Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

DOI: 10.37943/21EHLH9882

Almas Alzhanov
PhD student, Junior Researcher of Science and Innovation Center "Big Data and Blockchain Technologies"
almas.alzhanov01@gmail.com, orcid.org/0009-0007-8083-2366
Astana IT University, Kazakhstan
Aliya Nugumanova
PhD, Director of Science and Innovation Center "Big Data and Blockchain Technologies"
a.nugumanova@astanait.edu.kz, orcid.org/0000-0001-5522-4421
Astana IT University, Kazakhstan

FEATURE SELECTION METHODS FOR LSTM-BASED RIVER WATER LEVEL AND DISCHARGE FORECASTING

Abstract: Accurate forecasting of river discharge and water levels is essential for effective water resource management, flood mitigation, and public safety. This study compares correlation-based and PCA-based feature selection methods for LSTM forecasting models in the study area at Uba River basin, within Shemonaiha city in the East Kazakhstan region. The dataset spans from 1995 to 2021, with 1995 to 2019 used for training and validation and 2020 to 2021 for testing. Both feature selection methods reduced the original predictor set to 13 features while generally maintaining predictive accuracy. An ensemble of 10 LSTM models was trained using 60-day input sequences to forecast discharge and water levels over a 10-day horizon, reducing variance from random initialization and stabilizing predictions. Performance was evaluated using the Nash-Sutcliffe Efficiency. Results showed that correlation-based selection performed comparably to the full-feature baseline in 2020 test set, suggesting that removing highly correlated predictors did not decrease short-term forecasts capacity of the model. The model with PCA-based selected features, while slightly lagging at longer lead times in 2020, exhibited advantages in most lead times with 2021 forecasts. However, overall predictive performance declined in 2021 compared to 2020, indicating that the hydrological conditions deviate more from the historical training record, and suggesting the need for model updates with relevant historical training data. Both feature selection methods successfully reduced dimensionality, while preserving performance capacity, though neither was universally superior across all forecast lead times. These results emphasize the value of systematic feature selection in hydrological modeling and highlight the importance of model adaptability to evolving environmental conditions.

Keywords: LSTM; feature selection; water level forecasting; ERA5-Land; PCA; flood monitoring.

Introduction

Forecasting hydrological parameters such as river discharge and water level is an essential task for effective water resource management, flood monitoring and disaster prevention [1]. Accurate and timely predictions allow decision-makers to develop and implement strategies for reducing flood risks, optimizing reservoir operations, and ensuring public safety. Traditional approaches such as statistical or physically based models have been widely used in hydrological forecasting, however, they often require simplified assumptions and may fail to capture the complex, nonlinear relationships in hydrological systems [2]. With the advent of deep learning, neural networks, particularly Long Short-Term Memory or LSTM networks have emerged as powerful tools for time-series prediction [3].

Despite the widespread use of LSTM networks in hydrological forecasting [4], [5], identifying the most relevant predictors to ensure robust and efficient model performance remains a challenge. Including irrelevant or redundant features can lead to overfitting, increased computational cost, and reduced interpretability. To address these concerns, feature selection or dimensionality reduction techniques are employed to refine the input set. Both Principal Component Analysis (PCA)-based feature selection and correlation-based feature selection aim to identify key predictors by reducing the original set of variables. While PCA highlights features that contribute most to variance, correlation-based selection eliminates redundant variables by assessing their relationships, ensuring that the selected predictors are both informative and minimally correlated. A systematic comparison of these methods in the context of LSTM-based hydrological forecasting is therefore important for understanding the trade-offs in complexity, interpretability, and predictive performance.

In this study, we investigate how correlation-based and PCA-based feature selection influence LSTM forecasting of river discharge and water level. Specifically, we compare these approaches to a full-feature baseline across multiple lead times and varying hydrological conditions to assess predictive accuracy. We organize the remainder of this paper as follows. Section 2 reviews the relevant literature and presents an overview of LSTM networks along with the feature selection methods considered. Section 3 details the dataset and the experimental design, including the application of correlation-based and PCA-based approaches. Section 4 discusses the forecasting results, providing an in-depth comparison of the different feature subsets. Finally, Section 5 concludes by summarizing the main findings and offering directions for future work.

Literature review

Hydrological forecasting of discharge and water levels is crucial for water resources management and traditionally, forecasting relies on empirical or process-based models, but machine learning methods like LSTM networks have gained prominence for their ability to model complex, nonlinear temporal dependencies. LSTMs, a type of recurrent neural network, can learn long-term relationships in hydrological time series and have achieved accuracy on par with or exceeding that of conceptual hydrologic models [6], [7].

Not only have LSTMs improved accuracy, but they also enable leveraging large-scale datasets. Studies have built LSTM models using continental-scale data to predict flow in both gauged and ungauged basins [8]. By integrating static catchment attributes with dynamic inputs, regional LSTM models transfer learned hydrological behavior to ungauged locations, yielding highly accurate predictions for ungauged basins [9]. This shows that rather than relying solely on physical equations, we can learn the river response from data if we provide the right inputs.

The performance of LSTM models in hydrology is highly dependent on the input features provided. The selection and engineering of predictor variables can markedly influence model accuracy [10]. In the work [11], it is noted that determining the relevant environmental covariates like rainfall, evaporation, land use indices for streamflow modeling is essential to building accurate ML models. By excluding redundant variables, one can improve model interpretability and prevent the accuracy deterioration that occurs when irrelevant inputs introduce noise [12].

Two common feature selection approaches in hydrologic ML are correlation-based selection and PCA-based reduction. Correlation-based methods evaluate the relationship of each candidate predictor with the target or among predictors to select a subset, whereas PCA transforms the original variables into a smaller set of orthogonal components that explain most variance.

A comprehensive comparison of eight filter-based input variable selection methods for monthly streamflow forecasting was conducted in [13]. Among these methods were: Pearson correlation coefficient, partial correlation, mutual information, and gamma test. In operational forecasting contexts, correlation-based selection is often the first step. For instance, we might include only the gauges that show the highest correlation with a river’s flow when training LSTM. By doing so, the model focuses on a few strong signals rather than many weak ones. The authors of [14] took this a step further by using an attention mechanism in an LSTM to dynamically focus on informative inputs for flood forecasting.

As for the PCA, in hydrology it has been used to condense highly correlated variables, for example, readings from multiple gauge stations, or a suite of climate indices into a smaller set of components before feeding them to an LSTM. Alternatively, it can be used to assess the contribution of individual features to the PCA components. In a study on groundwater level prediction the authors of [15] employed a hybrid PCA-LSTM model to streamline the data fed into the LSTM network. Moreover, a study focusing on evaporation prediction integrated PCA with LSTM models to address issues related to correlated predictive factors [16]. The application of

PCA minimized multicollinearity among input variables, thereby enhancing the LSTM model's performance in forecasting evaporation rates. Another study [17] explored the impact of dimensionality reduction techniques on improving the generalization capability of LSTM models for streamflow prediction, and evaluated methods like PCA, Kernel PCA, t-SNE, and autoencoders. Authors of [18] applied improved Principal Component Analysis or i-PCA to reduce the dimensionality of daily and monthly rainfall data from 1901 to 2021, enhancing the predictive accuracy of a 1D-CNN for flood forecasting.

In summary, prior studies consistently demonstrate that feature selection techniques, including both correlation-based and PCA-based methods, can effectively reduce input dimensionality in LSTM models while preserving predictive accuracy. These approaches address high collinearity and complexity in hydrological data and improve model interpretability by focusing on the most informative predictors. However, the optimal selection strategy may vary across catchments, time scales, and target variables, highlighting the need for further systematic comparisons of correlation-based and PCA-based feature selection approaches for river discharge and water level forecasting.

Methods and Materials

Data collection

The study area is located in the Uba River basin, within Shemonaiha city in the East Kazakhstan region. The area experiences a continental climate where seasonal variations significantly contribute to river discharge and water level fluctuations. To capture the hydro-meteorological conditions affecting river discharge and water level, data from three principal sources covering the period 1995–2021 was collected:

- Gauging station measurements – the gauging station coordinates are 50.61°N, 81.87°E.
- Meteorological station records – the meteorological station coordinates are 50.6°N, 81.9°E.
- ERA5-Land [19] reanalysis data – extracted for a grid cell bounded by (50.55°N, 81.85°E) and (50.65°N, 81.95°E).

The visual representation of the study area is illustrated in Fig. 1.

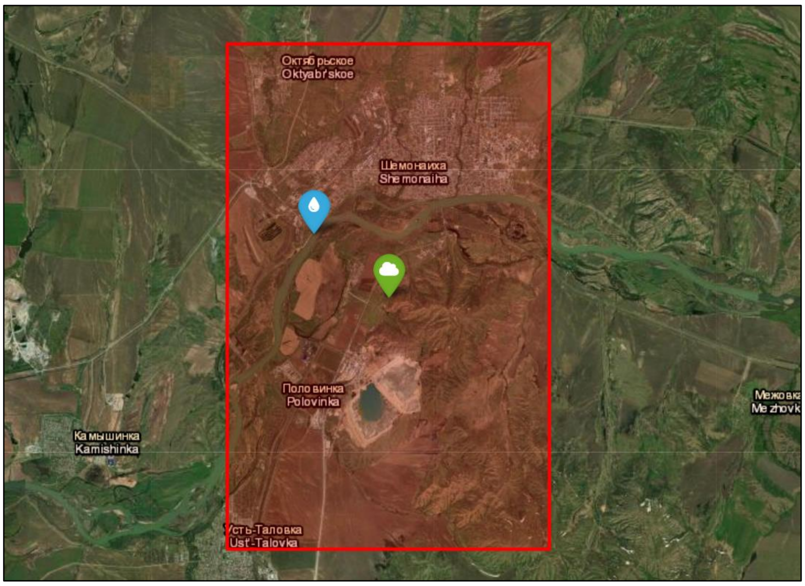

Figure 1. Visual representation of the study area, where blue marker – gauging station, green marker – meteostation, red bounding box – ERA5-Land grid cell.

Daily measurements of river discharge and water level were obtained from Kazhydromet, recorded at a gauging station located on the Uba River near Shemonaiha city. The observations of gauging station cover the 1995–2021 time period and the critical water level for this station is 430 cm. Local meteorological observations from the Shemonaiha city station were also provided by Kazhydromet. Although records date back to 1960, this study uses data from 1995–2021 to match the gauging station time span. The station measurements include air temperature (mean, maximum, minimum), relative humidity (mean, minimum), precipitation and soil surface temperature (mean, maximum, minimum).

ERA5-Land, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), is a global reanalysis product that offers daily and sub-daily meteorological variables at approximately 9 km resolution [19]. The hydrometeorological features were extracted from ERA5-Land using Google Earth Engine [20], a platform for geospatial analysis with the following steps:

- A polygon was defined to represent the bounding coordinates of ERA5-Land grid cell area, which cover both gauging station and meteorological station. This region of interest ensures that the extracted ERA5-Land data reflects the local hydro-meteorological conditions.
- Google Earth Engine was queried for ERA5-Land data and filtered by the time span of gauging station and by the region of interest. In total, 24 variables were selected to capture key meteorological and land-surface processes, including: air temperature (2 m, skin temperature), soil temperature and moisture (multiple depths), snow parameters (cover, density, depth, water equivalent, snowmelt), hydrological fluxes (precipitation, runoff, evaporation), other indicators (dew point temperature, lake temperature, skin reservoir content).
- The resulting daily time series of selected variables was exported from Google Earth Engine as a CSV file.

All datasets were synchronized to construct a time series of hydrometeorological features for LSTM-based forecasting. After collection and preprocessing, the final dataset comprised of 9862 rows and 36 columns including the date column.

Correlation-based feature selection

A straightforward yet effective approach to feature selection is to exclude predictors that exhibit strong pairwise correlations. High correlations among features often indicate redundancy, as multiple variables may be conveying essentially the same information. Retaining all such redundant features can unnecessarily increase model complexity, reduce interpretability, and potentially lead to overfitting. In this work, we computed a Pearson correlation matrix for all candidate predictors (Fig. 2).

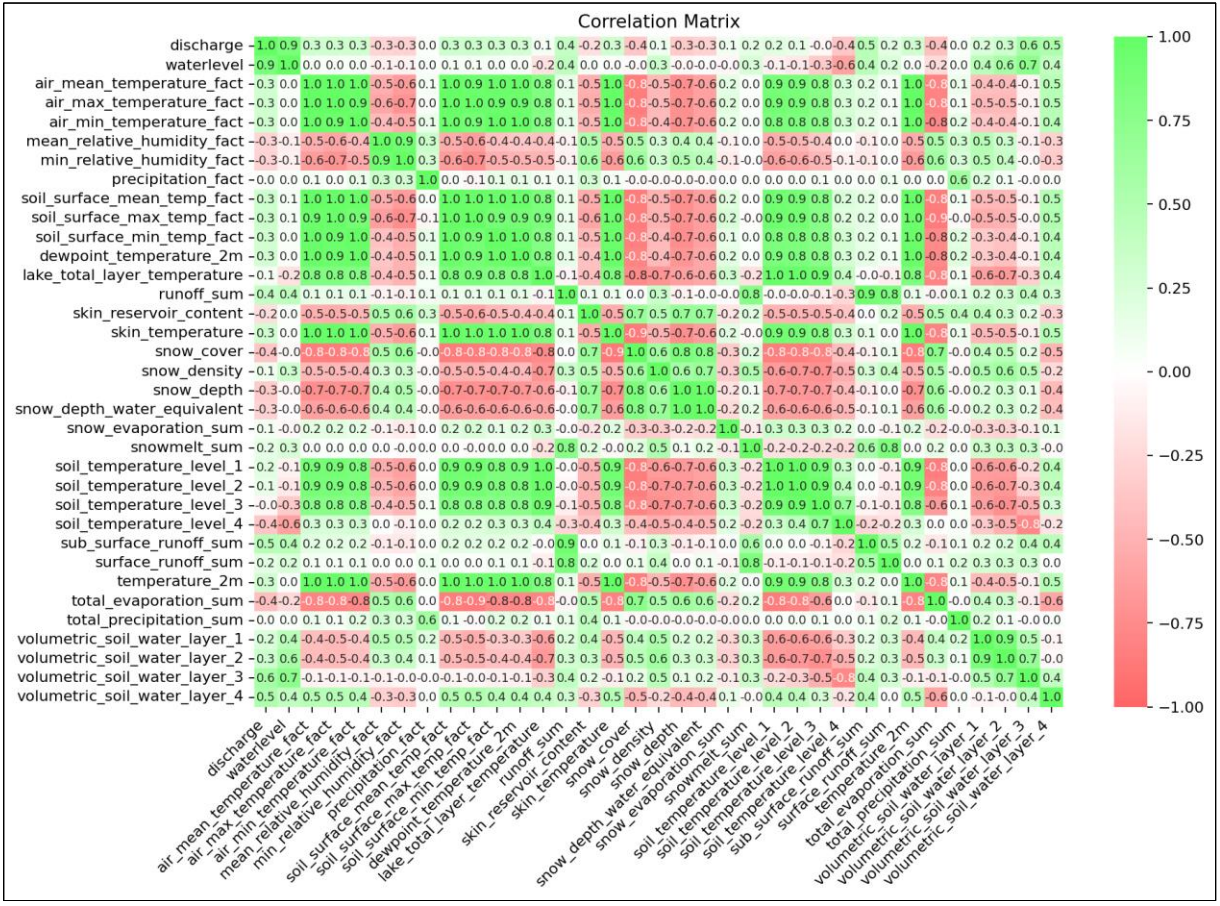

Figure 2. Pearson correlation matrix for all candidate predictors.

Any two features with an absolute correlation value exceeding a predetermined threshold were flagged as ‘highly correlated’. From each pair of highly correlated features, one was removed from the feature set. The correlation threshold is set to 0.7, although the choice is heuristic, it is widely used across various domains, as correlations above this value are typically considered strong [21][22]. Because discharge and water level are our primary target variables for LSTM forecasting and we aim to forecast their future values based on their past observations, we explicitly protected them from removal even if they appeared in a highly correlated pair.

By applying this threshold-based rule, we reduced our original feature set to 13 retained predictors and removed 22 redundant features. The final subset includes both observed hydrometeorological variables such as discharge, water level, mean air temperature, mean relative humidity, and precipitation, as well as reanalysis based variables: runoff sum, skin reservoir content, snow density, snow evaporation sum, soil temperature at 100 to 289 cm depth, total precipitation sum, volumetric soil water at 0 to 7 cm depth, and volumetric soil water at 100 to 289 cm depth.

PCA-based feature selection

PCA is a widely used dimensionality reduction technique that transforms a set of possibly correlated variables into a smaller number of uncorrelated variables called principal components. Unlike correlation-based methods, which focus on pairwise relationships between features, PCA identifies directions in the feature space that captures the maximum variance in the data. While PCA is traditionally used for projection onto a reduced subspace, it can also assist in feature selection by highlighting variables that contribute most strongly to the principal components with the highest variance. Feature selection using PCA was performed through the following steps:

- Data standardization
- Principal components computation

- Feature contribution analysis
- Ranking and Selection

PCA was applied to compute the principal components of the dataset. Each principal component represents a linear combination of the original features, with coefficients indicating their contribution. Before applying PCA, all numerical features excluding the date variable were standardized. Standardizing ensures that variables measured on different scales do not disproportionately influence the principal components due to their larger numeric range.

Parallel analysis was used to determine how many principal components to retain [23]. In this approach, random datasets matching the dimensions of the actual data were generated, and the eigenvalues of their correlation matrices were computed. The eigenvalues from the real dataset were compared against the 95th percentile of the random eigenvalue distribution, retaining only those components with eigenvalues exceeding this threshold. This ensures that retained components capture meaningful variance rather than random noise. In Fig. 3 the choice of the number of components is illustrated:

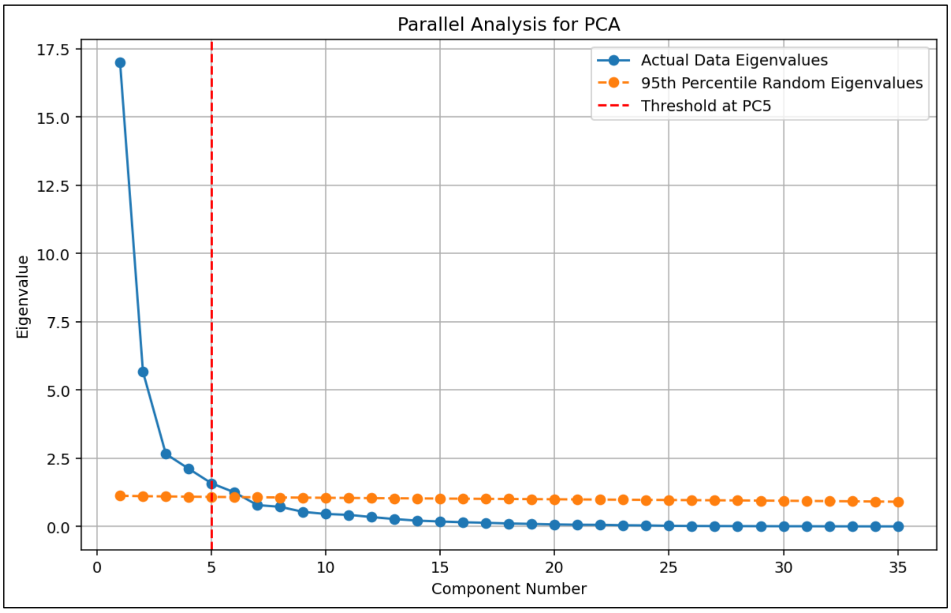

Figure 3. Parallel analysis to choose the number of components.

To identify the most influential features, the loadings of each feature on the principal components were examined. The absolute values of these loadings were computed and weighted by the fraction of variance explained by each principal component. This approach ensures that components with greater explanatory power contribute more to the feature selection process. The total weighted contribution of each feature across the first five principal components was aggregated, and features were ranked accordingly (Fig. 4).

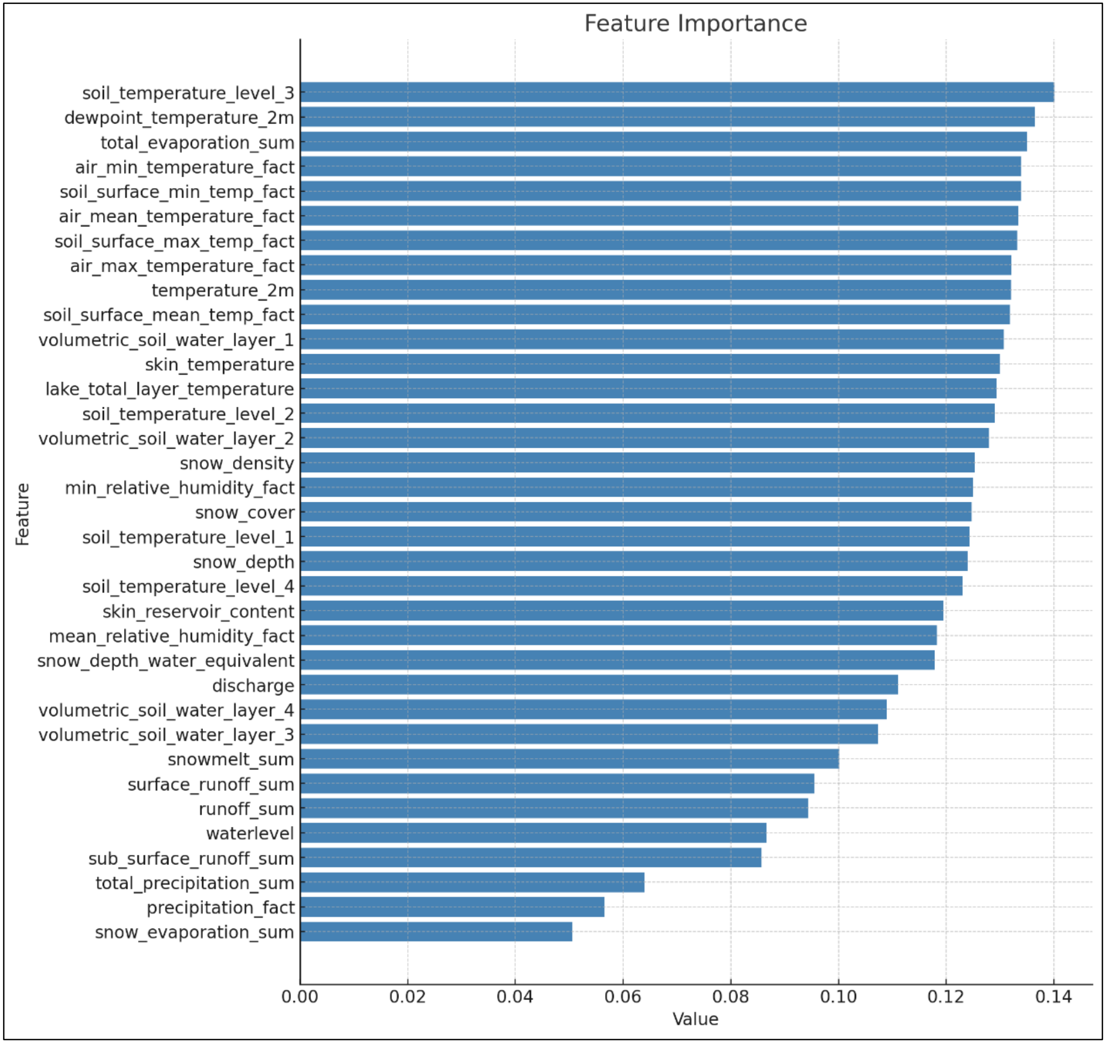

Figure 4. Feature importance by weighted contribution across the first five principal components.

The top 11 features from this ranking were selected. Additionally, we manually retained two target variables, discharge and water level, resulting in a final subset of 13 features. Retained features list includes observed variables like discharge, water level, air temperature (mean, min, max), soil surface temperature (mean, min, max), as well as reanalysis variables like soil temperature at 28-100 cm depth, dewpoint temperature at 2m, total evaporation sum, temperature at 2m, volumetric soil water at 0 - 7 cm.

Model training and testing

A series of LSTM neural networks was employed to forecast future river discharge and water level. LSTMs are particularly well suited for sequence modeling tasks with temporal dependencies and have demonstrated effectiveness in hydrological forecasting applications. To further enhance robustness, an ensemble of LSTMs was trained for each feature set, as ensemble have been shown to reduce variance produced by random initialization of weights and stochastic nature of optimizers, while improving predictive performance compared to a single model. Experiments were conducted using Google Colab, which provides a cloud-based environment with access to an NVIDIA Tesla T4 GPU.

All networks in the ensemble (Fig. 5) followed the same architecture, each model contained a single LSTM layer of 32 hidden units, followed by a dense layer outputting predictions for multiple future time steps for discharge and water level simultaneously. The LSTM layer employs a hyperbolic tangent activation function, while the dense layer is linear. The models were optimized via the mean squared error loss function and the Adam optimizer with learning rate set to 0.001.

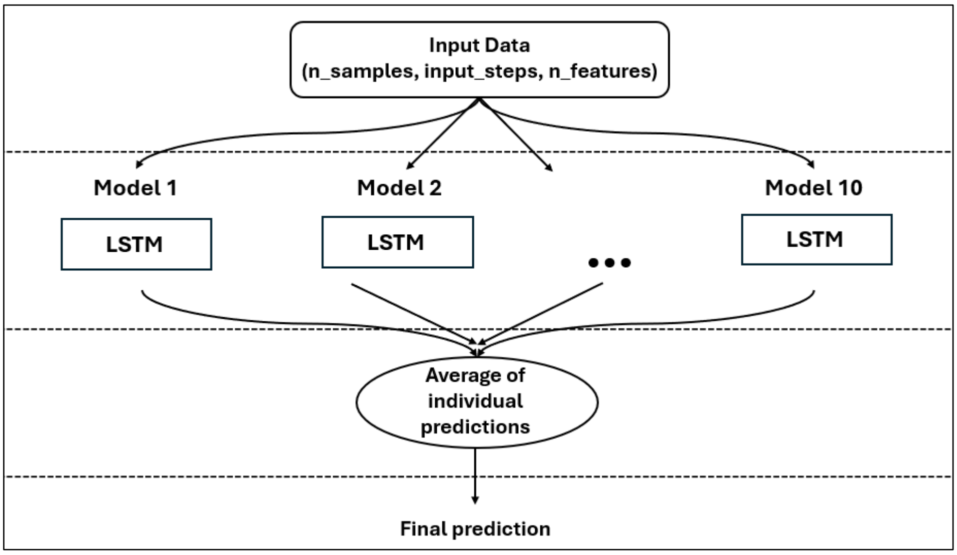
Input Data
(n_samples, input_steps, n_features)
Model 1
Model 2
Model 10
LSTM
LSTM
LSTM
•••
Average of individual predictions
Final prediction

Figure 5. Ensemble of LSTM models for discharge and water level prediction.

As a preprocessing step for the training of LSTMs, time series were organized into overlapping windows of 60 input time steps for each model, predicting the subsequent 10 time steps for both discharge and water level. For standardization, minimum and maximum values were computed on the training set, then applied to the validation and test sets to avoid data leakage. The full dataset was split by calendar years, retaining years not targeted for testing as the combined training and validation set, and reserving specific years which are 2020 and 2021 as independent test sets. The details of the training, validation, and test set splits are as follows:

- Training set (X_train): 8,155 samples of shape (60, num_features).
- Validation set (X_val): 907 samples of shape (60, num_features).
- Test set 2020 (X_test_2020): 297 samples of shape (60, num_features).
- Test set 2021 (X_test_2021): 296 samples of shape (60, num_features).

The number of features varies depending on the experimental setup:

- Full feature set: 35 features
- Reduced feature set: 13 features

For each of the three feature sets, 10 separate LSTM models were trained. Training was performed for 20 epochs with a batch size of 32. To evaluate performance, the models were applied to the two separate test years, and ensemble statistics were derived by aggregating predictions across the 10 trained models.

Evaluation metrics

To assess the performance of the time series discharge and water level forecasting models, we use the Nash-Sutcliffe Efficiency or NSE as the primary evaluation metric. NSE is widely used in hydrological modeling to measure how well simulated values match observed data. The NSE metric quantifies the predictive accuracy of a model by comparing its performance to that of a simple mean predictor. It is defined as:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Q_i^{obs} - Q_i^{sim})^2}{\sum_{i=1}^{n}(Q_i^{obs} - \overline{Q}^{obs})^2} \qquad (1)$$

where Q_i^{obs} is the observed discharge or water level at time step i, Q_i^{sim} is simulated discharge or water level, $\overline{Q}^{obs}$ is the mean of observed values, and n is the number of observations.

In the context of this study, higher NSE values indicate more accurate forecasts of river discharge or water level. Lead time refers to the time interval between when a forecast is made and when the predicted event occurs. We compute NSE separately for each lead time, specifically 1 through 10 days to reveal how predictive capacity evolves or reduces as the forecast horizon increases. Negative NSE values signify that the model predictions are less reliable than simply using the average of past observations. This makes NSE a robust and interpretable criterion for evaluating time-series forecasts in hydrological applications, as it directly relates model performance to baseline expectations derived from observed data variance.

Results

In this section, we present the results of our analysis of LSTM based river discharge and water level forecasts using different feature selection methods, including the full feature set, correlation reduced, and PCA reduced approaches, focusing on how these methods influence predictive performance and capture hydrological variability in the 2020 and 2021 test periods.

Fig. 6 presents the NSE scores for discharge and water level forecasts during 2020 test period at lead times of 1–10 days. In both discharge and water level predictions, NSE remains higher than 0.85 for short lead times of 1–3 days across all three models. Notably, the correlation-reduced model remains on par with the full-feature model for most lead times, indicating that removing highly collinear variables does not substantially degrade performance in 2020. The PCA-reduced model performs comparably at shorter horizons but begins to lag at longer lead times beyond day 4, indicating that certain event-specific factors may have been lost during the dimensionality reduction process.

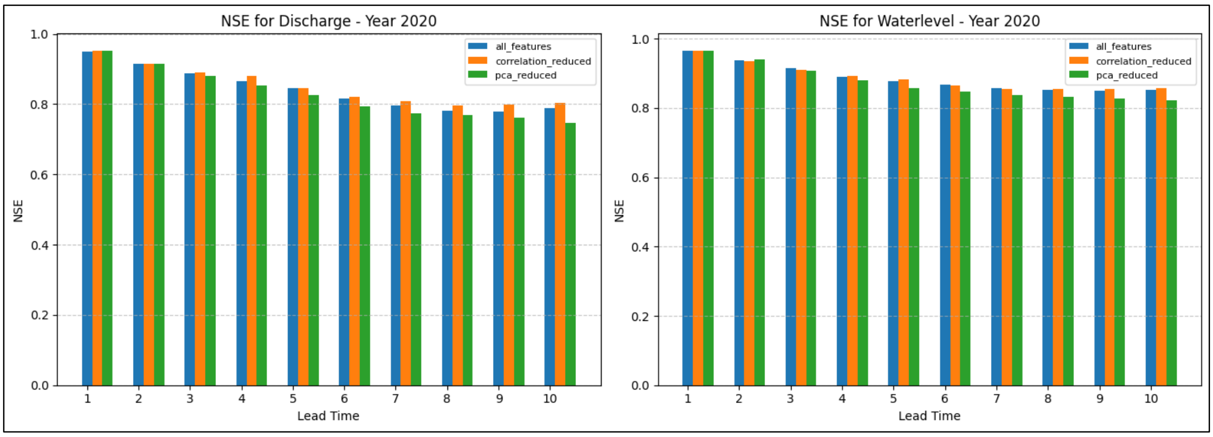

Figure 6. NSE scores comparison for discharge and water level forecasts for 2020 test year.

To illustrate model behavior on a single 10-day forecast sequence in 2020, Fig. 7 shows predicted discharge and water-level profiles starting on 2020-04-09. All three models produce similar trajectories, and the correlation-reduced forecasts align closely with the full-feature baseline. Under the hydrological conditions observed in 2020, it is demonstrated that feature selection approaches can optimize inputs without considerably compromising performance.

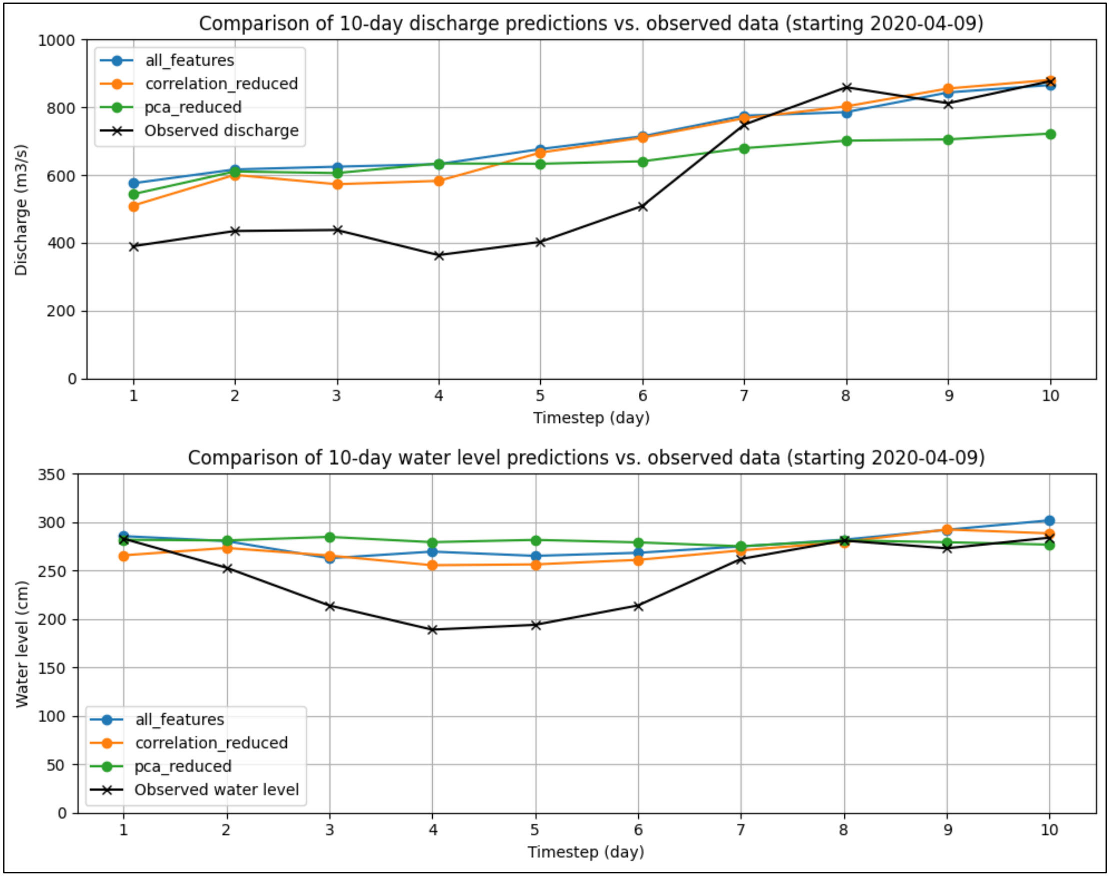

Figure 7. Comparison of 10-day discharge and water level predictions starting 2020-04-09.

Fig. 8 compares NSE scores for discharge and water level in 2021, revealing more pronounced model differences compared to 2020. The PCA-reduced model sometimes achieves equal or higher NSE than the other approaches at most lead times, while the correlation-reduced model experiences a noticeable drop at longer horizons. This suggests that PCA-based selection may isolate certain features that prove advantageous in specific flow regimes. Day-by-day forecasts provide further insight.

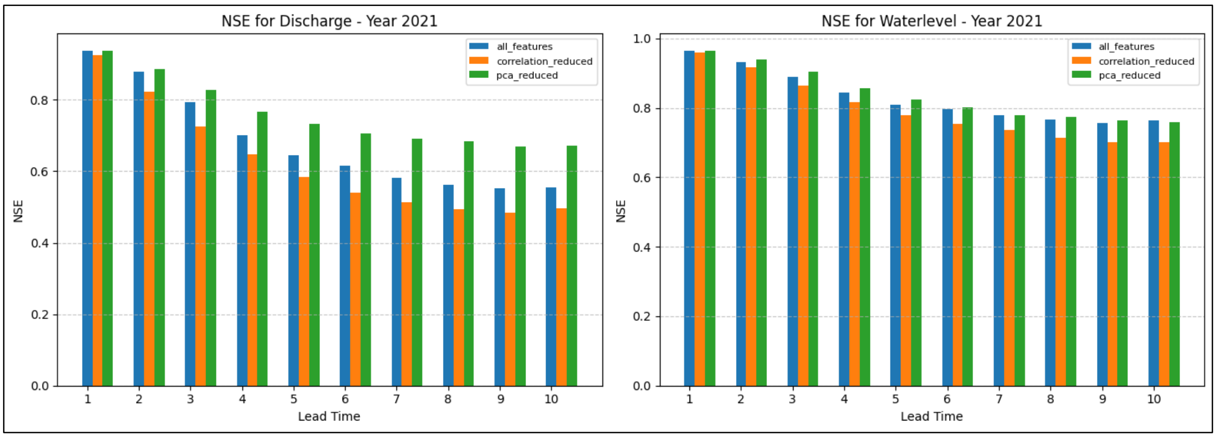

Figure 8. NSE scores comparison for discharge and water level forecasts for 2021 test year.

Fig. 9 presents a 10-day forecast window beginning on 2021-04-09. Here, the PCA-reduced model underestimates discharge for much of the sequence, even though it achieves high NSE overall when averaged across the entire year which highlights an aggregate vs. instance discrepancy: PCA excels on average, yet it can falter in specific short-term events. In contrast, the correlation-reduced model overestimates the peak, but variates lesser from observed values.

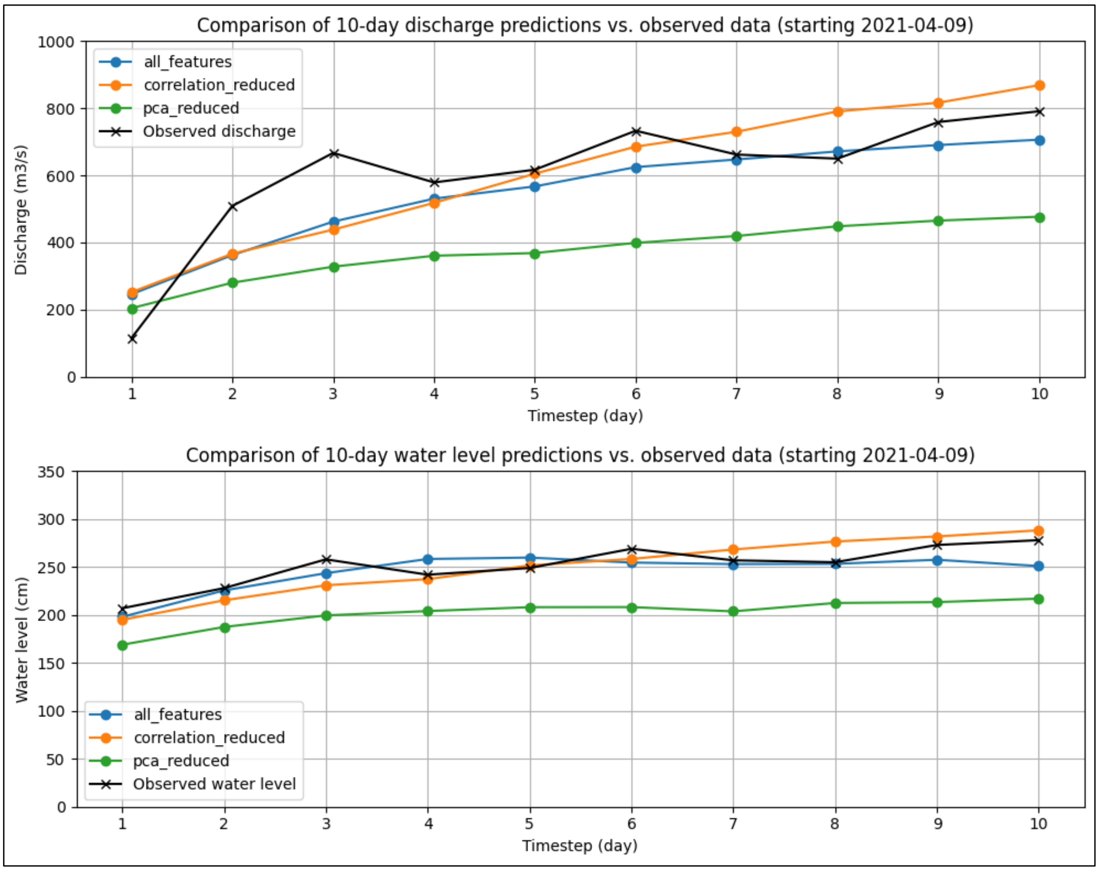

Figure 9. Comparison of 10-day discharge and water level predictions starting 2021-04-09.

Despite using the same 1995–2019 historical training set, forecast performance is generally lower in 2021 than in 2020 for all models. Even though 2020 was also excluded from training, its conditions may have been closer to the multi-decadal average, allowing the models to generalize more effectively which emphasize the need for ongoing retraining or adaptive techniques to maintain alignment with evolving hydroclimatic conditions.

Conclusion

In this study, we investigated the effects of correlation-based and PCA-based feature selection on LSTM forecasts of river discharge and water level. Both approaches substantially reduced the dimensionality of the original dataset, yielding input sets that generally preserved predictive accuracy across multiple lead times. In particular, the correlation-reduced model performed on par with the full-feature baseline in 2020 test set, indicating that discarding highly collinear predictors can simplify the model without sacrificing short-term forecast capacity. By contrast, the PCA-reduced model showed distinct advantages in certain 2021 test set forecasts, highlighting that capturing dominant variance structures through principal components could benefit under specific hydrological conditions.

Notably, we observed performance discrepancies between 2020 test set and 2021 test set, while the training data covered 1995–2019. The models attained higher NSE scores in 2020, suggesting that its hydro-climatological conditions more closely resembled the historical record used for training. Meanwhile, the somewhat lower performance in 2021 points to the potential non-stationary environmental factors and distribution shifts over time, emphasizing the need for continuous or adaptive model updates.

Overall, both correlation-based and PCA-based methods proved viable for feature reduction, reducing model complexity and potentially mitigating overfitting. However, neither approach was universally superior at every lead time or under all flow conditions. Future work could explore hybrid techniques that combine correlation filtering with principal component analysis or investigate data-driven feature selection methods like attention mechanisms to further refine predictor sets for the specific task of river discharge and water level forecasting. Additionally, alternative feature selection approaches such as Mutual Information and Recursive Feature Elimination could be explored to improve predictor selection and improve performance of deep learning models. Mutual Information, by capturing non-linear dependencies between variables, could identify hydrological features that provide the most distinct information. Recursive Feature Elimination, by systematically ranking and removing less relevant predictors, could offer a more structured way to optimize feature selection.

Furthermore, adaptive or continuous retraining, incorporating newly observed data, may enhance resilience against evolving hydroclimatic regimes and improve long-term forecast accuracy.

Acknowledgment

This work has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant Number BR24992899).

References

[1] Depetris, P. J. (2021). The importance of monitoring river water discharge. Frontiers in Water, 3, 745912. doi.org/10.3389/frwa.2021.745912.

[2] Özdoğan-Sarıkoç, G., & Dadaser-Celik, F. (2024). Physically based vs. data-driven models for streamflow and reservoir volume prediction at a data-scarce semi-arid basin. Environmental Science and Pollution Research, 1-22. doi.org/10.1007/s11356-024-33732-w.

[3] Hunt, K. M., Matthews, G. R., Pappenberger, F., & Prudhomme, C. (2022). Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. Hydrology and Earth System Sciences, 26(21), 5449-5472. doi.org/10.5194/hess-26-5449-2022.

[4] Anshuka, A., Chandra, R., Buzacott, A. J., Sanderson, D., & van Ogtrop, F. F. (2022). Spatio temporal hydrological extreme forecasting framework using LSTM deep learning model. Stochastic environmental research and risk assessment, 36(10), 3467-3485. doi.org/10.1007/s00477-022-02204-3.

[5] Kim, G. B., Hwang, C. I., & Choi, M. R. (2021). PCA-based multivariate LSTM model for predicting natural groundwater level variations in a time-series record affected by anthropogenic factors. Environmental Earth Sciences, 80(18), 657. doi.org/10.1007/s12665-021-09957-0.

[6] Nifa, K., Boudhar, A., Ouatiki, H., Elyoussfi, H., Bargam, B., & Chehbouni, A. (2023). Deep learning approach with LSTM for daily streamflow prediction in a semi-arid area: a case study of Oum Er-Rbia river basin, Morocco. Water, 15(2), 262. doi.org/10.3390/w15020262.

[7] Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. Hydrology and Earth System Sciences, 22(11), 6005-6022. doi.org/10.5194/hess-22-6005-2018.

[8] Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., ... & Matias, Y. (2024). Global prediction of extreme floods in ungauged watersheds. Nature, 627(8004), 559-563. doi.org/10.1038/s41586-024-07145-1.

[9] Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning

applied to large-sample datasets. Hydrology and Earth System Sciences, 23(12), 5089-5110. doi.org/10.5194/hess-23-5089-2019.

[10] Fang, W., Ren, K., Liu, T., Shang, J., Jia, S., Jiang, X., & Zhang, J. (2024). An evaluation of random forest based input variable selection methods for one month ahead streamflow forecasting. Scientific Reports, 14(1), 29766. doi.org/10.1038/s41598-024-81502-y.

[11] Reis, G. B., da Silva, D. D., Fernandes Filho, E. I., Moreira, M. C., Veloso, G. V., de Souza Fraga, M., & Pinheiro, S. A. R. (2021). Effect of environmental covariable selection in the hydrological modeling using machine learning models to predict daily streamflow. Journal of Environmental Management, 290, 112625. doi.org/10.1016/j.jenvman.2021.112625.

[12] Galelli, S., Humphrey, G. B., Maier, H. R., Castelletti, A., Dandy, G. C., & Gibbs, M. S. (2014). An evaluation framework for input variable selection algorithms for environmental data-driven models. Environmental Modelling & Software, 62, 33-51. doi.org/10.1016/j.envsoft.2014.08.015.

[13] Ren, K., Fang, W., Qu, J., Zhang, X., & Shi, X. (2020). Comparison of eight filter-based feature selection methods for monthly streamflow forecasting–three case studies on CAMELS data sets. Journal of Hydrology, 586, 124897. doi.org/10.1016/j.jhydrol.2020.124897.

[14] Liu, Z., Xu, W., Feng, J., Palaiahnakote, S., & Lu, T. (2018, August). Context-aware attention LSTM network for flood prediction. In 2018 24th international conference on pattern recognition (ICPR) (pp. 1301-1306). IEEE. doi.org/10.1109/ICPR.2018.8545385.

[15] Thakur, A., Chandel, A., & Shankar, V. (2025). Prediction of groundwater levels using a long short-term memory (LSTM) technique. Journal of Hydroinformatics, 27(1), 51-68. doi.org/10.2166/hydro.2024.239.

[16] Wang, C., Li, T., Xin, D., Wang, Q., Chen, R., & Cao, C. (2023). Pan Evaporation Prediction Using LSTM Models Based on PCA Factor Reduction and Firefly Optimization Algorithm. IEEE Journal on Miniaturization for Air and Space Systems. doi.org/10.1109/JMASS.2023.3319579.

[17] Ghobadi, F., Tayerani Charmchi, A. S., & Kang, D. (2023). Feature extraction from satellite-derived hydroclimate data: Assessing impacts on various neural networks for multi-step ahead streamflow prediction. Sustainability, 15(22), 15761. doi.org/10.3390/su152215761.

[18] John, T. J., & Nagaraj, R. (2023). Prediction of floods using improved pca with one-dimensional convolutional neural network. International Journal of Intelligent Networks, 4, 122-129. doi.org/10.1016/j.ijin.2023.05.004.

[19] Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., ... & Thépaut, J. N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth system science data, 13(9), 4349-4383. doi.org/10.5194/essd-13-4349-2021.

[20] Amani, M., Ghorbanian, A., Ahmadi, S. A., Kakooei, M., Moghimi, A., Mirmazloumi, S. M., ... & Brisco, B. (2020). Google earth engine cloud computing platform for remote sensing big data applications: A comprehensive review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 5326-5350. doi.org/10.1109/JSTARS.2020.3021052.

[21] McMillan, H. K., Gnann, S. J., & Araki, R. (2022). Large scale evaluation of relationships between hydrologic signatures and processes. Water Resources Research, 58(6), e2021WR031751. doi.org/10.1029/2021WR031751.

[22] Safeeq, M., Mauger, G. S., Grant, G. E., Arismendi, I., Hamlet, A. F., & Lee, S. Y. (2014). Comparing large-scale hydrological model predictions with observed streamflow in the Pacific

Northwest: Effects of climate and groundwater. Journal of Hydrometeorology, 15(6), 2501-2521. doi.org/10.1175/JHM-D-13-0198.1.

[23] Çokluk Bökeoğlu, Ö., & Koçak, D. (2016). Using Horn's parallel analysis method in exploratory factor analysis for determining the number of factors. Educational sciences-theory & practice, 16(2). doi.org/10.12738/estp.2016.2.0328.