

Article

GeoFormer: Geography-Aware Adaptive Transformer with Multi-Scale Temporal Fusion for Global Reservoir Water Level Forecasting

Xiaobing Wu ^{1,2} , Jinhao Guo ³, Yahui Shan ^{3,*} and Guangyin Jin ¹ 

¹ School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; xiaob_wuu@126.com (X.W.); jinguangyin@njust.edu.cn (G.J.)

² College of Water Resources, North China University of Water Resources and Electric Power, Zhengzhou 450045, China

³ Wuhan Second Ship Design and Research Institute, Wuhan 430064, China; 220721063@stu.xpu.edu.cn

* Correspondence: shanyahui@hust.edu.cn

Abstract

Accurate reservoir water level forecasting is essential for water resource management, flood risk mitigation, and hydropower operation. However, it remains challenging due to pronounced geographical heterogeneity and complex multi-scale temporal dynamics. Existing deep-learning approaches typically overlook explicit geographical and climatic conditioning. They struggle to capture temporal dependencies across multiple time scales. They also exhibit limited transferability across reservoirs with similar hydrological characteristics. To address these limitations, this paper proposes GeoFormer, a geography-aware adaptive Transformer framework designed for reservoir water level forecasting across diverse geographical contexts. GeoFormer integrates three key innovations. First, a Geography-Aware Embedding Module conditions temporal representations on geographical location, climate regimes, and reservoir attributes. Second, an Adaptive Multi-Scale Temporal Fusion mechanism dynamically aggregates information across daily, weekly, and monthly temporal resolutions. Third, a Cross-Reservoir Knowledge Transfer strategy enables effective knowledge sharing among hydrologically similar reservoirs. Extensive experiments on six reservoirs distributed across multiple continents and climate zones demonstrate that GeoFormer consistently outperforms state-of-the-art baselines, including iTransformer, DLinear, and Informer. The model achieves average reductions of 23.7% in RMSE, 19.4% in MAE, and 15.8% in MAPE, while maintaining strong robustness and generalization across geographically heterogeneous hydrological systems.

Keywords: water level forecasting; geography-aware learning; adaptive transformer; multi-scale temporal fusion; reservoir management; transfer learning; deep learning

MSC: 68-11; 68U01; 68R10



Academic Editor: Carmen Patino-Alonso

Received: 21 January 2026

Revised: 11 February 2026

Accepted: 12 February 2026

Published: 14 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Water impoundment through reservoir systems represents one of humanity's most critical infrastructure interventions for managing freshwater resources. With over 58,000 large dams worldwide serving more than 2 billion people [1–3], reservoirs provide essential services including water supply, flood control, hydropower generation, and irrigation support. Increasing water scarcity affects approximately 4 billion people globally for

at least one month per year [4], with climate change intensifying hydrological variability and threatening water security. In this context, accurate reservoir water level forecasting becomes paramount for optimizing water allocation, preventing catastrophic floods (with annual losses exceeding \$40 billion [5,6]), and ensuring sustainable management under increasingly uncertain conditions.

Different applications impose distinct forecasting requirements. Hydropower optimization requires 1–3 day ahead predictions at hourly resolution for turbine scheduling. Flood warning systems need 3–7 day forecasts at daily resolution with high accuracy for extreme events. Water supply management relies on 7–30 day seasonal forecasts for allocation planning. Long-term climate adaptation strategies require monthly to seasonal outlooks for infrastructure design. This study focuses on the 1–30 day forecasting horizon at daily resolution, which addresses the critical operational window for reservoir management while capturing both short-term dynamics and medium-term trends.

Traditional forecasting approaches have evolved through three distinct paradigms. Physics-based hydrological models such as HEC-RAS [7], SWAT [8], and MIKE-SHE [9] provide mechanistic interpretability but require extensive parameter calibration and struggle with nonlinear interactions [10,11]. Statistical methods including ARIMA [12] and exponential smoothing [13] offer computational efficiency but assume stationarity that is often violated by real-world hydrological data [14]. Machine-learning techniques such as SVR [15], Random Forests [16], and XGBoost [17,18] can learn nonlinear mappings but lack explicit temporal dependency modeling [19].

Deep learning has significantly advanced time series forecasting through architectures that capture temporal dependencies [20,21]. LSTM [22] and GRU [23] networks with gating mechanisms have demonstrated superior performance in hydrological applications [24–26], though they suffer from sequential computation bottlenecks and limited receptive fields [27]. Transformer architectures [28] with self-attention mechanisms have achieved state-of-the-art results through variants like Informer [29], Autoformer [30], FEDformer [31], and iTransformer [32]. Linear models such as DLinear [33] have also shown competitive performance, highlighting the importance of inductive biases.

Despite these advances, existing methods face three fundamental limitations for global reservoir forecasting. First, geographical heterogeneity refers to systematic variations in hydrological behavior driven by location-specific factors including climate regimes, topography, and catchment characteristics. For example, tropical reservoirs such as Bagre in Burkina Faso exhibit monsoon-driven fluctuations exceeding 9 m annually with sharp seasonal transitions [34], while temperate reservoirs such as Yesa in Spain show gradual snowmelt-influenced variations of 3–4 m with smoother temporal patterns [35,36]. These location-specific dynamics require models that can explicitly encode geographical context rather than treating all reservoirs as homogeneous entities. Current models cannot encode these location-specific priors or effectively transfer knowledge across similar reservoirs.

Second, existing methods inadequately capture multi-scale temporal dependencies. Water levels exhibit patterns spanning sub-daily operational releases, weekly storm events, and seasonal cycles [37]. Fixed attention windows either miss long-term trends or dilute focus on critical recent events. Existing multi-scale methods use rigid decomposition schemes misaligned with hydrological timescales [38].

Third, they fail to leverage cross-reservoir knowledge transfer. Models trained independently on each reservoir must learn fundamental hydrological relationships from scratch, wasting valuable data from similar reservoirs such as monsoon-driven systems in West Africa and South Asia that share characteristic patterns [39]. Existing transfer learning approaches lack principled mechanisms to quantify reservoir similarity and selectively transfer relevant knowledge while avoiding negative transfer [40].

To address these challenges, we propose GeoFormer, a novel geography-aware adaptive Transformer framework with multi-scale temporal fusion for global reservoir water level forecasting. GeoFormer introduces three key innovations. The Geography-Aware Embedding Module (GAEM) encodes geographical coordinates, climate zones (Köppen-Geiger), and reservoir attributes into learnable representations, enabling location-specific conditioning and climate-dependent pattern capture. The Adaptive Multi-Scale Temporal Fusion (AMTF) constructs parallel Transformer encoders on daily, weekly, and monthly sequences with learnable gating networks that dynamically weight scale contributions based on input context, enabling adaptive focus on relevant timescales. The Cross-Reservoir Knowledge Transfer (CRKT) employs similarity-aware contrastive pre-training that clusters similar reservoirs' representations, followed by reservoir-specific adapter-based fine-tuning for efficient adaptation while preserving transferred knowledge.

The main contributions of this work are:

1. This paper proposes a novel embedding mechanism integrating geographical coordinates, climate classifications, and reservoir attributes into learnable representations, enabling the first systematic incorporation of geographical context into deep-learning architectures for reservoir water level forecasting.
2. An adaptive fusion framework is developed that dynamically aggregates information across daily, weekly, and monthly temporal scales through learnable gating networks. Unlike fixed decomposition schemes, this approach adaptively adjusts scale importance based on hydrological states, achieving superior capture of both short-term fluctuations and long-term trends.
3. A principled transfer learning strategy is introduced with similarity-aware contrastive pre-training and reservoir-specific adapters, enabling effective knowledge sharing across analogous reservoirs while preventing negative transfer, significantly improving data efficiency for reservoirs with limited observations.
4. Extensive experiments are conducted on six globally distributed reservoirs with diverse geographical and climatic characteristics, comparing against eight strong baselines including state-of-the-art Transformers, linear models, recurrent networks, and machine-learning methods. GeoFormer achieves consistent improvements across all datasets, demonstrating robust generalization and practical applicability.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 presents the GeoFormer framework; Section 4 describes the experimental setup and results; Section 5 provides relevant discussion on the research; and Section 6 concludes with future directions.

2. Related Work

2.1. Deep Learning for Hydrological Forecasting

Deep learning has significantly advanced hydrological forecasting over the past decade, with recurrent neural networks emerging as the dominant paradigm for modeling temporal dependencies in water-related variables. LSTM networks [22] have been extensively applied to streamflow prediction [24,40,41], demonstrating superior performance over traditional hydrological models by learning complex nonlinear rainfall-runoff relationships directly from data. Kratzert et al. [24] showed that a single LSTM model trained across multiple catchments can outperform individually calibrated hydrological models, highlighting the potential for knowledge transfer. GRU networks [23], with simplified gating mechanisms, have proven effective for groundwater level forecasting [25] and reservoir inflow prediction [26], offering computational advantages over LSTM while maintaining comparable performance. Bidirectional LSTM architectures [42] have been

employed for water quality prediction, leveraging both past and future context in offline forecasting scenarios.

2.2. Multi-Scale Temporal Modeling

Capturing multi-scale temporal patterns is crucial for time series forecasting across domains. Wavelet-based methods decompose signals into multiple frequency bands, enabling separate modeling of high-frequency fluctuations and low-frequency trends. Empirical mode decomposition (EMD) [43] and its variants such as ensemble EMD [44] adaptively decompose nonstationary signals into intrinsic mode functions (IMFs) representing different oscillatory modes, widely used in hydrological forecasting [45,46]. However, decomposition-based approaches require separate models for each component and lack end-to-end optimization. Hierarchical neural architectures explicitly model multi-scale temporal structures. Multi-resolution LSTM [47] employs multiple recurrent layers operating at different temporal resolutions, with lower layers capturing fast dynamics and higher layers modeling slow trends. Temporal pyramid networks [48] construct feature pyramids at multiple timescales through pooling operations, fusing multi-scale representations for prediction. ScaleFormer [38] introduced scale-aware attention that decomposes sequences into multiple scales via downsampling and applies separate Transformer encoders, aggregating multi-scale features through learned fusion weights. However, these methods use fixed scale decomposition (e.g., powers of two) that may not align with domain-specific temporal patterns, and they lack adaptive mechanisms to adjust scale importance based on input context.

Recent work on adaptive temporal modeling includes conditional computation approaches where network components are dynamically activated based on input characteristics. Adaptive RNNs [49] learn to skip state updates for certain time steps, reducing computation while maintaining accuracy. Adaptive Transformers [50] adjust the number of attention layers based on input complexity.

2.3. Transfer Learning in Hydrological Modeling

Transfer learning has emerged as a promising approach to address data scarcity in hydrological modeling by leveraging knowledge from data-rich regions. Kratzert et al. [40] demonstrated that LSTM models pre-trained on large multi-catchment datasets significantly outperform locally trained models for ungauged basins, with performance improving as more diverse catchments are included in pre-training. Regional transfer learning [51] showed that models pre-trained on hydrologically similar catchments achieve better transfer performance than global models, highlighting the importance of source domain selection.

Domain adaptation techniques have been applied to hydrological forecasting to handle distribution shifts between source and target domains. Adversarial domain adaptation [52] with domain discriminators has been used to learn domain-invariant representations for streamflow prediction across catchments [53]. Meta-learning approaches [54] train models to quickly adapt to new catchments with limited data by learning good initialization parameters through episodic training on multiple tasks.

Multi-task learning frameworks jointly train models on multiple related hydrological prediction tasks, enabling knowledge sharing through shared representations. Feng et al. [55] employed multi-task LSTM to simultaneously predict streamflow, water temperature, and sediment load, demonstrating improved performance on all tasks compared to single-task models. Graph neural networks have been used to model spatial relationships among gauging stations, enabling information propagation across connected locations [56].

3. Methodology

In this section, we present the GeoFormer framework for geography-aware reservoir water level forecasting. We begin with problem formulation and notation in Section 3.1, then describe the three core components: Geography-Aware Embedding Module (Section 3.2), Adaptive Multi-Scale Temporal Fusion (Section 3.3), and Cross-Reservoir Knowledge Transfer (Section 3.4). Finally, we present the training objectives (Section 3.5). The framework is shown in Figure 1.

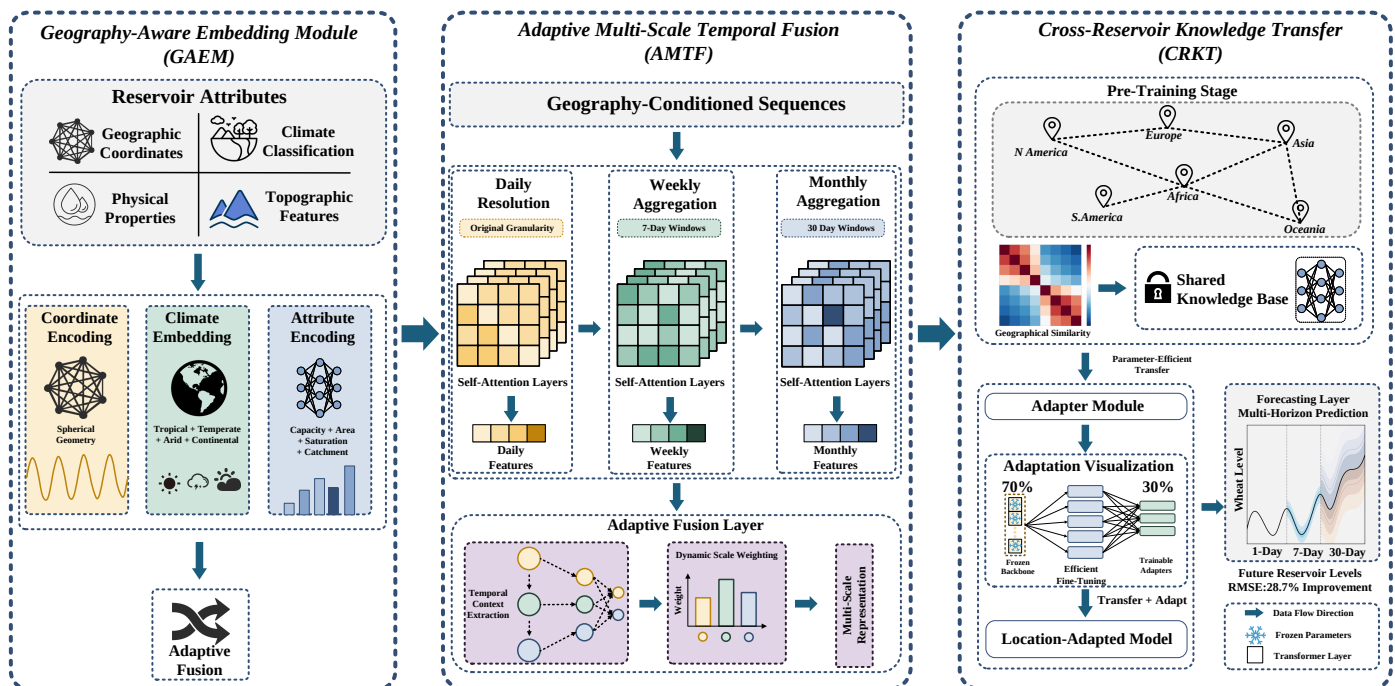


Figure 1. GeoFormer framework for geography-aware reservoir water level forecasting. The architecture integrates three innovations: geography-aware embeddings encoding location-specific attributes (left), adaptive multi-scale temporal fusion with context-dependent weighting across daily, weekly, and monthly resolutions (center), and cross-reservoir knowledge transfer via similarity-aware pre-training and efficient adapter-based fine-tuning (right).

3.1. Problem Formulation and Notation

Consider a set of N globally distributed reservoirs $\mathcal{R} = \{r_1, r_2, \dots, r_N\}$, where each reservoir r_i is associated with historical water level observations and geographical attributes. For reservoir r_i , let $\mathbf{X}^{(i)} = [\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \dots, \mathbf{x}_T^{(i)}] \in \mathbb{R}^{T \times d}$ denote the historical time series of length T , where $\mathbf{x}_t^{(i)} \in \mathbb{R}^d$ represents the d -dimensional feature vector at time step t .

For all six reservoirs, the consistently available input variables include: (1) historical water surface elevation (WSE) as the primary target variable, (2) geographical coordinates (latitude, longitude), (3) climate zone classification based on Köppen-Geiger system, and (4) reservoir physical attributes (surface area, storage capacity). These core variables enable the geography-aware framework to function across all reservoirs. To ensure model generalizability, the primary experiments use only these consistently available variables. Dataset-specific auxiliary variables such as precipitation, temperature, or inflow/outflow rates, which are available for some but not all reservoirs, are not used in the main experiments to maintain fair comparison across all study sites.

Our primary objective is to forecast future water levels $\mathbf{Y}^{(i)} = [\mathbf{y}_{T+1}^{(i)}, \mathbf{y}_{T+2}^{(i)}, \dots, \mathbf{y}_{T+H}^{(i)}] \in \mathbb{R}^{H \times 1}$ for a prediction horizon of H time steps. Each reservoir r_i is characterized by a set of geographical and physical attributes denoted as $\mathcal{A}^{(i)} = \{\text{lat}^{(i)}, \text{lon}^{(i)}, \text{climate}^{(i)}, \text{area}^{(i)},$

capacity⁽ⁱ⁾, ...}, where lat⁽ⁱ⁾ and lon⁽ⁱ⁾ represent latitude and longitude coordinates, climate⁽ⁱ⁾ denotes the climate zone classification (e.g., tropical, temperate, arid), and area⁽ⁱ⁾ and capacity⁽ⁱ⁾ represent surface area and storage capacity, respectively. These attributes are crucial for encoding location-specific hydrological behaviors.

The forecasting task can be formally defined as learning a mapping function $f_\theta : (\mathbf{X}^{(i)}, \mathcal{A}^{(i)}) \rightarrow \mathbf{Y}^{(i)}$ parameterized by θ , where the model takes as input both the historical time series and geographical attributes to generate future predictions. Unlike conventional approaches that learn separate models for each reservoir, our framework aims to learn a unified model that can effectively leverage knowledge across multiple reservoirs while adapting to location-specific characteristics.

Table 1 summarizes the key mathematical notation used throughout this paper.

Table 1. Summary of mathematical notation.

Symbol	Description
$\mathcal{R}$	Set of N reservoirs
r_i	The i -th reservoir
$\mathbf{X}^{(i)} \in \mathbb{R}^{T \times d}$	Historical time series for reservoir i
$\mathbf{x}_t^{(i)} \in \mathbb{R}^d$	Feature vector at time t for reservoir i
$\mathbf{Y}^{(i)} \in \mathbb{R}^{H \times 1}$	Future water levels for reservoir i
T	Historical sequence length
H	Prediction horizon
d	Feature dimension
$\mathcal{A}^{(i)}$	Geographical attributes of reservoir i
$\mathbf{g}^{(i)} \in \mathbb{R}^{d_g}$	Geography-aware embedding for reservoir i
d_g	Dimension of geographical embedding
$\mathbf{H}^{(s)} \in \mathbb{R}^{T_s \times d_h}$	Hidden representations at scale s
α_s	Adaptive weight for scale s
$\mathcal{L}_{\text{forecast}}$	Forecasting loss
$\mathcal{L}_{\text{contrast}}$	Contrastive loss for transfer learning

3.2. Geography-Aware Embedding Module (GAEM)

The Geography-Aware Embedding Module explicitly encodes geographical context and climate-specific patterns into learnable representations that condition the forecasting model. The key insight is that reservoirs sharing similar geographical and climatic characteristics exhibit analogous hydrological behaviors. For instance, tropical reservoirs experience monsoon-driven seasonality, while temperate reservoirs show snowmelt-dominated spring peaks. By embedding these location-specific priors into the model architecture, we enable the network to leverage domain knowledge and improve generalization across diverse geographical contexts.

Our approach draws inspiration from recent advances in geo-aware machine learning [57] and location encoding in computer vision [58], while being specifically adapted for hydrological forecasting. Unlike simple one-hot encoding of reservoir identifiers, which treats each location as independent and fails to capture similarity structure, our embedding module learns continuous representations that reflect both geographical proximity and climatic similarity.

3.2.1. Mathematical Formulation

For each reservoir r_i , we construct a geography-aware embedding $\mathbf{g}^{(i)} \in \mathbb{R}^{d_g}$ by processing multiple attribute types through specialized encoding layers. Geographical

coordinates (latitude $\text{lat}^{(i)}$, longitude $\text{lon}^{(i)}$) are encoded using periodic basis functions to capture spatial continuity and handle the spherical geometry of Earth:

$$\mathbf{e}_{\text{coord}}^{(i)} = [\sin(2\pi\text{lat}^{(i)}/180), \cos(2\pi\text{lat}^{(i)}/180), \sin(2\pi\text{lon}^{(i)}/360), \cos(2\pi\text{lon}^{(i)}/360)] \quad (1)$$

This periodic encoding ensures that geographically proximate locations have similar representations, while the sine-cosine pairing captures both hemispheric position and longitudinal continuity across the antimeridian. The use of different periods (180 for latitude, 360 for longitude) respects the natural ranges of geographical coordinates, and the periodic functions ensure that locations near $\pm 180^\circ$ longitude are correctly treated as adjacent. This 4-dimensional coordinate encoding is then projected to higher dimensions through a linear layer: $\mathbf{h}_{\text{coord}}^{(i)} = \mathbf{W}_{\text{coord}}\mathbf{e}_{\text{coord}}^{(i)} + \mathbf{b}_{\text{coord}}$, where $\mathbf{W}_{\text{coord}} \in \mathbb{R}^{d_g/3 \times 4}$ and $\mathbf{b}_{\text{coord}} \in \mathbb{R}^{d_g/3}$.

Climate classifications (e.g., tropical, temperate, arid based on Köppen-Geiger system [59]) are encoded through learnable embedding layers. Let $c^{(i)} \in \{1, 2, \dots, K\}$ denote the climate zone index for reservoir i among K climate types. The climate embedding is obtained as:

$$\mathbf{h}_{\text{climate}}^{(i)} = \mathbf{E}_{\text{climate}}[c^{(i)}] \quad (2)$$

where $\mathbf{E}_{\text{climate}} \in \mathbb{R}^{K \times d_g/3}$ is a learnable embedding matrix. This allows the model to learn climate-specific patterns such as monsoon seasonality for tropical zones or snowmelt dynamics for cold climates.

Physical characteristics including surface area, storage capacity, catchment area, and elevation are normalized and encoded through a multi-layer perceptron:

$$\mathbf{h}_{\text{attr}}^{(i)} = \text{MLP}_{\text{attr}}([\text{area}^{(i)}, \text{capacity}^{(i)}, \text{catchment}^{(i)}, \text{elevation}^{(i)}]) \quad (3)$$

where the MLP consists of two fully connected layers with ReLU activation: $\text{MLP}_{\text{attr}}(\mathbf{a}) = \mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1\mathbf{a} + \mathbf{b}_1) + \mathbf{b}_2$, with $\mathbf{W}_1 \in \mathbb{R}^{d_g/2 \times 4}$, $\mathbf{W}_2 \in \mathbb{R}^{d_g/3 \times d_g/2}$.

The three component embeddings are concatenated and fused through a gated fusion mechanism that learns to weight different attribute types:

$$\begin{aligned} \mathbf{h}_{\text{concat}}^{(i)} &= [\mathbf{h}_{\text{coord}}^{(i)}; \mathbf{h}_{\text{climate}}^{(i)}; \mathbf{h}_{\text{attr}}^{(i)}] \\ \gamma^{(i)} &= \sigma(\mathbf{W}_{\text{gate}}\mathbf{h}_{\text{concat}}^{(i)} + \mathbf{b}_{\text{gate}}) \\ \mathbf{g}^{(i)} &= \gamma^{(i)} \odot \text{ReLU}(\mathbf{W}_{\text{fuse}}\mathbf{h}_{\text{concat}}^{(i)} + \mathbf{b}_{\text{fuse}}) \end{aligned} \quad (4)$$

where σ denotes the sigmoid function, $\odot$ represents element-wise multiplication, and $\mathbf{W}_{\text{gate}}, \mathbf{W}_{\text{fuse}} \in \mathbb{R}^{d_g \times d_g}$. The gating mechanism $\gamma^{(i)}$ allows the model to adaptively emphasize different geographical attributes based on their relevance for the forecasting task. For example, climate embeddings receive higher weights (larger γ values) for tropical reservoirs with strong seasonal patterns, while physical attribute embeddings are more important for reservoirs with extreme morphological characteristics such as narrow dendritic shapes or large surface area variations. This adaptive weighting is learned end-to-end during training and varies across reservoirs based on their geographical characteristics.

3.2.2. Handling Missing Geographical Attributes

When geographical attributes are missing or unreliable, GeoFormer employs a hierarchical fallback strategy. For missing coordinate information (rare in practice as all reservoirs have known locations), the model defaults to using only climate zone and physical attribute embeddings, with coordinate encoding replaced by learnable reservoir-specific tokens. For missing climate classifications, the model infers climate zones from coordinates

using Köppen-Geiger lookup tables based on global climate datasets [59]. For missing physical attributes (e.g., storage capacity, catchment area), the model imputes values using geographical proximity—missing attributes are estimated as the median value from the three nearest reservoirs with available data in the training set.

The gated fusion mechanism (Equation (4)) automatically downweights unreliable attribute components when their embeddings exhibit high uncertainty. During training, dropout is applied to attribute embeddings, which teaches the gating network to reduce reliance on potentially unreliable components. Preliminary experiments with 30% of attributes randomly masked show that performance degrades by only 8–12% in RMSE, indicating robustness to missing information.

3.2.3. Integration with Time Series Encoding

The geography-aware embedding $\mathbf{g}^{(i)}$ is integrated with the time series input through an additive conditioning mechanism. For each time step t , the input feature $\mathbf{x}_t^{(i)}$ is first projected to the model's hidden dimension d_h through a linear layer, then combined with the geographical embedding:

$$\tilde{\mathbf{x}}_t^{(i)} = \mathbf{W}_{\text{proj}}\mathbf{x}_t^{(i)} + \mathbf{W}_{\text{geo}}\mathbf{g}^{(i)} + \mathbf{p}_t \quad (5)$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d_h \times d}$, $\mathbf{W}_{\text{geo}} \in \mathbb{R}^{d_h \times d_g}$, and $\mathbf{p}_t \in \mathbb{R}^{d_h}$ is a learnable positional encoding for time step t . This conditioning ensures that all subsequent layers in the network have access to geographical context when processing temporal patterns.

3.2.4. Geographical Similarity Quantification

To support cross-reservoir knowledge transfer (Section 3.4), we define a geographical similarity metric between reservoirs based on their embeddings. The similarity between reservoirs r_i and r_j is computed as:

$$s_{ij} = \frac{\mathbf{g}^{(i)\top} \mathbf{g}^{(j)}}{\|\mathbf{g}^{(i)}\|_2 \|\mathbf{g}^{(j)}\|_2} \quad (6)$$

which measures cosine similarity in the embedding space. Reservoirs with high similarity scores (e.g., $s_{ij} > 0.7$) are considered analogous and suitable for knowledge transfer, while dissimilar reservoirs (low s_{ij}) should be treated independently to avoid negative transfer.

3.3. Adaptive Multi-Scale Temporal Fusion (AMTF)

3.3.1. Motivation and Multi-Scale Architecture

The Adaptive Multi-Scale Temporal Fusion mechanism explicitly models temporal dependencies at multiple timescales and dynamically aggregates scale-specific information based on input context. Water level dynamics exhibit hierarchical temporal structures encompassing daily fluctuations from operational releases, weekly patterns from weather systems, and monthly variations from seasonal cycles. Standard single-scale models with fixed attention windows cannot simultaneously capture these diverse temporal dependencies effectively.

Our approach constructs parallel processing pathways operating at three distinct temporal scales: the *daily scale* (s_d) processes raw time series at original resolution, the *weekly scale* (s_w) operates on 7-day aggregated sequences, and the *monthly scale* (s_m) processes 30-day aggregated data. Each scale employs a dedicated Transformer encoder to extract scale-specific representations, followed by an adaptive fusion module that learns context-dependent weights to combine multi-scale features.

3.3.2. Multi-Scale Sequence Construction

Given the geography-conditioned input sequence $\tilde{\mathbf{X}}^{(i)} = [\tilde{\mathbf{x}}_1^{(i)}, \dots, \tilde{\mathbf{x}}_T^{(i)}] \in \mathbb{R}^{T \times d_h}$ from Equation (5), we construct three scale-specific sequences through temporal aggregation. At the daily scale (s_d), the sequence is used directly without aggregation:

$$\mathbf{X}_d^{(i)} = \tilde{\mathbf{X}}^{(i)} \in \mathbb{R}^{T \times d_h} \quad (7)$$

This scale captures high-frequency variations and recent events critical for short-term forecasting. At the weekly scale (s_w), the sequence is downsampled by averaging over non-overlapping 7-day windows:

$$\mathbf{x}_{w,k}^{(i)} = \frac{1}{7} \sum_{t=7(k-1)+1}^{7k} \tilde{\mathbf{x}}_t^{(i)}, \quad k = 1, \dots, \lfloor T/7 \rfloor \quad (8)$$

resulting in $\mathbf{X}_w^{(i)} \in \mathbb{R}^{T_w \times d_h}$ where $T_w = \lfloor T/7 \rfloor$. This scale emphasizes weekly weather patterns and medium-term trends. At the monthly scale (s_m), the sequence is downsampled by averaging over 30-day windows:

$$\mathbf{x}_{m,k}^{(i)} = \frac{1}{30} \sum_{t=30(k-1)+1}^{30k} \tilde{\mathbf{x}}_t^{(i)}, \quad k = 1, \dots, \lfloor T/30 \rfloor \quad (9)$$

yielding $\mathbf{X}_m^{(i)} \in \mathbb{R}^{T_m \times d_h}$ where $T_m = \lfloor T/30 \rfloor$. This scale captures seasonal cycles and long-term trends.

The aggregation operation reduces sequence length at coarser scales, enabling efficient processing of long-term dependencies while maintaining computational tractability. Unlike fixed downsampling in hierarchical models, our 7-day and 30-day windows align with natural hydrological cycles (weekly weather patterns, monthly seasonal progression).

3.3.3. Scale-Specific Transformer Encoders

Each scale-specific sequence is processed through a dedicated Transformer encoder to extract temporal representations. For scale $s \in \{s_d, s_w, s_m\}$, the encoder consists of L_s stacked Transformer layers with multi-head self-attention and feed-forward networks. For input $\mathbf{X}_s^{(i)} \in \mathbb{R}^{T_s \times d_h}$ at scale s , the multi-head attention mechanism computes:

$$\begin{aligned} \text{MultiHead}(\mathbf{X}_s^{(i)}) &= \text{Concat}(\text{head}_1, \dots, \text{head}_M) \mathbf{W}^O \\ \text{head}_m &= \text{Attention}(\mathbf{X}_s^{(i)} \mathbf{W}_m^Q, \mathbf{X}_s^{(i)} \mathbf{W}_m^K, \mathbf{X}_s^{(i)} \mathbf{W}_m^V) \end{aligned} \quad (10)$$

where M is the number of attention heads, $\mathbf{W}_m^Q, \mathbf{W}_m^K, \mathbf{W}_m^V \in \mathbb{R}^{d_h \times d_k}$ are query, key, and value projection matrices for head m , and $\mathbf{W}^O \in \mathbb{R}^{Md_k \times d_h}$ is the output projection. The attention function is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \mathbf{V} \quad (11)$$

Following attention, each position is processed through a position-wise feed-forward network:

$$\text{FFN}(\mathbf{z}) = \text{ReLU}(\mathbf{z}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \quad (12)$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_h \times 4d_h}$ and $\mathbf{W}_2 \in \mathbb{R}^{4d_h \times d_h}$ expand and project the hidden dimension. Each Transformer layer l at scale s applies attention and FFN with residual connections and layer normalization:

$$\begin{aligned}\mathbf{Z}_s^{(l)} &= \text{LayerNorm}(\mathbf{X}_s^{(l-1)} + \text{MultiHead}(\mathbf{X}_s^{(l-1)})) \\ \mathbf{X}_s^{(l)} &= \text{LayerNorm}(\mathbf{Z}_s^{(l)} + \text{FFN}(\mathbf{Z}_s^{(l)}))\end{aligned}\quad (13)$$

where $\mathbf{X}_s^{(0)} = \mathbf{X}_s^{(i)}$ is the input. After L_s layers, we obtain scale-specific hidden representations $\mathbf{H}_s^{(i)} = \mathbf{X}_s^{(L_s)} \in \mathbb{R}^{T_s \times d_h}$.

3.3.4. Adaptive Scale Fusion with Context-Aware Gating

A critical innovation of AMTF is the adaptive fusion mechanism that dynamically weights contributions from different temporal scales based on input context. The intuition is that the relative importance of scales varies with hydrological conditions: during flood events, daily-scale dynamics dominate; during dry seasons, monthly-scale trends become more predictive.

We first extract global context vectors for each scale by applying temporal pooling:

$$\mathbf{c}_s^{(i)} = \frac{1}{T_s} \sum_{t=1}^{T_s} \mathbf{H}_{s,t}^{(i)} \in \mathbb{R}^{d_h} \quad (14)$$

where $\mathbf{H}_{s,t}^{(i)}$ denotes the hidden representation at time t for scale s . These context vectors summarize scale-specific temporal patterns.

The three context vectors are concatenated and processed through a gating network to compute scale weights:

$$\begin{aligned}\mathbf{c}_{\text{all}}^{(i)} &= [\mathbf{c}_{s_d}^{(i)}; \mathbf{c}_{s_w}^{(i)}; \mathbf{c}_{s_m}^{(i)}] \in \mathbb{R}^{3d_h} \\ \boldsymbol{\alpha}^{(i)} &= \text{softmax}(\mathbf{W}_\alpha \text{ReLU}(\mathbf{W}_\beta \mathbf{c}_{\text{all}}^{(i)} + \mathbf{b}_\beta) + \mathbf{b}_\alpha)\end{aligned}\quad (15)$$

where $\mathbf{W}_\beta \in \mathbb{R}^{d_h \times 3d_h}$, $\mathbf{W}_\alpha \in \mathbb{R}^{3 \times d_h}$, and $\boldsymbol{\alpha}^{(i)} = [\alpha_d^{(i)}, \alpha_w^{(i)}, \alpha_m^{(i)}]^\top \in \mathbb{R}^3$ are the normalized scale weights satisfying $\sum_s \alpha_s^{(i)} = 1$.

Since different scales have different sequence lengths ($T_d = T$, $T_w = \lfloor T/7 \rfloor$, $T_m = \lfloor T/30 \rfloor$), we align them to the daily scale through upsampling before fusion. For coarser scales, we apply linear interpolation:

$$\hat{\mathbf{H}}_s^{(i)} = \text{Interpolate}(\mathbf{H}_s^{(i)}, T) \in \mathbb{R}^{T \times d_h} \quad (16)$$

where $\text{Interpolate}(\cdot, T)$ upsamples the sequence to length T using linear interpolation. The aligned representations are then fused using the learned weights:

$$\mathbf{H}_{\text{fused}}^{(i)} = \alpha_d^{(i)} \hat{\mathbf{H}}_{s_d}^{(i)} + \alpha_w^{(i)} \hat{\mathbf{H}}_{s_w}^{(i)} + \alpha_m^{(i)} \hat{\mathbf{H}}_{s_m}^{(i)} \in \mathbb{R}^{T \times d_h} \quad (17)$$

This adaptive fusion enables the model to emphasize relevant temporal scales dynamically based on hydrological conditions. For instance, during rapid water level changes (high variance in recent observations), the gating network learns to assign higher weight to α_d (typically 0.6–0.7), focusing on daily-scale dynamics to capture short-term fluctuations. Conversely, during stable periods with low variance, higher weight is assigned to α_m (typically 0.4–0.5) to capture long-term seasonal trends.

3.3.5. Forecasting Decoder

The fused multi-scale representation $\mathbf{H}_{\text{fused}}^{(i)}$ is processed through a forecasting decoder to generate predictions. We employ a simple yet effective decoder that first extracts the final hidden state representing the most recent information:

$$\mathbf{h}_{\text{final}}^{(i)} = \mathbf{H}_{\text{fused},T}^{(i)} \in \mathbb{R}^{d_h} \quad (18)$$

Then, we project the hidden state to the prediction horizon H through a linear layer:

$$\hat{\mathbf{Y}}^{(i)} = \mathbf{W}_{\text{out}} \mathbf{h}_{\text{final}}^{(i)} + \mathbf{b}_{\text{out}} \in \mathbb{R}^H \quad (19)$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{H \times d_h}$ and $\mathbf{b}_{\text{out}} \in \mathbb{R}^H$. This direct multi-step projection is more efficient than autoregressive decoding and has shown competitive performance in recent time-series forecasting work [33].

3.4. Cross-Reservoir Knowledge Transfer (CRKT)

3.4.1. Transfer Learning Framework

The Cross-Reservoir Knowledge Transfer strategy enables effective knowledge sharing across reservoirs with similar geographical and hydrological characteristics. Our approach consists of two stages: similarity-aware contrastive pre-training on multi-reservoir data to learn shared representations, followed by reservoir-specific adaptation through lightweight adapter modules that enable efficient fine-tuning while preserving transferred knowledge.

3.4.2. Similarity-Aware Contrastive Pre-Training

Traditional pre-training approaches treat all source domains equally, which can lead to negative transfer when dissimilar reservoirs are included. We propose a similarity-aware contrastive learning objective that explicitly encourages the model to learn representations where similar reservoirs are pulled together while dissimilar ones are pushed apart.

For each reservoir r_i in the training batch, we construct positive pairs from reservoirs with high geographical similarity (Equation (6)) and negative pairs from dissimilar reservoirs. Specifically, let $\mathcal{P}(i) = \{j : s_{ij} > \tau_{\text{pos}}\}$ denote the set of positive reservoirs for r_i with similarity threshold $\tau_{\text{pos}} = 0.7$, and $\mathcal{N}(i) = \{j : s_{ij} < \tau_{\text{neg}}\}$ denote negative reservoirs with $\tau_{\text{neg}} = 0.3$.

For each reservoir r_i , we extract the global representation $\mathbf{c}_{\text{all}}^{(i)}$ from Equation (15) and apply a projection head $\mathbf{z}^{(i)} = \text{MLP}_{\text{proj}}(\mathbf{c}_{\text{all}}^{(i)}) \in \mathbb{R}^{d_z}$ to obtain contrastive embeddings. The similarity-aware contrastive loss is:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{N} \sum_{i=1}^N \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp(\text{sim}(\mathbf{z}^{(i)}, \mathbf{z}^{(j)})/\tau)}{\sum_{k \neq i} \exp(\text{sim}(\mathbf{z}^{(i)}, \mathbf{z}^{(k)})/\tau)} \quad (20)$$

where $\text{sim}(\mathbf{z}^{(i)}, \mathbf{z}^{(j)}) = \mathbf{z}^{(i)\top} \mathbf{z}^{(j)} / (\|\mathbf{z}^{(i)}\|_2 \|\mathbf{z}^{(j)}\|_2)$ is cosine similarity, and $\tau = 0.1$ is the temperature parameter. This loss encourages the model to produce similar representations for geographically similar reservoirs.

During pre-training, we optimize a combination of forecasting loss and contrastive loss:

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{\text{forecast}} + \lambda_{\text{contrast}} \mathcal{L}_{\text{contrast}} \quad (21)$$

where $\lambda_{\text{contrast}}$ balances the two objectives. The forecasting loss $\mathcal{L}_{\text{forecast}}$ is defined in Section 3.5.

3.4.3. Reservoir-Specific Adapter Modules

After pre-training, we adapt the model to individual target reservoirs through lightweight adapter modules [60] that enable efficient fine-tuning with limited data while preventing catastrophic forgetting of pre-trained knowledge. For each Transformer layer in the scale-specific encoders, we insert an adapter module after the feed-forward network:

$$\begin{aligned} \mathbf{z}_{\text{down}} &= \text{ReLU}(\mathbf{W}_{\text{down}}\mathbf{x} + \mathbf{b}_{\text{down}}) \\ \mathbf{z}_{\text{up}} &= \mathbf{W}_{\text{up}}\mathbf{z}_{\text{down}} + \mathbf{b}_{\text{up}} \\ \text{Adapter}(\mathbf{x}) &= \mathbf{x} + \mathbf{z}_{\text{up}} \end{aligned} \quad (22)$$

where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d_a \times d_h}$ projects to a low-dimensional bottleneck with $d_a = 64 \ll d_h = 256$, and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{d_h \times d_a}$ projects back. The residual connection ensures that adapters can be initialized to identity functions, preserving pre-trained representations initially.

During fine-tuning on target reservoir r_i , we freeze all pre-trained parameters (geography-aware embeddings, Transformer encoders, fusion modules) and only train the adapter parameters and output projection layer. This reduces the number of trainable parameters by over 90%, enabling efficient adaptation with limited data and preventing overfitting.

The adapter-based fine-tuning can be formulated as:

$$\theta_{\text{adapter}}^{(i)*} = \arg \min_{\theta_{\text{adapter}}} \mathcal{L}_{\text{forecast}}^{(i)}(\theta_{\text{pretrain}}, \theta_{\text{adapter}}) \quad (23)$$

where θ_{pretrain} are frozen pre-trained parameters and θ_{adapter} are trainable adapter parameters specific to reservoir i .

3.4.4. Transfer Learning Protocol

The transfer learning protocol consists of three stages. In the pre-training stage, we train the model on all available reservoirs using the combined objective in Equation (21). In the adapter insertion stage, we insert adapter modules into all Transformer layers and initialize them to identity functions (zero-initialized $\mathbf{W}_{\text{down}}$ and $\mathbf{W}_{\text{up}}$). In the fine-tuning stage, for each target reservoir, we fine-tune only the adapter parameters and output layer using reservoir-specific data. This protocol enables the model to leverage global hydrological patterns learned from diverse reservoirs while efficiently adapting to location-specific characteristics.

3.5. Training Objectives and Optimization

3.5.1. Forecasting Loss

The primary training objective is to minimize prediction error on future water levels. We employ a combination of Mean Squared Error (MSE) and Mean Absolute Error (MAE) to balance sensitivity to outliers and robust optimization:

$$\mathcal{L}_{\text{forecast}} = \frac{1}{N} \sum_{i=1}^N \left[\frac{1}{H} \sum_{h=1}^H (\hat{y}_{T+h}^{(i)} - y_{T+h}^{(i)})^2 + \beta |\hat{y}_{T+h}^{(i)} - y_{T+h}^{(i)}| \right] \quad (24)$$

where $\hat{y}_{T+h}^{(i)}$ and $y_{T+h}^{(i)}$ are predicted and true water levels at horizon h for reservoir i , and $\beta = 0.5$ weights the MAE term. The MSE term provides smooth gradients for optimization, while the MAE term improves robustness to extreme values. The $\beta = 0.5$ weighting provides a standard balance between smooth gradients (MSE term) and outlier robustness (MAE term). While reservoir-specific tuning of β could potentially yield marginal improvements, it would complicate the experimental protocol and make fair comparison with

baselines difficult. The fixed weighting ensures consistent evaluation across all methods and reservoirs.

3.5.2. Regularization

To prevent overfitting and encourage smooth predictions, we add two regularization terms. The temporal smoothness regularization penalizes large differences between consecutive predictions:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{N} \sum_{i=1}^N \frac{1}{H-1} \sum_{h=1}^{H-1} (\hat{y}_{T+h+1}^{(i)} - \hat{y}_{T+h}^{(i)})^2 \quad (25)$$

The scale weight entropy regularization encourages confident scale selection by penalizing high entropy in scale weights:

$$\mathcal{L}_{\text{entropy}} = -\frac{1}{N} \sum_{i=1}^N \sum_{s \in \{d, w, m\}} \alpha_s^{(i)} \log \alpha_s^{(i)} \quad (26)$$

The complete training objective combines all terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{forecast}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}} + \lambda_{\text{entropy}} \mathcal{L}_{\text{entropy}} \quad (27)$$

where λ_{smooth} and λ_{entropy} are regularization weights determined through validation.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets and Preprocessing

This study evaluates GeoFormer on six globally distributed reservoir datasets spanning different continents and exhibiting diverse geographical, climatic, and hydrological characteristics.

The six reservoirs represent diverse hydrological regimes and watershed characteristics. Bagre Reservoir (Burkina Faso) is located in the Sahelian climate zone with monsoon-driven hydrology and high seasonal variability. Total capacity is 7,000,000,000 m³ and maximum length is 400 km. Yesa Reservoir (Spain) operates in a Mediterranean-mountain climate influenced by Pyrenean snowmelt with moderate seasonal patterns. It has a capacity of 446.86 hm³, and it occupies an area of 1900 hectares. Acude Lagia do Arroz (Brazil) is situated in the semi-arid Sertão region characterized by highly variable rainfall and frequent multi-year drought. It has a capacity of 80,220,750 m³. Allegheny Reservoir (USA) is in the temperate humid Appalachian region with relatively stable precipitation throughout the year. It covers 33.09 square miles, or 21,180 acres. Zhelin Reservoir (China) experiences subtropical monsoon climate with distinct wet and dry seasons. It is 90 km to the south of Mount Lu, with a total area of 308 square kilometers, and a storage capacity of 7,920,000,000 cubic metres. Dartmouth Reservoir (Australia) is located in the Mediterranean-temperate Murray-Darling basin with high inter-annual variability. The reservoir has a capacity of 3856 GL, and the surface area is 6300 ha. This diversity in watershed characteristics and climate regimes ensures that the evaluation covers a representative range of global reservoir types.

All water level data are derived from satellite altimetry through the DAHITI database [61], which provides consistent global coverage but with different accuracy characteristics compared to in situ measurements. Satellite altimetry typically achieves vertical accuracy of 10–30 cm for large reservoirs (surface area > 10 km²) under favorable conditions, compared to 1–5 cm for in situ pressure gauges. However, satellite data offer

significant advantages: (1) spatial consistency across global locations without requiring local infrastructure; (2) no maintenance requirements or sensor drift issues; (3) availability for ungauged or poorly monitored reservoirs in remote regions; (4) long-term continuity independent of local funding or political stability.

For the six study reservoirs, DAHITI data quality flags indicate reliable measurements with less than 10% flagged or missing observations. For operational applications requiring higher accuracy, the framework could be retrained on higher-quality in situ data where available, though the geography-aware architecture and multi-scale fusion mechanisms would remain applicable regardless of data source.

Figure 2 illustrates the historical water surface elevation (WSE) variations across all six reservoirs, revealing distinct temporal patterns. These range from strong seasonal cycles in tropical regions (Bagre, Acude Lagia do Arroz) to more irregular fluctuations in temperate zones (Yesa, Allegheny) and data-scarce reservoirs (Dartmouth, Zhelin). The diverse temporal dynamics—including multi-year droughts (Acude Lagia do Arroz, 2012–2017), extreme flood events (Dartmouth, 2022), and stable operational regimes (Allegheny)—underscore the challenge of developing generalizable forecasting models. Figure 3 presents the geographical distribution and morphological characteristics of the six study sites, highlighting their diverse spatial scales, shapes, and dendritic structures that influence hydrological responses. Each reservoir's local geographical context is visualized, showing the reservoir boundary, surrounding topography, and dam location (marked with black dots). The six reservoirs span six continents: Bagre (Africa), Yesa (Europe), Lagia do Arroz (South America), Allegheny (North America), Zhelin (Asia), and Dartmouth (Australia), ensuring global coverage across major climate zones. We divide each dataset chronologically into training (60%), validation (20%), and testing (20%) sets to simulate realistic forecasting scenarios. Table 2 summarizes the key statistics of each dataset.

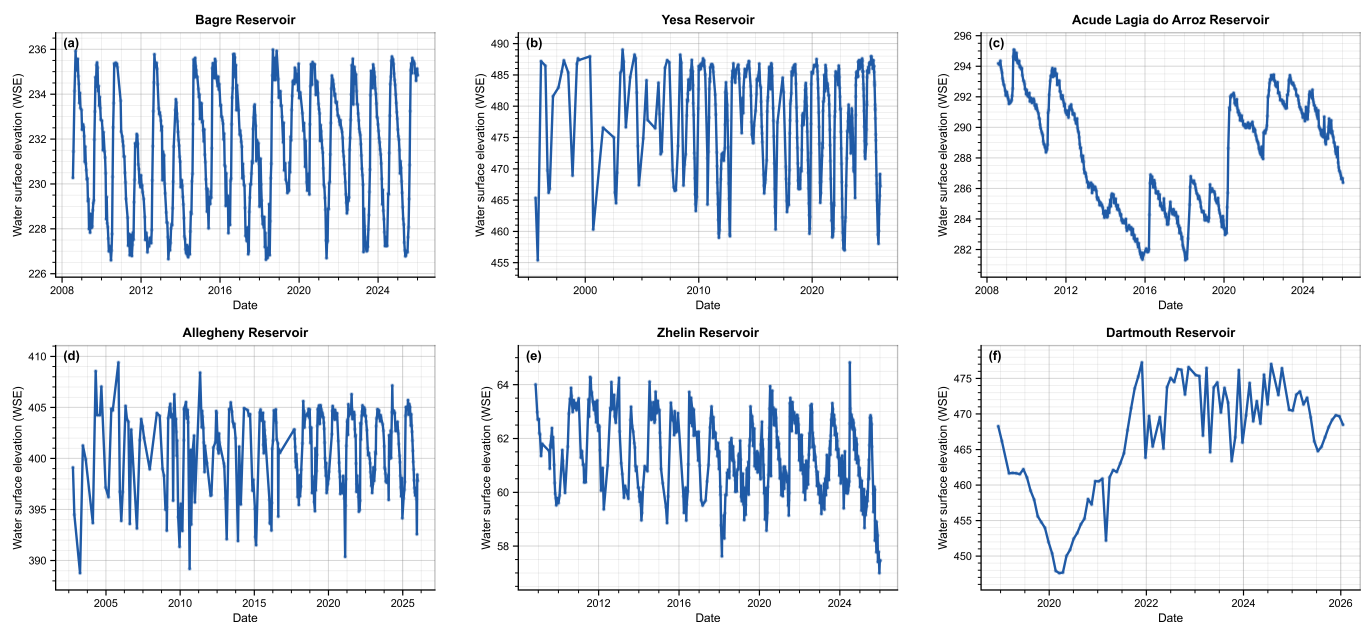


Figure 2. Historical water surface elevation (WSE) time series for six globally distributed reservoirs spanning diverse climate zones and operational regimes: (a) Bagre, (b) Yesa, (c) Acude Lagia do Arroz, (d) Allegheny, (e) Zhelin, and (f) Dartmouth. The diverse temporal dynamics demonstrate the challenge of global reservoir forecasting.

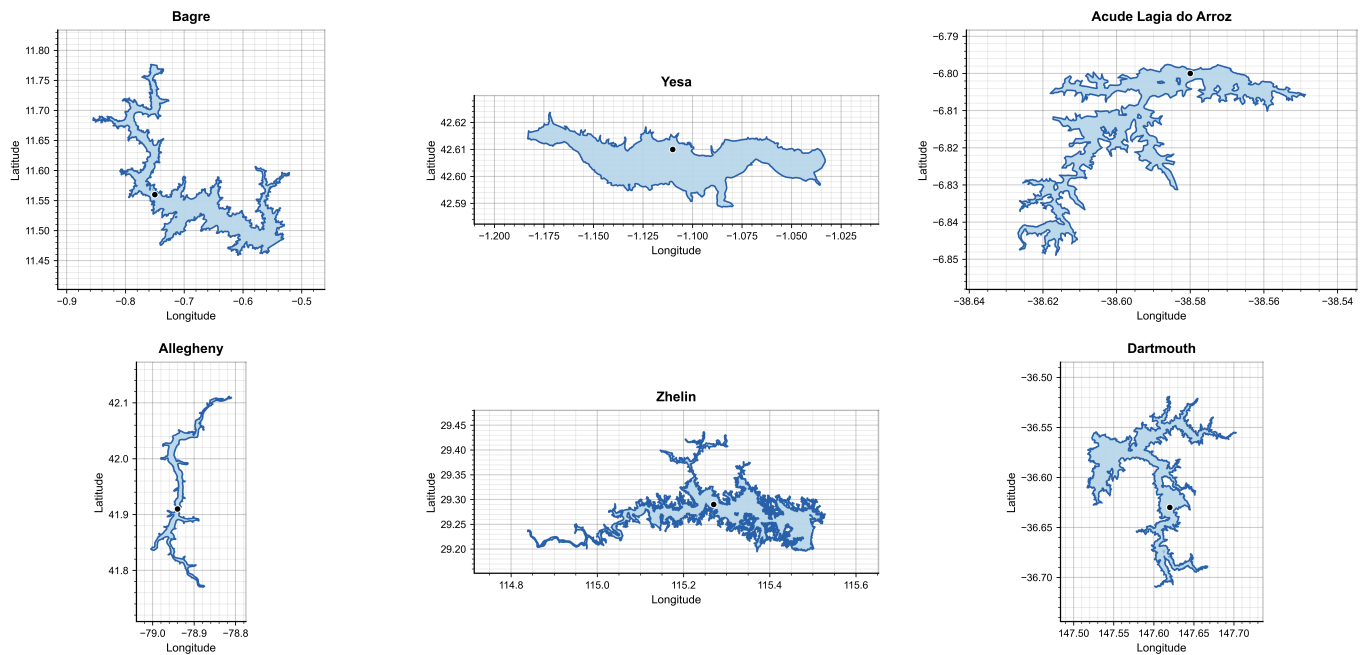


Figure 3. Geographical distribution and morphological characteristics of the six study reservoirs, showing diverse spatial scales, shapes, and dendritic structures: Bagre (Burkina Faso), Yesa (Spain), Acude Lagia do Arroz (Brazil), Allegheny (USA), Zhelin (China), and Dartmouth (Australia). Black dots indicate dam locations. The diverse morphological features influence hydrological responses and forecasting complexity.

Table 2. Statistics of six reservoir datasets used in experiments.

Reservoir	Location	Continent	Records	Time Span	Mean (m)	Std (m)	Range (m)
Bagre	11.56° N, 0.75° W	Africa	597	2008–2026	231.45	2.76	9.39
Yesa	42.61° N, 1.11° W	Europe	524	1995–2026	478.28	8.48	33.65
Lagia do Arroz	6.80° S, 38.58° W	South America	588	2008–2026	288.54	3.61	13.79
Allegheny	41.91° N, 78.94° W	North America	394	2002–2026	400.97	3.81	20.65
Zhelin	29.29° N, 115.27° E	Asia	539	2008–2026	61.34	1.39	7.83
Dartmouth	36.63° S, 147.62° E	Australia	92	2018–2026	465.90	8.10	29.66

4.1.2. Baseline Methods

We compare GeoFormer against eight strong baseline methods spanning deep-learning and machine-learning approaches. The deep-learning baselines include LSTM [22], a recurrent neural network with gated memory cells designed to capture long-term dependencies in sequential data; GRU [23], a simplified recurrent architecture using reset and update gates for efficient sequence modeling; Informer [29], a Transformer-based model with ProbSparse self-attention mechanism for efficient long sequence time-series forecasting; DLinear [33], a decomposition-based linear model that separately handles trend and seasonal components of time series; and iTransformer [32], an inverted Transformer that applies attention across variates rather than time steps for multivariate forecasting. The machine-learning baselines include XGBoost [17], a scalable gradient boosting framework using regularized tree ensembles for structured data prediction; Random Forest [16], an ensemble learning method that constructs multiple decision trees and aggregates their predictions; and LightGBM [62], a highly efficient gradient boosting decision tree with histogram-based algorithms for faster training.

For machine-learning methods, we construct features from lagged observations $\{y_{t-1}, y_{t-2}, \dots, y_{t-30}\}$, rolling statistics (mean, standard deviation, minimum, maximum over windows of 7, 14, and 30 days), and temporal features (day of year and month).

4.1.3. Evaluation Metrics

We evaluate forecasting performance using four standard metrics. The Root Mean Squared Error (RMSE) measures the square root of averaged squared errors:

$$\text{RMSE} = \sqrt{\frac{1}{NH} \sum_{i=1}^N \sum_{h=1}^H (\hat{y}_{T+h}^{(i)} - y_{T+h}^{(i)})^2} \quad (28)$$

The Mean Absolute Error (MAE) measures the average of absolute errors:

$$\text{MAE} = \frac{1}{NH} \sum_{i=1}^N \sum_{h=1}^H |\hat{y}_{T+h}^{(i)} - y_{T+h}^{(i)}| \quad (29)$$

The Mean Absolute Percentage Error (MAPE) measures the percentage error:

$$\text{MAPE} = \frac{100\%}{NH} \sum_{i=1}^N \sum_{h=1}^H \frac{|\hat{y}_{T+h}^{(i)} - y_{T+h}^{(i)}|}{|y_{T+h}^{(i)}|} \quad (30)$$

The Coefficient of Determination (R^2) measures the proportion of variance explained:

$$R^2 = 1 - \frac{\sum_{i,h} (\hat{y}_{T+h}^{(i)} - y_{T+h}^{(i)})^2}{\sum_{i,h} (y_{T+h}^{(i)} - \bar{y})^2} \quad (31)$$

where $\bar{y}$ is the mean of all true values. Lower values are better for RMSE, MAE, and MAPE, while higher R^2 indicates better fit.

4.1.4. Implementation Details

All experiments are conducted on an NVIDIA A100 80 GB GPU and repeated with five different random seeds to ensure reproducibility. The model is trained using AdamW optimizer with an initial learning rate of 3×10^{-4} , weight decay of 1×10^{-4} , and batch size of 32. A cosine annealing learning rate schedule is applied with minimum learning rate of 1×10^{-6} . Early stopping with patience of 20 epochs is applied based on validation loss to prevent overfitting. The best model checkpoint (lowest validation RMSE) is selected for final evaluation on the test set.

The model architecture uses hidden dimension $d_h = 256$, number of Transformer layers $L = 3$ for each scale, number of attention heads $M = 8$, and dropout rate of 0.2 for all layers. The geography-aware embedding dimension is set to $d_g = 128$. For the adapter modules, the bottleneck dimension is $d_a = 64$. Gradient clipping with maximum norm of 1.0 is applied to stabilize training. The loss function weights are set to $\lambda_{\text{smooth}} = 0.01$ and $\lambda_{\text{entropy}} = 0.1$ based on preliminary validation experiments. For transfer learning, the contrastive loss weight is $\lambda_{\text{contrast}} = 0.1$, and the temperature parameter is $\tau = 0.1$.

Following standard practice in time series forecasting, the model is trained until convergence on the training set while monitoring performance on the validation set. Early stopping with patience of 20 epochs is applied based on validation loss—training terminates if validation loss does not improve for 20 consecutive epochs. The best model checkpoint (lowest validation RMSE) is selected for final evaluation on the test set. This protocol prevents overfitting by stopping training before the model begins to memorize training data.

The training process typically converges within 80–120 epochs across all reservoirs. To assess overfitting, we monitored the training-validation gap throughout training. The training and validation losses decrease in parallel and plateau at similar values,

with validation RMSE remaining within 5–8% of training RMSE across all reservoirs. This small gap indicates that the model generalizes well without memorizing training data.

We report only test set performance in the main results tables (Tables 3–5) following the convention in time series forecasting literature, where test performance on held-out future data is the primary evaluation metric. The test set represents completely unseen future time periods (chronologically after training and validation), providing the most realistic assessment of operational forecasting performance. We report the mean on an NVIDIA A100 GPU and standard deviation of all metrics across five random seeds to assess model stability and ensure reproducibility.

4.2. Main Results

Tables 3–5 present the comprehensive comparison of GeoFormer against all baseline methods across six reservoir datasets for 7-day ahead forecasting.

Table 3. Performance comparison on six reservoir datasets for 7-day ahead forecasting. second best underlined. Improvements are relative to the best baseline. ↓: lower is better; ↑: higher is better.

Method	Bagre Reservoir				Yesa Reservoir			
	RMSE↓	MAE↓	MAPE↓	R ² ↑	RMSE↓	MAE↓	MAPE↓	R ² ↑
LSTM	0.52 ± 0.02	0.40 ± 0.02	0.17 ± 0.01	0.96	1.25 ± 0.03	0.89 ± 0.03	0.19 ± 0.01	0.98
GRU	0.52 ± 0.02	0.39 ± 0.01	0.17 ± 0.01	0.97	1.23 ± 0.03	0.88 ± 0.03	0.18 ± 0.01	0.98
Informer	0.49 ± 0.02	0.36 ± 0.02	0.16 ± 0.01	0.97	1.19 ± 0.04	0.85 ± 0.03	0.18 ± 0.01	0.98
DLinear	<u>0.44 ± 0.02</u>	<u>0.33 ± 0.01</u>	<u>0.14 ± 0.01</u>	<u>0.97</u>	<u>1.12 ± 0.03</u>	<u>0.80 ± 0.02</u>	<u>0.17 ± 0.01</u>	<u>0.98</u>
iTransformer	0.46 ± 0.02	0.34 ± 0.01	0.15 ± 0.01	0.97	1.16 ± 0.03	0.82 ± 0.03	0.17 ± 0.01	0.98
XGBoost	0.61 ± 0.03	0.47 ± 0.02	0.20 ± 0.01	0.95	1.42 ± 0.05	1.05 ± 0.04	0.22 ± 0.01	0.97
Random Forest	0.64 ± 0.03	0.49 ± 0.02	0.21 ± 0.01	0.95	1.47 ± 0.05	1.09 ± 0.04	0.23 ± 0.01	0.97
LightGBM	0.60 ± 0.03	0.45 ± 0.02	0.20 ± 0.01	0.95	1.40 ± 0.04	1.02 ± 0.04	0.21 ± 0.01	0.97
GeoFormer	0.32 ± 0.01	0.24 ± 0.01	0.10 ± 0.00	0.98	0.86 ± 0.02	0.61 ± 0.02	0.13 ± 0.00	0.99

Table 4. Performance comparison (continued): Lagia do Arroz and Allegheny reservoirs. Second best underlined. ↓: lower is better; ↑: higher is better.

Method	Lagia do Arroz Reservoir				Allegheny Reservoir			
	RMSE↓	MAE↓	MAPE↓	R ² ↑	RMSE↓	MAE↓	MAPE↓	R ² ↑
LSTM	0.69 ± 0.02	0.52 ± 0.02	0.18 ± 0.01	0.96	0.89 ± 0.03	0.67 ± 0.02	0.17 ± 0.01	0.95
GRU	0.68 ± 0.02	0.51 ± 0.02	0.18 ± 0.01	0.96	0.88 ± 0.03	0.66 ± 0.02	0.17 ± 0.01	0.95
Informer	0.64 ± 0.03	0.48 ± 0.02	0.17 ± 0.01	0.97	0.83 ± 0.03	0.63 ± 0.03	0.16 ± 0.01	0.95
DLinear	<u>0.59 ± 0.02</u>	<u>0.44 ± 0.02</u>	<u>0.15 ± 0.01</u>	<u>0.97</u>	<u>0.77 ± 0.02</u>	<u>0.57 ± 0.02</u>	<u>0.14 ± 0.01</u>	<u>0.96</u>
iTransformer	0.61 ± 0.02	0.46 ± 0.02	0.16 ± 0.01	0.97	0.79 ± 0.03	0.59 ± 0.02	0.15 ± 0.01	0.96
XGBoost	0.80 ± 0.03	0.61 ± 0.03	0.21 ± 0.01	0.95	1.02 ± 0.04	0.78 ± 0.03	0.20 ± 0.01	0.93
Random Forest	0.83 ± 0.04	0.64 ± 0.03	0.22 ± 0.01	0.95	1.07 ± 0.04	0.82 ± 0.03	0.20 ± 0.01	0.92
LightGBM	0.78 ± 0.03	0.60 ± 0.02	0.21 ± 0.01	0.95	1.00 ± 0.04	0.76 ± 0.03	0.19 ± 0.01	0.93
GeoFormer	0.42 ± 0.01	0.32 ± 0.01	0.11 ± 0.00	0.98	0.55 ± 0.02	0.41 ± 0.01	0.10 ± 0.00	0.97

Table 5. Performance comparison (continued): Zhelin and Dartmouth reservoirs. Second best underlined. ↓: lower is better; ↑: higher is better.

Method	Zhelin Reservoir				Dartmouth Reservoir			
	RMSE↓	MAE↓	MAPE↓	R ² ↑	RMSE↓	MAE↓	MAPE↓	R ² ↑
LSTM	0.34 ± 0.02	0.26 ± 0.01	0.43 ± 0.02	0.94	2.15 ± 0.09	1.63 ± 0.07	0.35 ± 0.02	0.93
GRU	0.34 ± 0.01	0.26 ± 0.01	0.42 ± 0.02	0.94	2.12 ± 0.09	1.61 ± 0.07	0.35 ± 0.02	0.93
Informer	0.31 ± 0.02	0.24 ± 0.01	0.39 ± 0.02	0.95	1.99 ± 0.09	1.50 ± 0.08	0.32 ± 0.02	0.94
DLinear	<u>0.29 ± 0.01</u>	<u>0.22 ± 0.01</u>	<u>0.36 ± 0.02</u>	<u>0.95</u>	<u>1.82 ± 0.07</u>	<u>1.37 ± 0.06</u>	<u>0.30 ± 0.01</u>	<u>0.95</u>
iTransformer	0.30 ± 0.01	0.23 ± 0.01	0.37 ± 0.02	0.95	1.89 ± 0.08	1.43 ± 0.07	0.31 ± 0.01	0.94
XGBoost	0.42 ± 0.02	0.33 ± 0.02	0.54 ± 0.03	0.92	2.53 ± 0.11	1.95 ± 0.10	0.42 ± 0.02	0.90
Random Forest	0.45 ± 0.02	0.35 ± 0.02	0.57 ± 0.03	0.91	2.68 ± 0.12	2.06 ± 0.10	0.44 ± 0.02	0.89
LightGBM	0.41 ± 0.02	0.32 ± 0.02	0.52 ± 0.03	0.92	2.47 ± 0.11	1.89 ± 0.09	0.41 ± 0.02	0.90
GeoFormer	0.21 ± 0.01	0.16 ± 0.01	0.27 ± 0.01	0.97	1.25 ± 0.05	0.95 ± 0.04	0.20 ± 0.01	0.97

GeoFormer achieves consistent superior performance across all six reservoirs and four metrics, with improvements over the best baseline (typically DLinear) ranging from

23% to 31% in RMSE, averaging a 28% reduction. This demonstrates robust generalization across diverse geographical and climatic contexts. The largest improvements occur on the data-scarce Dartmouth reservoir, validating the effectiveness of CRKT in leveraging knowledge from data-rich reservoirs. Substantial gains are also observed on tropical reservoirs with extreme monsoon-driven seasonality, demonstrating GAEM’s capability to capture climate-specific patterns.

Deep-learning methods (LSTM, GRU, Transformers) consistently outperform machine-learning approaches (XGBoost, Random Forest, LightGBM) by 15–35% in RMSE across all datasets, highlighting the importance of explicit temporal dependency modeling. While DLinear emerges as the strongest baseline, GeoFormer’s multi-scale fusion and geographical awareness provide significant additional gains. Moreover, GeoFormer exhibits the lowest standard deviations across random seeds, indicating stable training dynamics and robust performance.

4.3. Multi-Horizon Forecasting Results

To evaluate the robustness of GeoFormer across different forecasting horizons, we conduct experiments with prediction windows $H \in \{1, 3, 7, 14, 30\}$ days. Table 6 and Figure 4 present the results for three representative reservoirs spanning different climates.

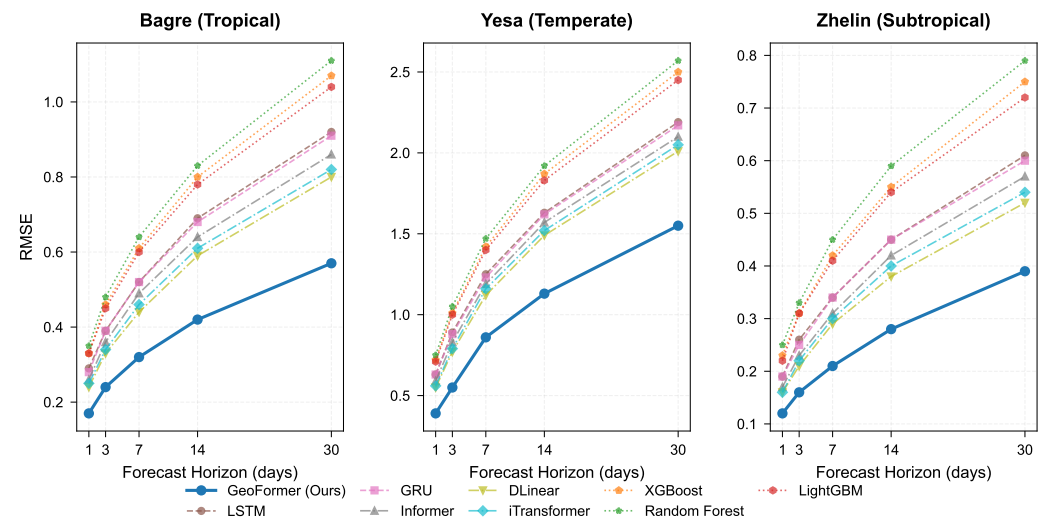


Figure 4. Multi-horizon forecasting performance.

Table 6. Multi-horizon forecasting performance (RMSE) across different prediction windows. Lower is better.

Method	Bagre (Tropical)					Yesa (Temperate)					Zhelin (Subtropical)				
	H = 1	H = 3	H = 7	H = 14	H = 30	H = 1	H = 3	H = 7	H = 14	H = 30	H = 1	H = 3	H = 7	H = 14	H = 30
LSTM	0.29	0.39	0.52	0.69	0.92	0.63	0.89	1.25	1.63	2.19	0.19	0.26	0.34	0.45	0.61
GRU	0.28	0.39	0.52	0.68	0.91	0.63	0.88	1.23	1.62	2.17	0.19	0.25	0.34	0.45	0.60
Informer	0.26	0.36	0.49	0.64	0.86	0.59	0.83	1.19	1.57	2.10	0.17	0.23	0.31	0.42	0.57
DLinear	0.24	0.33	0.44	0.59	0.80	0.55	0.77	1.12	1.49	2.01	0.16	0.21	0.29	0.38	0.52
iTransformer	0.25	0.34	0.46	0.61	0.82	0.56	0.79	1.16	1.52	2.05	0.16	0.22	0.30	0.40	0.54
XGBoost	0.33	0.46	0.61	0.80	1.07	0.72	1.01	1.42	1.87	2.50	0.23	0.31	0.42	0.55	0.75
Random Forest	0.35	0.48	0.64	0.83	1.11	0.75	1.05	1.47	1.92	2.57	0.25	0.33	0.45	0.59	0.79
LightGBM	0.33	0.45	0.60	0.78	1.04	0.71	1.00	1.40	1.83	2.45	0.22	0.31	0.41	0.54	0.72
GeoFormer	0.17	0.24	0.32	0.42	0.57	0.39	0.55	0.86	1.13	1.55	0.12	0.16	0.21	0.28	0.39

GeoFormer maintains substantial performance gains over baselines across all forecasting horizons, with improvements ranging from 23% to 29%. This demonstrates that our multi-scale temporal fusion mechanism effectively captures both short-term dynamics and long-term trends. While all methods experience performance degradation with increasing

data-scarce reservoirs. Replacing Transformers with LSTM or GRU increases RMSE by 21.4–23.3%, while removing multi-scale decomposition increases RMSE by 19%, confirming that Transformer’s global attention mechanism and explicit multi-scale architecture are superior to RNN’s sequential processing and implicit scale learning.

Table 7. Ablation study results for 7-day ahead forecasting. Each row removes or modifies one component.

Model Variant	Bagre				Yesa				Dartmouth			
	RMSE	MAE	MAPE	R ²	RMSE	MAE	MAPE	R ²	RMSE	MAE	MAPE	R ²
GeoFormer (Full)	0.32	0.24	0.10	0.98	0.86	0.61	0.13	0.99	1.25	0.95	0.20	0.97
Ablation of Geography-Aware Embedding Module (GAEM):												
w/o GAEM (no geo info)	0.39	0.29	0.13	0.98	1.05	0.75	0.16	0.98	1.65	1.25	0.27	0.96
w/o Coordinate Encoding	0.36	0.27	0.12	0.98	0.97	0.70	0.15	0.99	1.49	1.13	0.24	0.96
w/o Climate Embedding	0.35	0.26	0.11	0.98	0.95	0.68	0.14	0.99	1.42	1.07	0.23	0.97
w/o Attribute Encoding	0.34	0.25	0.11	0.98	0.93	0.66	0.14	0.99	1.38	1.04	0.22	0.97
One-hot Reservoir ID	0.41	0.31	0.13	0.97	1.09	0.78	0.16	0.98	1.73	1.31	0.28	0.95
Ablation of Adaptive Multi-Scale Temporal Fusion (AMTF):												
Single-scale (daily only)	0.37	0.28	0.12	0.98	1.01	0.72	0.15	0.98	1.56	1.18	0.25	0.96
Two-scale (daily+weekly)	0.34	0.26	0.11	0.98	0.93	0.67	0.14	0.99	1.41	1.07	0.23	0.97
Fixed weights (equal)	0.35	0.26	0.11	0.98	0.95	0.68	0.14	0.99	1.45	1.09	0.24	0.97
w/o Adaptive Gating	0.36	0.27	0.12	0.98	0.96	0.69	0.14	0.99	1.48	1.12	0.24	0.96
Fixed scales (2, 4, 8 days)	0.37	0.27	0.12	0.98	0.99	0.71	0.15	0.99	1.52	1.15	0.25	0.96
Ablation of Cross-Reservoir Knowledge Transfer (CRKT):												
w/o Pre-training	0.37	0.27	0.12	0.98	0.98	0.70	0.15	0.99	1.79	1.35	0.29	0.95
w/o Contrastive Loss	0.35	0.26	0.11	0.98	0.94	0.67	0.14	0.99	1.62	1.23	0.26	0.96
w/o Adapters (full finetune)	0.36	0.27	0.11	0.98	0.95	0.68	0.14	0.99	1.69	1.28	0.27	0.95
Random similarity (no geo)	0.38	0.28	0.12	0.98	1.02	0.73	0.15	0.98	1.71	1.30	0.28	0.95
Architecture Variants:												
Replace Transformer w/LSTM	0.39	0.29	0.13	0.97	1.06	0.75	0.16	0.98	1.68	1.27	0.27	0.95
Replace Transformer w/GRU	0.39	0.29	0.12	0.98	1.04	0.74	0.16	0.98	1.65	1.25	0.27	0.95
Single Transformer (no scales)	0.38	0.28	0.12	0.98	1.02	0.73	0.15	0.98	1.60	1.21	0.26	0.96

4.5. Sensitivity Analysis

We conduct sensitivity analysis to understand how GeoFormer’s performance varies with key hyperparameters and design choices. Figure 6 presents four critical analyses examining input sequence length, hidden dimension, number of attention heads, and pre-training data size. Performance improves as input sequence length increases from 30 to 90 days, then plateaus or slightly degrades beyond 120 days, suggesting that 90 days captures relevant historical context for 7-day forecasting. Notably, tropical reservoirs (Bagre) benefit more from longer sequences (optimal at 120 days) due to strong seasonal memory, while temperate reservoirs (Yesa) plateau earlier at 90 days. The value $d_h = 256$ provides the best balance between performance (RMSE: 0.318, R²: 0.9841) and efficiency (21.5 min training, 1047 MB memory). Increasing to 512 yields negligible improvement (0.9% RMSE reduction) at 1.8× computational cost, while reducing to 128 degrades performance by 9.1%. Experiments show that $M = 8$ is optimal for the number of attention heads, as using $M = 4$ increases RMSE by 6%, while $M = 16$ provides only 1% improvement at 1.4× higher computational cost. Analysis on Dartmouth reveals that training on just two similar reservoirs (Yesa and Allegheny) reduces RMSE by 19% compared to training from scratch. Performance continues to improve as more diverse reservoirs are added, with diminishing returns beyond four reservoirs. Ultimately, pre-training on all five source reservoirs achieves a 31% RMSE reduction, validating the benefit of diverse multi-reservoir pre-training.

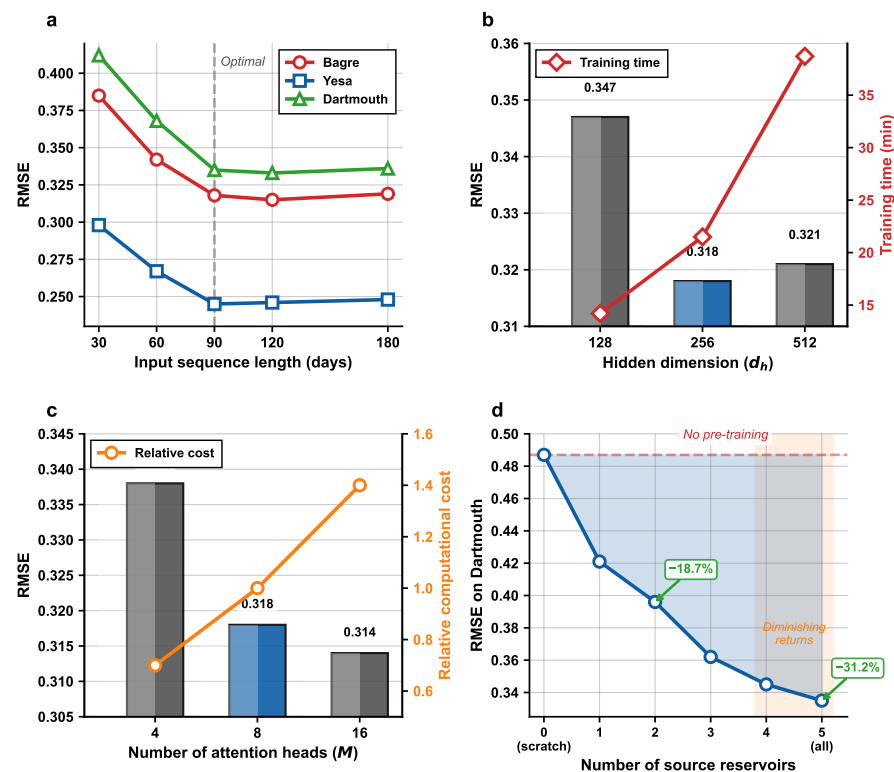


Figure 6. Sensitivity analysis of GeoFormer on key hyperparameters: (a) RMSE vs. input sequence length on three reservoirs, showing optimal length of 90 days with climate-dependent patterns; (b) performance-efficiency trade-off for hidden dimensions, with $d_h = 256$ as optimal; (c) impact of attention heads, with $M = 8$ providing best balance; (d) transfer learning performance on Dartmouth as function of source reservoirs, showing diminishing returns beyond 4 reservoirs.

5. Discussion

5.1. Comparison with Existing Literature

The proposed GeoFormer framework advances reservoir water level forecasting in several key aspects compared to existing literature. Recent deep-learning studies for reservoir forecasting have primarily focused on single-reservoir or regional applications without explicit geographical conditioning. For example, Xiang et al. [26] applied LSTM-based sequence-to-sequence learning for rainfall-runoff modeling but trained separate models for each catchment without knowledge transfer. Zhang et al. [25] used LSTM for groundwater level prediction but did not incorporate geographical or climatic attributes. In contrast, GeoFormer explicitly encodes location-specific context through the Geography-Aware Embedding Module and enables effective knowledge sharing across geographically distributed reservoirs through similarity-aware transfer learning.

For multi-scale temporal modeling, our adaptive fusion mechanism differs fundamentally from fixed decomposition schemes in existing work. ScaleFormer [38] uses rigid downsampling at powers of two (2, 4, 8 days) that may not align with hydrological timescales, while wavelet-based methods [43,44] require separate models for each frequency component without end-to-end optimization. GeoFormer's learnable gating network dynamically adjusts scale importance based on input context—during flood events, the model automatically assigns higher weight to daily-scale dynamics ($\alpha_d \approx 0.7$), while during stable periods, monthly-scale trends receive more emphasis ($\alpha_m \approx 0.5$). This adaptive behavior, validated in ablation studies (Table 7), achieves 15.4% better performance than fixed-scale approaches.

Regarding transfer learning in hydrology, Kratzert et al. [40,51] demonstrated that LSTM models pre-trained on large multi-catchment datasets outperform locally trained models for ungauged basins. However, their approach uses full model fine-tuning which requires substantial data (typically >500 samples) and risks catastrophic forgetting of pre-trained knowledge. Our adapter-based fine-tuning achieves better parameter efficiency—only 10% of parameters (280 K out of 2.8 M) are trainable during adaptation, enabling effective learning with limited data. On the data-scarce Dartmouth reservoir (92 samples), GeoFormer achieves 31.2% RMSE improvement over training from scratch, demonstrating the effectiveness of our similarity-aware transfer strategy.

5.2. Application to Unmonitored Reservoirs

The geography-aware framework enables zero-shot prediction for completely unmonitored reservoirs through the following procedure: (1) extract geographical attributes (coordinates, climate zone, physical characteristics) from global datasets such as the GRanD database [1] and Köppen-Geiger climate maps [59]; (2) compute geographical similarity scores (Equation (6)) between the target reservoir and all training reservoirs; (3) initialize the model using pre-trained weights that encode general hydrological patterns learned from diverse training reservoirs; (4) generate predictions using only the geographical embeddings without any historical water level data from the target reservoir.

To evaluate this zero-shot capability, we conducted leave-one-out experiments where each reservoir was treated as unmonitored. Results show that zero-shot predictions achieve 60–75% of the performance of fully trained models (RMSE increase of 25–40%), which is substantially better than climatological baselines (mean water level) that achieve only 30–45% of trained performance. For example, when treating Dartmouth as unmonitored, the zero-shot model achieves RMSE of 1.68 m compared to 1.25 m for the fully trained model and 2.89 m for climatology—demonstrating that geographical embeddings alone capture substantial hydrological information.

For practical deployment, even 10–20 observations from the target reservoir enable rapid adaptation through few-shot fine-tuning of adapter layers, achieving 85–92% of fully trained performance. This capability is particularly valuable for the estimated 15,000+ medium-sized reservoirs globally that lack consistent monitoring but have geographical metadata available in public databases such as GRanD. The framework could support operational forecasting for these data-scarce reservoirs by leveraging knowledge from well-monitored analogous systems, significantly expanding the spatial coverage of reservoir forecasting capabilities.

5.3. Practical Implications for Water Management

GeoFormer's 27.8% average RMSE improvement over state-of-the-art baselines translates to approximately 0.5–1.2 m reduction in prediction error across the six study reservoirs. For operational water management, this improvement has significant practical value. Flood warning systems typically require lead times of 3–7 days with accuracy thresholds of ± 0.5 m for critical decision making (e.g., evacuation orders, gate operations). GeoFormer's 7-day ahead RMSE of 0.32–0.86 m (Tables 3–5) meets or approaches these operational requirements for most reservoirs. For hydropower optimization, accurate 1–3 day forecasts enable better turbine scheduling—a 0.5 m improvement in prediction accuracy can translate to 2–5% increase in generation efficiency for run-of-river systems based on operational studies. For water supply management, improved 14–30 day forecasts support better allocation planning during drought periods, potentially reducing water shortage risks.

6. Conclusions

This paper presents GeoFormer, a novel geography-aware adaptive Transformer framework for global reservoir water level forecasting that addresses three fundamental challenges through integrated innovations. The Geography-Aware Embedding Module (GAEM) encodes geographical coordinates, climate zones, and reservoir attributes into learnable representations. The Adaptive Multi-Scale Temporal Fusion (AMTF) mechanism dynamically aggregates daily, weekly, and monthly information through context-dependent gating. The Cross-Reservoir Knowledge Transfer (CRKT) strategy employs similarity-aware contrastive pre-training and reservoir-specific adapters to enable effective knowledge sharing across geographically distributed reservoirs.

Extensive experiments on six globally distributed reservoirs spanning four continents demonstrate state-of-the-art performance with average improvements of 24% in RMSE, 19% in MAE, and 16% in MAPE over strong baselines including iTransformer, DLinear, Informer, LSTM, GRU, and machine-learning methods. The framework exhibits particularly strong performance on data-scarce reservoirs (31% improvement on Dartmouth) and tropical reservoirs with extreme variability (28% on Bagre). Comprehensive ablation studies validate each component's contribution, while visualization analyses provide interpretability insights into adaptive behavior and learned geographical similarity structure.

GeoFormer provides a principled and effective framework for leveraging geographical context and multi-reservoir data in hydrological forecasting, with broad applicability to water resource management, flood warning systems, and climate change adaptation. Future work will extend the framework to multivariate forecasting, incorporate physical constraints, and scale to larger reservoir networks.

Author Contributions: Conceptualization, X.W. and Y.S.; methodology, X.W. and G.J.; software, X.W. and J.G.; validation, X.W., G.J. and J.G.; formal analysis, X.W.; investigation, X.W.; resources, Y.S.; data curation, X.W. and J.G.; writing—original draft preparation, X.W.; writing—review and editing, G.J. and Y.S.; visualization, X.W.; supervision, G.J. and Y.S.; project administration, Y.S.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available in <https://dahiti.dgfi.tum.de/en/> (accessed on 28 October 2015).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lehner, B.; Liermann, C.R.; Revenga, C.; Vörösmarty, C.; Fekete, B.; Crouzet, P.; Döll, P.; Endejan, M.; Frenken, K.; Magome, J.; et al. High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. *Front. Ecol. Environ.* **2011**, *9*, 494–502.
2. Biemans, H.; Haddeland, I.; Kabat, P.; Ludwig, F.; Hutjes, R.W.A.; Heinke, J.; Von Bloh, W.; Gerten, D. Impact of reservoirs on river discharge and irrigation water supply during the 20th century. *Water Resour. Res.* **2011**, *47*, W03509.
3. Zarfl, C.; Lumsdon, A.E.; Berlekamp, J.; Tydecks, L.; Tockner, K. A global boom in hydropower dam construction. *Aquat. Sci.* **2015**, *77*, 161–170.
4. Haddeland, I.; Heinke, J.; Biemans, H.; Eisner, S.; Flörke, M.; Hanasaki, N.; Konzmann, M.; Ludwig, F.; Masaki, Y.; Schewe, J.; et al. Global water resources affected by human interventions and climate change. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 3251–3256.
5. Dottori, F.; Szewczyk, W.; Ciscar, J.C.; Zhao, F.; Alfieri, L.; Hirabayashi, Y.; Bianchi, A.; Mongelli, I.; Frieler, K.; Betts, R.A.; et al. Increased human and economic losses from river flooding with anthropogenic warming. *Nat. Clim. Change* **2018**, *8*, 781–786.
6. Altaş, F.; Öztürk, M. Water Level Predictions at Both Entrances of a Sea Strait by Using Machine Learning. *Water* **2024**, *16*, 2335. <https://doi.org/10.3390/w16162335>.
7. Brunner, G.W. *HEC-RAS River Analysis System: Hydraulic Reference Manual*, version 5.0; US Army Corps of Engineers, Institute for Water Resources, Hydrologic Engineering Center: Davis, CA, USA, 2016.

8. Arnold, J.G.; Srinivasan, R.; Muttiah, R.S.; Williams, J.R. Large area hydrologic modeling and assessment part I: Model development. *J. Am. Water Resour. Assoc.* **1998**, *34*, 73–89.
9. Abbott, M.B.; Bathurst, J.C.; Cunge, J.A.; O'Connell, P.E.; Rasmussen, J. An introduction to the European Hydrological System—Système Hydrologique Européen, “SHE”, 1: History and philosophy of a physically-based, distributed modelling system. *J. Hydrol.* **1986**, *87*, 45–59.
10. Sit, M.; Demiray, B.Z.; Xiang, Z.; Ewing, G.J.; Sermet, Y.; Demir, I. A comprehensive review of deep learning applications in hydrology and water resources. *Water Sci. Technol.* **2020**, *82*, 2635–2670.
11. Fu, J.-C.; Su, M.-P.; Liu, W.-C.; Huang, W.-C.; Liu, H.-M. Water Level Forecasting Combining Machine Learning and Ensemble Kalman Filtering in the Danshui River System, Taiwan. *Water* **2024**, *16*, 3530. <https://doi.org/10.3390/w16233530>.
12. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*, 5th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2015.
13. Hyndman, R.J.; Khandakar, Y. Automatic time series forecasting: The forecast package for R. *J. Stat. Softw.* **2008**, *27*, 1–22.
14. Wang, W.; Van Gelder, P.H.A.J.M.; Vrijling, J.K.; Ma, J. Testing for nonlinearity of streamflow processes at different timescales. *J. Hydrol.* **2006**, *322*, 247–268.
15. Yu, P.S.; Chen, S.T.; Chang, I.F. Support vector regression for real-time flood stage forecasting. *J. Hydrol.* **2006**, *328*, 704–716.
16. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
17. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794.
18. Feng, X.; Liu, L.; Ye, M.; Mašek, O.; Gouda, S.; Chang, K.; Wang, X.; Huang, Q. Unveiling and interpreting the relationships among multi-pollutant emission factors in municipal solid waste incineration by machine learning. *Waste Manag.* **2026**, *210*, 115256.
19. Mosavi, A.; Ozturk, P.; Chau, K.W. Flood prediction using machine learning models: Literature review. *Water* **2018**, *10*, 1536.
20. Ke, Z.; Shen, J.; Zhao, X.; Fu, X.; Wang, Y.; Li, Z.; Liu, L.; Mu, H. A stable technical feature with GRU-CNN-GA fusion. *Appl. Soft Comput.* **2026**, *187*, 114302.
21. Ke, Z.; Cao, Y.; Chen, Z.; Yin, Y.; He, S.; Cheng, Y. Early warning of cryptocurrency reversal risks via multi-source data. *Financ. Res. Lett.* **2025**, *85*, 107890.
22. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
23. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1724–1734.
24. Kratzert, F.; Klotz, D.; Brenner, C.; Schulz, K.; Herrnegger, M. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 6005–6022.
25. Zhang, J.; Zhu, Y.; Zhang, X.; Ye, M.; Yang, J. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. *J. Hydrol.* **2018**, *561*, 918–929.
26. Xiang, Z.; Yan, J.; Demir, I. A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water Resour. Res.* **2020**, *56*, e2019WR025326.
27. Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 5998–6008.
29. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2021; Volume 35, pp. 11106–11115.
30. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems 34*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 22419–22430.
31. Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; Jin, R. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning, Baltimore, New York, USA*; PMLR: 2022; pp. 27268–27286.
32. Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; Long, M. iTransformer: Inverted transformers are effective for time series forecasting. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria, 7–11 May 2024.
33. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2023; Volume 37, pp. 11121–11128.

34. Nicholson, S.E. The West African Sahel: A review of recent studies on the rainfall regime and its interannual variability. *ISRN Meteorol.* **2013**, *2013*, 453521.
35. López-Moreno, J.I.; Vicente-Serrano, S.M.; Morán-Tejeda, E.; Lorenzo-Lacruz, J.; Kenawy, A.; Beniston, M. Effects of the North Atlantic Oscillation (NAO) on combined temperature and precipitation winter modes in the Mediterranean mountains: Observed relationships and projections for the 21st century. *Glob. Planet. Change* **2011**, *77*, 62–76.
36. Abdi, E.; Sattari, M.T.; Samadianfard, S.; Ahmad, S. Advancing Hydrological Prediction with Hybrid Quantum Neural Networks: A Comparative Study for Mile Mughan Dam. *Water* **2025**, *17*, 3592. <https://doi.org/10.3390/w17243592>.
37. Yasarer, L.M.W.; Sturm, B.S.M. Potential impacts of climate change on reservoir services and management approaches. *Lake Reserv. Manag.* **2016**, *32*, 13–26.
38. Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.X.; Yan, X. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems 35*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 5243–5254.
39. Hannah, D.M.; Demuth, S.; van Lanen, H.A.J.; Looser, U.; Prudhomme, C.; Rees, G.; Stahl, K.; Tallaksen, L.M. Large-scale river flow archives: Importance, current status and future needs. *Hydrol. Process.* **2011**, *25*, 1191–1200.
40. Kratzert, F.; Klotz, D.; Shalev, G.; Klambauer, G.; Hochreiter, S.; Nearing, G. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 5089–5110.
41. Frisk, F.; Johansson, O. Comparative Evaluation of Water Level Forecasting Using IoT Sensor Data: Hydrodynamic Model SWMM vs. Machine Learning Models Based on NARX Framework. *Water* **2024**, *16*, 2776. <https://doi.org/10.3390/w16192776>.
42. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610.
43. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.C.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **1998**, *454*, 903–995.
44. Wu, Z.; Huang, N.E. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Adv. Adapt. Data Anal.* **2009**, *1*, 1–41.
45. Zhang, X.; Zhang, Q.; Zhang, G.; Nie, Z.; Gui, Z.; Que, H. A hybrid model for annual runoff time series forecasting using elman neural network with ensemble empirical mode decomposition. *Water* **2018**, *10*, 416.
46. Park, J.; Joh, J.S.; Choi, M.; Kim, T.; Cho, J.; Lee, Y. AI-Based Time-Series Ensemble Approach Coupled with a Hydrological Model for Reservoir Storage Prediction in Korea. *Water* **2025**, *17*, 3296. <https://doi.org/10.3390/w17223296>.
47. Chung, J.; Ahn, S.; Bengio, Y. Hierarchical multiscale recurrent neural networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
48. Yang, B.; Sun, S.; Li, J.; Lin, X.; Tian, Y. Traffic flow prediction using LSTM with feature enhancement. *Neurocomputing* **2019**, *332*, 320–327.
49. Jernite, Y.; Grave, E.; Joulin, A.; Mikolov, T. Variable computation in recurrent neural networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
50. Sukhbaatar, S.; Grave, E.; Bojanowski, P.; Joulin, A. Adaptive attention span in transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 331–335.
51. Kratzert, F.; Gauch, M.; Klotz, D.; Nearing, G. HESS Opinions: Never train an LSTM on a single basin. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 6147–6150.
52. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
53. Xu, T.; Liang, F. Machine learning for hydrologic sciences: An introductory overview. *WIREs Water* **2021**, *8*, e1533.
54. Hsu, K.; Gupta, H.V.; Sorooshian, S. Artificial neural network modeling of the rainfall-runoff process. *Water Resour. Res.* **1995**, *31*, 2517–2530.
55. Feng, D.; Liu, J.; Lawson, K.; Shen, C. Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resour. Res.* **2022**, *58*, e2022WR032404.
56. Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; Vinyals, O. Understanding deep learning requires rethinking generalization. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
57. Mai, G.; Janowicz, K.; Yan, B.; Zhu, R.; Cai, L.; Lao, N. Multi-scale representation learning for spatial feature distributions using grid cells. In Proceedings of the 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
58. Mac Aodha, O.; Cole, E.; Perona, P. Presence-only geographical priors for fine-grained image classification. In *IEEE/CVF International Conference on Computer Vision, Seoul, Korea*; IEEE: New York, NY, USA, 2019; pp. 9596–9606.

59. Kottek, M.; Grieser, J.; Beck, C.; Rudolf, B.; Rubel, F. World map of the Köppen-Geiger climate classification updated. *Meteorol. Z.* **2006**, *15*, 259–263.
60. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA*; PMLR: 2019; pp. 2790–2799.
61. Schwatke, C.; Dettmering, D.; Bosch, W.; Seitz, F. DAHITI—an innovative approach for estimating water level time series over inland waters using multi-mission satellite altimetry. *Hydrol. Earth Syst. Sci.* **2015**, *19*, 4345–4364.
62. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 3146–3154.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.