Торайғыров университетінің хабаршысы
ҒЫЛЫМИ ЖУРНАЛЫ

НАУЧНЫЙ ЖУРНАЛ
Вестник Торайгыров университета

Торайғыров университетінің ХАБАРШЫСЫ

Энергетикалық сериясы

1997 жылдан бастап шығады

ВЕСТНИК Торайгыров университета

Энергетическая серия

Издается с 1997 года

ISSN 2710-3420

№3 (2026)

ПАВЛОДАР

НАУЧНЫЙ ЖУРНАЛ
Вестник Торайгыров университета

Энергетическая серия
выходит 4 раза в год

СВИДЕТЕЛЬСТВО
о постановке на переучет периодического печатного издания,
информационного агентства и сетевого издания
№ 14310-Ж
выдано
Министерство информации и общественного развития
Республики Казахстан

Тематическая направленность
публикация материалов в области электроэнергетики,
электротехнологии, автоматизации, автоматизированных и
информационных систем, электромеханики и теплоэнергетики

Подписной индекс – 76136

---

https://doi.org/10.48081/BTOUEn3.2026/2710-3420

Бас редакторы – главный редактор
Талипов О. М.
доктор PhD, ассоц. профессор (доцент)

Заместитель главного редактора — Калтаев А.Г., *доктор PhD*
Ответственный секретарь — Сағындық Ә.Б., *доктор PhD*

Редакция алқасы – Редакционная коллегия

Клецель М. Я.,	*д.т.н., профессор*
Никифоров А. С.,	*д.т.н., профессор*
Новожилов А. Н.,	*д.т.н., профессор*
Алиферов А. И.,	*д.т.н., профессор (Российская Федерация)*
Кошеков К. Т.,	*д.т.н., профессор*
Приходько Е. В.,	*к.т.н., профессор*
Кислов А. П.,	*к.т.н., доцент*
Нефтисов А. В.,	*доктор PhD*
Шерьязов С. К.	т.ғ.д., *профессор (Российская Федерация)*
Искакова З. С.	*технический редактор*

За достоверность материалов и рекламы ответственность несут авторы и рекламодатели
Редакция оставляет за собой право на отклонение материалов
При использовании материалов журнала ссылка на «Вестник Торайгыров университета» обязательна

© Торайгыров университет

SRSTI 14.37.27

https://doi.org/10.48081/BTOUEn3.2026/91-113

A. M. Aubakirova[1], A. V. Neftissov[2], M. N. Orazbay[3], A. A. Biloshchytskyi[4], T. A. Honcharenko[5]

[1,3,4]Astana IT University,
Republic of Kazakhstan, Astana;
[2]Academy of Physical Education and Mass Sports,
Republic of Kazakhstan, Astana;
[4,5]Kyiv National University
of Construction and Architecture,
Ukraine, Kyiv.
[1]ORCID: https://orcid.org/0009-0004-6925-6714
[2]ORCID: https://orcid.org/0000-0003-4079-2025
[3]ORCID: https://orcid.org/0009-0001-4994-6484
[4]ORCID: https://orcid.org/0000-0001-9548-1959
[5]ORCID: https://orcid.org/0000-0003-2577-6916
*e-mail: Aliya.Aubakirova@astanait.edu.kz

DEVELOPMENT OF A DELTA-BASED XGBOOST MODEL FOR FORECASTING RESERVOIR STORAGE DYNAMICS

Accurately forecasting reservoir water volumes is crucial for climate change adaptation, drought mitigation, and provides a basis for energy planning and management, particularly in continental basins, where hydrological variability poses challenges to traditional modeling. In turn, reduced reservoir inflows limit the water available to hydroelectric power plants for optimal hydroelectric generation. To address this issue, we propose a novel delta-based learning framework using extreme gradient boosting (XGBoost) that forecasts daily volume increments (ΔV) rather than total water volume. The proposed model represents a novel solution where standard machine learning approaches often struggle to capture short-term dynamics due to non-stationary trends in absolute water volume data. The model integrates meteorological and hydrological characteristics of the ERA5 reanalysis lag, is trained on 2019–2022 data, tested on 2023–2025 data, and is adapted to regional specificities. The results demonstrate exceptional performance: the reconstructed volume achieved an R^2 of 0.983, a Nash–Sutcliffe efficiency (NSE) of 0.983, and a Kling–Gupta

efficiency (KGE) of 0.984. Most importantly, the delta-model-based approach reduced the root mean square error (RMSE) by 16 % compared to the baseline direct volume forecasting method, effectively eliminating systematic trend leakage. Performance analysis confirmed the physical dominance of delays in water volume and precipitation integrals. This study confirms that differential dynamics modeling significantly improves the stability and interpretability of forecasts, offering a reproducible and reliable tool.

Keywords: reservoir forecasting, delta-based learning, hydrological machine learning, XGBoost, ERA5 reanalysis, multi-horizon prediction, remote sensing, predictive modeling.

Introduction

Water resources currently play a crucial role in the development of the energy sector in the Republic of Kazakhstan, as the main sources are generated along rivers and reservoirs. Energy facilities such as combined heat and power plant (CHP), state district power plant (SDPP), and nuclear power plant (NPP) cannot exist without water, as it is used primarily in technological processes and cooling systems. Consequently, water resources play a crucial role in the overall energy system and have a significant impact on the region's energy security. Energy facilities require a nearby water source, as this allows for early assessment of water availability, informs subsequent energy planning, and requires monitoring and forecasting systems for uninterrupted operation. This relevance is reflected in the Concept for the Development of the Water Resources Management System of the Republic of Kazakhstan for 2024–2030, where effective water resources management is considered a key factor in ensuring the country's energy security and provides for the improvement of control systems. Furthermore, water resources are an essential component in maintaining the planet's ecological system, where drought is a priority environmental issue in Kazakhstan, particularly in arid and agriculturally intensive regions. Water shortages and depletion are common problems in the Zhambyl region of Kazakhstan.

Increased climate variability and limited water resources require the development of monitoring systems capable of providing rapid assessment of water shortages [1]. In these conditions, the integration of remote sensing, predictive models, and AI algorithms for sustainable water management is essential [2]. Traditional methods based on historical data and assumptions are unprepared for evolving climate changes, such as variable precipitation and the frequency of extreme weather events [3]. AI facilitates monitoring and analysis, which supports decision support systems and has the potential to revolutionize water resource management [4].

Machine learning algorithms allow computers to learn from data and are used to create predictive models, thereby improving decision-making processes [5]. Deep learning is a good tool that works based on huge data sets and is suitable for areas with a large historical data set, but difficulties can arise with a small amount of data [6]. AI is used for real-time monitoring and alerts to any changes, thereby assisting in informed decision-making [7]. When monitoring various water bodies, various space images are used and various methods are used to extract information from these images [8].

Research using GIS, remote sensing technologies, and Sentinel-3 satellite images has achieved accurate identification, extraction, and analysis of water data. However, this method has the disadvantage that cloud obscuration greatly affects the accuracy of the data, which is then considered invalid [9]. The image analysis method maximizes the enhancement of water bodies and attenuates other regions by obtaining Sentinel-3 satellite images with a 15-fold enhancement of spectral bands B21 and B03, but requires high computing resources and processing time. In contrast, the Normalized Difference Water Index (NDWI) method uses cloud mask processing, which improves accuracy when there are few clouds [10].

Considering all existing methods, machine learning approaches such as extreme gradient boosting have demonstrated good performance in reservoir water level forecasting over the past decade compared to traditional approaches [11; 12]. Machine learning methods are used to create models for specific systems based on existing modeling data [13]. Machine learning methods demonstrate good results in system modeling and provide more accurate results [14; 15].

This study focuses specifically on the Tasotkel Reservoir in the Zhambyl region of Kazakhstan. The reservoir's importance to regional socio-economic activity justifies its choice as a research object. Although some regional water resources studies have been conducted, comprehensive applications of machine learning-based reservoir water level forecasting remain rare. The study aims to identify a model capable of achieving high accuracy with a limited data set.

The scientific novelty of the study lies in the proposed adapted approach based on Delta-based XGBoost. Instead of directly forecasting absolute water volume, the model forecasts daily changes in water volume. This makes the model more robust, better able to cope with non-stationarity in hydrological series, and improves forecast quality in the Tasotkel reservoir.

A distinctive feature of the study is that the developed approach is oriented toward application in operational management systems under conditions of climate uncertainty and a shortage of hydrological observations.

The objective of this study is to evaluate and compare the predictive performance of linear regression and machine learning algorithms for water forecasting in the Tasotkel reservoir in the Zhambyl region of Kazakhstan.

The implementation of this goal includes the identification of significant forecast indicators characterizing water resources management processes, the creation of a data pre-processing sub-module using the NDVI vegetation index and surface temperature (LST) parameters, as well as the development of a main forecasting module focused on assessing the volume and dynamics of water resources in the considered.

Materials and Methods

The study was conducted at the Tasotkel Reservoir, located in the Shu district of the Zhambyl region, Republic of Kazakhstan. The reservoir regulates the flow of the transboundary Shu River and represents a critical component of the regional irrigation and water management system. Its hydrological regime directly influences irrigation planning, operational reliability, and the mitigation of critical hydrological events.

The analysis covers the period from 1 January 2017 to 2025 and integrates bathymetric data, satellite observations, and climate reanalysis products. Physical reservoir characteristics were obtained from the 2025 bathymetric survey report provided by UNI Project Group LLP. The digitized Area–Volume–Elevation (AVE) relationship curve derived from this survey formed the basis for reconstructing storage volume (million m^3) and water level (m) from satellite-derived surface area observations.

Surface water dynamics were assessed using data from the MultiSpectral Instrument (MSI) onboard Sentinel-2 satellites within the Copernicus program. Image processing and time-series extraction were performed in the Google Earth Engine (GEE) cloud platform. The Normalized Difference Water Index (NDWI) was applied to delineate water surfaces. A daily time series of reservoir surface area was generated and subsequently transformed into volume and water level estimates using the AVE curve.

Meteorological predictors were derived from the ERA5-Land reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) and accessed through GEE. ERA5-Land provides spatially continuous coverage at approximately 9 km resolution, offering improved representation of basin-scale processes compared to point-based meteorological observations.

The following daily variables were selected to represent key hydrological drivers: total precipitation, 2 m air temperature, potential evaporation, and snow water equivalent. All meteorological variables were spatially aggregated

by averaging across the entire upstream catchment of the Shu River above the Tasotkel Reservoir.

After harmonization of temporal resolution and spatial aggregation, all datasets were merged into a single daily time series. Each record contains the date, reconstructed reservoir volume and water level, and the corresponding catchment-averaged meteorological predictors. This integrated dataset served as the foundation for subsequent model development.

1 Study Area and Hydrological Context

The study focuses on the Tasotkel Reservoir (Tasotkel'skoye Vodokhranilishche), a strategic hydraulic infrastructure located in the Jambyl Region, Kazakhstan shown in Figure 1. The reservoir is situated on the Chu (Shu) River, centered at coordinates 43,3638° N, 73,9323° E.

Commissioned to regulate the flow of the Chu River, the Tasotkel Reservoir serves as a primary irrigation node for the agricultural zones of the Chu-Talas basin, supplying water to approximately 35,000 hectares of arable land. The reservoir has a design full storage capacity of approximately 620×10^6 m^3 and a surface area of 77 km^2 at the normal retention level.

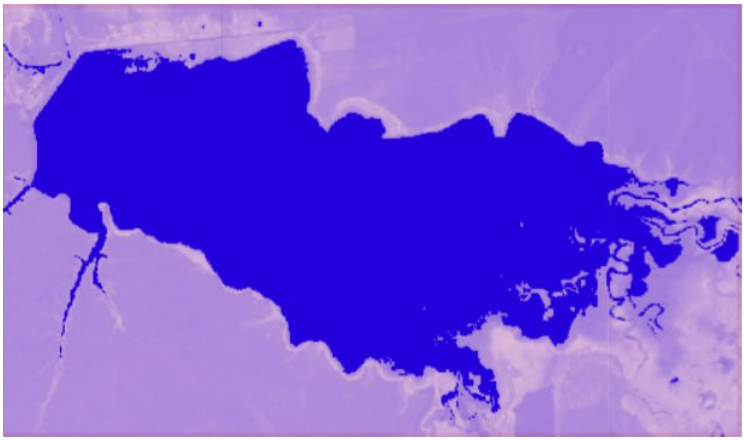


Figure 1 – Tasotkel region

The region is characterized by a sharp continental semi-arid climate, exhibiting significant seasonal thermal amplitude. Winters are cold with stable snow cover, while summers are hot and dry, with high evaporation rates. The hydrological regime of the Chu River is predominantly nival-glacial fed, originating from the Northern Tian Shan mountain ranges (Kyrgyz Ala-Too and Trans-Ili Alatau).

Consequently, approximately 65-70 % of the annual runoff occurs during the spring-summer flood season (typically April to July), driven by snowmelt and glacial ablation. This distinct seasonality introduces high variance in inflow rates, making the system prone to non-stationary dynamics where rapid filling events during warm springs contrast with water scarcity risks during the vegetative period.

Understanding these climatic and morphological characteristics is essential for the modeling strategy, as the "memory" of the system–represented by lag features–and the delay between precipitation/snowmelt and inflow are physically determined by the catchment's topography and the distance from the glacial headwaters to the reservoir inlet.

The dataset constructed for this study spans a period of six years, from January 1, 2019, to September 30, 2025, comprising a total of 2,465 daily observations. The data pipeline integrates local hydrological telemetry with global meteorological reanalysis products.

Daily records of reservoir water levels (H, meters above sea level), water surface area (S, km^2), and total storage volume (V, $10^6 m^3$). The raw data underwent a rigorous quality control (QC) procedure. Missing values, which constituted less than 0,5 % of the dataset, were imputed using linear interpolation, a valid approach given the high autocorrelation of reservoir storage dynamics. Outliers caused by sensor transmission errors were detected using the Interquartile Range (IQR) method and replaced with the rolling median of the adjacent 3-day window.

2 The Delta-Based Feature Engineering Framework

Due to the low density of in-situ meteorological stations in the upper Chu basin, we utilized the ERA5-Land reanalysis dataset produced by the ECMWF. ERA5-Land provides a consistent view of land variables at a horizontal resolution of 9 km.Using the Google Earth Engine (GEE) API, we aggregated daily mean values for the following variables over the catchment area upstream of the Tasotkel dam: Total Precipitation (P, mm): The sum of large-scale and convective precipitation. 2m Air Temperature : Essential for determining the phase of precipitation (rain vs. snow) and melt rates. Snow Depth Water Equivalent (SDWE, mm): A critical proxy for the accumulated snowpack available for melting in the Tian Shan mountains. Potential Evaporation (E_{pot}, mm): Used to estimate water losses from the reservoir surface, which is critical in the semi-arid Jambyl region.

A core methodological contribution of this study is the Delta-Based Approach. Traditional ML models in hydrology often map input features (X_t) directly to absolute volume (V_t). However, reservoir volume time series are inherently non-stationary.

To address this, we defined our target variable as the daily change in volume (ΔV_t):

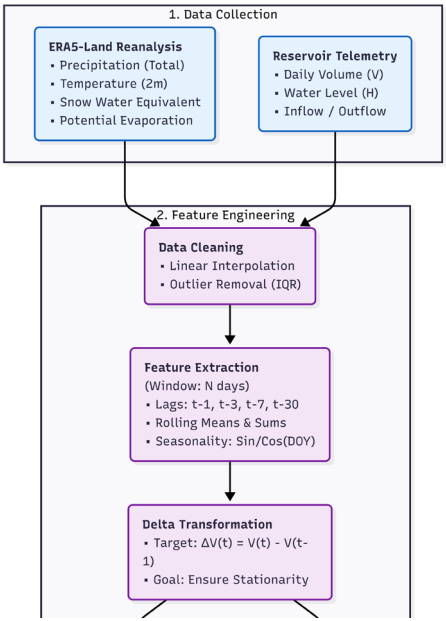
$$\Delta V_t = V_t - V_{t-1} \quad (1)$$

where V_t denotes the reconstructed reservoir storage volume (10^6 m^3) at time step t, and ΔV_t represents the net daily storage increment. Physically, ΔV_t represents the net balance of the reservoir system (Inflow – Outflow – Losses). This variable is quasi-stationary and fluctuates around zero. We engineered a feature space of 32 predictors: Autoregressive Terms (Volume Lags): V_{t-1}, V_{t-2}, V_{t-3}, V_{t-7}, V_{t-30}. Meteorological Lags: Precipitation and Temperature lags (t-1, t-3, t-7). Accumulated Indices: Rolling sums of precipitation ($\sum P_{7d}$, $\sum P_{30d}$) and rolling means of temperature. Cyclical Encoding: To explicitly model seasonality:

$$\phi_{\sin}(t) = \sin\left(\frac{2\pi\, d(t)}{365}\right), \qquad \phi_{\cos}(t) = \cos\left(\frac{2\pi\, d(t)}{365}\right) \quad (2)$$

where $d(t) \in \{1,2,\ldots,365\}$ is the day-of-year index at time t, and φ_{sin}, φ_{cos} form an orthogonal pair encoding annual seasonality without discontinuity at the year boundary. To address this, we defined our target variable as the daily change in volume (ΔV_t).

The complete data processing and modeling pipeline is illustrated in Figure 2.



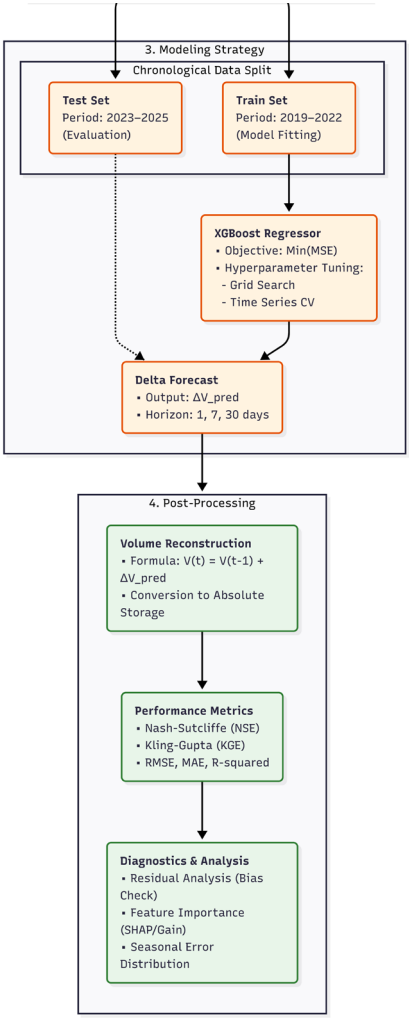


Figure 2 – Schematic representation of the proposed forecasting framework

The workflow integrates ERA5-Land reanalysis and telemetry data, applies the delta-transformation to remove non-stationarity, utilizes an XGBoost regressor with chronological validation, and reconstructs absolute volumes for final evaluation.

3 Machine Learning Algorithm: Extreme Gradient Boosting

The forecasting framework was implemented using the Extreme Gradient Boosting (XGBoost) algorithm. The model represents predictions as an additive ensemble of Kregression trees:

$$\widehat{\Delta V}_i = \sum_{k=1}^{K} f_k(\mathrm{x}_i), \qquad f_k \in \mathcal{F} \tag{3}$$

where $\widehat{\Delta V}_i$ is the predicted daily volume increment for observation ; is the feature vector of 32 engineered predictors; is the total number of regression trees; is the -th tree; and denotes the space of regression tree functions.

$$\mathcal{L}^{(m)} = \sum_{i=1}^{n} \ell\left(\Delta V_i, \widehat{\Delta V}_i^{(m-1)} + f_m(\mathrm{x}_i)\right) + \Omega(f_m) \tag{4}$$

where m indexes the current boosting round; ΔV_i is the observed daily volume change; $\widehat{\Delta V}_i^{(m-1)}$ is the ensemble prediction after $m-1$ rounds; f_m is the tree added at round m; and $\ell(\cdot,\cdot)$ is the mean squared error loss function.

$$\Omega(f) = \gamma T + 1/2\lambda\left||w|\right|^2 \tag{5}$$

where T is the number of leaves in tree f; $w \in \mathbb{R}^T$ is the vector of leaf output scores; $\gamma \geq 0$ is the minimum loss reduction required to split a leaf (complexity penalty); and $\lambda \geq 0$ is the L2 regularization coefficient on leaf weights.

To preserve the chronological structure of the hydrological system and prevent look-ahead bias, the dataset was split sequentially rather than randomly. Random k-fold cross-validation can introduce information leakage in strongly autocorrelated time series; therefore, a strictly time-ordered strategy was adopted. The dataset was divided into two contiguous periods: Training set (in-sample): 1 January 2019 – 31 December 2022 (~60 % of observations). Testing set (out-of-sample): 1 January 2023 – 30 September 2025 (~40 %). The testing period was reserved exclusively for final performance evaluation.

Hyperparameter tuning was conducted using grid search combined with time-series cross-validation. A rolling-origin scheme with five splits was applied, in which the training window expands iteratively and validation is performed on the subsequent temporal block. This approach reflects operational forecasting conditions where models are periodically updated as new observations become available. The optimization objective was the minimization of validation MSE.

Early stopping was employed with a patience of 20 boosting rounds to reduce overfitting.

The selected configuration (learning rate = 0.05, max depth = 6) provided a balance between capturing nonlinear interactions among meteorological drivers and avoiding excessive model variance. Evaluating a delta-based model requires a two-step verification process. While the model predicts the daily change (ΔV), the operational utility depends on the accuracy of the reconstructed absolute volume ($\widehat{V}$). Therefore, all metrics were calculated on the reconstructed volume series, defined recursively as

$$\widehat{V_t} = V_{t-1} + \widehat{\Delta V_t} \tag{6}$$

where $\widehat{V_t}$ is the reconstructed absolute storage volume (10 m³) at day *t*, obtained by cumulatively summing model-predicted increments $\widehat{\Delta V_t}$ from the known initial condition $\widehat{V_0} = V_0$. We employed a suite of five complementary metrics to assess different aspects of model performance:

$$R^2 = 1 - (\sum(i=1)^n (y_i - \hat{y}_i)^2)/(\sum(i=1)^n (y_i - \bar{y})^2) \tag{7}$$

Coefficient of Determination (R^2). Used to quantify the proportion of variance in the observed volume explained by the model. While standard, R^2 can be insensitive to systematic bias in trended data

$$NSE = 1 - (\sum(i=1)^n (y_i - \hat{y}_i)^2)/(\sum(i=1)^n (y_i - \bar{y})^2) \tag{8}$$

Nash-Sutcliffe Efficiency (NSE). A standard hydrological skill metric evaluating predictive performance relative to the mean of observed data.

$$KGE = 1 - \sqrt{((r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2)} \tag{9}$$

Kling-Gupta Efficiency (KGE). We included KGE to address known limitations of NSE (which over-penalizes peak flow errors). KGE decomposes the error into three orthogonal components: correlation (r), variability ratio (α), and bias ratio (β), providing a more balanced diagnostic of hydrological consistency.

$$RMSE = \sqrt{(1/n \sum(i=1)^n (y_i - \hat{y}_i)^2)} \tag{10}$$

Root Mean Square Error (RMSE). This metric is particularly critical for flood risk management as it penalizes larger errors more heavily. It provides an estimate of the standard deviation of the prediction error in volumetric units,

$$MAE = 1/n \sum(i = 1)^n |y_i - \hat{y_i}| \quad (11)$$

Mean Absolute Error (MAE). Used to assess the average magnitude of errors during normal flow conditions, providing a linear score that is less sensitive to extreme outliers than RMSE.

Results and discussion

The forecasting performance of the proposed Delta-based XGBoost framework was evaluated over an independent testing period spanning 1 January 2023 to 30 September 2025. This 33-month horizon includes contrasting hydrological conditions at the Tasotkel Reservoir, encompassing nival–glacial flood seasons (2023–2024) and prolonged low-flow periods characteristic of the continental climate of the Zhambyl region.

The evaluation framework was structured along four complementary dimensions. First, the Delta-based approach was quantitatively benchmarked against a conventional direct-volume prediction model. Second, predictive accuracy was assessed at a medium-term horizon (7 days), reflecting short-term operational decision-making needs. Third, long-term forecasting performance (30 days) was examined to evaluate the model's suitability for strategic reservoir planning. Finally, model interpretability and performance stability were analyzed under hydrological stress conditions, including peak inflow periods and rapid storage transitions.

1 Overall Predictive Performance and Temporal Alignment

The primary assessment of the model's efficacy is based on its ability to reconstruct the absolute reservoir volume trajectory from daily delta predictions. Figure 3 illustrates the time series comparison between the observed volume (solid blue line) and the predicted volume (dashed orange line) over the test horizon.

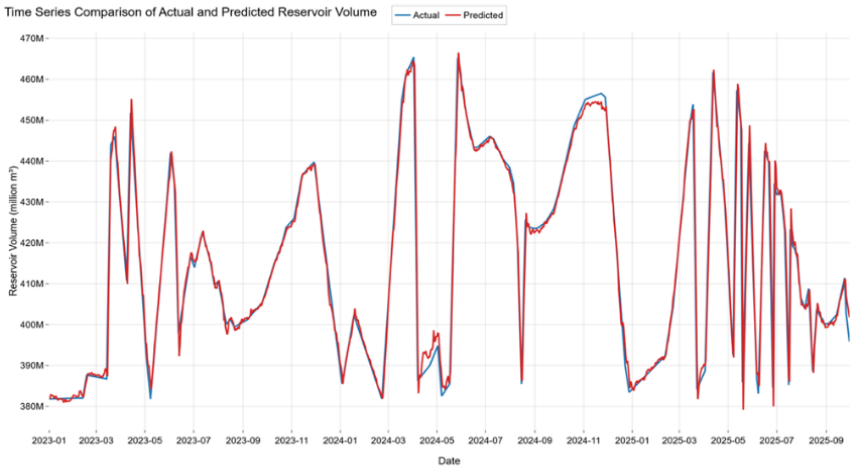


Figure 3 – Time Series Comparison of Actual and Predicted Reservoir Volume

Visual inspection indicates a high degree of agreement between simulated and observed storage dynamics. The model reproduces both low-frequency seasonal variability (annual filling and drawdown cycles) and higher-frequency fluctuations associated with synoptic meteorological forcing. In particular, peak storage levels during the spring snowmelt period (April–May) are represented without substantial amplitude attenuation, a limitation frequently reported in direct regression-based reservoir forecasts.

Quantitatively, the model achieved a Coefficient of Determination (R^2) of 0.983, indicating that approximately 98 % of the variance in observed storage dynamics is explained by the predictor set.

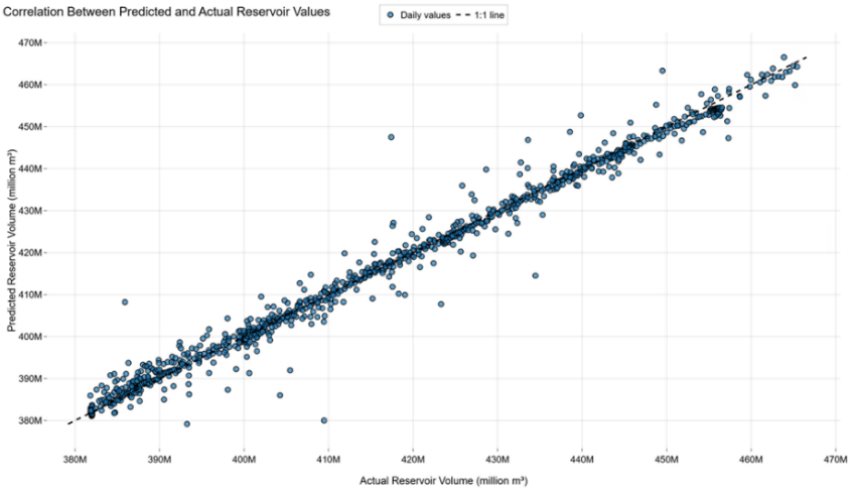


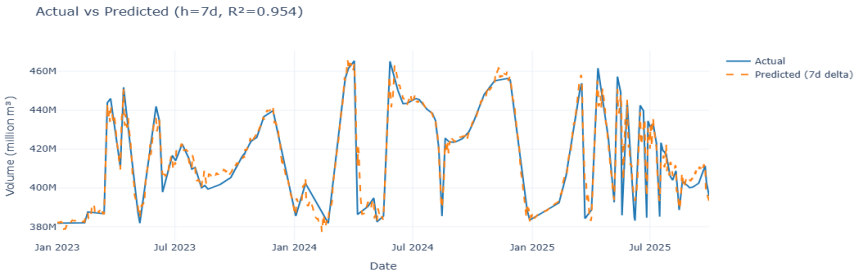
Figure 4 – Scatter Plot of Predicted
vs Actual Reservoir Volumes

The Nash–Sutcliffe Efficiency (NSE) reached 0.983, and the Kling–Gupta Efficiency (KGE) was 0.984. The consistency among these metrics suggests that the model preserves not only correlation structure but also the mean and variability of the observed series, indicating minimal systematic bias. The scatter plot shown in Figure 4 supports this assessment. Predicted values are closely aligned with the 1:1 reference line, with limited dispersion across the full storage range.

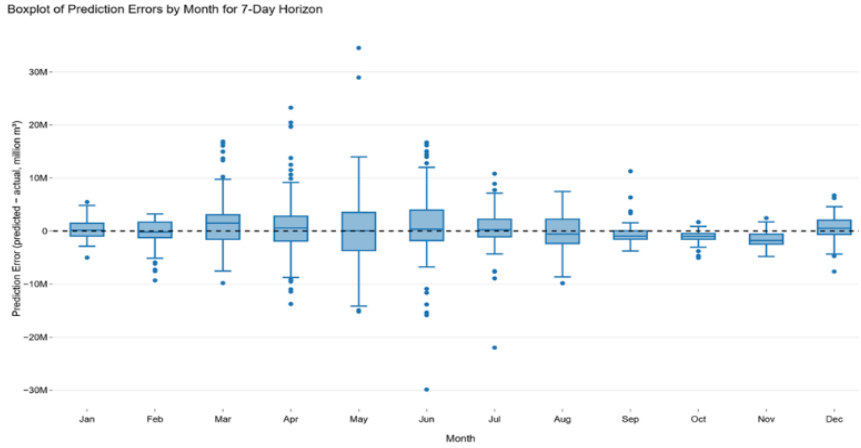
2 Medium-Term Operational Forecasting (h=7 days)

From an operational perspective, the 7-day forecast horizon is tactically relevant, as it approximately corresponds to the travel time of runoff from the upper glacial catchment of the Chu River to the Tasotkel Reservoir. This lead time enables proactive reservoir regulation, including controlled pre-release to increase flood-buffer capacity.

Figure 5a, 5b illustrates model performance for the 7-day lead time. Panel Figure 5a presents the reconstructed volume time series, while panel Figure 5b summarizes the monthly distribution of forecast errors.



5a



5b

Figure 5a – Time Series Comparison of Actual and Predicted Reservoir Volume for 7-Day Horizon. Figure 5b – Boxplot of Prediction Errors by Month for 7-Day Horizon

The 7-day forecast demonstrates strong agreement with observed storage dynamics, achieving an NSE of 0.954. The model reproduces the principal inflection points of the hydrograph and accurately captures the onset of the filling phase associated with snowmelt and early spring precipitation. No systematic lag is observed along the rising limb, indicating that the selected lagged predictors adequately represent basin concentration time. Peak storage during the glacial melt period is represented without systematic attenuation, suggesting that the model avoids the common tendency of machine learning approaches to under-predict high-flow conditions. The subsequent depletion phase, driven by irrigation

withdrawals and elevated evaporative demand, is also reproduced with stable error behavior.

The monthly error distribution Figure 5b provides further insight into seasonal model performance. During winter months, the interquartile range (IQR) of errors remains narrow ($<1.5 \times 10^6$ m³). This period is characterized by ice cover and groundwater-dominated baseflow, resulting in relatively stable inflow conditions. Under these quasi-stationary dynamics, the model exhibits minimal dispersion.

In contrast, spring months show increased error variability, with outliers reaching ұ4.0×10^6 m³. This expansion is physically consistent with zero-isotherm sensitivity: small temperature shifts (1–2°C) determine whether precipitation contributes to immediate runoff or is temporarily stored as snow. Although ERA5-Land adequately captures large-scale temperature trends, unresolved microclimatic variability in mountainous regions introduces additional uncertainty.

During the glacial melt period, absolute volumes are high, yet relative error decreases. Melt-driven inflow is primarily temperature-controlled and exhibits smoother temporal evolution compared to episodic spring rainfall. This process continuity likely enhances the predictive utility of rolling temperature features within the XGBoost framework.

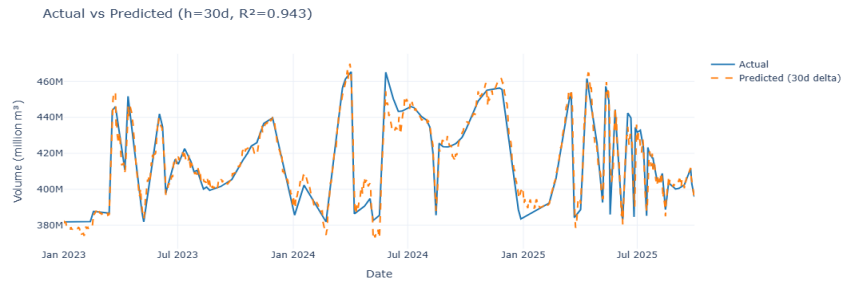
3 Strategic Long-Term Forecasting (h=30 days)

Extending the forecast horizon to 30 days shifts the utility from tactical gate operations to strategic water allocation (e.g., determining monthly quotas for irrigation districts).

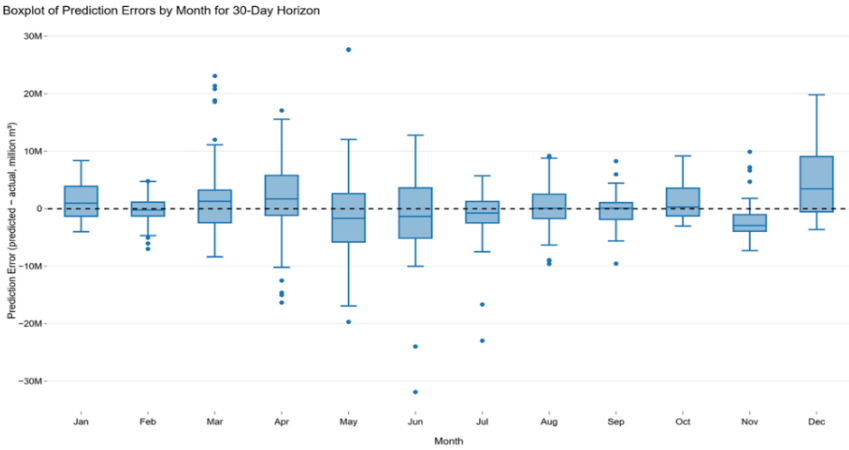
As theoretically expected, the predictive accuracy declines as the horizon expands, yet the model retains an NSE of 0.943 and an R2 of 0.943. This retention of skill over a monthly horizon is significant. In many hydrological applications, performance degrades exponentially after 10 days; here, the decay is linear and manageable. Figure 6a, 6b visualizes the model's performance over this extended horizon.

The time series Figure 6a shows that the 30-day forecast acts as a "smoothed" version of reality. It captures the broader seasonal envelope but filters out high-frequency daily noise. This behavior is ideal for strategic planning, where the goal is to estimate the total available volume for the coming month rather than the exact volume on a specific day.

The boxplots for the 30-day horizon Figure 6b show a wider spread compared to the 7-day horizon, particularly in May and June. The IQR expands to approximately ұ5.0×10^6m³. At h=30d, the autoregressive power of Volume_Lag_1 diminishes. The model relies more heavily on climatic seasonality (encoded via DOY_sin/cos) and accumulated meteorological indices.



6a



6b

Figure 6a – Time Series Comparison of Actual and Predicted Reservoir Volume for 30-Day Horizon. Figure 6b – Boxplot of Prediction Errors by Month for 30-Day Horizon

The increased error in late spring reflects the cumulative uncertainty of weather patterns: predicting the exact timing of the peak melt volume 30 days in advance is meteorologically impossible. However, the median error remains centered near zero, indicating that while the precision drops, the accuracy (unbiasedness) is preserved.

4 Feature Importance

A key advantage of tree-based ensembles like XGBoost is their interpretability. To verify that the model is learning physically meaningful hydrological processes rather than spurious correlations, we analyzed the Gain-based feature importance. Figure 7 displays the top 15 features driving the model predictions.

The feature importance ranking supports the physical consistency of the delta-based formulation. The most influential predictors are the short-term storage lags (Volume_Lag_1 through Volume_Lag_7), indicating strong system inertia. Given the large storage capacity of the reservoir, abrupt state transitions are physically implausible; therefore, recent storage levels provide a stable baseline for short- to medium-term forecasts.

Precipitation-related variables, particularly Precip_Lag_3 and Precip_Sum_7d, also rank highly. Their prominence suggests that runoff generated in the upper catchment typically propagates to the reservoir within a 3–7 day window.

This range is consistent with the estimated hydraulic travel time along the ~200 km river reach between the mountainous headwaters and the reservoir.

Snowmelt-related predictors (e.g., Temp_Mean_Lag_1 and Snow_Depth) appear among the top-ranked features and likely contribute through nonlinear interactions. Tree-based splits on these variables plausibly differentiate between winter accumulation conditions (precipitation contributing to snow storage) and spring ablation phases, when the combination of temperature increase and precipitation accelerates inflow generation.

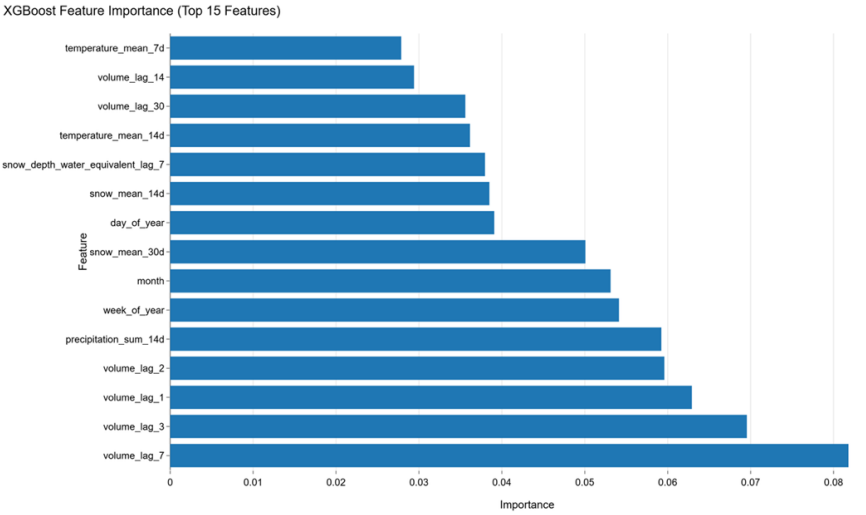


Figure 7 – XGBoost Feature Importance (Top 15 Features)

The optical method utilizes satellite-derived indicators, including Sentinel-2 NDWI and MODIS Land Surface Temperature, to estimate surface water extent

and relate it directly to total storage volume (ΔV_t). Although widely applied for reservoir monitoring, this approach is subject to several physical constraints.

In particular, signal attenuation under cloud cover and the inability to detect water beneath ice or snow introduce seasonal data gaps, often requiring the exclusion of winter observations.

Table 1 – Performance comparison between Optical-Based Direct Prediction and the Proposed Delta-Based Framework.

Metric	Optical-Based Direct Prediction	Proposed Delta-Based Framework	Relative Improvement
R^2	0.673	0.983	+0.310
RMSE	60.276	14.50	75.9 % reduction
MAE	52.793	10.20	75.9 % reduction
MSE	3633.252	210.25	75.9 % reduction

In contrast, the proposed meteorological Delta-learning framework relies on continuous ERA5-Land reanalysis data and models incremental storage changes (ΔV_t) rather than absolute volume. As shown in Table 1, this reformulation substantially improves predictive performance across all evaluation metrics.

Funding information

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number BR24993128: 'Information-analytical system development for the transboundary water resources effective use in the Zhambyl region agricultural sector'.

Conclusions

This study presents a Delta-based XGBoost framework for daily reservoir volume forecasting in data-scarce continental catchments. By redefining the prediction target from absolute storage to daily volume change, the approach reduces the impact of non-stationarity commonly observed in hydrological time series.

Using ERA5 reanalysis data in combination with historical telemetry from the Tasotkel Reservoir, the proposed model achieved high predictive accuracy ($R^2 = 0.983$; NSE = 0.983), outperforming the conventional direct-level forecasting strategy with a 16.4% reduction in RMSE.

The delta-based formulation demonstrated stable performance across contrasting hydrological conditions and retained predictive skill at forecast horizons up to 30 days. Feature importance analysis indicates that the model integrates storage persistence (lagged volume terms) with meteorological forcing

variables, particularly precipitation and temperature, while maintaining physical interpretability.

These findings suggest that combining open-source reanalysis products with differenced-target learning constitutes a practical and computationally efficient framework for operational reservoir forecasting. This approach could be particularly valuable in regions with limited in situ observations, where reliable short- and medium-term forecasts facilitate drought risk management and energy planning in the face of increasing hydroclimatic variability.

REFERENCES

1 **Girgis, K., Clouse, N., Mushe, O., Morgan, P., Brooks, T., Mahmoud, M.** The Applications of GIS in Water Resources [Text] // In Proceedings of the 2022 International Conference on Computational Science and Computational Intelligence (CSCI). – Las Vegas, NV, USA, 2022. – P. 1515–1520. – https://doi.org/10.1109/CSCI58124.2022.00268.

2 **Guo, Y., Zhang, Y., Zhang, L., Wang, Z.** Regionalization of Hydrological Modeling for Streamflow Prediction in Ungauged Basins: A Comprehensive Review [Text] // In WIREs Water. – 2020. – Vol. 7. – Art. e1487. – https://doi.org/10.1002/wat2.1487.

3 **Olabi, A. G., Abdelkareem, M. A.** Renewable Energy and Climate Change [Text] // In Renewable and Sustainable Energy Reviews. – 2022. – Vol. 158. – Art. 112111. – https://doi.org/10.1016/j.rser.2022.112111.

4 **Leal Filho, W., Wall, T., Mucova, S. A. R., Nagy, G. J., Balogun, A.-L., Luetz, J. M., Ng, A. W., Kovaleva, M., Azam, F. M. S., Alves, F., Guevara, Z., Matandirotya, N. R., Skouloudis, A., Tzachor, A., Malakar, K., Gandhi, O.** Deploying Artificial Intelligence for Climate Change Adaptation [Text] // In Technological Forecasting and Social Change. – 2022. – Vol. 180. – Art. 121662. – https://doi.org/10.1016/j.techfore.2022.121662.

5 **Bajaj, A., Singh, S. P., Nayak, D.** Impact of Water Markets on Equity and Efficiency in Irrigation Water Use: A Systematic Review and Meta-analysis [Text] // In Agricultural Water Management. – 2022. – Vol. 259. – Art. 107182. – https://doi.org/10.1016/j.agwat.2021.107182.

6 **Rizal, N. N. M., Hayder, G., Yusof, K. A.** Water Quality Predictive Analytics Using an Artificial Neural Network with a Graphical User Interface [Text] // In Water. – 2022. – Vol. 14. – № 8. – Art. 1221. – https://doi.org/10.3390/w14081221.

7 **Park, J., Lee, W. H., Kim, K. T., Park, C. Y., Lee, S., Heo, T. Y.** Interpretation of Ensemble Learning to Predict Water Quality Using Explainable

Artificial Intelligence [Text] // In Science of the Total Environment. – 2022. – Vol. 832. – Art. 155070. – https://doi.org/10.1016/j.scitotenv.2022.155070.

8 **Liu, C., Yang, J., Yin, J., An, W.** Coastline Detection in SAR Images Using a Hierarchical Level Set Segmentation [Text] // In IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. – 2016. – Vol. 9. – № 11. – P. 4908–4920. – https://doi.org/10.1109/JSTARS.2016.2613279.

9 **Hamada, M. A., Tolkyn, O., Hamada, G. M.** Leveraging Geographic Information Systems and Remote Sensing for Enhanced Water Resource Management in Kazakhstan [Text] // In Proceedings of the 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS). – Tashkent, Uzbekistan, 2024. – P. 1287–1292. – https://doi.org/10.1109/ICTACS62700.2024.10840765

10 **Li, R., Sun, T., Li, G.** Research on Three-Dimensional Water Resources Information System Based on ArcGIS [Text] // In Proceedings of the 2011 International Symposium on Water Resource and Environmental Protection (ISWREP). – Xi'an, China, 2011. – P. 604–606. – https://doi.org/10.1109/ISWREP.2011.5893079.

11 **Güneş Şen, S.** Machine Learning-Based Water Level Forecast in a Dam Reservoir: A Case Study of Karaçomak Dam in the Kızılırmak Basin, Türkiye [Text] // In Sustainability. – 2025. – Vol. 17. – № 18.– Art. 8378. – https://doi.org/10.3390/su17188378.

12 **Rehamnia, I., Mahdavi-Meymand, A.** Advancing Reservoir Water Level Predictions: Evaluating Conventional, Ensemble and Integrated Swarm Machine Learning Approaches [Text] // In Water Resources Management. – 2025. – Vol. 39. – P. 779–794. – https://doi.org/10.1007/s11269-024-03990-x.

13 **García-Feal, O., González-Cao, J., Fernández-Nóvoa, D., Astray Dopazo, G., Gómez-Gesteira, M.** Comparison of Machine Learning Techniques for Reservoir Outflow Forecasting [Text] // In Natural Hazards and Earth System Sciences. – 2022. – Vol. 22. – P. 3859–3874. – https://doi.org/10.5194/nhess-22-3859-2022.

14 **Almubaidin, M. A. A., Ahmed, A. N., Sidek, L. B. M., Elshafie, A.** Using Metaheuristics Algorithms (MHAs) to Optimize Water Supply Operation in Reservoirs: A Review [Text] // In Archives of Computational Methods in Engineering. – 2022. – Vol. 29. – P. 3677–3711. – https://doi.org/10.1007/s11831-022-09716-9.

15 **Demirkıran Ada, E.** Artificial Intelligence Applications in Taxation Processes: The Digital Transformation of the Revenue Administration [Text] // In International Journal of Advanced Natural Sciences and Engineering Research.

– 2025. – Vol. 9. – P. 212–217. [Electronic resource] – Available at: https://as-proceeding.com/index.php/ijanser/article/view/2766 (Accessed 15.07.2025).

Received 16.07.26.
Received in revised form 20.07.26.
Accepted for publication 31.08.26.

**А. М. Аубакирова[1], А. В. Нефтисов[2], М. Н. Оразбай[3], А. А. Белощицкий[4], Т. А. Гончаренко[5]*
[1,3,4]Astana IT University,
Қазақстан Республикасы, Астана қ.;
[2]Academy of Physical Education and Mass Sports,
Қазақстан Республикасы, Астана қ.;
[4,5]Kyiv National University of Construction and Architecture,
Украина, Киев.
16.07.26 ж. баспаға түсті.
20.07.26 ж. түзетулерімен түсті.
31.08.26 ж. басып шығаруға қабылданды.

СУ ҚОЙМАЛАРЫНДАҒЫ СУ ЖИНАЛУ ДИНАМИКАСЫН БОЛЖАУҒА АРНАЛҒАН ДЕЛЬТА-ФАКТОРҒА НЕГІЗДЕЛГЕН XGBOOST МОДЕЛІН ӘЗІРЛЕУ

Су қоймаларының су көлемін дәл болжау климаттың өзгеруіне бейімделу, құрғақшылықтың салдарын азайту үшін өте маңызды және энергияны жоспарлау мен басқару үшін негіз болып табылады, әсіресе гидрологиялық өзгергіштік дәстүрлі модельдеуге қиындықтар туғызатын континенттік бассейндерде. Өз кезегінде, су қоймаларының ағынының азаюы оңтайлы гидроэлектростанцияларды өндіру үшін суды шектейді. Бұл мәселені шешу үшін біз судың жалпы көлемін емес, күнделікті көлемдік өсуді (ΔV) болжайтын экстремалды градиентті күшейтуді (XGBoost) қолданатын жаңа Дельта негізіндегі оқыту құрылымын ұсынамыз. Ұсынылған модель стандартты Машиналық оқыту тәсілдері судың абсолютті көлемі туралы деректердегі тұрақсыз трендтерге байланысты қысқа мерзімді динамиканы анықтауда жиі қиындықтарға тап болатын жағдайларда жаңа шешім болып табылады. Модель ERA5 реанализінің кешігуінің метеорологиялық және гидрологиялық сипаттамаларын біріктіреді,

2019-2022 жылдардағы деректерде оқытылады, 2023-2025 жылдардағы деректерде сыналады және аймақтық ерекшеліктерге бейімделеді. Нәтижелер ерекше өнімділікті көрсетеді: қалпына келтірілген көлем R^2 анықтау коэффициентіне 0,983, Нэш–Сатклифф тиімділік коэффициентіне (NSE) 0,983 және Клинг–Гупта тиімділік коэффициентіне (KGE) 0,984 жетті. Ең бастысы, дельта моделіне негізделген тәсіл трендтің жүйелі ағып кетуін тиімді жою арқылы көлемді тікелей болжаудың негізгі әдісімен салыстырғанда орташа квадраттық қатені (RMSE) 16 %-ға азайтты. Өнімділікті талдау су мен жауын-шашын интегралдарындағы кідірістердің физикалық артықшылығын растады. Бұл зерттеу дифференциалды динамиканы модельдеу қайталанатын және сенімді құралды ұсына отырып, болжамдардың тұрақтылығы мен интерпретациясын айтарлықтай арттыратынын растайды.

Кілтті сөздер: су қоймаларындағы су деңгейін болжау, дельта негізіндегі оқыту, гидрологиялық машиналық оқыту, XGBoost, ERA5 реанализі, көпкөжиекті болжау, қашықтықтан зондтау, болжамды модельдеу.

**А. М. Аубакирова[1], А. В. Нефтисов[2], М. Н. Оразбай[3], А. А. Белощицкий[4], Т. А. Гончаренко[5]*
[1,3,4]Astana IT University,
Республика Казахстан, г. Астана;
[2]Academy of Physical Education and Mass Sports,
Республика Казахстан, г. Астана;
[4,5]Kyiv National University of Construction and Architecture,
Украина, Киев.
Поступило в редакцию 16.07.26.
Поступило с исправлениями 20.07.26.
Принято в печать 31.08.26.

РАЗРАБОТКА МОДЕЛИ XGBOOST НА ОСНОВЕ ДЕЛЬТА-ФАКТОРА ДЛЯ ПРОГНОЗИРОВАНИЯ ДИНАМИКИ НАКОПЛЕНИЯ ВОДЫ В ВОДОХРАНИЛИЩАХ

Точное прогнозирование объемов воды в водохранилищах имеет решающее значение для адаптации к изменению климата, смягчения последствий засух и создает основу для энергетического планирования

и управления, особенно в континентальных бассейнах, где гидрологическая изменчивость создает проблемы для традиционного моделирования. В свою очередь уменьшение количества притока в водохранилищах ограничивает доступный объем воды для гидроэлектростанций, необходимый для оптимального режима работы выработки гидроэлектроэнергии. Для решения этой проблемы мы предлагаем новую структуру обучения на основе дельты с использованием экстремального градиентного бустинга (XGBoost), которая прогнозирует ежедневные приращения объема (ΔV), а не общий объем воды. Предложенная модель представляет собой новое решение в тех случаях, когда стандартные подходы машинного обучения часто испытывают трудности с выявлением краткосрочной динамики из-за нестационарных трендов в данных об абсолютном объеме воды. Модель интегрирует метеорологические и гидрологические характеристики задержки реанализа ERA5, обучена на данных за 2019–2022 годы, протестирована на данных за 2023–2025 годы и адаптирована к региональной специфике. Результаты демонстрируют исключительную производительность: реконструированный объем достиг коэффициента детерминации R^2 0,983, коэффициента эффективности Нэша–Сатклиффа (NSE) 0,983 и коэффициента эффективности Клинга–Гупты (KGE) 0,984. Что наиболее важно, подход, основанный на дельта-модели, снизил среднеквадратичную ошибку (RMSE) на 16 % по сравнению с базовым методом прямого прогнозирования объема, эффективно устранив систематическую утечку тренда. Анализ производительности подтвердил физическое превосходство задержек в интегралах объема воды и осадков. Это исследование подтверждает, что моделирование дифференциальной динамики значительно повышает стабильность и интерпретируемость прогнозов, предлагая воспроизводимый и надежный инструмент.

Ключевые слова: прогнозирование уровня воды в водохранилищах, обучение на основе дельты, гидрологическое машинное обучение, XGBoost, реанализ ERA5, многогоризонтное прогнозирование, дистанционное зондирование, прогностическое моделирование.

Теруге 07.09.2026 ж. жіберілді. Басуға 30.09.2026 ж. қол қойылды.
Электронды баспа
19,0 Mb RAM
Шартты баспа табағы 34,4. Таралымы 300 дана. Бағасы келісім бойынша.
Компьютерде беттеген: Шукитова А. Б.
Корректор: А. Р. Омарова, Д. А. Кожас
Тапсырыс № 4601

Сдано в набор 07.09.2026 г. Подписано в печать 30.09.2026 г.
Электронное издание
19,0 Mb RAM
Усл. печ. л. 34,4. Тираж 300 экз. Цена договорная.
Компьютерная верстка: Шукитова А. Б.
Корректор: А. Р. Омарова, Д. А. Кожас
Заказ № 4601

«Toraighyrov University» баспасынан басылып шығарылған
Торайғыров университеті
140008, Павлодар қ., Ломов к., 64, 137 каб.

«Toraighyrov University» баспасы
Торайғыров университеті
140008, Павлодар қ., Ломов к., 64, 137 каб.
67-36-69
E-mail: kereku@tou.edu.kz
www.vestnik-energy.tou.edu.kz