malaysia. *Eng. Appl. Comput. Fluid Mech.* **2022**, *16*, 422–440. [CrossRef]
29. Brakenridge, G.R.; Cohen, S.; Kettner, A.J.; De Groeve, T.; Nghiem, S.V.; Syvitski, J.P.; Fekete, B.M. Calibration of satellite measurements of river discharge using a global hydrology model. *J. Hydrol.* **2012**, *475*, 123–136. [CrossRef]
30. Bustami, R.; Bessaih, N.; Bong, C.; Suhaili, S. Artificial Neural Network for Precipitation and Water Level Predictions of Bedup River. *IAENG Int. J. Comput. Sci.* **2007**, *34*, 2.
31. Khan, M.; Hasan, F.; Panwar, S.; Chakrapani, G.J. Neural network model for discharge and water-level prediction for Ramganga River catchment of Ganga Basin, India. *Hydrol. Sci. J.* **2016**, *61*, 2084–2095. [CrossRef]
32. Jung, S.; Lee, D.; Lee, K. Prediction of river water level using deep-learning open library. *J. Korean Soc. Hazard Mitig.* **2018**, *18*, 1–11. [CrossRef]
33. Jung, S.; Cho, H.; Kim, J.; Lee, G. Prediction of water level in a tidal river using a deep-learning based LSTM model. *J. Korea Water Resour. Assoc.* **2018**, *51*, 1207–1216.
34. Tao, H.; Al-Bedyry, N.K.; Khedher, K.M.; Shahid, S.; Yaseen, Z.M. River water level prediction in coastal catchment using hybridized relevance vector machine model with improved grasshopper optimization. *J. Hydrol.* **2021**, *598*, 126477. [CrossRef]

35. Hussain, D.; Khan, A.A. Machine learning techniques for monthly river flow forecasting of Hunza River, Pakistan. *Earth Sci. Inform.* **2020**, *13*, 939–949. [CrossRef]
36. Ditthakit, P.; Pinthong, S.; Salaeh, N.; Binnui, F.; Khwanchum, L.; Pham, Q.B. Using machine learning methods for supporting GR2M model in runoff estimation in an engaged basin. *Sci. Rep.* **2021**, *11*, 19955. [CrossRef] [PubMed]
37. Thanh, H.V.; Binh, D.V.; Kantoush, S.A.; Nourani, V.; Saber, M.; Lee, K.K.; Sumi, T. Reconstructing daily discharge in a megadelta using machine learning techniques. *Water Resour. Res.* **2022**, *58*, e2021WR031048. [CrossRef]
38. Sahoo, D.P.; Sahoo, B.; Tiwari, M.K.; Behera, G.K. Integrated remote sensing and machine learning tools for estimating ecological flow regimes in tropical river reaches. *J. Environ. Manag.* **2022**, *322*, 116121. [CrossRef]
39. Bjerklie, D.M.; Birkett, C.M.; Jones, J.W.; Carabajal, C.; Rover, J.A.; Fulton, J.W.; Garambois, P.-A. Satellite remote sensing estimation of river discharge: Application to the Yukon River Alaska. *J. Hydrol.* **2018**, *561*, 1000–1018. [CrossRef]
40. Fok, H.S.; Chen, Y.; Zhou, L. Daily runoff and its potential error sources reconstructed using individual satellite hydrological variables at the basin upstream. *Front. Earth Sci.* **2022**, *10*, 821592. [CrossRef]
41. Hirpa, F.A.; Hopson, T.M.; De Groeve, T.; Brakenridge, G.R.; Gebremichael, M.; Restrepo, P.J. Upstream satellite remote sensing for river discharge forecasting: Application to major rivers in South Asia. *Remote Sens. Environ.* **2013**, *131*, 140–151. [CrossRef]
42. Koblinsky, C.J.; Clarke, R.T.; Brenner, A.; Frey, H. *Measurement of River Level Variations with Satellite Altimetry*; Wiley Online Library: Hoboken, NJ, USA, 1993.
43. Tarpanelli, A.; Camici, S.; Nielsen, K.; Brocca, L.; Moramarco, T.; Benveniste, J. Potentials and limitations of Sentinel-3 for river discharge assessment. *Adv. Space Res.* **2021**, *68*, 593–606. [CrossRef]
44. Jason-3 Altimetry Mission. Available online: https://www.eoportal.org/satellite-missions/jason-3#mission-capabilities (accessed on 4 September 2023).
45. Lebedev, S.; Kostyanoy, A.; Popov, S. Satellite altimetry of the Barents Sea. *Sovrem. Probl. Distantsionnogo Zondirovaniya Zemli Iz Kosmosa* **2021**, *12*, 194–212. [CrossRef]
46. Vittucci, C.; Guerriero, L.; Ferrazzoli, P.; Rahmoune, R.; Barraza, V.; Grings, F. River water level prediction using passive microwave signatures—A case study: The Bermejo Basin. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 3903–3914. [CrossRef]
47. Verma, K.; Nair, A.S.; Jayaluxmi, I.; Karmakar, S.; Calmant, S. Satellite altimetry for Indian reservoirs. *Water Sci. Eng.* **2021**, *14*, 277–285. [CrossRef]
48. Grimaldi, S.; Li, Y.; Pauwels, V.R.; Walker, J.P. Remote sensing-derived water extent and level to constrain hydraulic flood forecasting models: Opportunities and challenges. *Surv. Geophys.* **2016**, *37*, 977–1034. [CrossRef]
49. Göttl, F.; Dettmering, D.; Müller, F.L.; Schwatke, C. Lake level estimation based on CryoSat-2 SAR altimetry and multi-looked waveform classification. *Remote Sens.* **2016**, *8*, 885. [CrossRef]
50. Kleinherenbrink, M.; Naeije, M.; Slobbe, C.; Egido, A.; Smith, W. The performance of CryoSat-2 fully-focussed SAR for inland water-level estimation. *Remote Sens. Environ.* **2020**, *237*, 111589. [CrossRef]
51. Ahmed, A.M.; Deo, R.C.; Ghahramani, A.; Feng, Q.; Raj, N.; Yin, Z.; Yang, L. New double decomposition deep learning methods for river water level forecasting. *Sci. Total Environ.* **2022**, *831*, 154722. [CrossRef]
52. Mohsen, A.; Kovács, F.; Kiss, T. Remote Sensing of Sediment Discharge in Rivers Using Sentinel-2 Images and Machine-Learning Algorithms. *Hydrology* **2022**, *9*, 88. [CrossRef]
53. Terekhov, A. Satellite monitoring of the river bed of the transboundary Ili River in the task of water discharge estimation. In Proceedings of the Sixteenth All-Russian Open Conference "Modern Problems of Remote Sensing of the Earth from Space", Moscow, Russia, 12–16 November 2018; p. 115.
54. Abayev, N.N.; Terekhov, A.G.; Sagatdinova, G.N.; Mukhamediev, R.I.; Amirgaliyev, E.N. Satellite monitoring of the river shoals of the transboundary Ili River (Central Asia) in the task of the water level estimation. *Mod. Probl. Remote Sens. Earth Space* **2023**, *20*, 170–181.
55. Gizatullin, A.; Sharafutdinov, R. Distinctive features of modeling the zones of possible flooding during the passage of floods on the plain and mountainous territory. In *Geoinformation Technologies in Projecting and Constructing the Corporate Information Systems*; Springer: Ufa, Russia, 2010; pp. 154–160.
56. McFeeters, S.K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Int. J. Remote Sens.* **1996**, *17*, 1425–1432. [CrossRef]
57. Xu, H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *Int. J. Remote Sens.* **2006**, *27*, 3025–3033. [CrossRef]
58. Sentinel-2 Bands. Available online: https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/bands/ (accessed on 4 September 2023).
59. Mukhamediev, R.I.; Popova, Y.; Kuchin, Y.; Zaitseva, E.; Kalimoldayev, A.; Symagulov, A.; Levashenko, V.; Abdoldina, F.; Gopejenko, V.; Yakunin, K. Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges. *Mathematics* **2022**, *10*, 2552. [CrossRef]
60. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
61. Mukhamediev, R.I.; Merembayev, T.; Kuchin, Y.; Malakhov, D.; Zaitseva, E.; Levashenko, V.; Popova, Y.; Symagulov, A.; Sagatdinova, G.; Amirgaliyev, Y. Soil Salinity Estimation for South Kazakhstan Based on SAR Sentinel-1 and Landsat-8, 9 OLI Data with Machine Learning Models. *Remote Sens.* **2023**, *15*, 4269. [CrossRef]

62. Yu, H.-F.; Huang, F.-L.; Lin, C.-J. Dual coordinate descent methods for logistic regression and maximum entropy models. *Mach. Learn.* **2011**, *85*, 41–75. [CrossRef]
63. Santosa, F.; Symes, W.W. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comput.* **1986**, *7*, 1307–1330. [CrossRef]
64. Goncharsky, A.; Stepanov, V.; Tikhonov, A.; Yagola, A. *Numerical Methods for the Solution of Ill-Posed Problems*; Springer: Berlin/Heidelberg, Germany, 1995.
65. Hoerl, A.E.; Kennard, R.W. Ridge regression: Applications to nonorthogonal problems. *Technometrics* **1970**, *12*, 69–82. [CrossRef]
66. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
67. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
68. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3149–3157.
69. Al Daoud, E. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *Int. J. Comput. Inf. Eng.* **2019**, *13*, 6–10.
70. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2021**, *54*, 1937–1967. [CrossRef]
71. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
72. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
73. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [CrossRef]
74. Galushkin, A.I. The Back Propagation Error Method and Russian Works on Neural Networks Theory. *Inf. Technol.* **2014**, *7*, 66–76. Available online: http://novtex.ru/IT/it2014/It714_web.pdf#page=66 (accessed on 23 November 2023).
75. Mukhamediev, R.I.; Kuchin, Y.; Amirgaliyev, Y.; Yunicheva, N.; Muhamedijeva, E. Estimation of Filtration Properties of Host Rocks in Sandstone-Type Uranium Deposits Using Machine Learning Methods. *IEEE Access* **2022**, *10*, 18855–18872. [CrossRef]
76. Mukhamediev, R.; Amirgaliyev, Y.; Kuchin, Y.; Aubakirov, M.; Terekhov, A.; Merembayev, T.; Yelis, M.; Zaitseva, E.; Levashenko, V.; Popova, Y. Operational Mapping of Salinization Areas in Agricultural Fields Using Machine Learning Models Based on Low-Altitude Multispectral Images. *Drones* **2023**, *7*, 357. [CrossRef]
77. Kuchin, Y.; Mukhamediev, R.; Yunicheva, N.; Symagulov, A.; Abramov, K.; Mukhamedieva, E.; Zaitseva, E.; Levashenko, V. Application of Machine Learning Methods to Assess Filtration Properties of Host Rocks of Uranium Deposits in Kazakhstan. *Appl. Sci.* **2023**, *13*, 10958. [CrossRef]
78. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290. [CrossRef]
79. Moriasi, D.N.; Arnold, J.G.; Van Liew, M.W.; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* **2007**, *50*, 885–900. [CrossRef]
80. Borshch, S.V.; Simonov, Y.A.; Khristoforov, A.V.; Yumina, N.V. Forecasting the inflow into the Tsimlyansk Reservoir. *Hydrometeorological studies and forecasts* **2022**, *4*, 47–189. [CrossRef]
81. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 17301.
82. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **2020**, *23*, 18. [CrossRef] [PubMed]
83. Raschka, S. MLxtend: Providing Machine Learning and Data Science Utilities and Extensions to Python's Scientific Computing Stack. *Open Source Softw.* **2018**, *3*, 638. [CrossRef]
84. MLxtend Documentation. Available online: https://rasbt.github.io/mlxtend/ (accessed on 3 May 2023).