

УДК 504.064.2.001.18

DOI: 10.46698/VNC.2025.82.17.001

Оригинальная статья

Оценка применимости моделей машинного обучения для прогноза гидроэкологического состояния малых рек

В.П. Долгих 

Донбасский государственный технический университет,
Россия, 294204, г. Алчевск, пр. Ленина, 16,
e-mail: vidoscience@mail.ru

Статья поступила: 02.07.2025, доработана: 09.09.2025, принята к публикации: 15.09.2025

Резюме: Актуальность работы. Водосборная территория р. Белая вблизи г. Луганска подвергается природным и антропогенным факторам, которые связаны с закрытием угольных шахт. Использование современных методов машинного обучения позволит спрогнозировать расход воды в реке и наметить пути по достаточному обеспечению промышленности и населения. **Цель исследования.** Месячное прогнозирование расхода воды на основе нескольких методов машинного обучения (ARIMA, SARIMAX, LSTM, CatBoost). **Методы исследования.** Рассмотрены данные по расходу воды, средней температуре воздуха, средней скорости ветра, суточной сумме осадков и рассчитанному в результате химического анализа сухому остатку из проб воды на уровнемерном посту, который находится ближе всего к Исаковскому водохранилищу. На оригинальных данных построены прогностические модели, рассчитаны соответствующие средние абсолютные ошибки и показана возможность построения эффективной системы для месячного прогнозирования расхода воды. С целью оценки взаимовлияния задаваемых данных предлагается использовать факторный анализ. **Результаты.** В результате исследования выявлена достаточная для практических целей применимость алгоритма градиентного бустинга (CatBoost), использующего задаваемую гидрометеорологическую информацию на водосборе. Реализация моделей выполнена с использованием открытых программных библиотек на языке Python, что подтверждает возможность широкого использования описанных методик для научных исследований и решения прикладных задач. Дополнительно сформированы группы взаимовлияния факторов и дана их оценка. Полученные результаты являются одним из важных составляющих для программно-технического инструментария, необходимого для выявления закономерностей многофакторного процесса формирования стока поверхностных водных объектов и прогнозирования расхода воды для рассматриваемой водосборной территории.

Ключевые слова: модели речного стока, модели машинного обучения, прогнозирование речного стока, река Белая.

Благодарности: Исследования выполнены за счет средств федерального бюджета (код темы: FRRU-2024-0004 в ЕГИСУ НИОКТР).

Для цитирования: Долгих В.П. Оценка применимости моделей машинного обучения для прогноза гидроэкологического состояния малых рек. *Геология и геофизика Юга России*. 2025. 15(3): 148–161. DOI: 10.46698/VNC.2025.82.17.001

DOI: 10.46698/VNC.2025.82.17.001

Original paper

Assessment of the applicability of machine learning models for forecasting the hydroecological state of small rivers

V.P. Dolgikh 

Donbass State Technical University,
16 Lenin Ave., Alchevsk 294204, Russian Federation,
e-mail: vidoscience@mail.ru

Received: 02.07.2025, revised: 09.09.2025, accepted: 15.09.2025

Abstract: Relevance. The catchment area of the Belaya River near Luhansk is exposed to natural and anthropogenic factors associated with the closure of coal mines. The use of modern machine learning methods will make it possible to predict the flow of water in a river and identify ways to provide sufficient supplies to industry and the public. **Aim.** Monthly prediction of water consumption based on several machine learning methods (ARIMA, SARIMAX, LSTM, CatBoost). **Methods.** The data of water flow rate, average air temperature, average wind speed, daily precipitation and dry residue, calculated as a result of chemical analysis from water samples at the level measuring station, which is located closest to the Isakovskoye Reservoir, are considered. Prognostic models are built on the original data, the corresponding average absolute errors are calculated and the possibility of building an effective system for monthly forecasting of water flow rate is shown. In order to assess the mutual influence of the specified data, it is proposed to use factor analysis. **Results.** As a result of the study, the applicability of the gradient boosting algorithm (CatBoost), which uses preset hydrometeorological information in the catchment area, has been revealed. The models are implemented using open software libraries in Python, which confirms the possibility of widespread use of the described techniques for scientific research and solving applied problems. Additionally, groups of mutual influence of factors were formed and their assessment was given. The obtained results are one of the important components for software and hardware tools necessary for identifying patterns of the multifactorial process of formation of runoff of surface water bodies and forecasting water consumption for the considered catchment area.

Keywords: river flow models, machine learning models, river flow forecasting, Belaya River.

Acknowledgements: The research was carried out at the expense of the federal budget (subject code: FRRU-2024-0004 at EGISU NIOKTR).

For citation: Dolgikh V.P. Assessment of the applicability of machine learning models for forecasting the hydroecological state of small rivers. *Geologiya i Geofizika Yuga Rossii = Geology and Geophysics of Russian South*. (in Russ.). 2025. 15(3): 148-161. DOI: 10.46698/VNC.2025.82.17.001

Введение

Методы расчета и прогнозирования речного стока, основанные на физико-математической интерпретации взаимосвязей между компонентами в моделях, получили широкое распространение для решения различных научных и прикладных задач [Мотовилов, 2016а, б; Beven, 2020]. Их главное достоинство – возможность оценки реакции составляющих гидрологического цикла на изменения характеристик климатической системы. Однако наличие сложной параметризации описываемых в моделях гидродинамических процессов не позволяет в полной мере проводить различные оценки влияния отдельных факторов на поверхностный речной сток [Борщ и др., 2023; Погорелов и др., 2023; Borsch et al., 2022; Долгих и др., 2024].

Решение указанной проблемы может быть достигнуто при использовании высокопроизводительных вычислительных систем. Новые расчетные алгоритмы, которые «обучаются» на основании накопленных баз данных, получили название «модели машинного обучения» (ML). Интерес практикующих исследователей объясняется доступностью открытого исходного кода, обилием подгружаемых в программы библиотек, возможностями доработки существующих проектов под требования конкретных задач. ML характеризуются неприязнательностью к задаваемым массивам данных, что является весомым преимуществом, потому что при построении математических моделей формирования стока отсутствует возможность максимально детализировать все процессы и учесть неопределенности. Приведенное обстоятельство дает возможность научиться строить модели и оперативно получать результаты, что способствует интересу исследователей к решению достаточно сложных задач, в частности, гидроэкологической оценке состояния речного стока.

В настоящем исследовании предпринята попытка продемонстрировать возможность построения алгоритма прогнозирования речного стока на период до месяца [Свод правил ..., 2023] с помощью ряда моделей: интегрированной модели авторегрессии-скользящего среднего Auto Regressive Integrated Moving Average (ARIMA), сезонной авторегрессионной интегрированной скользящей средней с экзогенными регрессорами Seasonal ARIMA with eXogenous regressors (SARIMAX), модели с долгосрочной кратковременной памятью Long short-term memory (LSTM) и алгоритма градиентного бустинга (CatBoost). В качестве объекта исследования используется бассейн малой реки Белая, истоки которой находятся в районе с. Тимирязево Перевальского района. Прогнозы речного стока заблаговременностью до месяца строятся для равномерного поста реки Белая, являющегося замыкающим перед Исаковским водохранилищем. В статье приводятся описания архитектур моделей, характеристики рядов наблюдений, расчеты ошибок прогнозирования расхода воды в реке. Дополнительно оценивается взаимовлияние прогностических факторов, что имеет отдельную научную и практическую ценность в связи с изменяющимися гидрологическими процессами в бассейне водосбора р. Белая после закрытия угольных шахт.

Объект исследования

Малая река Белая является правым притоком реки Лугань и относится к бассейну реки Северский Донец. Протекает по территории Перевальского, Лутугинского и Славяносербского районов. Река Белая впадает в реку Лугань между п. Сабовкой Славяносербского района и г. Александровском (Артемовский район г. Луганска). Общая протяженность реки составляет 88 км, площадь водосбора – 755 км² [Смирнова, Вознюк, 2023]. Климат в бассейне реки Белая характеризуется как умеренно континентальный с довольно жарким засушливым летом и сравнительно холодной зимой с неустойчивым снежным покровом. Слой осадков по водосбору реки не одинаков и в зависимости от зон, располагаемых севернее и южнее г. Перевальска, колеблется в среднем от 525 до 354 мм осадков в год.

Особенностью территории является то, что площадь водосбора реки от истока до плотины Исаковского водохранилища подработана крупными закрытыми шахтами: «Фащевская», имени С. В. Косиора, «Селезнёвская», «Перевальская». Также обнаружены выработки 63-х старых мелких шахт: на левой стороне водосбора – 24 шахты, на правой – 23 шахты и на р. Чернухина, левом притоке р. Белая – 16 шахт [Крамаренко и др., 2022; Коптева, 2023]. С учетом климатических изменений, связанных с возобновившимся периодом засухи с 2021 года, питание реки осуществляется в основном подземными водами, которые поступают в поверхностный сток с шахтными. Данное обстоятельство вносит свой вклад в экологическую безопасность территории и должно учитываться для ведения хозяйственной деятельности населения и планирования работы промышленных предприятий.

Суммарный модуль стока реки Белая составляет в устье 2,2 л/с с км² [Правила эксплуатации ..., 2003]. Для данного исследования был выбран уровнемерный пост реки Белая возле поссовета пгт. Бугаевки, который является аккумулярующим звеном перед Исаковским водохранилищем. Наблюдения именно в этой зоне замеров являются наиболее важными, потому что с речной водой, в которую поступают шахтные воды, в водохранилище попадают различные химические компоненты (ионы кальция, гидрокарбонат-ионы, сульфат-ионы и др. компоненты), которые и формируют общее представление о состоянии искусственного водоема.

Материалы и методы исследований

Наборы исходных данных для уровнемерного поста созданы на основе проведения полевых исследований сотрудниками Молодежной научно-исследовательской лаборатории геоэкологии и прикладной химии в 2024 году (от 04.04.2024 г. до 24.12.2024 г.). Исходная информация представлена тремя группами данных: данные о состоянии поверхностного стока, метеоданные (от информационных сервисов погоды) и данные, полученные в результате проведения химического анализа. В качестве задаваемого массива были выбраны параметры, фрагмент которых приведен в таблице 1.

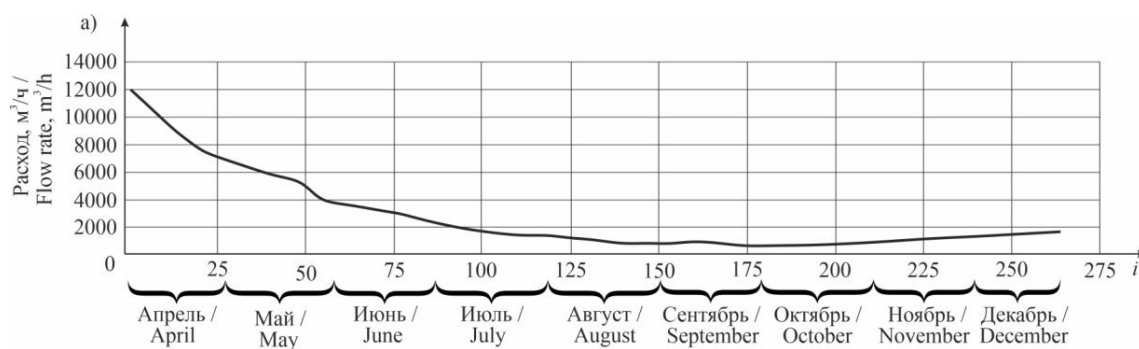
Таблица 1 / Table 1

**Массив исходных данных после предварительной обработки (фрагмент) /
Array of initial data after preliminary processing (fragment)**

Номер замера (i) / Measurement number (i)	Дата замера / Date of measurement	Данные о состоянии поверхностного стока / Surface runoff data	Метеоданные / Weather data			Химический анализ / Chemical analysis
		Расход (Q), м ³ /ч / Flow rate (Q), m ³ /h	Средняя температура воздуха [https://global-weather.ru], °C / Average air temperature [https://global-weather.ru], °C	Средняя скорость ветра [https://www.meteoblue.com], м/с / Average wind speed [https://www.meteoblue.com], m/s	Количество осадков [https://www.meteoblue.com], мм / Amount of precipitation [https://www.meteoblue.com], mm	Сухой остаток, мг/дм ³ / Dry residue, mg/dm ³
1	04.04.24	12171	14,5	5,3	0	648
2	05.04.24	11924	10	3,3	0	662
3	06.04.24	11677	10	2,2	0	677
...	...	...	...	...	...	...
263	22.12.24	1612	3,5	4,4	1,6	2529
264	23.12.24	1623	5,5	6,1	1,2	2532
265	24.12.24	1635	5	2,8	0	2534

Фактические замеры проводились не реже одного раза в неделю, но для более полной картины изменения уровня воды и сопутствующих параметров использовался метод линейной интерполяции [Данилов, Гарькина, 2014].

Распределение параметров на рассматриваемом временном диапазоне приведено на рисунке 1.



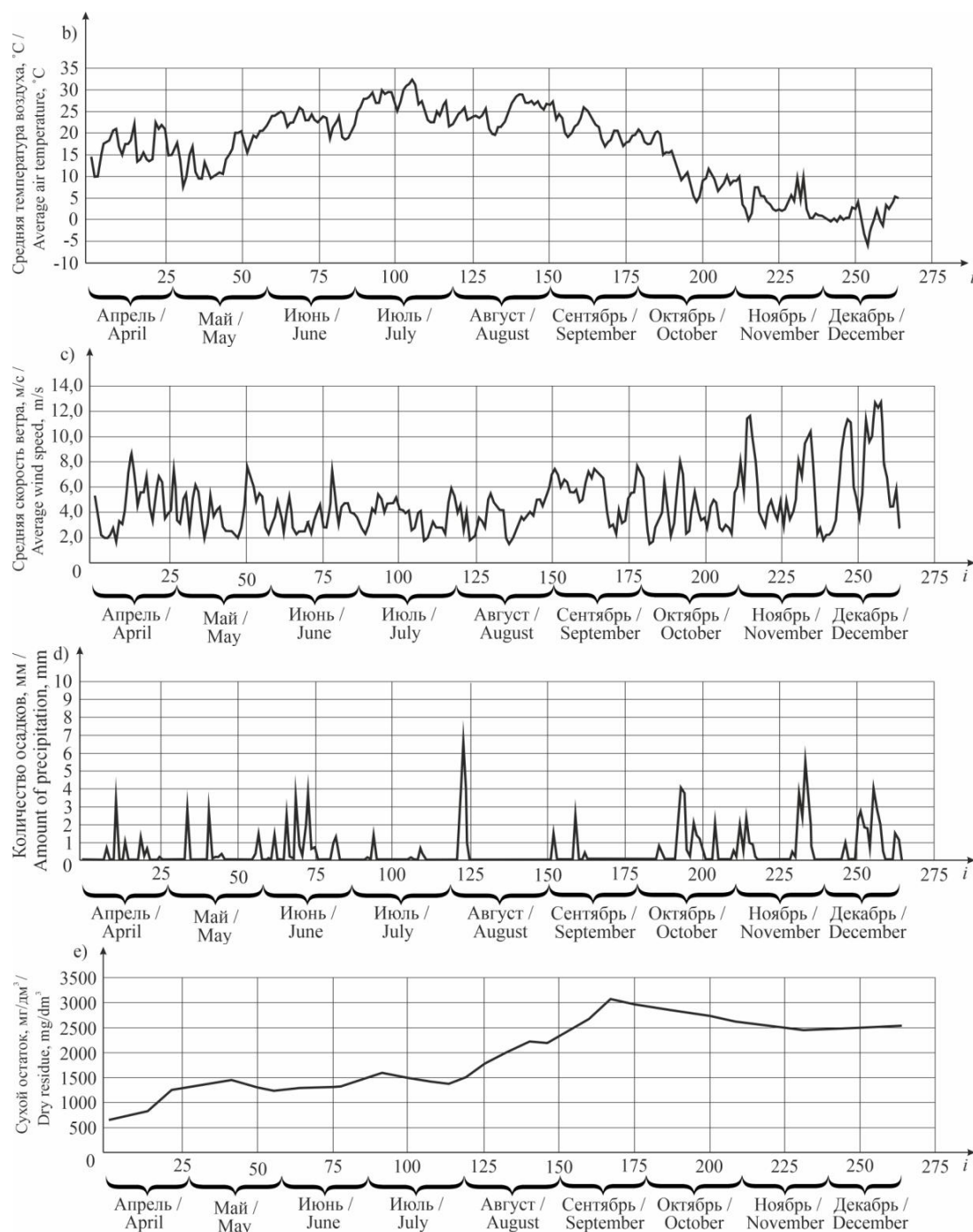


Рис. 1. Распределение исходных данных за исследуемый период /

Fig. 1. Distribution of initial data for the period under study

Для прогнозирования расхода воды в р. Белая были построены несколько моделей машинного обучения. Программная реализация выполнена на языке Python [Daniel, 2023; Deepali, Gresha, 2023; John, 2023; Stewart, 2023] в интерактивной оболочке Scientific Python Development Environment (Spyder), специальной IDE для научных исследований, машинного обучения и анализа данных. В качестве стандартных библиотек машинного обучения использовались scikit-learn и Tensorflow.

Исследования проводились с использованием следующих программных архитектур:

– интегрированной модели авторегрессии-скользящего среднего (ARIMA), в которой заложен алгоритм прогнозирования будущих значений только на основании предыдущих значений временного ряда [Артамонов и др., 2021];

– сезонной авторегрессионной интегрированной скользящей средней с экзогенными регрессорами (SARIMAX), которая расширяет возможности традиционной модели ARIMA путем учета сезонности и ряда внешних факторов [Trancart et al., 2013];

– модели с долгосрочной кратковременной памятью (LSTM), которая является модификацией рекуррентной нейронной сети и используется для решения задач глубокого обучения с дополнительными переменными состояния, где необходимо сохранять веса отдельных нейронов и передавать их между расчетными шагами при обучении алгоритма, если их значение приводит к снижению ошибки модели [Vennerod et al., 2021];

– алгоритма градиентного бустинга, специально разработанного для эффективной обработки категориальных признаков и расчета регрессии CatBoost (категориальное бустинг-обучение) [Prokhorenkova et al., 2018; Hancock, Khoshgoftaar, 2020].

Модели ARIMA и SARIMAX достаточно подробно представлены в научной литературе [Артамонов и др., 2021; Douglas et al., 2015]. С целью ознакомления с принципами построения алгоритмов LSTM и CatBoost далее приводится их краткое описание.

Модель с долгосрочной кратковременной памятью сохраняет информацию о состояниях между расчетными шагами и на основании параметров регуляторов памяти определяет, как долго сохранять эту информацию и когда она больше не понадобится. Общий вид архитектуры сети LSTM представлен на рисунке 2.

Работа модели LSTM и ее компонентов заключается в следующем. Ячейка имеет три внутренних блока, называемых регуляторами входа (input gate), выхода (output gate) и памяти (forget gate). Эти регуляторы управляют внутренним состоянием ячейки c_t и скрытым состоянием h_t . В оранжевых блоках отображено, как входной вектор $x_t[n-1]$ и предыдущее скрытое состояние $h_{t-1}[m; 1]$ умножаются соответственно на четыре различные матрицы $W_j[m; m+n]$ и суммируются соответственно с четырьмя различными смещениями $b_j[m; 1]$.

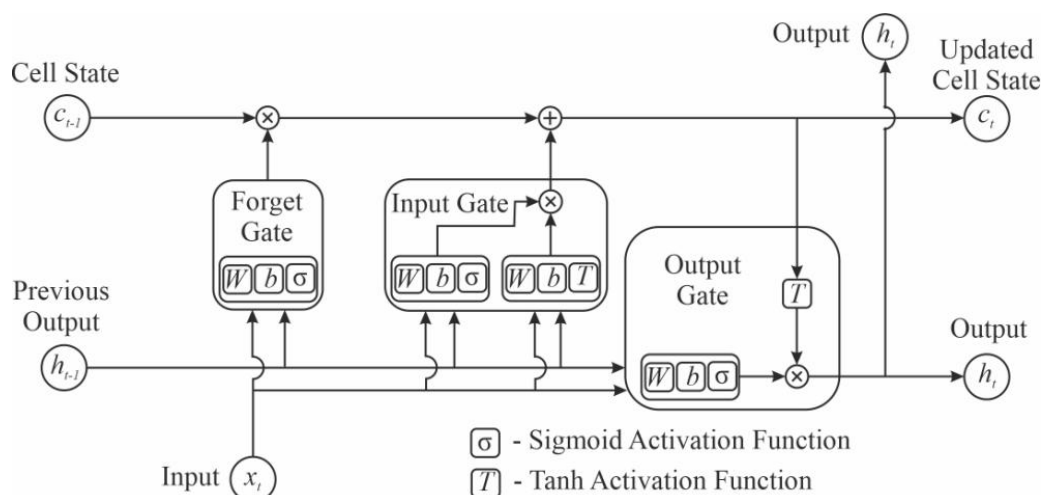


Рис. 2. Архитектура модели LSTM [Vennerod et al., 2021] /

Fig. 2. LSTM model architecture [Vennerod et al., 2021]

Выходы этих регуляторов в дальнейшем используются как для обновления состояния самой модели с целью формирования $s_t[m-1]$, так и для вычисления нового скрытого состояния $h_t[m; 1]$. Обозначения $[m; 1]$ и $[m; m + n]$ – размеры матриц, «крестик» в регуляторах – это скалярное произведение, а «плюсик» – сложение. Четыре комбинации матрицы W , смещения b и функции активации, которая является сигмой или гиперболическим тангенсом, можно рассматривать как однослойную нейронную сеть.

Для программной реализации алгоритма LSTM выборка значений делилась на обучающую (80 % от данных с апреля по июль), валидационную (20 % от данных с апреля по июль) и тестовую (данные за август).

Алгоритм градиентного бустинга CatBoost представляет собой эффективный метод ансамблевого обучения, в котором объединяются слабые модели прогнозирования для создания надежной прогностической модели [Prokhorenkova et al., 2018; Hancock, Khoshgoftaar, 2020]. Алгоритм работает путем итеративного добавления новых моделей в ансамбль, каждая из которых обучается тому, чтобы исправлять ошибки предыдущих моделей. Особенностью алгоритма является то, что обработка признаков происходит без необходимости их предварительной обработки, что делает его эффективным инструментом для работы с реальными наборами данных. Дополнительным преимуществом выступает добавление коэффициента регуляризации в функцию потерь, используемую в процессе обучения, что предотвращает переобучение и улучшает способности модели к обобщению данных.

Алгоритм CatBoost начинает с предположения, часто представляющего собой среднее значение целевой переменной. Далее он итеративно создает ансамбль деревьев, минимизируя функцию потерь с помощью градиентного спуска. На каждой итерации он вычисляет отрицательный градиент функции потерь по отношению к текущим прогнозам и «подгоняет» новое дерево под отрицательный градиент. Скорость обучения определяет размер шага при градиентном спуске. Процесс повторяется до тех пор, пока не будет добавлено заданное количество деревьев или не будет достигнут критерий сходимости. При составлении прогнозов CatBoost объединяет прогнозы всех деревьев в ансамбле. Такое объединение прогнозов приводит к созданию высокоточных и надежных моделей.

Математическое описание категориального бустинг-обучения сводится к следующему. Учитывая набор обучающих данных, состоящий из N выборок и M признаков, где каждая выборка обозначается как (x_i, y_i) , где x_i – это вектор из M признаков, а y_i – соответствующая целевая переменная, CatBoost стремится изучить функцию $F(x)$, которая предсказывает целевую переменную y .

$$F(x) = F_0(x) + \sum_{m=1}^M \sum_{i=1}^N f_m(x_i), \quad (1)$$

где $F(x)$ – общая функция прогнозирования, которую CatBoost стремится изучить. Она принимает входной вектор x и прогнозирует соответствующую целевую переменную y ;

$F_0(x)$ – начальное предположение (базовый прогноз), который часто устанавливается как среднее значение целевой переменной в обучающем наборе данных и отражает общее среднее поведение целевой переменной;

$\sum_{m=1}^M$ – обозначает суммирование по ансамблю деревьев (M – общее количество деревьев в ансамбле);

$\sum_{i=1}^N$ – обозначает суммирование по обучающим выборкам (N – общее количество обучающих выборок);

$f_m(x_i)$ – прогноз m -го дерева для i -й обучающей выборки.

Из уравнения (1) следует, что общий прогноз $F(x)$ получается путем суммирования начального предположения $F_0(x)$ с прогнозами каждого дерева $f_m(x_i)$ для каждой обучающей выборки. Это суммирование выполняется для всех деревьев (m) и всех обучающих выборок (i).

Принцип построения и реализации модели для прогнозирования расхода воды с использованием CatBoost заключался в выделении целевого признака (расход воды), задании смещения (5 шагов до начала расчета с целью улучшения сходимости дальнейших результатов) и выборе диапазона для обучения (апрель 2024 – июль 2024). Следующим шагом было нормирование данных, корреляция параметров и выбор значимости факторов (все факторы оказались значимыми). После обучения модели добавлялись данные за август, проводилось прогнозирование и сравнение с реальными данными.

Результаты исследований и их обсуждение

Задача прогнозирования ставится следующим образом. На основании имеющихся исходных данных требуется осуществить прогноз расхода малой реки на месяц вперед. В качестве примера выбран август 2024 года, т.к. в этот период наблюдалось снижение количества воды в реке, однако в дальнейших исследованиях при наличии дополнительных данных планируется выполнить проверку алгоритма для произвольных временных диапазонов.

Для оценки качества прогноза предлагается использовать ошибку между фактическими и прогнозными значениями рассматриваемой переменной, определяемую функцией потерь MAPE (средняя абсолютная ошибка в процентах). Метрика MAPE получила широкое распространение в ML. Обеспечение минимальной ошибки прогнозирования проводится в соответствии со следующим условием:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{Q_i - \tilde{Q}_i}{Q_i} \right| \cdot 100\% \rightarrow \min, \quad (2)$$

где Q_i – значение расхода в i -й момент времени наблюдения;

i – номер наблюдения ($i=1 \dots n$);

n – количество наблюдений ($n=31$);

$\tilde{Q}_i$ – прогнозное значение расхода для i -го наблюдения.

Результаты расчетов

Результаты расчетов функции потерь MAPE представлены в таблице 2, а графическая интерпретация на рисунке 3.

Таблица 2 / Table 2

Результаты расчета функции потерь MAPE для прогнозирования расхода воды в реке различными алгоритмами /
Results of calculating the MAPE loss function for predicting water flow in a river using different algorithms

Функция потерь / Loss function	Расчетный алгоритм / Calculation algorithm			
	ARIMA	SARIMAX	LSTM	CatBoost
MAPE, %	18,31	12,72	14,76	4,92

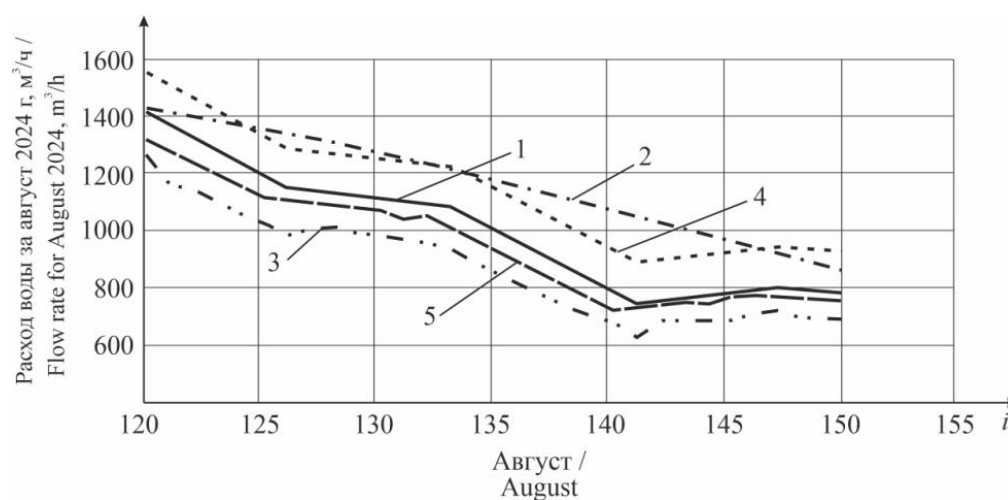


Рис. 3. Результаты прогнозирования (август 2024 г.)

1 – реальные данные; 2 – ARIMA; 3 – SARIMAX; 4 – LSTM; 5 – CatBoost /

Fig. 3. Forecasting results (August 2024)

1 – real data; 2 – ARIMA; 3 – SARIMAX; 4 – LSTM; 5 – CatBoost

Сравнение результатов расчетов расхода воды в реке Белая за август 2024 года, полученных различными методами прогнозирования, с реальными данными дает основание утверждать, что алгоритм градиентного бустинга CatBoost показал наименьшую погрешность (4,92 %). Несмотря на небольшую выборку значений, предлагаемый алгоритм машинного обучения показал свою эффективность. Полученный результат учел сезонность рассматриваемого процесса, что полностью не удалось ARIMA. Модель с долгосрочной кратковременной памятью повторила вид кривой, но оказалась выше реальной в среднем на 12,75 %, однако с удалением от начального значения расхода наблюдалось увеличение разности (15,75 %). SARIMAX была точнее, чем ARIMA, однако наблюдалось запаздывание результатов, что отразилось на MAPE (12,72 %).

С целью выявления степени влияния исследуемых данных, а также формирования групп параметров, тесно связанных между собой, проводился факторный анализ. Алгоритм расчета выполнен на языке Python и учитывает данные таблицы 1, в которой отсутствуют пропуски. На первом шаге расчета ставилась задача определения количества групп факторов, результаты которых схожи между собой и имеют взаимосвязь. Результаты изображены на рисунке 4.

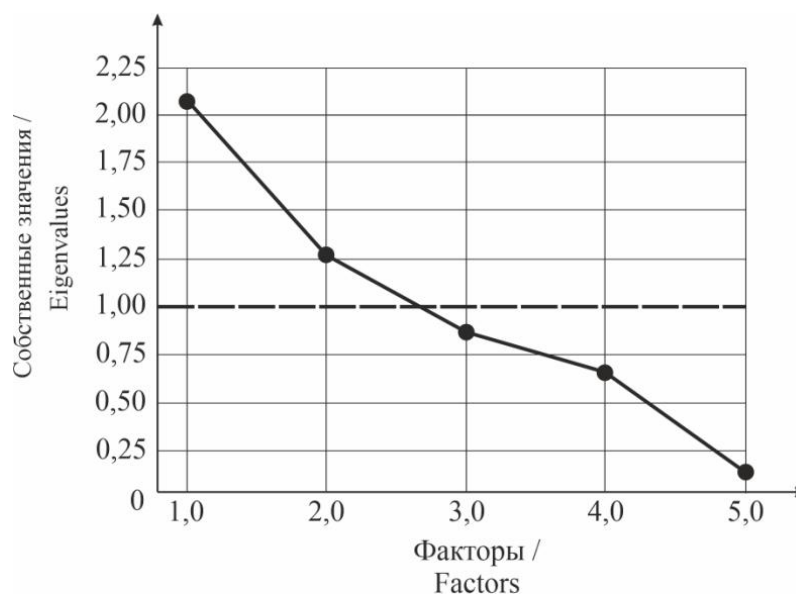


Рис. 4. Графическое представление количества факторов, рекомендуемых программой /
Fig.4. Graphical representation of the number of factors recommended by the program

Несмотря на установленную границу на уровне 1 (собственные значения) в дальнейшем принимались 4 группы факторов, для которых была рассчитана корреляционная матрица между пятью переменными (таблица 1) и полученными четырьмя факторами. При использовании специальной библиотеки для создания статистических графиков Seaborn результаты исследований могут быть представлены в виде тепловой карты. В данном случае полученная информация, которая отражает корреляции переменных с каждым из выделенных факторов и позволяет выделить закономерности взаимовлияния отдельных составляющих друг на друга, представлена в таблице 3.

Таблица 3 / Table 3

**Представление группировки переменных в факторы согласно
рассчитанной корреляции между ними /
Representation of the grouping of variables into factors according
to the calculated correlation between them**

Переменные / Variables	Факторы / Factors			
	Фактор 1 / Factor 1	Фактор 2 / Factor 2	Фактор 3 / Factor 3	Фактор 4 / Factor 4
Расход / Flow rate	92,9 %	9,5 %	10,0 %	0,9 %
Средняя температура воздуха / Average air temperature	6,4 %	71,4 %	24,8 %	0,4 %
Сухой остаток / Dry residue	87,5 %	46,7 %	2,9 %	5,8 %
Средняя скорость ветра / Average wind speed	10,8 %	32,1 %	45,8 %	13,8 %
Количество осадков / Amount of precipitation	1,3 %	7,8 %	45,5 %	3,1 %

При анализе таблицы 3 сделаны следующие выводы. Для фактора 1 характерно, что изменение расхода воды (92,9 %) тесно связано с появлением в воде сухого остатка (87,5 %). Вполне очевидно, что уменьшение уровня воды в реке приводит к повышению концентрации растворенных веществ и повсеместному накоплению вдоль водотока, что в дальнейшем приводит к попаданию их в технологический цикл работы предприятий и к использованию для хозяйственных нужд воды с повышенным содержанием минеральных солей. Данное обстоятельство подтверждается регулярными полевыми наблюдениями и отборами проб по водосборной территории р. Белая.

Фактор 2 показывает высокую взаимосвязь между средней температурой воздуха (71,4 %), средней скоростью ветра (32,1 %) и сухим остатком (46,7 %). Такой результат также не вызывает сомнений, потому что повышенная температура и сильный ветер приводят к интенсификации испарения жидкости и, тем самым, ускоряют процессы минерализации солей (кальция, магния, натрия, калия), слипания органических частиц, металлов и нерастворимых элементов.

Согласно фактору 3 необходимо учитывать взаимосвязь средней скорости ветра (45,8 %), средней температуры воздуха (24,8 %) и количества осадков (45,5 %). Полученные взаимосвязи также вполне объяснимы, потому что насыщенные влагой воздушные массы могут переноситься ветром на значительные расстояния.

В сравнении с тремя предыдущими факторами четвертый фактор не вносит никакой дополнительной информации. Такая проверка полезна с целью использования всех возможностей программного анализа данных, однако требует качественного понимания полученных результатов.

Выводы

Выполненные расчеты позволили оценить применимость моделей (ARIMA, SARIMAX, LSTM, CatBoost) в качестве основы для системы месячных прогнозов расходов воды в речной системе. Расчеты подтвердили общее представление о том, что наличие гидрометеорологической информации (данных о предыдущих расходах воды в р. Белая, средней температуре воздуха на водосборе, средней скорости ветра и количестве осадков) и рассчитанного в результате химанализа сухого остатка может быть использовано для прогнозов. Данное обстоятельство позволяет создавать более устойчивые алгоритмы, которые учитывают сезонные колебания, а также взаимовлияния между прогнозирующими факторами.

Программная реализация всех моделей показала свою эффективность, однако наилучшие результаты получены с использованием алгоритма градиентного бустинга CatBoost (средняя абсолютная ошибка составила 4,92 %, что в 2–3 раза меньше, чем у других моделей). Факторный анализ позволил выделить группы (факторы), в которых расчетным путем получены степени влияния исследуемых данных друг на друга. Полученные результаты станут основой для создания программно-технического инструментария для выявления закономерностей многофакторного процесса формирования стока поверхностных водных объектов, который в том числе позволит прогнозировать расходы воды для рассматриваемой водосборной территории.

Литература

1. Артамонов Н.В., Ивин Е.А., Курбацкий А.Н., Фантаццини Д. Введение в анализ временных рядов. Вологда: ВолНИЦ РАН, 2021. – 134 с.
2. Борщ С.В., Симонов Ю.А., Христофоров А.В. Прогнозирование стока рек России. М.: Гидрометцентр России, 2023. – 200 с.

3. Данилов А.М., Гарькина И.А. Интерполяция, аппроксимация, оптимизация: анализ и синтез сложных систем. Пенза: ПГУАС, 2014. – 168 с.
4. Долгих В.П., Рыбалка С.В., Боблева И.С. Краткосрочное прогнозирование водного режима рек и водохранилищ методом Брауна. // Экологический вестник Донбасса. – 2024. – № 12. – С. 9–15.
5. Коптева А.К. О влиянии подземной добычи угля на водные ресурсы реки Белая и наполняемость Исаковского водохранилища в Луганской народной Республике. // Экологический вестник Донбасса. – 2023. – № 9. – С. 47–55.
6. Крамаренко А.А., Коптева А.К., Лысенко И.Л. О формировании стока малых рек в условиях углеработок и о выборе рек-аналогов. // Экологический вестник Донбасса. – 2022. – № 7. – С. 37–50.
7. Мотовилов Ю.Г. Гидрологическое моделирование речных бассейнов в различных пространственных масштабах. 1. Алгоритмы генерализации и осреднения. // Водные Ресурсы. – 2016а. – Т. 43. № 3. – С. 243–253. DOI: 10.7868/S0321059616030111.
8. Мотовилов Ю.Г. Гидрологическое моделирование речных бассейнов в различных пространственных масштабах. 2. Результаты испытаний. // Водные Ресурсы. – 2016б. – Т. 43. № 5. – С. 467–475. DOI: 10.7868/S0321059616050096.
9. Погорелов А.В., Липилин Д.А., Яроцкая Е.В. Геоинформационный анализ современных изменений гидрографических характеристик речной сети на Азово-Кубанской равнине (бассейн реки Ея). // Геология и геофизика Юга России. – 2023. – Т. 13. № 2. – С. –161–179. DOI: 10.46698/VNC.2023.35.64.001.
10. Правила эксплуатации Исаковского гидроузла. Луганск: ОКП «Луганская областная топливно-энергетическая компания», 2003. – 17 с.
11. Свод правил по проектированию и строительству. СП 529.1325800.2023. Определение основных расчетных гидрологических характеристик. М.: Госстрой России, 2023. – 103 с.
12. Смирнова И.В., Вознюк Ю.С. Расчёт подземного стока в реку Белая гидрохимическими методами. // Экологический вестник Донбасса. – 2023. – № 9. – С. 41–46.
13. Beven K. Deep learning, hydrological processes and the uniqueness of place. // Hydrological Processes. – 2020. – Vol. 34. Issue 16. – pp. 3608–3613. DOI: 10.1002/hyp.13805.
14. Borsch S., Simonov Y., Khristoforov A., Semenova N., Koliy V., Ryseva E., Krovotyntsev V., Derugina V. Russian Rivers Streamflow Forecasting Using Hydrograph Extrapolation Method. // Hydrology. – 2022. – Vol. 9. DOI: 10.3390/hydrology9010001.
15. Daniel C. Pandas for everyone: Python data analysis, 2nd edition. Addison-Wesley Data & Analytics Series. – 2023. – 512 p.
16. Douglas C.M., Cheryl L.J., Kulahci M. Introduction to time series analysis and forecasting. – John Wiley & Sons, Inc., Hoboken, New Jersey, 2015 – 671 p.
17. Deepali R.V., Gresha S. Bhatia python machine learning projects. – BPB Publications. – 2023. – 260 p.
18. Hancock J., Khoshgoftaar T. CatBoost for big data: an interdisciplinary review. // Journal of Big Data. – 2020. – Vol. 7(1). DOI: 10.1186/s40537-020-00369-8.
19. Prokhorenkova L., Gusev G., Vorobev A. CatBoost: unbiased boosting with categorical features. // Advances in Neural Information Processing Systems. – 2018. – Vol. 31.
20. Stewart J. Python for Scientists, 3rd Edition. – Cambridge University Press. – 2023. – 228 p.
21. Trancart T., Acou A., De Oliveira E. Forecasting animal migration using SARIMAX: An efficient means of reducing silver eel mortality caused by turbines. // Endangered Species Research. – 2013. – Vol. 21(2). – pp. 181–190. DOI: 10.3354/esr00517.
22. Vennerod Ch.B., Kjaran A., Bugge E.S. Long short-term memory RNN. – 2021. – 16 p.

References

1. Artamonov N.V., Ivin E.A., Kurbatsky A.N., Fantazzini D. Introduction to time series analysis. Vologda: VolSC RAS. 2021. 134 p. (In Russ.)

2. Borsch S.V., Simonov Yu.A., Khristoforov A.V. Forecasting the flow of Russian rivers. Moscow. Hydrometcenter of Russia. 2023. 200 p. (In Russ.)
3. Danilov A.M., Garkina I.A. Interpolation, approximation, optimization: analysis and synthesis of complex systems. Penza. PGUAS. 2014. 168 p. (In Russ.)
4. Dolgikh V.P., Rybalka S.V., Bobleva I.S. Short-term forecasting of the water regime of rivers and reservoirs by the Brown method. Environmental Bulletin of Donbass. 2024. No. 12. pp. 9-15. EDN AWPNZW. (In Russ.)
5. Kopteva A.K. On the impact of underground coal mining on the water resources of the Belaya River and the filling capacity of the Isakov reservoir in the Lugansk People's Republic. Environmental Bulletin of Donbass. 2023. No. 9. pp. 47–55. (In Russ.)
6. Kramarenko A.A., Kopteva A.K., Lysenko I.L. On the formation of the flow of small rivers in the conditions of coal mining and on the choice of analog rivers. Environmental Bulletin of Donbass. 2022. No. 7. pp. 37–50. (In Russ.)
7. Motovilov Yu.G. Hydrological modeling of river basins on various spatial scales. 1. Generalization and averaging algorithms. Water Resources. 2016a. Vol. 43. No. 3. pp. 243–253. DOI: 10.7868/S0321059616030111. (In Russ.)
8. Motovilov Yu.G. Hydrological modeling of river basins on various spatial scales. 2. Test results. Water Resources. 2016b. Vol. 43. No. 5. pp. 467–475. DOI: 10.7868/S0321059616050096. (In Russ.)
9. Pogorelov A.V., Lipilin D.A., Yarotskaya E.V. GIS analysis of modern changes in the hydrographic characteristics of the river network in the Azov-Kuban plain (Eya river basin). Geology and Geophysics of Russian South. 2023. Vol. 13. No. 2. pp. 161–179. DOI: 10.46698/VNC.2023.35.64.001. (In Russ.)
10. Operating rules of the Isakov hydroelectric system. Lugansk. OKPD Lugansk Regional Fuel and Energy Company. 2003. 17 p. (In Russ.)
11. A rule book for design and construction. SP 529.1325800.2023. Determination of the main calculated hydrological characteristics. Moscow. State Construction Committee of Russia. 2023. 103 p. (In Russ.)
12. Smirnova I.V., Voznyuk Yu.S. Calculation of underground flow into the Belaya River by hydrochemical methods. Ecological Bulletin of Donbass. 2023. No. 9. pp. 41–46. (In Russ.)
13. Beven K. Deep learning, hydrological processes and the uniqueness of place. Hydrological Processes. 2020. Vol. 34. Issue 16. pp. 3608–3613. DOI: 10.1002/hyp.13805.
14. Borsch S., Simonov Y., Khristoforov A., Semenova N., Koliy V., Ryseva E., Krovotyntsev V., Derugina V. Russian Rivers Streamflow Forecasting Using Hydrograph Extrapolation Method. Hydrology. 2022. Vol. 9. DOI: 10.3390/hydrology9010001.
15. Daniel C. Pandas for everyone: Python data analysis, 2nd edition. Addison-Wesley Data & Analytics Series. 2023. 512 p.
16. Douglas C.M., Cheryl L.J., Kulahci M. Introduction to time series analysis and forecasting. John Wiley & Sons, Inc., Hoboken, New Jersey. 2015 – 671 p.
17. Deepali R.V., Gresha S. Bhatia python machine learning projects. BPB Publications. 2023. – 260 p.
18. Hancock J., Khoshgoftaar T. CatBoost for big data: an interdisciplinary review. Journal of Big Data. 2020. Vol. 7(1). DOI: 10.1186/s40537-020-00369-8.
19. Prokhorenkova L., Gusev G., Vorobev A. CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems. 2018. Vol. 31.
20. Stewart J. Python for Scientists, 3rd Edition. Cambridge University Press. 2023. 228 p.
21. Trancart T., Acou A., De Oliveira E. Forecasting animal migration using SARIMAX: An efficient means of reducing silver eel mortality caused by turbines. Endangered Species Research. 2013. Vol. 21(2). pp. 181–190. DOI: 10.3354/esr00517.
22. Vennerod Ch.B., Kjarran A., Bugge E.S. Long short-term memory RNN. 2021. 16 p.