Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

DOI: 10.37943/22NEJN3212

Aliya Aubakirova
PhD student, Junior Researcher, Research and Innovation Center "Industry 4.0"
aliya.aubakirova@astanait.edu.kz, orcid.org/0009-0004-6925-6714
Astana IT University, Kazakhstan

Andrii Biloshchytskyi
Doctor of Technical Sciences, Professor, Vice-Rector for Science and Innovations
a.b@astanait.edu.kz , orcid.org/0000-0001-9548-1959
Astana IT University, Kazakhstan
Professor Department of Information Technologies,
Kyiv National University of Construction and Architecture, Ukraine

Mukhtar Orazbay
MS, AI developer, Research and Innovation Center "Industry 4.0"
m.orazbay@astanait.edu.kz , orcid.org/0009-0001-4994-6484
Astana IT University, Kazakhstan

Ilyas Kazambayev
Junior Researcher, Research and Innovation Center "Industry 4.0"
ilyaskazambayev@gmail.com, orcid.org/0000-0003-0850-7490
Astana IT University, Kazakhstan

Alexandr Neftissov
PhD, Associate Professor, Research and Innovation Center "Industry 4.0"
alexandr.neftissov@astanait.edu.kz, orcid.org/0000-0003-4079-2025
Astana IT University, Kazakhstan

DEVELOPMENT OF THE INTEGRATED WATER RESOURCES MONITORING AND FORECASTING MODULE FOR DECISION SUPPORT SYSTEMS AT HYDROTECHNICAL STRUCTURES

Abstract: Nowadays, it is necessary to use monitoring and forecasting technologies for effective water resources management at water management facilities. The objective of this study is to develop and verify an integrated approach to water resources forecasting with the task of identifying features for forecasting, designing a data preprocessing submodule and a forecasting module. The workflow diagram of the water forecasting system includes sequential stages of data collection, preprocessing, filtering, feature extraction, and training. Sentinel-2 and MODIS satellite sources were used for data preprocessing. Predictors for the formation of time series by normalized difference water index (NDWI) and water surface temperature (LST) were selected in the feature engineering stage. The XGBoost Regressor algorithm was chosen due to its ability to model nonlinear relationships and feature interactions. Excluding winter months improved the model performance for all metrics, which demonstrates the importance of seasonal filtering when working with optical satellite data. The machine learning algorithm takes into account the analysis of satellite data (NDWI and LST indices) through the Google Earth Engine (GEE) platform. Both seasonal and long-term dynamics of water volumes in the Tasotkel reservoir are monitored for the period from 2020 to 2024. In practice, image initial filtering submodules were developed using linear regression and the XGBoost model. Model trained without winter data shows high performance using Metrics Mean Absolute Error (MAE)

Received: 26.05.2025 Accepted: 24.06.2025 Published: 30.06.2025

DOI: 10.37943/22NEJN3212

of 52.793, Root Mean Squared Error (RMSE) of 60.276, coefficient of determination (R^2) of 0.673 and Mean Squared Error (MSE) of 3633.252 metrics. However, a decrease in clarity was observed due to snow and ice on reflective properties in winter. For the purpose of rational water resources management, the combination of satellite images and machine learning algorithms in this study shows the prospects for development.

Keywords: machine learning, remote sensing, water resources monitoring, predictive modeling, reservoir

Introduction (Literary review)

According to the Concept for the Development of the Water Resources Management System of the Republic of Kazakhstan for 2024-2030, the main problems include irrational and inefficient use of water resources, lack of proper information and analytical support and monitoring of the water resources management system [1]. According to forecasts, by 2040, Kazakhstan will be among the 20 countries with the highest level of water deficit in the world if current conditions persist [2]. Wasteful use of water resources is mostly visible in the agricultural sector, where, according to the results of 2022, the share of losses in irrigated agriculture is 65%. According to forecasts, in 2029, there is a risk of a decrease in the rate of socio-economic development due to water shortages. Water accounting is organized at a low level, water measurement is carried out using outdated methods and not all devices are properly tested. Insufficient control leads to the drying up of small rivers and their tributaries. There is no proper information and analytical support and monitoring of the water resources management system. The methodological basis of forecasts based on outdated methods and technologies has led to a steady trend towards deterioration in the quality of hydrological data.

The existing problem of acute shortage of water resources is complicated by discrepancies in the indicators of water resources data provided by various sources. This leads to the lack of reliable and truthful information on the real state of the water sphere. Discrepancies in accounting indicators complicate the planning process and make it difficult to make effective decisions. This leads to an increase in deficit, a decrease in the sustainability of water supply and an increase in conflicts between sectors dependent on water resources. For example, the river flow was 181 km^3 (2010), 160 km^3 (2016) according to the Water Management Committee of the Ministry of Water Resources and Irrigation of the Republic of Kazakhstan, in parallel with this, the republican state enterprise "Kazhydromet" indicates other data 143.6 km^3 (2010) and 146.0 km^3 (2016). The data on the number of rivers and temporary watercourses also have discrepancies [1], [3], [4].

With proper water management, the amount and location of available water resources in each geographic region, especially in certain mountains and glaciers, can be effectively monitored to prevent damaging events [5]. To prevent and reduce droughts, water shortages and their impacts in the future, be more prepared and resilient by knowing where natural disasters have occurred in the past. It is also necessary to select the right combination of satellite platforms, water parameters and machine learning models when optimizing water monitoring. Models such as convolutional neural network (CNN), recurrent neural networks (RNN) and generative adversarial network (GAN) provide high accuracy, but their applicability depends on the quality and resolution of satellite data.

In the paper [6], RNN and GAN models demonstrated high efficiency in eliminating distortions in remote sensing images, which allowed accurate prediction of water quality indicators. In deep learning, RNN methods are used for forecasting, but require large amounts of data [7]. The RNN-based time series model is effective for intelligent forecasting of water levels in reservoirs but has fewer parameters and is suitable for a small dataset [8]. The combined ARIMA-RNN model takes into account the linear and nonlinear components of the data when

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

forecasting the water level and demonstrates accuracy but requires further optimization taking into account the spatial relationship of the sites [9]. In small reservoirs, combining satellite data and CNN to obtain the water level shows stable and accurate results [10]. In addition, convolutional neural networks CNN use semantic segmentation methods for more accurate detection of reservoirs [11]. Combining satellite data with machine learning makes it possible to develop real-time monitoring systems, promoting the rational use of resources [12]. An information portal with a high-speed storage of satellite images recording information on surface waters in both spatial and temporal terms helps to use maps of the Water Observation from Space (WOfS) product for visualization, analysis and risk forecasting. The method used showed that one algorithm is operationally applicable in many environmental and climatic conditions, but known errors remain due to cloud shadows and relief [13]. The database using ICESat/ICESat-2 altimetry to determine demonstrate and Landsat images to determine the exact boundaries of surface water bodies demonstrates the general trend of water levels in rivers or reservoirs [14].

Monitoring changes in water volumes in rivers and reservoirs is of great importance in predicting water disasters, assessing the impact of human activities on water security, and analyzing climate and environmental changes. The KazRivDyn platform, developed on Google Earth Engine (GEE), implements an automatic algorithm for identifying rivers on a map in real time. However, no dependencies (correlations) between water volume and glacier volume have been identified, and no forecasting model using the regression method has been built [15]. The study used SARIMAX and RNN models to predict water levels, and the use of GEE to analyze satellite images provided an accurate calculation of the water surface, but RNN-based models require large amounts of data and significant computing resources [16].

Automatic detection of water bodies plays an important role in water resources management. One of the important sensors for monitoring surface water change is synthetic aperture radar (SAR). A semi-supervised algorithm for detecting water and land in SAR images is presented using lognormal density mixtures as a probability model for pixel intensities, which are estimated using the expectation–maximization (EM) algorithm simultaneously with the contour evolution [17]. However, for more accurate SAR image extraction, a water extraction algorithm based on a modified Markov random field and a CNN model is needed. In order to extract and compile reliable features of water and non-water pixels, a CNN-based method is used and tested on images using scenes with different characteristics of water resources based on Sentinel-1 SAR. This method shows more accurate overall classifications compared to OTSU thresholding and support vector machines (SVM) methods [18]. Extraction of parameters from various types of remotely sensed data is carried out by geographic information systems Sentinel Application Platform (SNAP), Multi-national Geographic Information System (Arc-GIS), Quantum Geographic Information System (QGIS), American [19].

At the moment, the water resources management system of Kazakhstan lacks modern digital technologies with a single digital database that could quickly consolidate relevant information. In these conditions, the introduction of GIS, automation and digitalization are effective tools for improving the efficiency of management and technological processes.

The aim of this study is to develop and verify an integrated approach to forecasting water resources at hydraulic structures by creating a forecasting module integrated into a decision support system.

The set goal can be achieved by defining features for forecasting water resources at hydraulic structures, followed by developing a submodule for preliminary data processing based on NDVI and LST methods and further developing a module for forecasting water resources at hydrotechnical structures.

DOI: 10.37943/22NEJN3212

Methods and Materials

Data and preprocessing

The dataset combines satellite observations with ground-based measurements to model the daily water volume in a reservoir. Multispectral Sentinel-2 Level-1C images were used to derive the Normalized Difference Water Index (NDWI) over the Tasotkel Reservoir, a widely used indicator for surface water detection. The Normalized Difference Water Index (NDWI) was calculated using the green (B3) and near-infrared (B8) bands according to McFeeters [20]:

$$NDWI = \frac{B3 - B8}{B3 + B8} \quad (1)$$

where B3 represents reflectance in the green band, and B8 represents reflectance in the near-infrared band.

All available Sentinel-2 images from 2020 to 2024 covering the reservoir were processed. Scenes with excessive cloud cover were removed using a cloud filter (metadata field cloudy_pixel_percentage < 20%). After cloud filtering, NDWI was calculated for each remaining image. These NDWI values were then spatially averaged over the reservoir's region of interest (ROI) to obtain a single representative water index value for the image date.

In parallel, land surface temperature (LST) data were obtained from the daily MODIS Terra MOD11A1.061 product, which provides daytime land surface temperature at a 1 km resolution. Each MODIS LST pixel value was converted to degrees Celsius (°C) according to Wan et al. [21]:

$$LST_C = (LST_k * 0.02) - 273.15 \quad (2)$$

where is the temperature in Kelvin provided by MODIS.

LST values within the reservoir ROI were averaged to produce a daily mean LST value for the reservoir (denoted as LST_C). To focus on periods when the reservoir is not frozen and NDWI is reliable, days with LST_C < 0 °C were excluded from the analysis (typically winter months when the reservoir surface may be covered with ice).

Daily water volume data for the reservoir from 2020 to 2024 were provided by the regional division of Kazhydromet and expressed in million cubic meters (Million m^3). These values were used as the target variable.

The satellite-derived NDWI and land surface temperature (LST_C) data were synchronized by date with the observational water volume data. Since Sentinel-2 does not provide daily imagery (its revisit period is 5 days), and some images were filtered due to cloud cover, linear interpolation was applied to restore continuity in the NDWI time series.

Thus, NDWI was filled in for days with missing values based on the nearest available observations. This enabled the formation of a continuous daily NDWI time series synchronized with the water volume data. Each final daily record contained two features (NDWI and LST_C) and the corresponding water volume value. Sentinel-2 and MODIS data are publicly available via Google Earth Engine.

Importantly, volume lags (past values of the target variable) were not included in the model to make the forecast entirely based on external satellite-derived indicators, without using the history of the target variable itself. This approach allows testing the predictive power of only remotely sensed parameters, although it excludes potential model improvements from the autocorrelation of past volume values.

The geographic setting of the Tasotkel Reservoir, situated in the southern region of Kazakhstan, straddling the administrative boundary between the Zhambyl and Zhetysu provinces (Figure 1). The reservoir is primarily fed by the Shu River, which flows through its central basin and serves as the main hydrological inflow source.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

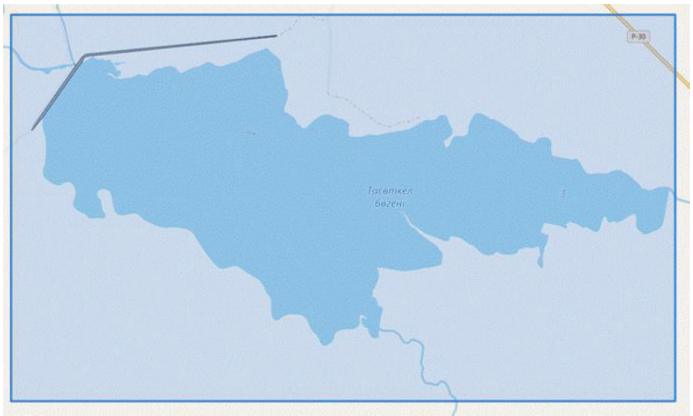

Figure 1. Area of interest of the Tasotkel reservoir

The image provides a sharply delineated view of the water surface, highlighted in a vivid cyan hue to distinguish it from surrounding terrain. The region of interest (ROI) used in the satellite-based analysis is delineated with a rectangular bounding box encompassing the entire reservoir along with adjacent land areas. This spatial configuration ensures comprehensive coverage of both the reservoir's open water body and its littoral zones. Such inclusion is particularly critical for the accurate computation of water-related indices, such as the Normalized Difference Water Index (NDWI) and land surface temperature (LST_C), using data processed within the Google Earth Engine environment. The seasonal dynamics of the reservoir's surface water extent, as derived from NDWI time series data (Figure 2).

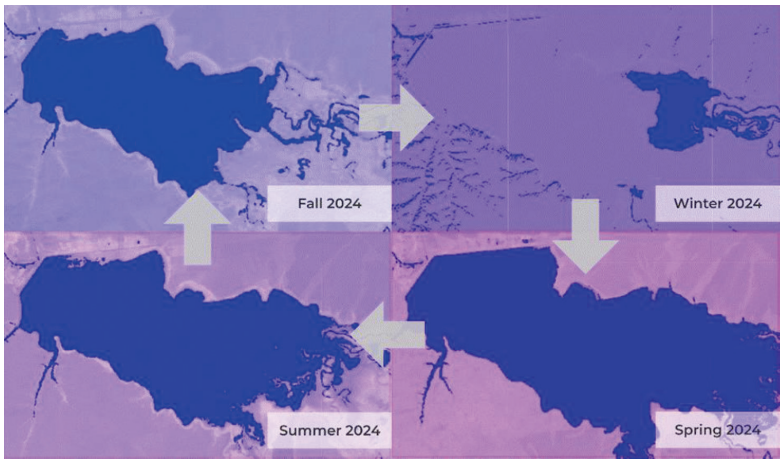

Figure 2. Seasonal dynamics of water surface in 2024

A pronounced seasonal pattern emerges, closely aligned with regional hydrological and climatic cycles.

Feature engineering

The selected predictors for the forecasting model are NDWI and LST_C, which represent complementary physical characteristics of reservoir conditions. NDWI captures the spatial extent of surface water coverage: higher NDWI values indicate a broader water surface area, which typically correlates with increased water volume. This index provides a direct proxy for the dynamics of the reservoir's water body, as NDWI effectively distinguishes water pixels in satellite imagery and serves as a robust estimator of surface water area.

DOI: 10.37943/22NEJN3212

In this study, a threshold of NDWI > 0 was employed to delineate surface water, consistent with widely adopted practices in remote sensing. NDWI values above zero are interpreted as indicative of water presence, reflecting water's strong reflectance in the green band (B3) and low reflectance in the near-infrared band (B8), which together yield positive index values. Pixels with NDWI ⩽ 0 are classified as land, vegetation, or other non-water surfaces. Hence, NDWI serves not only as a continuous measure of water coverage but also as the basis for generating binary water masks.

LST_C functions as a meteorological indicator capturing thermal influences on the reservoir. Elevated surface temperatures can intensify evaporative losses from the water surface, leading to volume reductions, while simultaneously accelerating snowmelt in the upstream basin during spring and summer, thereby increasing inflow. Thus, LST_C encapsulates two opposing hydrological drivers:

- evaporative demand, associated with high temperatures and arid conditions that promote water loss
- meltwater contribution, driven by seasonal warming that enhances runoff

Incorporating LST_C alongside NDWI is physically justified, as it introduces sensitivity to temperature-dependent processes that influence reservoir volume variability. Prior to model training, both features were standardized to zero mean and unit variance to prevent any single feature from disproportionately influencing the model and to ensure efficient convergence during optimization. Each feature x was subjected to z-score normalization, defined as:

$$x' = \frac{x - \mu}{\sigma} \tag{3}$$

where μ is the mean and σ is the standard deviation computed over the training period.

The target variable, daily water volume (in million cubic meters), was not normalized, given that regression models can handle target values on their original physical scale. Error metrics are likewise reported in physical units to facilitate interpretation and relevance for decision-making.

The combination of NDWI and LST_C represents a minimalist yet informative feature set, balancing model interpretability and generalization capability under data-scarce conditions. NDWI provides insights into surface water availability, while LST_C captures climatic pressures affecting the reservoir's hydrological regime.

It is important to acknowledge that other influential variables, such as precipitation and river inflow, were excluded due to the unavailability of high-temporal-resolution data. Nonetheless, their effects are indirectly reflected: NDWI responds to water level changes following precipitation, and LST_C mediates processes such as evaporation and snowmelt. Previous research underscores precipitation and evapotranspiration as critical drivers in hydrological modeling. While evapotranspiration is partially addressed through temperature, the lack of explicit precipitation input remains a recognized limitation of the present framework.

Machine learning model

The primary model selected for forecasting the daily reservoir volume was the XGBoost Regressor algorithm. XGBoost (eXtreme Gradient Boosting) is an ensemble learning technique based on decision trees that constructs an additive model by sequentially building trees and optimizing the error via gradient boosting. This algorithm was chosen due to its capacity to model nonlinear relationships and feature interactions, its robustness to outliers and feature scaling, and its strong track record in regression tasks with limited data. Ensemble methods

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

like boosting have demonstrated superior accuracy and reliability in hydrological forecasting contexts.

In this application, XGBoost naturally captures the nonlinear dependence of reservoir volume on NDWI, which may exhibit plateauing or nonlinear behavior relative to surface area, as well as temperature thresholds, such as LST effects manifesting primarily above freezing. Furthermore, XGBoost incorporates built-in regularization mechanisms to control model complexity, thereby reducing the risk of overfitting given the relatively small dataset.

The model is trained by minimizing a regularized objective function. Given a training set $\{(x_i, y_i)\}_{i=1}^{n}$, where represents the feature vector and y_i the corresponding target (water volume), the ensemble comprises K regression trees $f_i(x)$, each belonging to a functional space of tree-based learners. The objective function is expressed as:

$$£ = \sum_{i=1}^{n} l(y_i, \bar{y}_i) + \sum_{k=1}^{k} \Omega(f_k) \tag{4}$$

where

$$\bar{y}_i = \sum_{k=1}^{k} f_k(x_i) \tag{5}$$

and the loss function lll is the mean squared error:

$$I(y, \bar{y}) = (y - \bar{y})^2 \tag{6}$$

The regularization term $\Omega(f)$ penalizes model complexity and is formulated as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda |w|^2 \tag{7}$$

Here, T is the number of leaves in the tree, w is the vector of leaf weights, and γ and λ are hyperparameters governing the strength of regularization. This formulation balances fitting accuracy against model complexity, minimizing training error while constraining overcomplex trees.

XGBoost implements gradient boosting by iteratively adding new trees f_k trained on the residuals of previous predictions. Each added tree minimizes the gradient of the loss function, with the process continuing until a predefined number of trees is reached or model improvements plateau.

Hyperparameter tuning was performed using grid search combined with cross-validation on training data spanning 2020–2023. Parameters optimized included the number of trees (n_estimators), maximum tree depth, learning rate, and L2 regularization coefficient (lambda). To respect the temporal dependence of the data, cross-validation was conducted with chronological splits: training on data from 2020–2022 and validation in 2023, preventing data leakage from future periods. The optimal configuration balanced bias and variance with a tree depth of 3, learning rate of 0.1, approximately 100 boosting iterations, and moderate regularization. This setting achieved the lowest RMSE on the validation subset.

This strategy enhances model generalizability and mitigates overfitting risks tied to specific years within the training timeframe. The training period covered January 1, 2020, through December 31, 2023.

For evaluating predictive performance, the entire year 2024 (January 1 to December 31) was reserved as an independent test set, simulating real-world forecasting of future periods.

DOI: 10.37943/22NEJN3212

Data were neither shuffled nor randomly partitioned, as preserving temporal order is critical for valid time series model validation.

A baseline model was implemented using multiple linear regression with ordinary least squares. The target water volume was modeled as a linear function of the two predictors, NDWI and LST_C:

$$Volume = \beta_0 + \beta_1 * NDWI + \beta_2 * LST\_C \quad (8)$$

By assuming linear correlations between predictors and reservoir volume, the baseline model offers a simple, straightforward approximation. The benefits of nonlinear ensemble learning can be quantified by comparing XGBoost to this benchmark.

A linear model is unable to capture the subtleties of volume dynamics when nonlinear effects or feature interactions are present, such as temperature effects that only occur above freezing, or NDWI saturation at high reservoir levels. XGBoost can do this. Threshold behaviors and index saturation during reservoir overflow cannot be modeled by linear regression.

XGBoost's ability to model complex nonlinear hydrological processes affecting reservoir storage led to its selection. While XGBoost provides the sophisticated modeling flexibility needed to more accurately model natural system behavior in water management, linear regression serves as a low-complexity baseline. As a boosting model, XGBoost gradually reduces ensemble errors while maintaining robustness to overfitting using built-in regularization. Compared to neural network-based approaches, XGBoost requires less data, which is an advantage given the limited feature set.

Results

After training the models on data from 2020 to 2023, their predictive ability was assessed on daily observations for 2024. The XGBoost model demonstrated significantly higher accuracy compared to the baseline linear regression. Table 1 presents the error metrics on the test set for both models.

The workflow (Figure 3) integrates Sentinel-2 NDWI and MODIS LST data with ground-truth volumes, filtered to exclude winter months (Dec–Feb) and subzero temperatures Temporal features (month/year, winter flags) were engineered, and linear regression and XGBoost models (with/without winter data) were trained using standardized features (seed = 42).

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

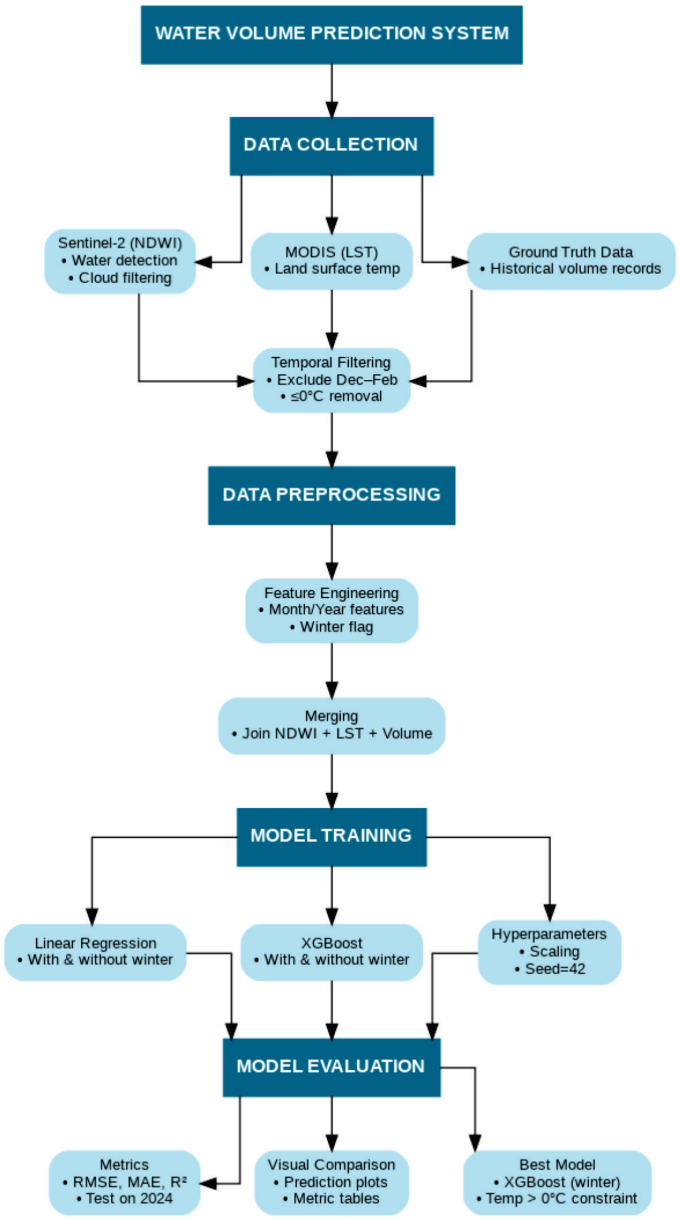

Figure 3. Workflow for water volume prediction system

Models were evaluated on 2024 holdout data via RMSE, MAE, and R^2, with XGBoost (winter-inclusive) selected as optimal.

Evaluation Setup

Model performance was assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2), and Adjusted R-squared.

MAE measures the average absolute deviation of predicted house prices from actual values, providing an interpretable estimate of typical prediction errors. Lower MAE values indicate higher precision; in the study, Polynomial Regression achieved the lowest MAE (99,958), demonstrating superior accuracy compared to other models [22].

RMSE amplifies the impact of large errors by squaring residuals before averaging, making it sensitive to outliers. A lower RMSE reflects stable predictions across the dataset. Polynomial Regression again outperformed with an RMSE of 161,044, while Support Vector Regression (SVR) showed the highest errors (RMSE: 390,627), revealing poor fit [23].

DOI: 10.37943/22NEJN3212

R^2 quantifies the proportion of variance in housing prices explained by the model. Values closer to 1 indicate stronger predictive power. Polynomial Regression achieved the highest R^2 (0.82), suggesting it captured 82% of price variability, whereas SVR's negative R^2 (-0.056) implied it performed worse than a baseline mean model [24].

Baseline Comparison

To evaluate the quality of nonlinear modeling, the performance of XGBoost was compared to a linear regression baseline. The baseline is a standard linear modeling approach that assumes a direct proportional relationship between the predictors and the target variable. Despite its simplicity and interpretability, linear regression is often insufficient to capture complex hydrological dynamics governed by multiple interacting physical processes.

Across all tested configurations, XGBoost significantly outperformed the linear model. The mean absolute error (MAE) for XGBoost was 52.793 Mm^3 when excluding winter months, compared to 80.105 Mm^3 for linear regression trained under the same conditions. Similarly, the root mean square error (RMSE) was 60.276 million m^3 for XGBoost versus 93.792 million m^3 for linear regression, reflecting a significant reduction in large forecast errors. The superiority of XGBoost was also evident in the coefficient of determination (R^2), which reached 0.673, while the linear regression model explained only 20.7% of the reservoir volume variance. In addition, the mean square error (MSE) was reduced by more than 58% in the nonlinear model. These differences highlight the ability of ensemble learning algorithms to model nonlinear relationships between remotely sensed metrics and hydrological responses. Such a significant performance gap justifies the use of advanced learning methods, especially in applications where system behavior depends on multiple overlapping physical factors. The nonlinear interactions between NDWI and LST_C and their different influences across seasons are more effectively captured by the flexible XGBoost structure compared to the rigid linear assumptions.

Metric Comparison and Interpretation

Model evaluation was conducted using four complementary metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Relative Absolute Error (RAE). These metrics were selected to jointly assess the models' accuracy, robustness to outliers, ability to capture variance, and comparative utility against naive baselines (Table 1).

MAE provides an interpretable measure of average prediction error in the same units as the target variable (million m^3). Values below 60 indicate high predictive reliability, while those above 80 reflect limited utility in decision-making contexts. RMSE, which penalizes large deviations more heavily due to its quadratic nature, further differentiates model robustness. An RMSE under 65 denotes consistent performance, whereas higher values signal instability under extreme conditions.

The R^2 statistic quantifies the proportion of variance in the observed data that is explained by the model. A value of 0.673, achieved by the best-performing XGBoost model, indicates that nearly two-thirds of the fluctuations in daily reservoir volume are captured by the predictors. In contrast, the linear model explained less than a quarter of the variance, suggesting underfitting and the inability to capture critical hydrological patterns.

Relative Absolute Error (RAE) contextualizes the MAE by comparing it to a naive model that always predicts the historical meaning. The RAE of 0.35 for XGBoost (no winter) confirms a 65% reduction in error relative to the naive baseline, while the RAE of 0.48 for linear regression reveals substantially poorer performance.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

Table 1. Performance comparison of reservoir volume prediction models across evaluation metrics

Model	RMSE	MAE	R^2	MSE
XGBoost (no winter)	60.276	52.793	0.673	3633.252
XGBoost (with winter)	61.486	54.635	0.659	3780.555
Linear Regression (with winter)	92.031	77.810	0.237	8469.707
Linear Regression (no winter)	93.792	80.105	0.207	8796.996

Notably, the exclusion of winter months improved model performance across all metrics, particularly for XGBoost. The model trained without winter data showed a decrease in MAE by 1.8 million m^3 and an increase in R^2 by 0.014, confirming the detrimental impact of cold-season NDWI anomalies. These results highlight the importance of seasonal filtering when working with optical satellite data in hydrological contexts.

Visual Analysis of Predictions

A comparison of the performance of the models with and without winter data (Figure 4).

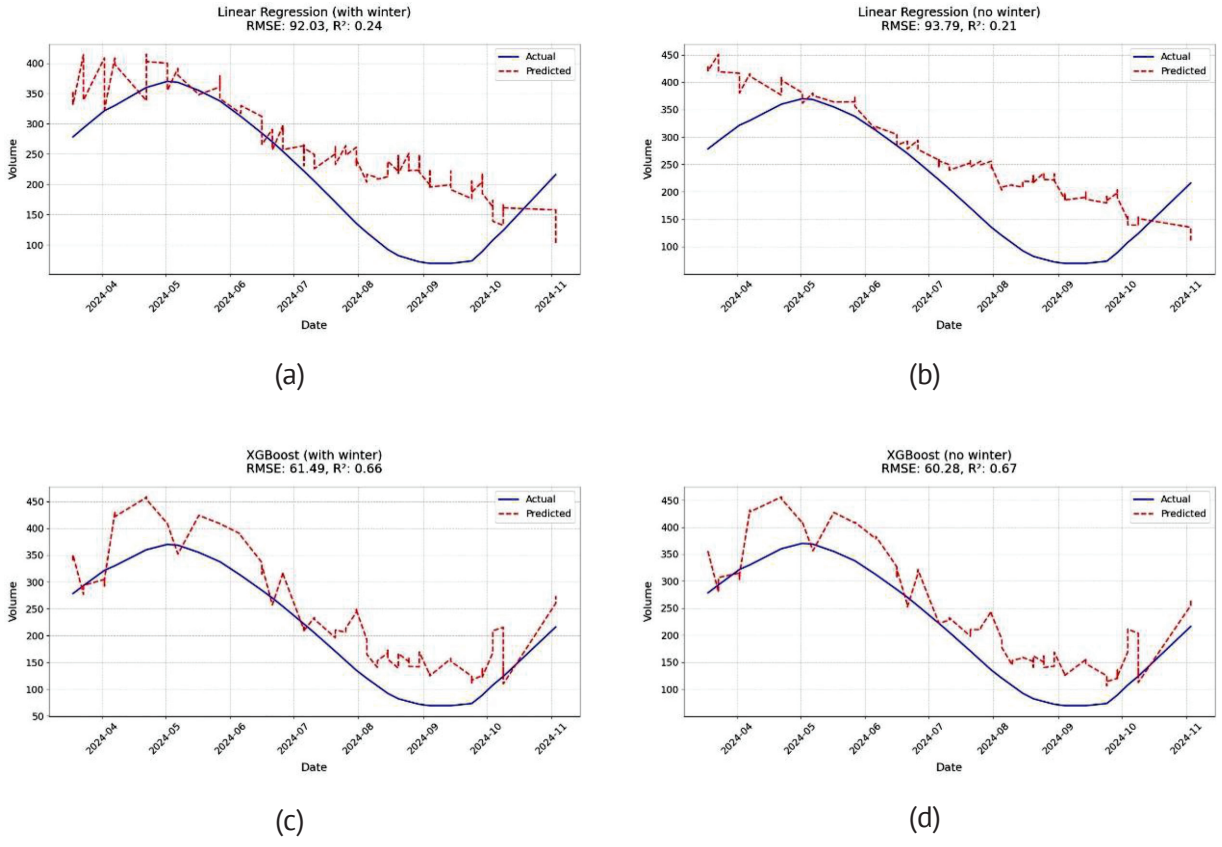

Figure 4. Performance comparison of models with and without winter data inclusion:
(a) Linear Regression (winter), (b) Linear Regression (no winter),
(c) XGBoost (winter), (d) XGBoost (no winter).

The XGBoost model trained without winter data reduced prediction errors by 1.8% (RMSE) and improved explained variance by 2.1% (R^2) compared to its winter-inclusive counterpart, indicating minimal added value from winter observations in this basin's hydrological regime.

DOI: 10.37943/22NEJN3212

Discussion

In this study developed a satellite-based machine learning model for reservoir volume forecasting and evaluated its performance across seasons. The results demonstrate both the potential and limitations of using satellite data for hydrological monitoring. Several important insights emerged from this work that can inform future research in this area.

First, it was found that careful data filtering is essential to obtain reliable forecasts. Although satellite data provide broad coverage, not all observations are equally useful for modeling. Visualizing the seasonal dynamics of NDWI (Figure 3) not only allows us to track spatiotemporal changes in the water surface but also to identify factors that distort remote sensing data, especially during the cold season. The winter decrease in NDWI, accompanied by a decrease in the water mask area, is likely due not so much to the actual decrease in water volume as to the presence of ice, snow, and low atmospheric transparency. Therefore, using winter data without correcting for spectral and atmospheric distortions may reduce the reliability of monitoring.

Spring expansion of the water surface and increase in NDWI indicate high sensitivity of the index to flood events, making it a valuable indicator for early warning systems and reservoir assessment during snowmelt periods. Summer stabilization and autumn decrease in NDWI reflect the seasonal cyclicity of the water balance. These data confirm that NDWI can act not only as an indicator of a static state, but also as an input variable in forecasting models.

The simulation results presented in Figure 4 revealed a significant impact of the quality of training data on the forecast accuracy. The model trained without winter data showed more stable behavior, better following the seasonal trajectory of water volume. This is confirmed by both the visual agreement with actual measurements and the lower variance of forecasts. At the same time, the inclusion of winter data containing spectral distortions of NDWI led to the appearance of fluctuations and erroneous peaks in the forecasts, especially during transition periods. This demonstrates the model's inability to handle seasonally compromised data and highlights the need for input feature cleaning or selection, especially in the presence of snow, ice, and cloud cover.

In terms of model performance, our comparison showed clear advantages of the XGBoost approach, especially when excluding winter data. The practical implications of this work have important implications for water resources management. Our best-performing model achieved sufficient accuracy for operational monitoring during ice-free periods, providing a cost-effective complement to ground-based measurements.

Despite the demonstrated performance of the XGBoost model, several limitations need to be considered when interpreting the results and formulating future research directions.

The model relies solely on two exogenous predictors extracted from remote sensing data. Although NDWI and LST_C are physically interpretable indicators, they do not capture the full range of hydrological factors affecting reservoir dynamics. Variables such as daily precipitation, inflow, evaporation, controlled emissions, reservoir management regimes, and other meteorological parameters were excluded due to the lack of high-quality publicly available data with sufficient temporal and spatial resolution. Another critical limitation concerns the size of the training and testing dataset. For supervised machine learning problems, especially those involving non-stationary environmental processes, this represents a relatively small sample size. However, the study shows that even simple but physically sound adjustments to the training configuration can yield significant performance gains. In particular, excluding cold season data from the training set resulted in consistent improvements in model accuracy and stability. The model without winter not only showed lower average forecast errors but also showed reduced variance and greater generalization ability. This finding highlights the importance of considering seasonal effects and the reliability of satellite input data when developing forecasting models for water resources management.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

Future work

While the initial results look promising, there is still much room for further development of the approach. Several directions seem particularly worth exploring.

First, from a feature engineering perspective, it makes sense to consider using a modified version of the Normalized Difference Water Index, specifically MNDWI. This index appears to provide better robustness to urban and vegetation noise compared to the classic NDWI. Thus, including MNDWI could improve the accuracy of water surface retrieval, especially in complex shoreline environments, which tend to be challenging. Then there is the idea of expanding the feature set by introducing more meteorological data. Another promising avenue is to integrate lagged features, such as NDWI from the previous time step or lagged land surface temperature (e.g. NDWI(t−1), LST_C(t−1)). These kinds of features help to capture the inertia in how water area and temperature evolve, which is important to better model temporal dynamics in hydrological systems. Therefore, time dependencies cannot be ignored. Given how well ensemble methods and hybrid models have performed in related hydrological problems, it seems reasonable to explore stacking models, maybe something like combining XGBoost with neural networks or plugging Random Forest into a larger ensemble. This may squeeze out some additional performance, especially in edge cases. There is also the issue of seasonality and long-term climate trends. These are non-trivial, and ignoring them may limit the quality of the forecast over longer horizons. Using tools like STL decomposition or even something simpler like a Fourier transform can help reveal these cyclic patterns.

Conclusion

This article demonstrates reservoir volumes changes over time using satellite indices, primarily NDWI and LST, obtained from Google Earth Engine, and coupled them with machine learning models built and run in Google Colab. The main focus was on the Tasotkel Reservoir, with about 250 weekly data points spanning the period from 2020 to 2024. This gave us a decent temporal resolution to track both seasonal trends and longer-term patterns.

In terms of performance, the difference between the two models was quite striking. For the full dataset, including winter, XGBoost gave us an RMSE of 61.49, an MAE of 54.63, and an R^2 of 0.659. Without winter, those numbers improved slightly, with the RMSE dropping to 60.28, the MAE to 52.79, and the R^2 rising to 0.673. So it looks like the winter data degrades the model's quality a bit, which makes sense given how ice and snow interfere with the optical and thermal readings. LST is particularly sensitive at sub-zero temperatures, and NDWI tends to level off when surface reflectivity is dominated by snow cover.

In contrast, the linear regression model really struggled. Including winter months, it showed an RMSE of 92.03, an MAE of 77.81, and a very weak R^2 of 0.237. Excluding winter did not help either: RMSE of 93.79, an MAE of 80.10, and an R^2 of 0.207. This suggests that simple linear approaches cannot truly capture the more complex, non-linear dynamics of reservoir volume change, especially under variable seasonal conditions.

Acknowledgement

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."

References

[1] Respublika Kazakhstan. (2024). Postanovlenie Pravitel'stva Respubliki Kazakhstan ot 19 yanvarya 2024 goda № 66 «Ob utverzhdenii Kontseptsii upravleniya vodnymi resursami Respubliki Kazakhstan do 2040 goda» [Resolution of the Government of the Republic of Kazakhstan No. 66 dated January 19, 2024 "On approval of the Water Resources Management Concept of the Republic of Kazakhstan until 2040"]. Adilet.zan.kz. Retrieved from https://adilet.zan.kz/rus/docs/P2400000066/

[2] Luo, T., Young, R., & Reig, P. (2015). Aqueduct projected water stress country rankings. World Resources Institute. Retrieved from https://www.wri.org/data/aqueduct-projected-water-stress-country-rankings

[3] Dankova, R., Burton, M., Salman, M., Clark, A. K., & Pek, E. (2022). Modernizing irrigation in Central Asia: Concept and approaches. Directions in Investment, No. 6. FAO and The World Bank. https://doi.org/10.4060/cb8230en

[4] Imanbayeva, Z., Abuselidze, G., Bukharbayeva, A., Jrauova, K., Oralbayeva, A., & Kushenova, M. (2024). State regulation of the digital transformation of agribusiness in the context of the climate crisis intensification. *Economies, 12*(10), 270. https://doi.org/10.3390/economies12100270

[5] Girgis, K., Clouse, N., Mushe, O., Morgan, P., Brooks, T., & Mahmoud, M. (2022). The applications of GIS in water resources. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 1515–1520). IEEE. https://doi.org/10.1109/CSCI58124.2022.00268

[6] Deng, Y., Zhang, Y., Pan, D., Yang, S. X., & Gharabaghi, B. (2024). Review of recent advances in remote sensing and machine learning methods for lake water quality management. *Remote Sensing, 16*(22), 4196. https://doi.org/10.3390/rs16224196

[7] Amireddy, R., & Dileep, P. (2024). A comparative study on water quality prediction using machine learning and deep learning techniques. In *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICDCECE60827.2024.10548555

[8] Zhang, J., Chen, W., Li, X., Wang, Y., Huang, Z., & Zhang, H. (2020). Using recurrent neural network for intelligent prediction of water level in reservoirs. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)* (pp. 1125–1126). IEEE. https://doi.org/10.1109/COMPSAC48688.2020.0-108

[9] Xu, G., Cheng, Y., Liu, F., Ping, P., & Sun, J. (2019). A water level prediction model based on ARIMA-RNN. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 221–226). IEEE. https://doi.org/10.1109/BigDataService.2019.00038

[10] Wu, J., Huang, X., Xu, N., Zhu, Q., Zorn, C., Guo, W., Wang, J., Wang, B., Shao, S., & Yu, C. (2023). Combining satellite imagery and a deep learning algorithm to retrieve the water levels of small reservoirs. *Remote Sensing, 15*(24), 5740. https://doi.org/10.3390/rs15245740

[11] Yuan, K., Zhuang, X., Schaefer, G., Feng, J., Guan, L., & Fang, H. (2021). Deep-learning-based multispectral satellite image segmentation for water body detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14*, 7422–7434. https://doi.org/10.1109/JSTARS.2021.3098678

[12] Zhang, L., & Zhang, L. (2022). Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geoscience and Remote Sensing Magazine, 10*(2), 270–294. https://doi.org/10.1109/MGRS.2022.3145854

[13] Pickens, A. H., Hansen, M. C., Hancher, M., Stehman, S. V., Tyukavina, A., Potapov, P., Marroquin, B., & Sherani, Z. (2020). Mapping and sampling to characterize global inland water dynamics from 1999 to 2018 with full Landsat time-series. Remote Sensing of Environment, 243, 111792. https://doi.org/10.1016/j.rse.2020.111792

[14] Xu, N., Ma, Y., Zhang, W., & Wang, X. H. (2021). Surface-water-level changes during 2003–2019 in Australia revealed by ICESat/ICESat-2 altimetry and Landsat imagery. *IEEE Geoscience and Remote Sensing Letters, 18*(7), 1129–1133. https://doi.org/10.1109/LGRS.2020.2996769

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

[15] Ospan, A., Mansurova, M., Kakimzhanov, E., & Aldakulov, B. (2021). KazRivDyn: Toolkit for measuring the dynamics of Kazakhstan rivers with a graphics based on Google Earth Engine. In *2021 IEEE International Conference on Smart Information Systems and Technologies (SIST)* (pp. 1–4). IEEE. https://doi.org/10.1109/SIST50301.2021.9465902

[16] Reyes-Baeza, P., Trujillo-Guiñez, R., & Vidal, M. (2023). Long-term water level forecasting for El Yeso Reservoir using time-series data and satellite images. In *2023 42nd IEEE International Conference of the Chilean Computer Science Society (SCCC)* (pp. 1–7). IEEE. https://doi.org/10.1109/SCCC59417.2023.10315730

[17] Wang, Z., Zhang, R., Zhang, Q., Zhu, Y., Huang, B., & Lu, Z. (2019). An automatic thresholding method for water body detection from SAR image. *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, 1–4. https://doi.org/10.1109/ICSIDP47821.2019.9172964

[18] Xie, C., Zhang, X., Zhang, X., Shen, S., Zhuang, L., & Chen, K. (2023). Delineation of surface water bodies from SAR imagery based on improved MRF and CNN model. In *2023 8th International Conference on Image, Vision and Computing (ICIVC)* (pp. 478–482). IEEE. https://doi.org/10.1109/ICIVC58118.2023.10270690

[19] Samanta, B., Banerjee, S., Gupta, R. K., Moulik, S., & Chaudhuri, S. S. (2024). In-depth survey of automated shoreline alteration identification methods. In *2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 61–67). IEEE. https://doi.org/10.1109/ICICV62344.2024.00016

[20] Rad, A.M., Kreitler, J., & Sadegh, M. (2021). Augmented Normalized Difference Water Index for improved surface water monitoring. Environmental Modelling & Software, 140, 105030.

[21] Reiners, P., Asam, S., Frey, C., Holzwarth, S., Bachmann, M., Sobrino, J., & Kuenzer, C. (2021). Validation of AVHRR Land Surface Temperature with MODIS and in situ LST–A timeline thematic processor. Remote Sensing, 13(17), 3473.

[22] Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv:1809.03006*.

[23] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science*, *7*, e623.

[24] Tatachar, A.V. (2021). Comparative assessment of regression models based on model evaluation metrics. *International Research Journal of Engineering and Technology (IRJET)*, *8*(09), 2395-0056.