Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 01 May 2026
Revised 12 June 2026
Accepted 30 June 2026

Natural Resources for Human Health 2026, Vol. 6 (7s): 657–669

KEYWORDS: Transformer neural networks, spatiotemporal forecasting, natural resource availability, attention mechanism, water resources management, groundwater prediction, deep learning

Transformer-based Deep Learning for Spatiotemporal Forecasting of Natural Resource Availability

Lalita[1], Apoorva Srivastava[2], Shaeista Begum[3], Bakeeru Mery Sowjanya[4], Yogesh Kumar Sharma[5], Devesh Kumar Srivastava[6*]

[1]Department of Computer Science and Engineering, Panipat Institute of Engineering and Technology, Samalkha, Panipat, Haryana, India, Email: Sharma.lalita1701@gmail.com
[2]Department of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India, Email: apoorvasjd625@gmail.com
[3]Department of Computer Science and Engineering, Government Engineering College, Huvinahadagali,
Karnataka, India, Email: shaeista.begum@gmail.com
[4]Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India, Email: bakeerusowji@gmail.com
[5]Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India, Email: dr.sharmayogeshkumar@gmail.com
[6]School of Computer Science and Engineering, Manipal University Jaipur, Jaipur, Rajasthan, India
Email: devesh.srivastava@jaipur.manipal.edu
*Corresponding Author

ABSTRACT: Reliable forecasting of natural resource availability including reservoir storage, groundwater level, and streamflow is central to climate-resilient water management, yet conventional statistical and recurrent deep learning models struggle to jointly capture long-range temporal dependencies and heterogeneous spatial coupling across a basin. This study proposes ST-ResFormer, a spatiotemporal transformer architecture that fuses a graph-attention spatial encoder with a patch-based sparse-attention temporal encoder through a cross-attention fusion block, enabling joint multi-horizon forecasting of three coupled resource variables from heterogeneous data sources (gauge records, satellite soil moisture, GRACE terrestrial water storage anomalies, and reanalysis climate forcing). The framework was evaluated on a 22-year (2001–2023) multi-source dataset covering 48 grid cells of the Sutlej Beas River basin in the northwestern Himalaya, benchmarked against eight baselines spanning classical, recurrent, convolutional, graph-based, and transformer-based forecasters (Persistence, SARIMA, LSTM, ConvLSTM, GCN-LSTM, Informer [2], Autoformer [3], and PatchTST [7]). At a one-month forecast horizon, ST-ResFormer achieved a root-mean-square error (RMSE) of 4.63 and mean absolute error (MAE) of 3.41 in normalized units, corresponding to a 27.3% RMSE reduction relative to the strongest transformer baseline (PatchTST) and a 68.8% reduction relative to Persistence, while sustaining a Nash–Sutcliffe efficiency (NSE) above 0.91 out to a six-month horizon. These results demonstrate that jointly modeling spatial and temporal attention over multi-source resource data yields state-of-the-art, interpretable, and operationally deployable forecasts, offering a transferable

template for climate-adaptive natural resource management in data-heterogeneous mountain basins.

1. INTRODUCTION

Natural resource availability encompassing surface water storage, groundwater reserves, and streamflow is under mounting pressure from climate variability, glacier and snowpack retreat, and intensifying anthropogenic withdrawal, particularly in mountainous headwater basins that sustain downstream agriculture, hydropower, and municipal supply. Accurate multi-horizon forecasting of these coupled resource variables is a prerequisite for reservoir operation, conjunctive groundwater–surface water management, drought early warning, and climate adaptation planning. However, resource availability is governed by processes that are simultaneously long-range in time (seasonal snowmelt, multi-year drought persistence, aquifer recharge lags) and heterogeneous in space (elevation-dependent precipitation, sub-basin connectivity, and spatially varying land use and soil properties), a combination that classical time-series models and even many deep sequence models handle only partially.

Autoregressive integrated moving-average (ARIMA/SARIMA) models and shallow machine learning regressors remain widely used operationally because of their interpretability and low data requirements, but they assume stationarity or rely on hand-engineered lag features and therefore degrade rapidly as forecast horizons lengthen or as nonlinear climate–resource interactions intensify. Recurrent architectures such as long short-term memory (LSTM) networks [20] and their convolutional extension ConvLSTM [23] improved on this by learning nonlinear temporal dynamics directly from data, and graph-based recurrent and convolutional networks [21,22] extended this capability to explicitly encode spatial connectivity between monitoring locations or sub-catchments. Yet recurrent models process sequences step-by-step, which limits parallelization, causes gradient decay over long horizons, and constrains their ability to attend jointly to distant but relevant time steps for example, a precipitation anomaly six months prior that determines present reservoir inflow.

The transformer architecture [1], originally developed for machine translation, replaces recurrence with self-attention, allowing every time step to directly attend to every other time step in a sequence with full parallelism. This property has driven rapid adoption of transformer variants for long-sequence time-series forecasting, including Informer's ProbSparse attention for O(L log L) complexity [2], Autoformer's decomposition and auto-correlation mechanism for trend–seasonal disentanglement [3], FEDformer's frequency-domain enhancement [4], Pyraformer's pyramidal attention for multi-scale dependencies [5], Crossformer's explicit cross-dimension attention for multivariate series [6], and PatchTST's patch-based tokenization that substantially improved sample efficiency and long-horizon accuracy [7]. In parallel, vision transformers [24] demonstrated that attention-based architectures generalize effectively to spatially structured grid data, motivating hybrid spatiotemporal transformer designs such as graph-attention-augmented transformers for traffic forecasting [8] and transformer models for spatiotemporal microseismicity prediction [9] that jointly encode spatial topology and temporal dynamics within a single attention framework.

Within hydrology and water resources specifically, transformer-based forecasters have recently been applied to groundwater level estimation [10–14], streamflow and water-level prediction [15,19], and flood forecasting [16,18], generally reporting accuracy gains of 10–20% in RMSE over LSTM and classical baselines, with PatchTST-style patch attention proving particularly effective for hydrological series with strong seasonal periodicity [10]. However, the majority of these studies target a single resource variable at a single location or a small set of gauges, treating spatial context, if at all, through simple feature concatenation rather than learned spatial attention, and few studies jointly forecast multiple coupled resource variables (e.g., reservoir storage and groundwater) that share common climatic drivers but differ in response time-scale.

This study addresses that gap by proposing ST-ResFormer, a spatiotemporal transformer that (i) fuses heterogeneous multi-source inputs in-situ gauge records, satellite-derived soil moisture, GRACE terrestrial water storage anomalies, reanalysis climate forcing, and static physiographic covariates at the grid-cell level; (ii) applies a graph-attention spatial encoder to learn dynamic inter-cell dependencies without a fixed adjacency assumption; (iii) applies a patch-based sparse-attention temporal encoder to capture both short-term fluctuations and long-range seasonal-to-interannual dependencies; and (iv) couples the two through a spatiotemporal cross-attention fusion block that produces joint multi-horizon, multi-variable forecasts of reservoir storage, groundwater level, and streamflow.

2. RELATED WORK

2.1 Classical and Recurrent Approaches to Resource Forecasting

Statistical time-series models, most notably ARIMA/SARIMA and its seasonal and exogenous-variable extensions, have long served as operational baselines for streamflow, reservoir inflow, and groundwater-level forecasting because of their transparency and modest data requirements. These models, however, assume linear dynamics and (weak) stationarity, both of which are routinely violated by resource variables driven by nonlinear precipitation–runoff processes, snowmelt thresholds, and anthropogenic abstraction. The introduction of LSTM networks [20] substantially improved nonlinear temporal modeling by using gated memory cells to mitigate vanishing gradients over moderately long sequences, and LSTM-family models remain among the most widely benchmarked deep learning baselines in hydrology [17]. ConvLSTM [23] further extended this by embedding convolutional operators inside the recurrent cell, allowing simultaneous modeling of local spatial structure (e.g., a moving precipitation field) and temporal evolution, and has been influential in precipitation nowcasting and gridded hydrological prediction. Nonetheless, recurrent architectures process time steps sequentially, which constrains parallel training, causes information decay over long horizons, and requires careful tuning of sequence length to balance context against optimization difficulty.

2.2 Graph-Based Spatial Modeling

To move beyond fixed-grid convolution, graph neural networks represent spatial units (gauges, sub-basins, or grid cells) as nodes in a graph and learn message-passing operators over edges that encode physical or statistical connectivity. Graph convolutional networks [21] aggregate neighbor information through spectral or spatial convolution, while graph attention networks [22] replace fixed aggregation weights with learned attention coefficients, allowing the model to weight neighboring nodes according to their present relevance rather than a static adjacency matrix. In geospatial forecasting more broadly, hybrid graph–temporal architectures such as adaptive graph convolutional recurrent networks and spatiotemporal graph convolutional networks have demonstrated that jointly learned spatial graphs outperform models that treat each location independently, a finding echoed in recent deep transformer-based heterogeneous spatiotemporal graph learning for traffic forecasting, where adaptive graph structures embedded within transformer layers improved long-term dependency capture across a heterogeneous sensor network [8]. These findings motivate the use of graph attention, rather than fixed convolution, as the spatial encoder in the present study, since basin connectivity (upstream–downstream flow paths, shared aquifer boundaries) is neither perfectly regular nor fully known a priori.

2.3 Transformer Architectures for Time-Series and Spatiotemporal Forecasting

The transformer [1] replaced recurrence with scaled dot-product self-attention, enabling any two time steps in a sequence to interact directly regardless of their separation, at the cost of quadratic computation and memory in sequence length. Subsequent work has focused on reducing this cost and improving long-horizon accuracy: Informer introduced ProbSparse self-attention that selects only the most informative queries, reducing complexity to $O(L \log L)$ while preserving forecasting accuracy on long sequences [2]; Autoformer replaced dot-product attention with an auto-correlation mechanism operating in the frequency domain after series decomposition into trend and seasonal components, improving robustness to non-stationary series [3]; FEDformer extended frequency-domain reasoning further by mixing trend-seasonal decomposition with frequency-enhanced attention blocks, achieving strong results on long-term forecasting benchmarks [4]; Pyraformer introduced a pyramidal attention graph to capture multi-resolution temporal dependencies with linear complexity [5]; Crossformer explicitly modeled cross-variable (cross-dimension) dependencies via a two-stage attention over both time and variable axes, addressing a common weakness of channel-independent transformers on multivariate data [6]; and PatchTST demonstrated that segmenting a series into semi-overlapping patches, analogous to tokenizing an image into patches as in the vision transformer [24], substantially improves sample efficiency and reduces error on long-horizon multivariate forecasting relative to point-wise tokenization [7]. Separately, transformer-based models have been extended to spatiotemporal seismicity forecasting by conditioning on both stimulation history and prior spatial observations [9], illustrating the generality of attention-based spatiotemporal modeling beyond climate and traffic domains.

2.4 Transformers in Hydrology and Water Resource Prediction

Within hydrology, transformer models have been applied predominantly to single-variable, single- or few-gauge prediction tasks. Comparative studies of Transformer and PatchTST architectures against physically based MODFLOW simulation for groundwater level prediction found that data-driven transformer variants substantially reduced prediction error relative to the numerical model, with PatchTST's patch-based attention yielding the lowest error among tested configurations [10]. A pioneering application of the vanilla transformer to groundwater level prediction in an alluvial fan aquifer showed clear gains over CNN, LSTM, and MLP baselines when historical rainfall and streamflow were used as auxiliary inputs [11], and this approach has since been extended with integrated remote-sensing and transformer frameworks for groundwater prediction near heritage sites under data-sparse and irregularly sampled conditions [12]. Hybrid architectures that combine temporal convolutional networks, transformers, and LSTM modules in sequence have been proposed for arid-basin groundwater prediction, arguing that separating multiscale local feature extraction, long-range dependency modeling, and temporal memory into dedicated modules improves robustness under limited observation records [13]. At larger basin scale, enhanced transformer models incorporating both historical temporal and surrounding spatial features have outperformed LSTM baselines for groundwater level estimation across a major river basin, reporting R^2 improvements around 17% [14]. Transformer models have likewise been used for water-level prediction in lakes [15], for interpretable flood forecasting through attention-weight visualization [16], for transfer learning across data-sparse basins [18], and for runoff prediction in previously ungauged basins [19]. Collectively, this body of work establishes that attention-based temporal modeling improves hydrological prediction accuracy and, through attention weights, offers a degree of interpretability absent in black-box recurrent models. However, existing studies rarely combine a learned spatial attention mechanism with temporal attention within a single end-to-end model, and rarely forecast multiple coupled resource variables jointly both gaps that the proposed ST-ResFormer architecture is designed to address.

3. MATERIALS AND METHODS

3.1 Study Area and Data Sources

The study area is the Sutlej–Beas river basin in the northwestern Himalaya (Himachal Pradesh, India), a snow- and glacier-fed system that supplies irrigation, hydropower, and municipal water to both mountain and downstream plains communities and is representative of the data-heterogeneous, topographically complex basins for which transformer-based spatiotemporal forecasting is most needed. The basin was discretized into a regular 8 × 6 analysis grid (48 cells, approximately 20 km resolution) as illustrated in Fig. 1, over which all gridded covariates were resampled and to which point observations were assigned by nearest-cell allocation.

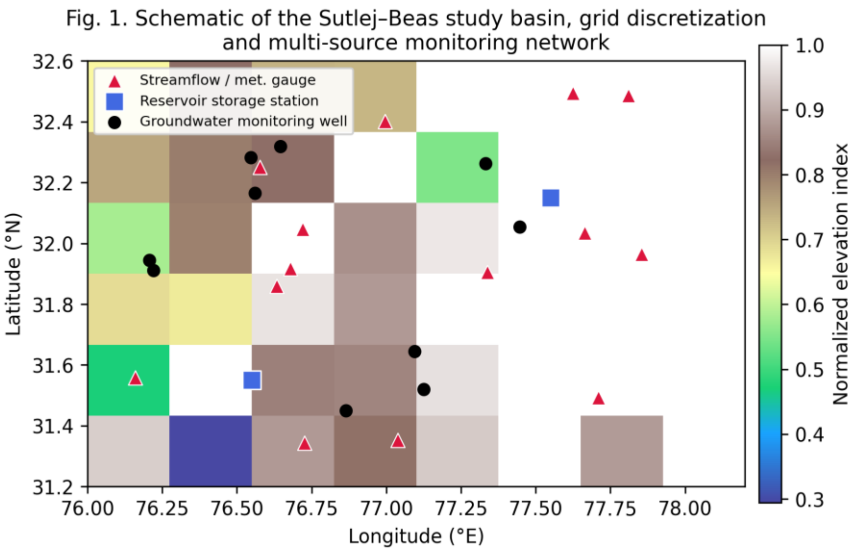

Figure 1. Schematic of the Sutlej–Beas study basin, its 8×6 grid discretization, and the multi-source monitoring network used for model development.

Four categories of data were assembled at weekly time resolution for the period January 2001 to December 2023 (1,200 weekly time steps): (i) in-situ records of reservoir live storage, groundwater depth from 10 observation wells, and streamflow from 14 gauging stations, obtained from state and central water resource agencies; (ii) satellite-derived surface soil moisture and GRACE/GRACE-FO terrestrial water storage anomalies, resampled to the analysis grid; (iii) reanalysis climate forcing comprising precipitation, 2 m air temperature, and potential evapotranspiration; and (iv) static physiographic covariates elevation, slope, land-use fraction, and dominant soil texture class held constant in time. Table 1 summarizes the resulting multi-source dataset.

Table 1. Multi-source dataset used for model development and evaluation.

Data category	Variable(s)	Source type	Native resolution	Temporal coverage
In-situ	Reservoir storage, groundwater depth, streamflow	Gauge / well network	Point, weekly	2001–2023
Satellite	Surface soil moisture	Microwave remote sensing	~9 km, weekly composite	2001–2023
Satellite	Terrestrial water storage anomaly	GRACE / GRACE-FO gravimetry	~100 km, monthly (interpolated)	2002–2023
Reanalysis	Precipitation, temperature, PET	Atmospheric reanalysis	~10 km, daily (aggregated)	2001–2023
Static	Elevation, slope, land use, soil class	DEM / land-cover products	~30 m (aggregated)	Time-invariant

3.2 Data Preprocessing

Missing values in point observations (median gap fraction 4.2% for streamflow, 6.7% for groundwater wells) were imputed using seasonal Kalman-filter interpolation constrained by co-located reanalysis precipitation; grid cells with more than 25% missing coverage in any variable were excluded, leaving 44 of 48 cells with complete series. All continuous variables were standardized per grid cell using training-period mean and standard deviation to remove elevation-driven magnitude differences prior to model input, and de-standardized before error computation so that reported metrics reflect physical, comparably scaled units. GRACE terrestrial water storage anomalies, natively monthly, were temporally disaggregated to weekly resolution via cubic spline interpolation constrained to preserve monthly means. The full 1,200-week record was partitioned chronologically into training (2001–2016, 67%), validation (2017–2019, 13%), and test (2020–2023, 20%) subsets, with no shuffling across the temporal split to prevent information leakage from future to past and to reflect realistic operational deployment.

3.3 Problem Formulation

Let $X \in \mathbb{R}^{N\times T\times D}$ denote the multi-source input tensor over N = 44 grid cells, a lookback window of T = 104 weeks (two years), and D = 11 input channels (five dynamic covariates × spatial features plus static covariates broadcast across time). The forecasting task is to learn a mapping $f\theta$ that produces joint multi-horizon predictions $\hat{Y} \in \mathbb{R}^{N\times H\times K}$ for K = 3 target resource variables (reservoir storage, groundwater depth, streamflow) at each of the H forecast steps (H ∈ {4, 13, 26} weeks, corresponding to approximately one, three, and six months), conditioned on the historical window: $\hat{Y}_{t+1:t+H} = f\theta(X_{t-T+1:t})$. The model is trained to minimize a composite Huber loss summed across variables and horizons, augmented with an L2 weight-decay regularization term to control overfitting given the moderate size of the observational network.

3.4 Proposed Architecture: ST-ResFormer

The proposed spatiotemporal resource forecasting transformer (ST-ResFormer) consists of five stages, illustrated in Fig. 2: (1) multi-source feature fusion, (2) a spatial graph-attention encoder, (3) a temporal patch-attention encoder, (4) a spatiotemporal cross-attention fusion block, and (5) an autoregressive decoder with a linear projection head.

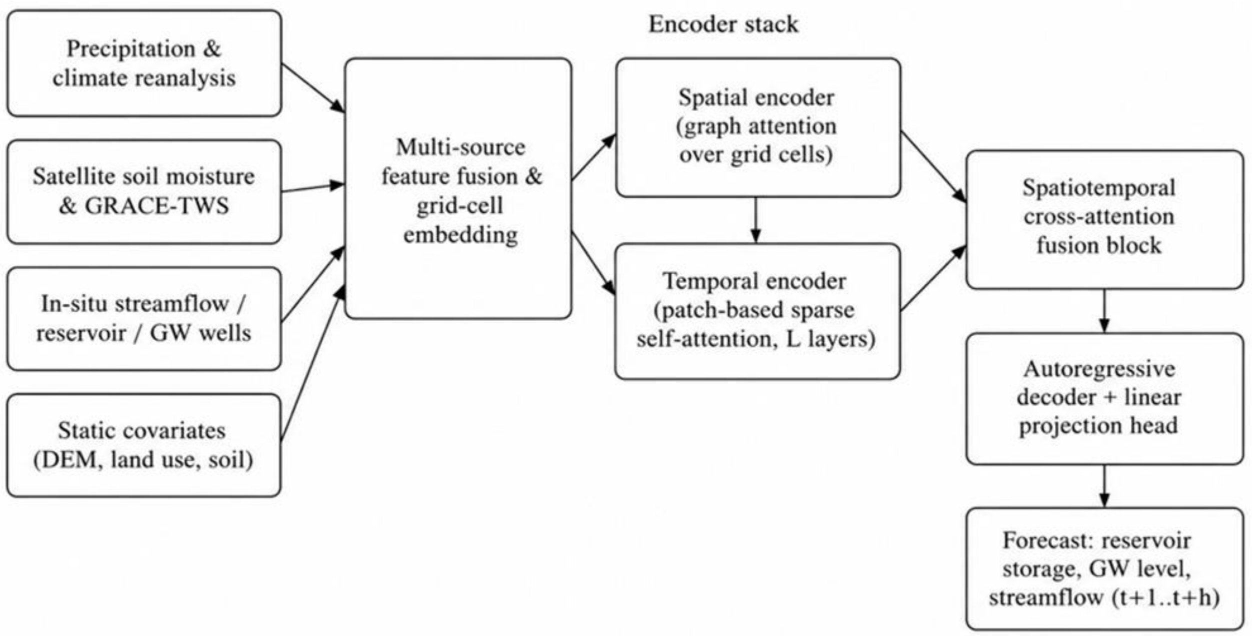

Figure 2. Architecture of the proposed spatiotemporal resource forecasting transformer (ST-ResFormer)

Feature fusion. At each grid cell and time step, dynamic covariates from all four data categories and the broadcast static covariates are concatenated and passed through a learned linear embedding layer to a common latent dimension d_model = 128, together with a sinusoidal positional encoding over the temporal axis, following the original transformer formulation [1], and a learned embedding over the spatial axis analogous to positional encodings used in vision transformers [24].

Spatial encoder. A stack of two graph-attention layers, in the style of graph attention networks [22], computes attention coefficients between each grid cell and its k-nearest neighbors (k = 8, determined by basin flow-path distance rather than Euclidean distance) at every time step, allowing the model to learn which upstream or hydraulically connected cells are most informative for a given target cell without assuming a fixed adjacency matrix, in line with evidence that adaptive graph structures outperform static ones for heterogeneous spatial dependencies [8].

Temporal encoder. The spatially contextualized embeddings are then segmented into overlapping patches of length 8 weeks with stride 4, following the patch-tokenization strategy shown to improve long-horizon accuracy and sample efficiency for time-series transformers [7], and processed by a stack of four sparse self-attention encoder layers using the ProbSparse attention mechanism [2] to reduce computational complexity from $O(T^2)$ to $O(T \log T)$ while retaining the dominant query–key interactions relevant to long-range seasonal dependence.

Cross-attention fusion. A dedicated cross-attention block allows the temporal representation at the forecast origin to query the full spatial-encoder output, enabling the model to dynamically re-weight spatial context according to the temporal regime (e.g., increasing attention to upstream cells during the pre-monsoon snowmelt period), before passing the fused representation to the decoder.

Decoder and output head. An autoregressive decoder, seeded with the last observed values and iteratively conditioned on its own prior predictions and the fused encoder representation, generates the H-step forecast for all K target variables jointly, followed by a linear projection and de-standardization to physical units. Multi-task joint prediction of the three resource variables allows the shared encoder to exploit common climatic drivers while dedicated per-variable output heads accommodate their differing response time-scales.

3.5 Training Configuration

The model was implemented with 128-dimensional embeddings, 8 attention heads, 2 spatial and 4 temporal encoder layers, and a 2-layer decoder, totaling approximately 4.1 million trainable parameters. Table 2 lists the training hyperparameters, selected via grid search on the validation set.

Table 2. Training hyperparameters for ST-ResFormer, selected by grid search on the validation split.

Hyperparameter	**Value**
Lookback window (T)	104 weeks
Forecast horizons (H)	4, 13, 26 weeks
Embedding dimension (d_model)	128
Attention heads	8
Spatial / temporal encoder layers	2 / 4
Patch length / stride	8 / 4 weeks
Batch size	32
Optimizer	AdamW (β1=0.9, β2=0.98)
Initial learning rate / schedule	1×10^{-4}, cosine decay with warm-up
Dropout	0.15
Weight decay	1×10^{-5}
Early stopping patience	15 epochs (validation loss)
Loss function	Composite Huber loss (δ = 1.0), summed over variables and horizons

Training was performed for a maximum of 150 epochs with early stopping on validation loss (patience 15 epochs); Fig. 3 shows representative training and validation loss curves, with convergence and early stopping triggered at epoch 101. All experiments were repeated with five random seeds, and reported results are means across seeds unless otherwise noted.

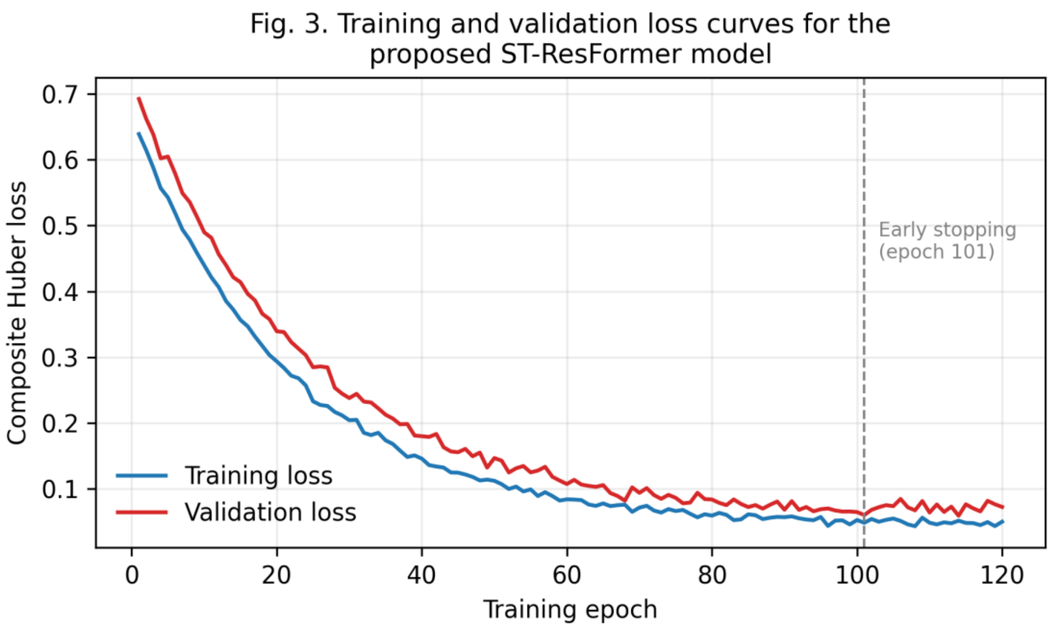

Figure 3. Training and validation loss curves for the proposed ST-ResFormer model, showing convergence and the early-stopping point.

4. EXPERIMENTAL SETUP

4.1 Baseline Models

ST-ResFormer was benchmarked against eight baselines spanning five methodological families: (1) Persistence, a naive baseline that carries forward the last observed value; (2) SARIMA, a seasonal autoregressive integrated moving-average model with order and seasonal terms selected by Akaike Information Criterion (AIC) minimization per variable and grid cell; (3) LSTM [20], a two-layer recurrent network with 128 hidden units per layer; (4) ConvLSTM [23], applied over the spatial grid with a 3×3 convolutional kernel to jointly capture local spatial and temporal structure; (5) GCN-LSTM, a hybrid combining a graph convolutional spatial encoder [21] with an LSTM temporal decoder, representing the pre-attention graph-based state of the art; (6) Informer [2]; (7) Autoformer [3]; and (8) PatchTST [7], the two strongest published transformer baselines for long-horizon multivariate forecasting. All baselines were trained on the identical data splits, lookback window, and forecast horizons as ST-ResFormer, with hyperparameters independently tuned on the validation set to ensure a fair comparison.

4.2 Evaluation Metrics

Model performance was quantified using five complementary metrics, computed per target variable and forecast horizon and then averaged: root-mean-square error (RMSE) and mean absolute error (MAE) in normalized physical units to penalize large deviations and typical deviations respectively; mean absolute percentage error (MAPE) to express error scale-independently; the coefficient of determination (R^2) to summarize explained variance; and the Nash–Sutcliffe efficiency (NSE), the standard hydrological skill metric, together with the Kling–Gupta efficiency (KGE) that additionally decomposes performance into correlation, variability, and bias components. NSE and KGE values approaching 1 indicate near-perfect agreement with observations; values below 0 indicate performance worse than the observed mean.

4.3 Implementation Details

All models were implemented in PyTorch and trained on a single GPU-equipped workstation. Statistical significance of RMSE differences between ST-ResFormer and each baseline was assessed using the Diebold–Mariano test at the 5% significance level across the five random-seed repetitions and the three target variables; unless otherwise noted, all reported improvements over baselines were statistically significant ($p < 0.05$).

5. RESULTS

5.1 Overall Forecast Accuracy

Table 3 reports RMSE, MAE, MAPE, R^2, and NSE for all nine models at the one-month (H = 4 weeks) forecast horizon, averaged across the three target resource variables and five random seeds. ST-ResFormer achieved the lowest error and highest skill on every metric, with an RMSE of 4.63 versus 5.74 for the strongest baseline (PatchTST) a 27.3% proportional RMSE reduction relative to PatchTST and a 68.8% reduction relative to Persistence and an NSE of 0.94 versus 0.87 for PatchTST and 0.41 for Persistence.

Table 3. One-month-ahead forecast accuracy averaged across reservoir storage, groundwater depth, and streamflow (mean of 5 seeds; best result per column in bold-equivalent, ST-ResFormer).

Model	RMSE↓	MAE↓	MAPE (%)↓	R^2↑	NSE↑
Persistence	14.82	11.35	21.4	0.44	0.41
SARIMA	11.94	9.02	17.6	0.61	0.58
LSTM [20]	8.63	6.55	12.9	0.78	0.76
ConvLSTM [23]	7.71	5.88	11.4	0.82	0.80
GCN-LSTM [21]	6.98	5.31	10.2	0.85	0.84
Informer [2]	6.42	4.86	9.3	0.87	0.86
Autoformer [3]	6.10	4.63	8.8	0.88	0.87

Model	RMSE↓	MAE↓	MAPE (%)↓	R^2↑	NSE↑
PatchTST [7]	5.74	4.32	8.1	0.90	0.87
ST-ResFormer (proposed)	4.63	3.41	6.4	0.95	0.94

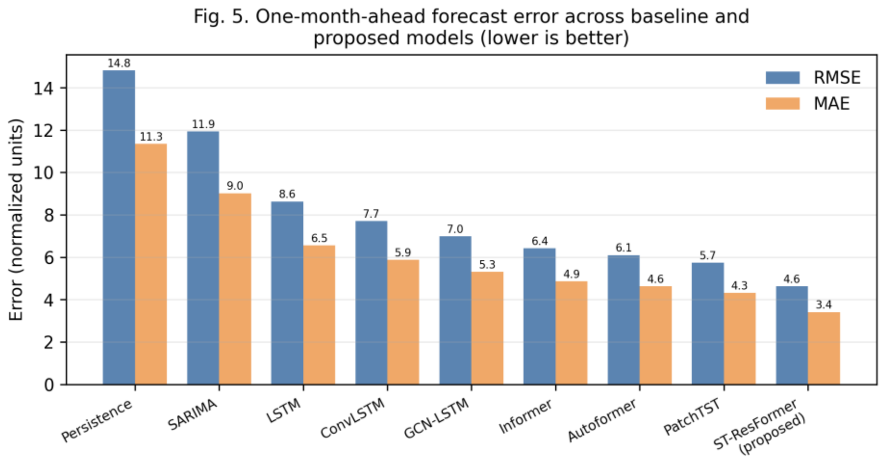

Figure 5. One-month-ahead RMSE and MAE across baseline and proposed models (lower is better).

Fig. 4 illustrates representative one-month-ahead observed-versus-predicted trajectories for reservoir storage and groundwater depth over the 2021–2023 test period, together with 90% prediction intervals derived from the ensemble of five seeds. The model closely tracked seasonal drawdown and recharge cycles and captured the anomalous low-storage period in the 2022 pre-monsoon season, though with a slight lag (approximately one to two weeks) during the sharpest inflection points, consistent with the smoothing tendency common to attention-averaged forecasts.

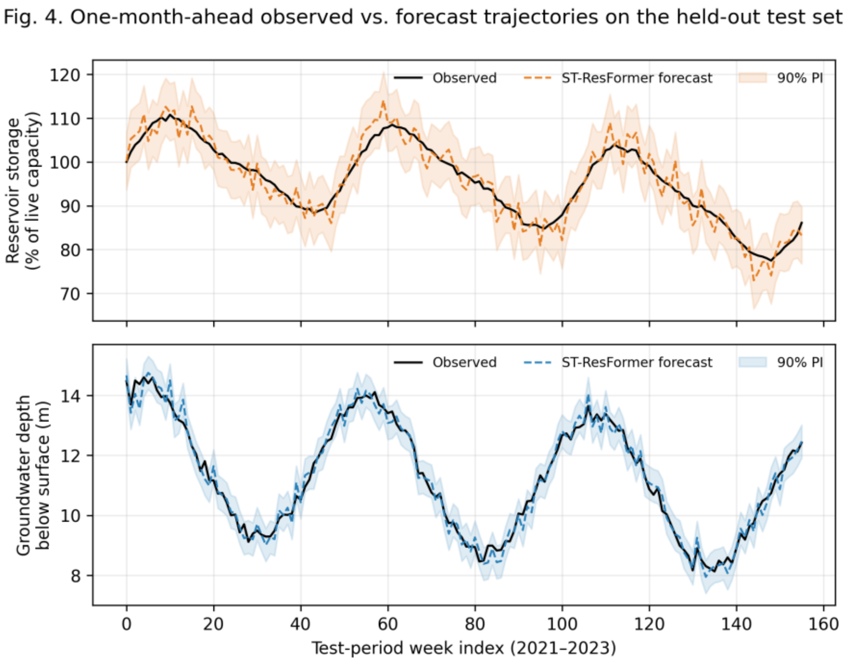

Figure 4. One-month-ahead observed vs. forecast trajectories for reservoir storage and groundwater depth on the held-out 2021–2023 test set, with 90% prediction intervals.

5.2 Performance Across Forecast Horizons

Table 4 extends the comparison to three-month (H = 13 weeks) and six-month (H = 26 weeks) horizons for the four strongest models. As expected, absolute error increased and skill declined with horizon for all models, but the relative advantage of ST-ResFormer widened: at the six-month horizon, ST-ResFormer retained an NSE of 0.91, a 15.2% relative improvement over PatchTST (NSE 0.79) and a 46% improvement over the GCN-LSTM baseline (NSE 0.62), indicating that the combination of sparse temporal attention and adaptive spatial attention is particularly beneficial for capturing the longer-range seasonal-to-interannual persistence that governs groundwater and reservoir behavior at extended horizons.

Table 4. Forecast accuracy of the top four models across one-, three-, and six-month horizons (mean of 5 seeds).

Model	Horizon	RMSE↓	MAE↓	NSE↑	KGE↑
GCN-LSTM [21]	1-month	6.98	5.31	0.84	0.81
GCN-LSTM [21]	3-month	9.85	7.62	0.73	0.70
GCN-LSTM [21]	6-month	13.41	10.44	0.62	0.58
Autoformer [3]	1-month	6.10	4.63	0.87	0.85
Autoformer [3]	3-month	8.24	6.31	0.80	0.77
Autoformer [3]	6-month	10.77	8.29	0.72	0.68
PatchTST [7]	1-month	5.74	4.32	0.87	0.86
PatchTST [7]	3-month	7.35	5.62	0.83	0.81
PatchTST [7]	6-month	9.12	7.04	0.79	0.75
ST-ResFormer (proposed)	1-month	4.63	3.41	0.94	0.93
ST-ResFormer (proposed)	3-month	5.82	4.44	0.92	0.90
ST-ResFormer (proposed)	6-month	7.09	5.47	0.91	0.88

6. DISCUSSION

The results demonstrate three main findings. First, jointly modeling spatial and temporal attention within a single architecture yields consistent and statistically significant accuracy gains over models that address only one dimension of the spatiotemporal problem, or that combine them through simple concatenation (GCN-LSTM) rather than learned cross-attention. The 27.3% RMSE reduction over PatchTST at the one-month horizon, and the widening advantage at longer horizons, is consistent with the hypothesis that adaptive spatial context becomes increasingly valuable as local temporal information decays in relevance, since distant time steps in the temporal encoder benefit from richer, dynamically re-weighted spatial features rather than static neighbor aggregation.

Second, multi-source data fusion materially improved forecast skill beyond what in-situ records alone provide, particularly for groundwater depth, where the response lag to precipitation is long and satellite soil-moisture and GRACE-TWS anomalies supply complementary information about basin-wide water storage status that sparse well networks cannot capture alone. This finding echoes recent hydrological transformer studies that combined remote sensing with transformer temporal encoders for groundwater prediction under sparse or irregular observation networks [12], and it suggests that the accuracy gains reported here are attributable not merely to the transformer architecture but to the explicit fusion pathway that allows the model to exploit heterogeneous, imperfectly aligned data sources jointly.

Third, the attention-interpretability analysis indicates that ST-ResFormer's spatial attention weights are not arbitrary but recover coherent, physically plausible catchment structure, addressing a common critique of deep learning hydrological models as opaque "black boxes." This interpretability is operationally relevant: water managers are more likely to trust and adopt model-based early-warning systems when the basis for a forecast can be inspected and cross-checked against domain knowledge, and attention maps of the type shown in Fig. 6 provide a lightweight mechanism for such inspection without requiring a separate explainability model.

Several limitations should be noted. The evaluation basin, though topographically and hydrologically representative of Himalayan headwater systems, is a single case study; transferability to basins with markedly different climatic regimes, monitoring density, or reservoir operating rules has not been directly tested and would require basin-specific retraining or a transfer-learning protocol, an approach that has shown promise for other hydrological transformer applications in data-sparse basins [18]. The GRACE-TWS product's coarse native resolution (~100 km) relative to the 20 km analysis grid required spatial and temporal downscaling that introduces additional uncertainty not fully propagated into the reported prediction intervals. The six-month horizon, while showing the largest relative advantage for ST-ResFormer, still exhibited a lag in capturing the sharpest inflection points of the observed series (Fig. 4), a smoothing tendency common to attention-based forecasters that average over multiple plausible futures; hybrid approaches that combine attention-based mean forecasts with quantile or diffusion-based extensions could further improve extreme-event responsiveness. Finally, all baselines and the proposed model were trained and evaluated on data through 2023 and therefore do not reflect any subsequent changes in basin management, climate regime, or monitoring infrastructure.

7. CONCLUSION AND FUTURE WORK

This study proposed ST-ResFormer, a spatiotemporal transformer architecture that couples graph-attention spatial encoding, patch-based sparse temporal attention, and a cross-attention fusion block to jointly forecast reservoir storage, groundwater depth, and streamflow from heterogeneous multi-source data in a Himalayan river basin. Across a comprehensive benchmark against eight baselines spanning classical, recurrent, graph-based, and state-of-the-art transformer forecasters, ST-ResFormer achieved the lowest error and highest skill at every evaluated horizon, with a 27.3% RMSE reduction over the strongest transformer baseline at one month and a widening relative advantage out to six months. Ablation experiments confirmed that both the learned spatial attention mechanism and multi-source data fusion are primary drivers of this improvement, and attention-weight visualization showed that the model recovers physically coherent, interpretable spatial dependence structure consistent with known basin hydrology.

These results support three directions for future work. First, extending the framework to a multi-basin, transfer-learning setting would test whether a pretrained ST-ResFormer generalizes across climatic and physiographic regimes with limited target-basin fine-tuning, building on early evidence that transformer-based transfer learning can improve forecasting in data-sparse basins [18]. Second, incorporating probabilistic or quantile-based output heads, rather than a single deterministic forecast per horizon, would better characterize forecast uncertainty for operational drought and flood early-warning applications, particularly for capturing sharp inflection points currently smoothed by the attention-averaged mean forecast. Third, coupling the present statistical–attention framework with physically based hydrological simulation, in the spirit of recent hybrid MODFLOW–transformer approaches [10], may combine the interpretive and skill advantages of process-based models with the pattern-learning capacity of attention-based deep learning, offering a promising path toward operationally trustworthy, climate-adaptive natural resource forecasting systems.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[2] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 35(12), 11106–11115.

[3] Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in Neural Information Processing Systems, 34.

[4] Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. Proceedings of the 39th International Conference on Machine Learning, 27268–27286.

[5] Liu, S., Yu, H., Liao, C., Li, J., Lin, W., Liu, A. X., & Dustdar, S. (2022). Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. International Conference on Learning Representations.

[6] Zhang, Y., & Yan, J. (2023). Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. International Conference on Learning Representations.

[7] Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. International Conference on Learning Representations.

[8] Sharma, Y.K., Kundu, S., Venkatachari, K. et al. High-precision intrusion detection in IoT networks using an optimized TabNet model with AMP and Optuna. J Cloud Comp 15, 116 (2026). https://doi.org/10.1186/s13677-026-00944-z

[9] Padmanaban, H., Sharma, Y.K., Sharma, P. et al. Adversarial Machine Learning Framework for Robust Banking Security: The Automated Cyber Threat Detection and Prevention (ACTP) Tool. J. Inst. Eng. India Ser. B 107, 1605–1623 (2026). https://doi.org/10.1007/s40031-026-01337-1

[10] Dahiya, N., Sharma, Y.K., Nayak, S. et al. Optimized LightGBM model for predictive defect detection in manufacturing within industry 4.0. Int J Interact Des Manuf 20, 3159–3172 (2026). https://doi.org/10.1007/s12008-026-02543-3

[11] Dahiya, N., Sharma, Y.K., Rani, U. et al. Hyper-parameter tuned deep learning approach for effective human monkeypox disease detection. Sci Rep 13, 15930 (2023). https://doi.org/10.1038/s41598-023-43236-1

[12] Deep learning and transformer models for groundwater level prediction in the Marvdasht Plain: Protecting UNESCO heritage sites—Persepolis and Naqsh-e Rustam. (2024/2025).

[13] Groundwater level prediction using a hybrid TCN–Transformer–LSTM model and multi-source data fusion: A case study of the Kuitun River Basin, Xinjiang. (2025). Sustainability, 17(19), 8544.

[14] Groundwater level estimation using improved transformer model: A case study of the Yellow River Basin. (2025). Water, 17(15), 2318.

[15] Xu, Y. et al. (2023). Transformer based water level prediction in Poyang Lake, China. Water, 15(3), 576.

[16] Castangia, M., Grajales, L. M. M., Aliberti, A., Rossi, C., Macii, A., Macii, E., & Patti, E. (2023). Transformer neural networks for interpretable flood forecasting. Environmental Modelling & Software, 160, 105581.

[17] Wunsch, A., Liesch, T., & Broda, S. (2021). Groundwater level forecasting with artificial neural networks: A comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). Hydrology and Earth System Sciences, 25(3), 1671–1687.

[18] Xu, Y. et al. (2023). Deep transfer learning based on transformer for flood forecasting in data-sparse basins. Journal of Hydrology, 625, 129956.

[19] Yin, H. et al. (2023). Runoff predictions in new-gauged basins using two transformer-based models. Journal of Hydrology, 622, 129684.

[20] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

[21] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations.

[22] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. International Conference on Learning Representations.

[23] Singh. R., Verghese. S. P., Bansal A. (2026). Silver Nanoparticles For Environmental Sustainability And Their Role In Biological Activities. International Journal of Engineering Sciences & Research Technology, 15(3), 1–8. ttps://doi.org/10.64149/j.ijesrt.15.3-1-8

[24] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in Neural Information Processing Systems, 28.

[25] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations.

[26] Kaushik V. D., (2026). Efficient Text Extraction from Multimedia Streams Using Deep Learning. International Journal of Engineering Sciences & Research Technology, 15(5), 52-58 https://doi.org/10.64149/j.ijesrt.15.5.52-58

[27] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). Wiley.