JGR Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2024JH000519

Key Points:

- An innovative Dual-Transformer deep learning framework is developed for seasonal forecasting of water levels in the Great Lakes
- The input features are widely accessible as standard outputs from most atmospheric and climate models, enhancing model applicability
- The model achieves unprecedented accuracy at very low computational cost in predicting lake levels with a lead time of up to 6 months

Correspondence to:
P. Xue,
pexue@mtu.edu

Citation:
Chen, Y., & Xue, P. (2025). Dual-transformer deep learning framework for seasonal forecasting of Great Lakes water levels. *Journal of Geophysical Research: Machine Learning and Computation*, *2*, e2024JH000519. https://doi.org/10.1029/2024JH000519

Received 13 NOV 2024
Accepted 19 APR 2025

Dual-Transformer Deep Learning Framework for Seasonal Forecasting of Great Lakes Water Levels

Yi Chen[1] and Pengfei Xue[1,2,3]

[1]Great Lakes Research Center, Michigan Technological University, Houghton, MI, USA, [2]Department of Civil, Environmental and Geospatial Engineering, Michigan Technological University, Houghton, MI, USA, [3]Environmental Science Division, Argonne National Laboratory, Lemont, IL, USA

Abstract The Great Lakes of North America form one of the largest freshwater systems on Earth, and their lake-wide average water levels (lake levels) can fluctuate by more than 0.5 m on a seasonal scale. These fluctuations pose substantial challenges for coastal resilience, flood risk management, and navigation planning. Accurate seasonal forecasting of lake levels using traditional mechanistic models is challenging due to the complex physical mechanisms and coupled hydroclimatic processes involved. Recently, deep learning has gained prominence in geoscience applications for its ability to recognize intricate patterns within multiphysical data sets. Here, we introduce a novel Dual-Transformer deep learning framework, tested on the Great Lakes. This architecture integrates two modified Transformer models: the Prophet, which predicts underlying trends, and the Critic, which refines the Prophet's predictions. The final lake level prediction is derived by weighting the outputs of both models through a multi-layer perceptron, jointly trained with the Prophet and Critic to enhance overall accuracy. Our results demonstrate that the innovative learning framework achieves the highest prediction accuracy compared to established deep learning models when using identical input features. It attains a root mean square error of 4–7 cm in predicting lake levels up to 6 months in advance across the lakes. Additionally, the Dual-Transformer model runs six orders of magnitude faster than conventional mechanistic models, producing results in less than one second on a typical personal computer. These findings suggest that our deep learning framework has strong potential to advance lake level prediction and carries important implications for water management and disaster mitigation, thereby enhancing the quality of life in coastal regions.

Plain Language Summary The Great Lakes, one of the world's largest freshwater systems, experience natural seasonal water level fluctuations exceeding half a meter. These fluctuations pose challenges for coastal communities in water management and coastal resilience. Traditional forecasting methods struggle to capture these changes accurately due to the complex interactions between water and atmospheric conditions. To address this, we developed a new deep learning model, the Dual-Transformer, which we tested on the Great Lakes. This model comprises two main components: the Prophet, which forecasts general trends, and the Critic, which adjusts these forecasts based on recent weather data. A third component combines the two predictions to improve accuracy. Our results show that the Dual-Transformer, using seven key weather and lake variables, can predict lake levels with very high accuracy up to 6 months into the future. The model is also computationally efficient, running up to a million times faster than traditional methods and producing results in less than a second on a standard computer. This deep learning model has significant potential to enhance water management, safety, and community planning along the coasts of the Great Lakes.

1. Introduction

The Great Lakes of North America encompass five distinct bodies of water: Lake Superior (the largest Great Lake), Lake Michigan, Lake Ontario, Lake Huron, and Lake Erie. Collectively, they constitute the most extensive unfrozen freshwater system in the world, containing 84% of North America's surface freshwater. The deepest point within the Great Lakes exceeds 400 m, located in the central basin of Lake Superior. Their total surface area spans approximately 244,000 square kilometers, collectively making the system similar in size to the United Kingdom and about 2.6 times larger than the Gulf of Maine. The vast expanses and substantial depths of the Great Lakes impart many characteristics similar to those of seas, and they are thus often recognized as “inland seas” in the context of hydrodynamics.

10.1029/2024JH000519

Water levels in the Great Lakes are highly dynamic, with lake-wide average water level variations reaching over 0.5 m on a seasonal scale and nearly 2 m on a decadal scale. Lake-wide average water level (hereafter referred to as lake level) is a crucial water management parameter, it provides a system-wide perspective, smoothing out short-term variations caused by storms, wind-driven fluctuations, or barometric pressure changes. This ensures that water management decisions—such as outflow regulation at control structures—are based on long-term hydrological trends rather than transient local events. This information has been utilized by international regulatory bodies and government agencies for decision-making, supporting consistency, accuracy, and long-term sustainability in Great Lakes water management. Lake level fluctuations are primarily driven by the hydrological cycle, with the highest levels typically occurring in late summer and the lowest in early spring (Gronewold & Stow, 2014). The main factors influencing lake level changes are precipitation, basin runoff, and lake evaporation (Gronewold et al., 2021), which together shape the lakes' hydroclimatic conditions. As a critical variable in the Great Lakes system, lake levels play a pivotal role in coastal resilience and risk management. High lake levels increase the risk of coastal inundation, threatening communities, infrastructure, and ecosystems along shorelines, and often causing substantial economic losses and displacement of residents (Melillo et al., 2014). Huang, Anderson, et al. (2021) demonstrated that a meteotsunami in 2018 would have caused much more severe coastal inundation if it had occurred during the high lake levels of 2020, potentially resulting in extensive damage to coastal properties and resident displacement. Conversely, low lake levels can significantly affect navigation, making some areas difficult or even impossible for vessels to traverse, thereby disrupting commercial shipping routes and raising transportation costs (Millerd, 2005). Low water levels also pose challenges to hydropower production, which requires a consistent water supply for efficient electricity generation (Stakhiv & Stewart, 2010). The Great Lakes have experienced significant fluctuations over recent years, with record lows observed from 1999 to 2013, followed by a rapid rise of nearly 2 m by 2020, and a return to average levels within two years (Gronewold et al., 2021). These fluctuations impact not only coastal inundation and navigation but also accelerate coastal erosion, leading to the loss of valuable land and habitats and negatively affecting aquatic life (Trebitz & Hoffman, 2015).

Accurately predicting lake levels is crucial. Early predictions help mitigate the harmful effects of abnormal levels, reducing economic losses and risks to public safety. A better understanding of lake level fluctuations also supports effective water resource management, minimizes impacts on human and natural systems, and promotes sustainable development and conservation in the region. In a mechanistic modeling approach, the primary method for predicting lake levels involves estimating Net Basin Supply (NBS), which is the sum of over-lake precipitation, evaporation, and basin runoff (Gronewold et al., 2021; Kayastha et al., 2022). In various studies, researchers have used Regional Climate Models (RCMs) and hydrological models to estimate these three NBS components (Kayastha et al., 2022; MacKay & Seglenieks, 2013; Notaro et al., 2021). Earlier efforts featured RCMs two-way coupled with a one-dimensional (1D) lake model, but this approach introduced significant biases in estimating evaporation because it did not account for three-dimensional (3D) lake circulation and associated turbulent mixing processes (Xue et al., 2024; Ye et al., 2020). These factors influence lake surface temperature and ice cover, which interact with the atmosphere to determine over-lake evaporation (Bennington et al., 2014; Notaro et al., 2021). Recent advancements have achieved a two-way coupling of RCMs with a 3D lake model (Xue et al., 2017; Xue, Ye, et al., 2022), leading to improved estimates of lake evaporation and precipitation and providing more reliable projections of lake levels under a changing climate (Kayastha et al., 2022; Xue et al., 2024). The NBS is then applied to the Coordinated Great Lakes Regulating and Routing Model (CGLRRM), which further accounts for inter-lake flow and regulation, enabling dynamic water level modeling across the Great Lakes system (Kayastha et al., 2022).

However, the mechanistic modeling approach described above faces several challenges. Lake level prediction requires a suite of complex models that simulate multiple physical processes (Kayastha et al., 2022). Each model requires different data sets for calibration and validation, which are sometimes limited or even unavailable. Additionally, uncertainties in model parameters and input data can propagate and accumulate among the models, affecting prediction reliability (Pringle et al., 2024). Running a suite of complex mechanistic models is also computationally intensive, requiring significant computational resources and time for real-time applications where stakeholders need timely, actionable information. As a result, there is growing interest in deep learning (DL) as an alternative approach for predicting lake levels in the Great Lakes for several compelling reasons:

First, DL models can bypass complex physical processes by using well-designed neural networks to replace traditional methods of discretizing and integrating governing equations. Instead of explicitly modeling Net Basin

Supply (NBS) components—precipitation, evaporation, and runoff—using separate atmospheric, lake, and land models, DL models in this study directly establishes the relationship between climate variables and future lake levels. In other Great Lakes applications, the DL approach has shown potential to reduce errors arising from biases in individual process-based models, particularly when one or more of these mechanistic models introduces significant uncertainties (Xue, Wagh, et al., 2022). Second, the DL approach often improves computational efficiency over mechanistic modeling, allowing for faster lake level predictions. Furthermore, Furthermore, by using methods such as explainable AI (Doshi-Velez & Kim, 2017; Lundberg & Lee, 2017; Samek, 2017), DL can identify the relative contributions of input features to predictions, providing additional insight into system dynamics and complementing insights from mechanistic models.

Many studies have successfully integrated deep learning (DL) into geophysical research, achieving significant advancements. Reichstein et al. (2019) provided a comprehensive overview of DL's potential to advance Earth system science, emphasizing its capabilities in predictive accuracy and process understanding within complex, nonlinear systems. For example, DL neural networks have been employed to develop emulators of general circulation and weather models (Nguyen et al., 2023; Rasp et al., 2018; Scher, 2018), detect extreme weather events in large climate data sets (Liu et al., 2016), and predict uncertainties in weather forecasting (Scher & Messori, 2018). More recently, Rasp and Thuerey (2021) introduced a ResNet model for medium-range weather predictions, pretrained on outputs from the MPI-ESM-HR. By incorporating time as an explicit input feature, the model effectively learns temporal patterns among atmospheric variables, enabling accurate forecasts up to 5 days ahead. Adaptive Fourier Neural Operators (AFNOs) have also been applied in weather forecasting, utilizing Fourier transforms for spatial mixing to capture long-range dependencies and complex interactions among atmospheric variables (Pathak et al., 2022). This approach allows models like FourCastNet to handle time-series data collectively, generating simultaneous predictions for multiple variables up to 14 days ahead and operating approximately 45,000 times faster than traditional prediction systems. Nguyen et al. (2023) employed a Transformer model to enhance medium-range weather forecasting capabilities, extending predictions up to 14 days. A key feature of their approach is the integration of multiple time interval combinations, which significantly improves forecasting accuracy. Additionally, they conducted ablation studies to gain deeper insights into the specific contributions of various techniques embedded within the model. Additionally, DL has been integrated into physics-based models to approximate subgrid-scale processes—such as clouds, convection, and turbulence—resulting in enhanced simulation accuracy compared to traditional empirical methods (Bolton & Zanna, 2019; Brenowitz & Bretherton, 2018; Rasp et al., 2018; Yuan et al., 2024).

In recent years, DL has also been utilized in modeling the Great Lakes system. Feng et al. (2020) implemented a multi-layer perceptron (MLP) network as an emulator for wave prediction in Lake Michigan, significantly enhancing computational efficiency compared to traditional mechanistic models. Xue, Wagh, et al. (2022) improved lake surface temperature forecasts by statistically integrating a Long Short-Term Memory (LSTM) neural network with a hydrodynamic model. Building on this work, Kayastha et al. (2023a) employed an LSTM network to generate a comprehensive daily lake surface temperature data set for all five Great Lakes, spanning 42 years at a spatial resolution of approximately 1 km.

In particular, there have been recent efforts focusing on predicting lake levels using DL. Wang and Wang (2020) applied multiple machine learning models to predict daily water levels in Lake Erie. Their approach used meteorological variables and the previous day's water level to forecast the next day's level. While this method demonstrated good predictive accuracy on a short-term, daily scale, it does not address longer-term, seasonal lake level predictions, which are far more challenging and critical for effective water management, especially for forecasts up to 6 months in advance.

A more recent study by Kurt (2024) utilized two components of NBS—evaporation and precipitation—along with air temperature and the previous month's lake levels as input features for next-month lake level predictions. This method provides a relatively straightforward estimation of water levels, as lake levels are primarily influenced by the NBS mass balance. Therefore, when accurate NBS components are available, predicting water level changes becomes relatively simple. However, it does not address the significant challenge of obtaining accurate input data, particularly for lake evaporation. Long-term evaporation data for the Great Lakes are limited, with direct measurements only available since 2008, restricted to single-point observations for each lake. In Kurt (2024), long-term evaporation data from 1950 to 2020 were derived from the one-dimensional Large Lake Thermodynamics Model (LLTM) (Hunter et al., 2015), limiting its utility for present-day predictions. Additionally, the model's predictive time scale is restricted to 1 month ahead, which constrains its applicability for seasonal water management.

An intriguing study by Barzegar et al. (2021) took a different approach by excluding meteorological conditions as inputs. The study first applied the Boundary Corrected Maximal Overlap Discrete Wavelet Transform (BC-MODWT) as a preprocessing step to decompose the historical monthly water level time series of Lake Michigan and Lake Ontario over the past 12 months into wavelet and scaling coefficients, effectively addressing the data's nonstationarity and noise. The denoised time series generated by BC-MODWT was then fed into a hybrid deep learning model comprising a Convolutional Neural Network (CNN) layer followed by a LSTM layer. The CNN layer automatically extracted relevant features from the past 12 months of water level data, while the LSTM layer made water level forecasts for one, two, and 3 months ahead. This approach offered the advantage of longer-term, 3-month predictions while also achieving higher accuracy than Kurt (2024), with a root mean square error (RMSE) of approximately 10 cm for the 3-month forecast.

To achieve the highest accuracy and extend prediction lead times up to 6 months, this paper presents an innovative Dual-Transformer DL framework to predict the lake levels. We first provide a detailed analysis to evaluate and understand the model's performance on Lake Superior, the largest and deepest of the Great Lakes. Next, we expand the framework to the other Great Lakes and evaluate its performance across the lakes, demonstrating the framework's robustness and adaptability. Our model directly captures the relationship between historical lake levels and atmospheric conditions to project future water levels 6 months ahead. We emphasize the utilization of fundamental and widely available atmospheric variables as input features, which are standard outputs from virtually all atmospheric and climate models. To ensure practicality, we select variables that are readily accessible from model outputs or reanalysis data sets, including surface air temperature, surface wind speed, and precipitation. By incorporating these commonly available atmospheric features, along with lake surface temperature (LST)—the most readily available lake feature from remote sensing, in-situ observations, or hydrodynamic models—our method builds a robust representation of the key processes driving lake level changes without relying on limited hydrological data. Achieving this requires an innovative DL framework capable of effectively learning the complex, multi-scale relationships between atmospheric conditions and lake level dynamics, as described below.

2. The Data Set

Atmospheric data were sourced from the Climate Forecast System Reanalysis (CFSR), available from January 1979 to March 2011, and its upgraded version, the Climate Forecast System Version 2 (CFSv2) analysis product, a continuation of the CFSv2 Reanalysis, available from 2011 onward (Saha et al., 2014). For simplicity, both data sets are collectively referred to as CFSR. This data set represents a global reanalysis, incorporating satellite radiance data alongside conventional and satellite observations. It has been extensively used for Great Lakes modeling (Feng et al., 2020; Huang, Zhu, et al., 2021; Lv et al., 2019; Xue et al., 2015). From CFSR, we obtained five spatially averaged daily atmospheric features. A lake basin refers to both the lake water body and its surrounding watershed—the land area where all water drains into the lake. For example, for Lake Superior, we obtained five spatially averaged daily atmospheric features: air temperature at 2-m height averaged over the lake, air temperature at 2-m height averaged over the land (i.e., Lake Superior watershed), wind speed at 10-m height averaged over the lake, wind speed at 10-m height averaged over the land, and precipitation averaged over the Lake Superior basin.

In addition, the lake surface temperature (LST) data set was retrieved from the satellite-based Great Lakes Surface Environmental Analysis (GLSEA) by the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL). As GLSEA data only dates back to 1995, the LST data from 1981 to 1994 was retrieved from the reconstructed LST data set using an LSTM model by Kayastha et al. (2023a).

Please note that the lake levels of each Great Lake have an official definition and calculation method, ensuring consistency in lake level data released by federal agencies and scientific communities. Lake levels are defined as the average of selected U.S. and Canadian water level monitoring stations, with these stations selected by the Coordinating Committee on Great Lakes Basic Hydraulic and Hydrologic Data. Further details are available on NOAA's website: https://www.glerl.noaa.gov/data/wlevels/#monitoringNetwork.

Both atmospheric and lake data sets used in this study span 43 years, from 1981 to 2023, comprising 15,705 entries. The DL models utilized these seven feature vectors (i.e., five atmospheric features and two lake features)

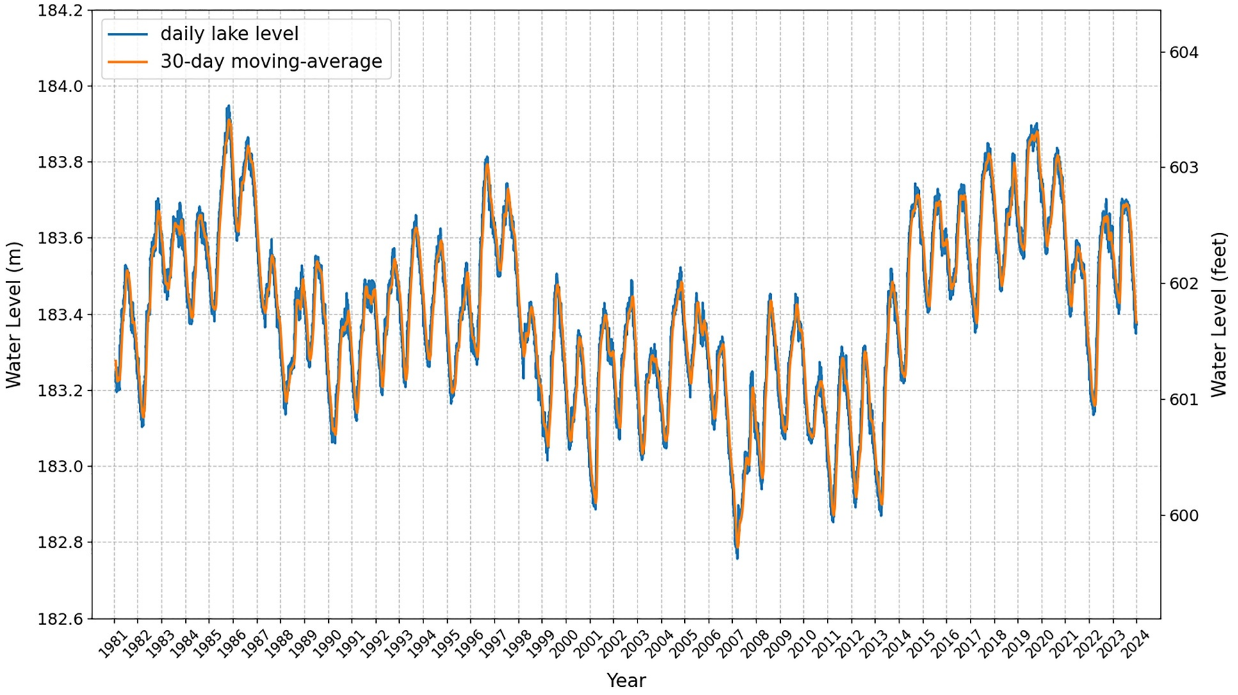

Figure 1. The daily lake level of Lake Superior and its 30-day moving average.

to predict the lake levels up to 6 months into the future. Our study utilizes the data from 1981 to 2021 (14,975 entries) as the training data set for DL models to identify underlying patterns. The data from 2022 to 2023 (730 entries) is used as the testing data set to evaluate the model's performance. Since water management practices mainly focus on monthly lake level changes, we applied a 30-day trailing moving average on the entire data set to smooth short-term fluctuations and highlight seasonal trends. This approach preserves the overall data set size (losing only 30 data samples) while still effectively smoothing short-term variations. Directly using monthly averages would significantly reduce the number of available training samples by a factor of 30, making it unsuitable for this study.

Notably, if one is only interested in the monthly average lake level, it can be easily approximated by the 30-day trailing moving average of the first day of the next month. This operation was conducted on all seven features across the entire data set (comprising both the training and testing data sets). Figure 1 illustrates the contrast between the original daily data and the smoothed data after applying the moving average, highlighting the reduction in daily variability and the enhancement of discernible seasonal trends. The moving average method calculates the average for each day based on the preceding 30 days, which reduced the training data set by 30 entries to 14,945 (14,975–30). The testing data set remained at 730 entries. By smoothing the data, we aimed to improve the model's ability to capture seasonal patterns and reduce noise, ultimately enhancing the models' predictive performance for forecasting lake levels 6 months ahead.

To ensure that the model treats different features equally during training, each feature's values in the input vector should be distributed within the same range. The seven features used in this study varied in scale, which could lead the model to prioritize features with larger numerical values and potentially skew the learning process. Therefore, it was essential to normalize the data before feeding it into the model. Before applying scaling, all data sets first go through an outlier detection process to identify and remove unrealistic values, if present. Additionally, recall that the 30-day moving average smooths fluctuations caused by short-term weather events and hydrodynamic processes. This preprocessing step ensures that the data captures seasonal and long-term trends rather than transient variations. We then applied the Min-Max Scaler for normalization, adjusting all variables in the data set to a range of 0–1. This approach enables the model to effectively learn the evolutionary characteristics of each variable without bias toward features with larger magnitudes.

10.1029/2024JH000519

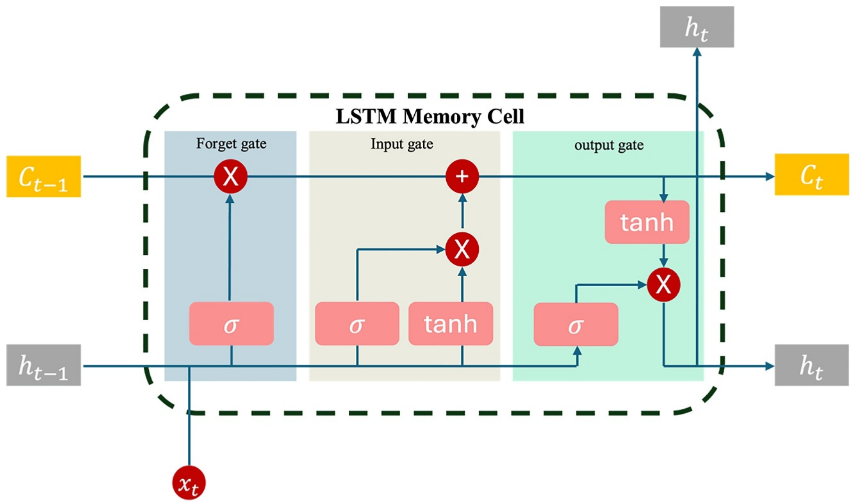

Figure 2. The structure of a Long Short-Term Memory (LSTM) memory cell. The diagram illustrates the key components of an LSTM unit, including the forget gate (blue section), input gate (beige section), and output gate (green section). The previous cell state (c_{t-1}) and hidden state (h_{t-1}) are processed alongside the current input (x_t) to generate the updated cell state (c_t) and hidden state (h_t). The sigmoid function (σ) regulates the flow of information by outputting values between 0 and 1, while the tanh function scales values between -1 and 1. Multiplication ($\times$) and addition ($+$) operations are indicated by red circles. The dashed boundary represents the memory cell, encapsulating these operations to facilitate sequential data processing. The arrows indicate the flow of data within the model.

3. Deep Learning Approach

In this study, we present an innovative DL framework: the Dual-Transformer. To evaluate its performance and necessity, we also tested several conventional deep learning models. In this section, we describe the structure of each model.

3.1. Long Short-Term Memory (LSTM)

Time-series prediction problems, such as forecasting lake levels, can be effectively addressed using neural networks capable of retaining information from past events. Recurrent Neural Networks (RNNs) and their variants are widely used for such tasks due to their ability to model temporal dependencies (Yu et al., 2019). Among these, LSTM networks have demonstrated superior performance over other RNN variants because of their unique architecture, which includes gating mechanisms specifically designed to manage the flow of information, allowing the network to maintain and leverage long-term dependencies effectively (Hochreiter & Schmidhuber, 1997).

As a well-known DL model, the structure of an LSTM cell is concisely illustrated in Figure 2. The gating mechanisms in LSTMs—comprising the input, forget, and output gates—enable the network to retain significant information while discarding less relevant data across time steps. This capability addresses the challenges associated with learning from long sequences and enhances the model's overall performance. Similar LSTM networks were applied to predict Great Lakes water temperature by Xue, Wagh, et al. (2022) and Kayastha et al. (2023a). Two LSTM models with varying numbers of layers (i.e., two-layer or four-layer) were tested in comparison with single Transformer and Dual-Transformer.

3.2. Modified Transformer

The traditional Transformer model, introduced by Vaswani et al. (2017), has revolutionized natural language processing (NLP) by leveraging a self-attention mechanism to effectively capture long-range dependencies within sequential data. Originally designed for NLP tasks, the Transformer architecture has demonstrated remarkable versatility, extending its applicability to diverse domains (Dosovitskiy, 2020; Li et al., 2019).

While the original Transformer architecture consists of both an encoder and a decoder, not all tasks require both components. The decoder is primarily designed for sequence-to-sequence tasks such as machine translation and

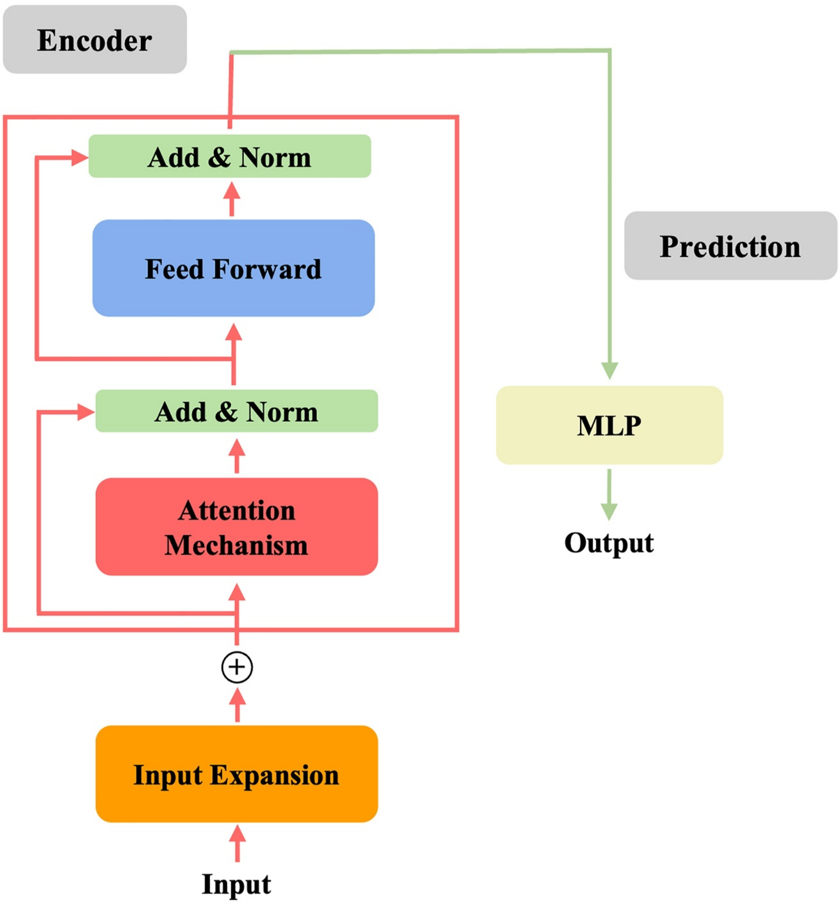

Figure 3. The structure of the modified Transformer model used for lake level projection.

text generation, where generating an output sequence is essential (Vaswani et al., 2017). However, our problem is fundamentally different: lake level prediction is a sequence-to-one forecasting task, where the goal is to predict a single scalar value rather than an entire sequence. In such cases, the full Transformer architecture is unnecessarily complex. Instead, we modify the model by removing the decoder and directly connecting a MLP to the encoder's output (Figure 3). This simplification improves computational efficiency and reduces training time, making the model more practical for time series forecasting.

To better understand the structure and functionality of our modified Transformer, we first define the input representation. In our lake level prediction model, the preprocessed input time series data is represented as $X(t,f)$, where t is the sequence length (number of time steps) and f is the number of features. Since the Transformer does not inherently capture positional dependencies, positional encoding is added to the input embeddings to provide temporal order information. In our implementation, we employ sinusoidal positional encoding, defined as follows:

$$\mathrm{PE}_{(\mathrm{pos},2i)} = \sin\left(\frac{\mathrm{pos}}{10000^{2i/f}}\right); \mathrm{PE}_{(\mathrm{pos},2i+1)} = \cos\left(\frac{\mathrm{pos}}{10000^{2i/f}}\right) \tag{1}$$

where pos represents the position index in the sequence; i is the dimension index of the feature vector. The denominator $1000^{2i/f}$ serves as a frequency scaling factor, ensuring that different time scales are encoded effectively. This positional encoding is then added to the input features, allowing the model to differentiate positions within the sequence before processing them in the self-attention module.

Once positional encoding is applied, the enhanced sequential data is fed into the self-attention module, which dynamically assesses the importance of each time step relative to others (Vaswani et al., 2017). This mechanism enables the model to identify key dependencies within the input sequence that might be overlooked by traditional models. Unlike recurrent neural networks (RNNs) such as LSTMs, which process inputs sequentially and often struggle with long-range dependencies due to vanishing gradients, Transformers handle entire sequences simultaneously (Dai, 2019). This parallelization not only enhances the model's ability to capture dependencies between distant time steps but also significantly improves computational efficiency.

In our modified Transformer architecture, the self-attention mechanism plays a crucial role in capturing dependencies within different time steps. To efficiently process the output from positional encoding,the input data is first projected into a higher-dimensional space (d_{model}) using learnable weight matrices:

$$Q = XW^Q, K = XW^K, V = XW^V \tag{2}$$

where W^Q, W^K, and W^V are learnable weight matrices that transform the input into query (Q), key (K), and value (V) matrices, each dimensions $[t, d_{\text{model}}]$. These matrices represent distinct feature-extended transformations of the input time series, incorporating both atmospheric and lake information. Q (Query) serves as a reference for computing attention scores, interacting with K (Key) to assess the similarity (or relevance) between different time steps, ultimately forming the attention score matrix. V (Value) is served to compute a weighted sum with the attention scores, ensuring that the final representation captures the most relevant contextual information from the input data set.

Specifically, the attention module in this study employs a multi-head attention mechanism, which consists of h independent attention heads. Each head processes a portion of the feature space by computing separate $Q_i,\ K_i, V_i$ where each has a dimension of $d_k = d_{\text{model}}/h$. Here, $Q_i,\ K_i, V_i$ are uniformly partitioned sub-matrices derived from the original Q, K, V, and d_k represents the feature dimension of the key (K_i) vectors in each attention head.

Each attention head i independently computes attention scores using the following formula:

$$\text{Attention}(Q_i,\ K_i, V_i) = \text{softmax}\left(\frac{Q_i\ K_i{}^T}{\sqrt{d_k}}\right) V_i \tag{3}$$

where the attention mechanism first computes the attention score matrix as $\alpha = \text{softmax}\left(\frac{Q_i\ K_i{}^T}{\sqrt{d_k}}\right)$. The dot product $Q_i\ K_i{}^T$ determines the relevance between different time steps; The scaling factor $\sqrt{d_k}$ prevents excessively large dot-product values, stabilizing the gradient updates during training; The softmax function normalizes these scores so that they sum to 1 across all time steps, effectively converting them into attention weights. At this stage, α is a matrix that encodes how much each time step should attend to every other time step.

Once the attention scores are obtained, the model computes a weighted sum of the value matrix (V_i) using αV_i. Since each attention head captures different aspects of the sequence independently, their outputs need to be combined to form a unified representation. To achieve this, the outputs from all h attention heads are concatenated along the feature dimension. This step ensures that information from multiple representation subspaces is retained, allowing the model to incorporate diverse dependencies across the sequence. As a result, the output from the self-attention module primarily encodes temporal dependencies by dynamically weighting different time steps based on their relevance.

After self-attention, the encoded features pass through a feed-forward network (FFN) to refine representations and introduce non-linearity. To stabilize training and enhance feature propagation, Add & Norm layers are applied before and after the FFN. The residual connections in Add & Norm help mitigate vanishing gradient issues by preserving earlier information, while layer normalization ensures stable feature distributions throughout training.

Once the input sequence has been processed through the encoder, the model outputs a transformed representation that encodes both local patterns and long-range dependencies. Since our task is sequence-to-one forecasting, rather than generating another sequence, the encoder output is passed through an MLP prediction module consisting of fully connected layers, which further refine the extracted features and map them to the final scalar prediction—the predicted lake level.

By leveraging a self-attention-based modified Transformer, the model effectively captures both short-term fluctuations and long-term dependencies, ensuring robust and accurate lake level predictions.

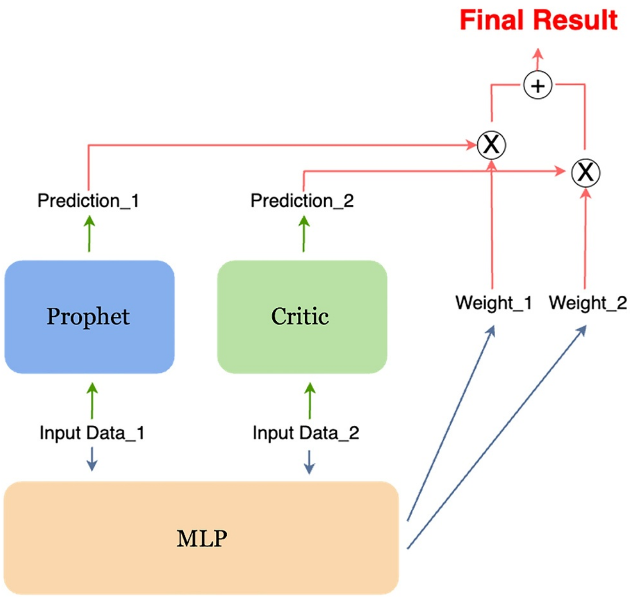

Figure 4. The architecture of the Dual-Transformer model. The framework comprises two parallel prediction modules: Prophet (blue) and Critic (green), each processing distinct input data streams—Input Data_1 (historical data set) and Input Data_2 (future data set). Prophet generates Prediction_1, which serves as the coarse prediction, while the Critic produces Prediction_2, refining Prophet's output. These predictions are then weighted by Weight_1 and Weight_2, respectively. The weights are learned through a multi-layer perceptron (MLP, orange), which processes input features to determine the optimal combination of both predictions. The final output (Final Result, red) is obtained by computing the weighted sum of the two predictions.

3.3. Innovative Dual-Transformer

Most importantly, we developed an enhanced architecture that significantly improved prediction accuracy and achieved unprecedented high accuracy in seasonal forecasting of lake levels, as illustrated in Figure 4. This architecture comprises two Transformer models: the Prophet and the Critic. The Prophet estimates lake levels based on historical water levels and historical atmospheric information. The Critic then refines the Prophet's predictions by incorporating future atmospheric conditions over the 6 months leading up to the prediction date. Finally, an MLP weighting model determines the appropriate weighting between the Prophet and Critic outputs, dynamically adjusting the intensity of the Critic's corrections based on the integrated information from both data sets. The three models are trained collectively, using a loss function defined by the mean square error (MSE). The training process utilizes backpropagation and the Adam optimization algorithm.

The Prophet and the Critic consist of six stacked encoder layers, each utilizing an eight-head multi-head self-attention mechanism. Each layer comprises multiple sub-layers, including multi-head self-attention and position-wise feed-forward neural networks. In the multi-head self-attention mechanism, each of the eight heads independently computes attention scores, and their outputs are concatenated and linearly transformed to produce the final attention output for the layer. This six-layer structure enables the model to progressively extract and refine information from the input sequence through successive transformations. Additionally, the eight-head self-attention mechanism allows the model to capture diverse patterns and relationships within the input data by attending to different subspaces of the feature representation.

The input configurations for the Dual-Transformer include three data sets. Input data set for the Prophet (Input Data_1) includes 12 historical data entries up to the current day, each separated by 30-day intervals and containing all 7 features. Input data set for the Critic (Input Data_2) consists of the 6 features, excluding any future lake level information, sampled at five uniformly distributed time points between the current day and the prediction date. Input data set for the MLP weighting model (Input Data_3) combines Input Data_1 and Input Data_2.

For example, if the current date is 1 January 2021, and we aim to forecast the 30-day moving average lake level 180 days (referred to as 6 months for simplicity) ahead on 30 June 2021 (Note: the 30-day moving average on June 30 is also June 2021's monthly mean lake level), Input Data_1 would include the seven features recorded on 1 January 2021; 2 December 2020; 2 November 2020; 3 October 2020; 3 September 2020; 4 August 2020; 5 July 2020; 5 June 2020; 6 May 2020; 6 April 2020; 7 March 2020; 6 February 2020, reflecting 30-day intervals leading up to the current date. Input Data_2 would include only the six atmospheric features recorded on 31 January 2021; 2 March 2021; 1 April 2021; 1 May 2021; and 31 May 2021, without using any future water level information.

4. Results and Discussion

4.1. Evaluation of DL Models Performance on Lake Superior

As described above, the Dual-Transformer represents a more sophisticated and intricate framework. Justifying the necessity of such a model requires demonstrating its superiority over other deep learning models under identical input formation. To this end, we conduct a comparative analysis evaluating the performance of various models, including a two-layer LSTM, a four-layer LSTM, a single Transformer, a Dual-Transformer without MLP (using a fixed empirical average), and a Dual-Transformer with MLP weighting.

To ensure that the comparison objectively reflects each model's best performance under identical conditions, all models utilize the same amount of input feature information—i.e., the combined Input_Data_1 and Input_Data_2 for training and testing. Due to their architectural design, the LSTMs and the single Transformer directly use the combined data set, whereas within the Dual-Transformer framework, Prophet utilizes Input_Data_1, and Critic

Table 1
Hyperparameters and Comparative Results of Different DL Models

Model	Type	Feature input and hyper-parameters			RMSE(cm) on testing data set
LSTM	**Two-layer**	**Feature input: Input Data_1 and Input Data_2** Optimizer: Adam Hidden size: 512 Sequence length: 17 Activation functions: ReLU Dropout: 0.5 Learning rate: 0.0001 Epochs: 100 BatchsSize: 256			**10.8**
	Four-layer	**Feature input: Input Data_1 and Input Data_2** Optimizer: Adam Hidden size: 512 Sequence length: 17 Activation functions: ReLU Dropout: 0.5 Learning rate: 0.0001 Epochs: 100 Batch size: 256			**11.0**
Transformer	**Single**	**Feature input: Input Data_1 and Input Data_2** Optimizer: Adam $d_k = d_v$: 128 MLP hidden layer: 128 Sequence length: 17 Activation functions: ReLU Dropout: 0.1 Learning rate: 0.0001 Epochs: 100 Batch dize: 256			**8.9**
	Dual (without MLP)	**Prophet**	**Critic**	**Empirical weighted average**	**7.7**
	Dual (with MLP)	**Feature input: Input Data_1** Optimizer: Adam $d_k = d_v$ (one head): 128 MLP hidden layer: 128 Sequence length: 12 Heads: 8 Activation functions: ReLU Dropout: 0.1 Learning rate: 0.0001 Epochs: 100 Batch size: 256	**Feature input: Input Data_2** Optimizer: Adam $d_k = d_v$ (one head): 128 MLP hidden layer: 128 Sequence length: 5 Heads: 8 Activation functions: ReLU Dropout: 0.1 Learning rate: 0.0001 Epochs: 100 Batch size: 256	**MLP weighting** Optimizer: Adam Hidden layers: [512, 256, 128] activation functions: ReLU learning rate: 0.0001 Epochs: 100 Batch size: 256	**4.2**

Note. Key differences in the models are highlighted in bold.

utilizes Input_Data_2, as described in Section 3.3. Each model's performance is evaluated based on its optimized training using the same amount of data.

Under these conditions, the Dual-Transformer with MLP achieves the best performance, with an RMSE of 4.2 cm, followed by the Dual-Transformer without MLP (RMSE: 7.7 cm), the single Transformer (RMSE: 8.9 cm), and the two-layer and four-layer LSTM models (RMSE: 10.8 and 11 cm, respectively). The model configurations and comparative results are summarized in Table 1.

The upper panel of Figure 5 shows the training loss for all models. It is evident from the zoomed-in portion that the Dual-Transformer with an MLP weighing model achieves a significantly lower loss compared to the other models. Additionally, the loss curve between epochs 90 and 100 reveals the stability of the model, with minimal

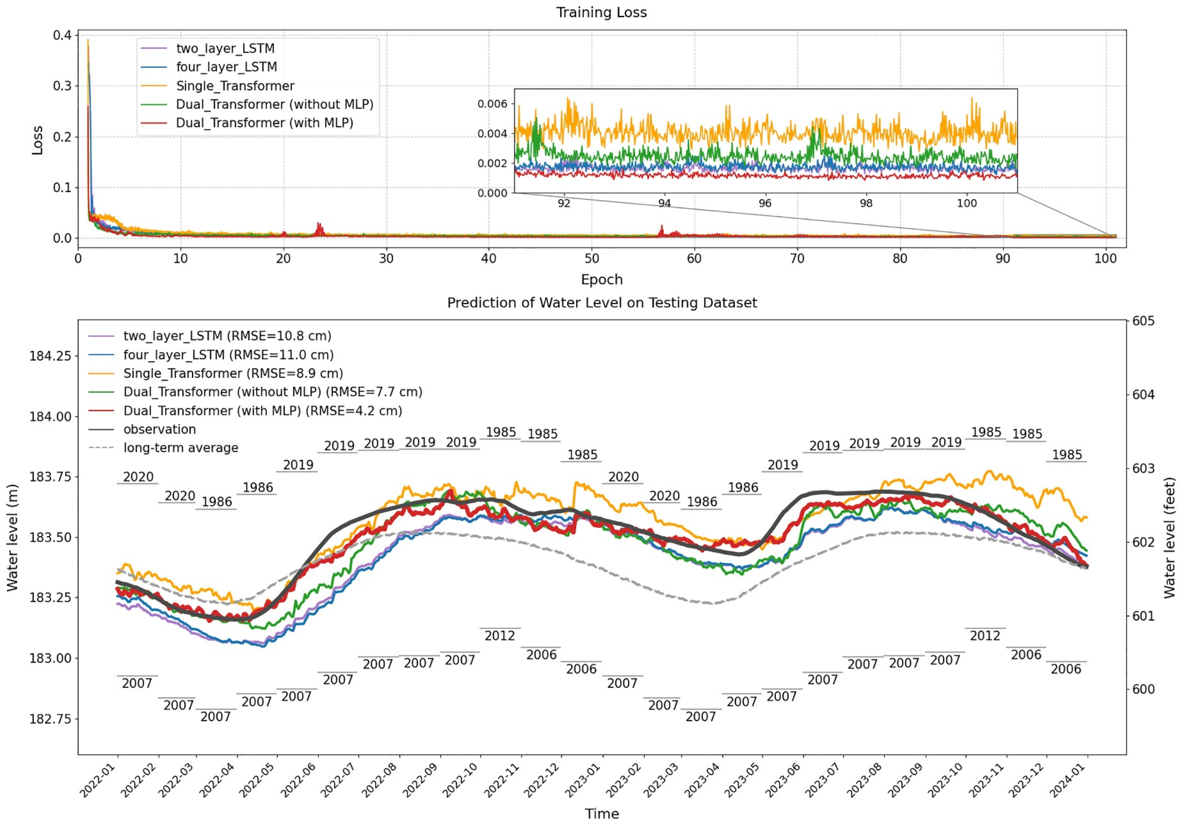

Figure 5. Comparison of five DL models in predicting lake level of Lake Superior. Upper panel: Training loss function values for each model. Lower panel: Model performance in predicting lake levels 6 months ahead, compared with observed values. The long-term average observed lake levels, as well as years with recorded monthly highs and lows, are also shown. All models utilize the same amount of input feature information.

fluctuations around a very low loss value (approximately 0.001). This indicates that the model has effectively learned the underlying patterns in the data and is less prone to overfitting.

However, the loss values alone do not fully reflect the generalization performance of each model. Although the Two-Layer and Four-Layer LSTMs achieve relatively low loss values during training, their performance on the testing data set is notably poor, suggesting that these models have overfitted to the training data. In contrast, the Single Transformer model, despite having a higher and more fluctuating loss curve, produces significantly better predictions on the testing data set, demonstrating improved generalization. The Dual-Transformer without an MLP weighting model further refines the results by reducing both training loss and testing error.

A closer look at the testing results is shown in the lower panel of Figure 5, which presents 6-month-ahead lake level predictions from 2022 to 2023 for all five models after training, compared to the observed lake levels. Predictions from each model are evaluated using Root Mean Squared Error (RMSE) as the accuracy metric, with results summarized in Table 1. The Two-Layer LSTM (RMSE = 10.8 cm) and Four-Layer LSTM (RMSE = 11.0 m) show significant deviations from observed values and fail to capture the seasonality of lake level changes, highlighting their limited ability to generalize beyond the training data. In comparison, the Single Transformer model exhibits a more accurate representation of seasonal fluctuations, achieving a lower RMSE (8.9 cm) than both LSTM models.

Applying the Dual-Transformer model without MLP significantly improves results, with an RMSE of 7.7 cm. However, the absence of an MLP weighting model limits the model to fixed empirical weighting (0.5 and 0.5 for both the Prophet and the Critic) between the two Transformer components. When the MLP is integrated into the Dual-Transformer, it learns to dynamically adjust the weighting between the outputs of the two distinct Transformer components. As a result, the model achieves the lowest RMSE of 4.2 cm and demonstrates excellent

10.1029/2024JH000519

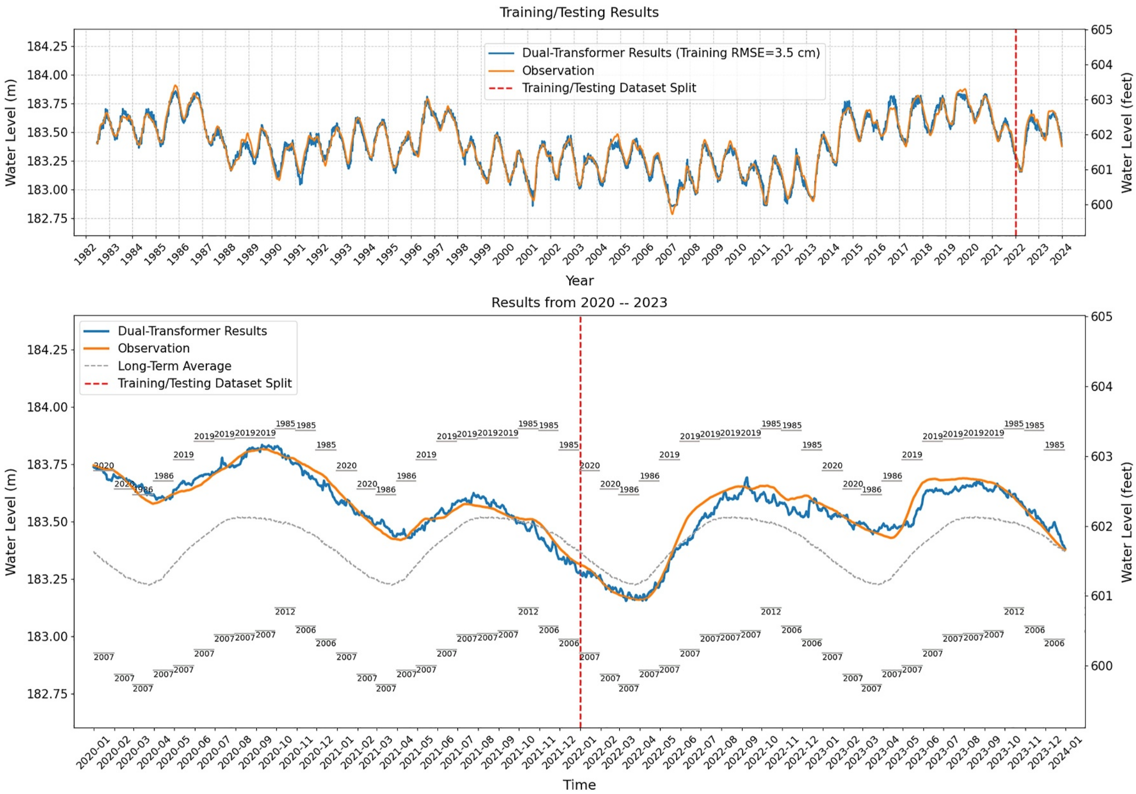

Figure 6. The performance of the Dual Transformer with MLP for Lake Superior on the training data set and testing data set, separated with a vertical line.

performance in predicting lake level changes 6 months ahead, accurately capturing both phase and magnitude throughout the entire test period. It is noteworthy that training the Prophet, the Critic, and the MLP weighting model simultaneously is critical to achieve the best results rather than training the three models individually. Simultaneous training allows for concurrent optimization, enabling the three models to progress cooperatively.

A closer examination of the Dual-Transformer's training process for six-month-ahead predictions reveals that the model not only learns regular seasonal patterns in lake level changes but also captures significant variations over years, including stable periods, gradual and rapid rises to record highs, and sharp declines (Figure 6, upper panel). This indicates that the Dual-Transformer has effectively learned the core dynamics of lake level fluctuations and can model complex relationships between lake levels and both historical and current atmospheric conditions, achieving an RMSE of 3.5 cm and a correlation coefficient of 0.99 on training data set. More importantly, the model maintains such high accuracy on the 2022–2023 testing data set, with an RMSE of 4.2 cm and a correlation coefficient of 0.97. The consistent performance across the training and testing data sets ensures the robustness of the Dual-Transformer's predictive performance (Figure 6).

4.2. Role of the Prophet and the Critic

To examine the role of each component in the Dual-Transformer, this section presents a detailed discussion and analysis of the learning outcomes for each deep learning structure. The final predictions from the Prophet and Critic are flexibly coordinated by an MLP, which automatically calculates the optimal weights for the Prophet and Critic under varying prediction conditions. The resulting weight curves are shown in Figure 7, revealing that the weights vary significantly in response to different lake level conditions. The weight assigned to the Prophet is higher during years or seasons with elevated water levels. Conversely, during years or seasons with lower water levels, the weight assigned to the Critic increases. This suggests that predictions from the Prophet are more reliable during periods of higher water levels, while more significant corrections from the Critic are needed during years and seasons of lower water levels.

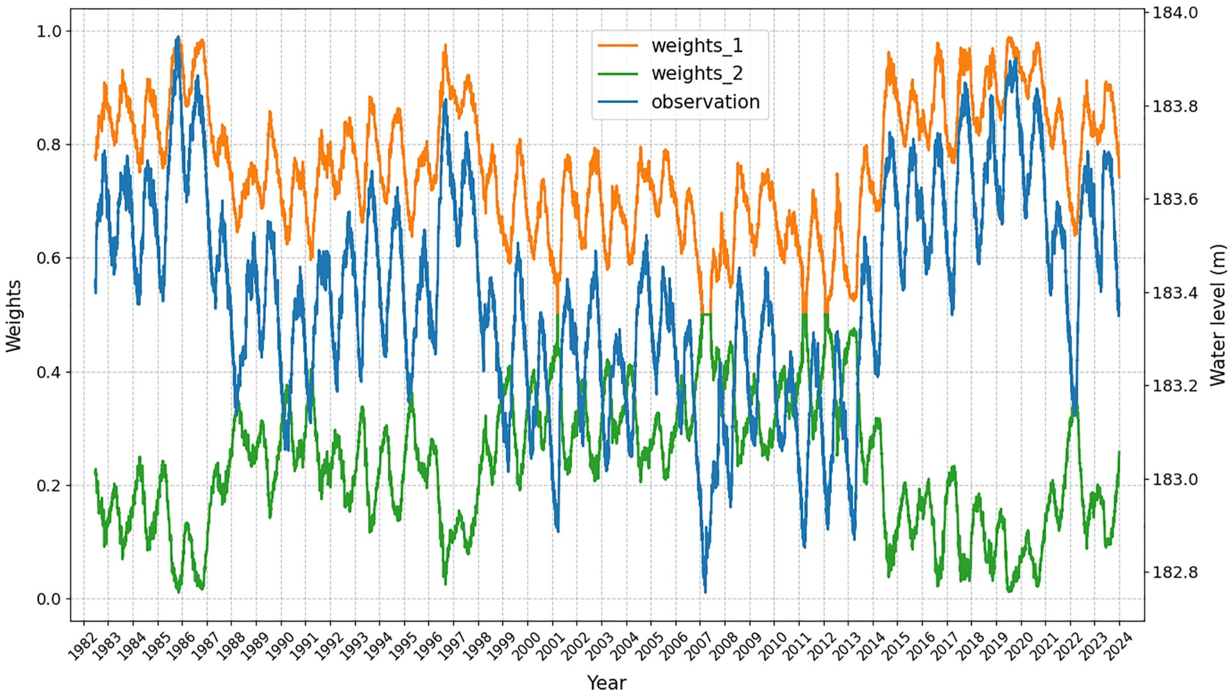

Figure 7. The weights of the Prophet (weights_1) and the Critic (weights_2) are estimated by the MLP along with the lake levels (blue line) for Lake Superior.

To further illustrate how the two components form the final prediction, the predictions from the Prophet, the corrections from the Critic, and the normalized observations of lake level are shown in Figure 8. The results from the Prophet closely follow the target trend, although all predictions from the Prophet are consistently higher than the targets. Meanwhile, the Critic's corrections are negative, offsetting the overestimated lake levels from the

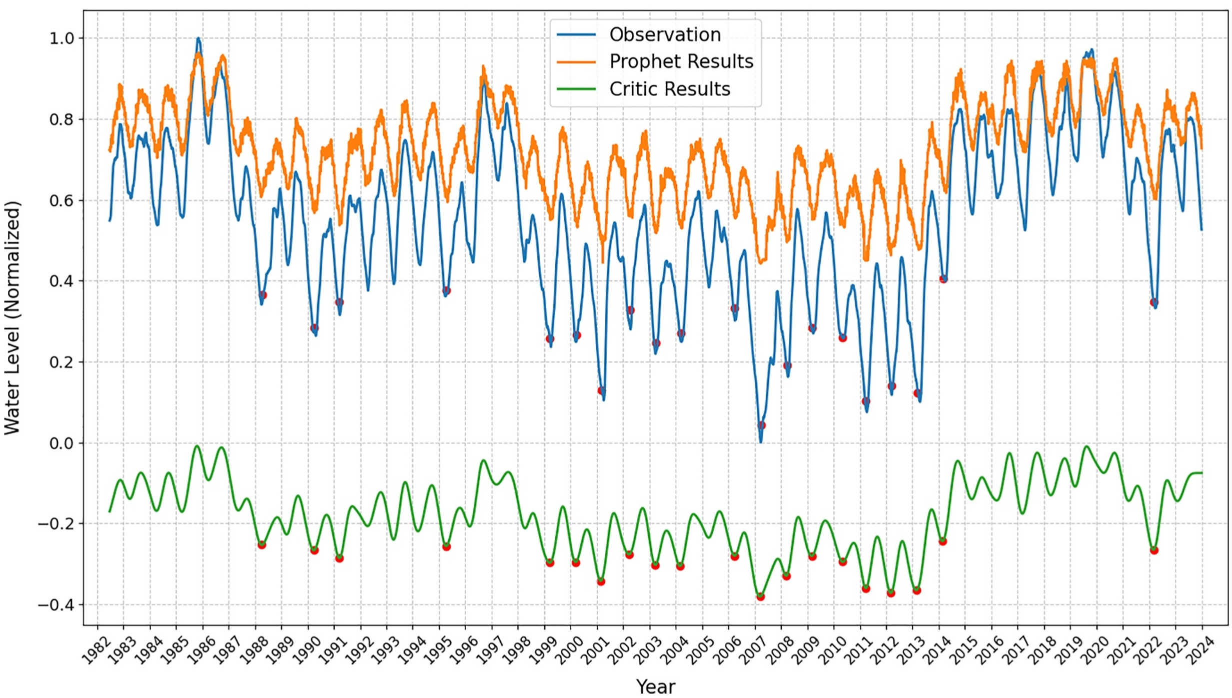

Figure 8. The results from the Prophet and the Critic estimated by the MLP along with the normalized lake levels for Lake Superior. Red dots mark the 20 most significant adjustments made by the Critic to the Prophet's results corresponding to the 20 lowest lake levels, respectively.

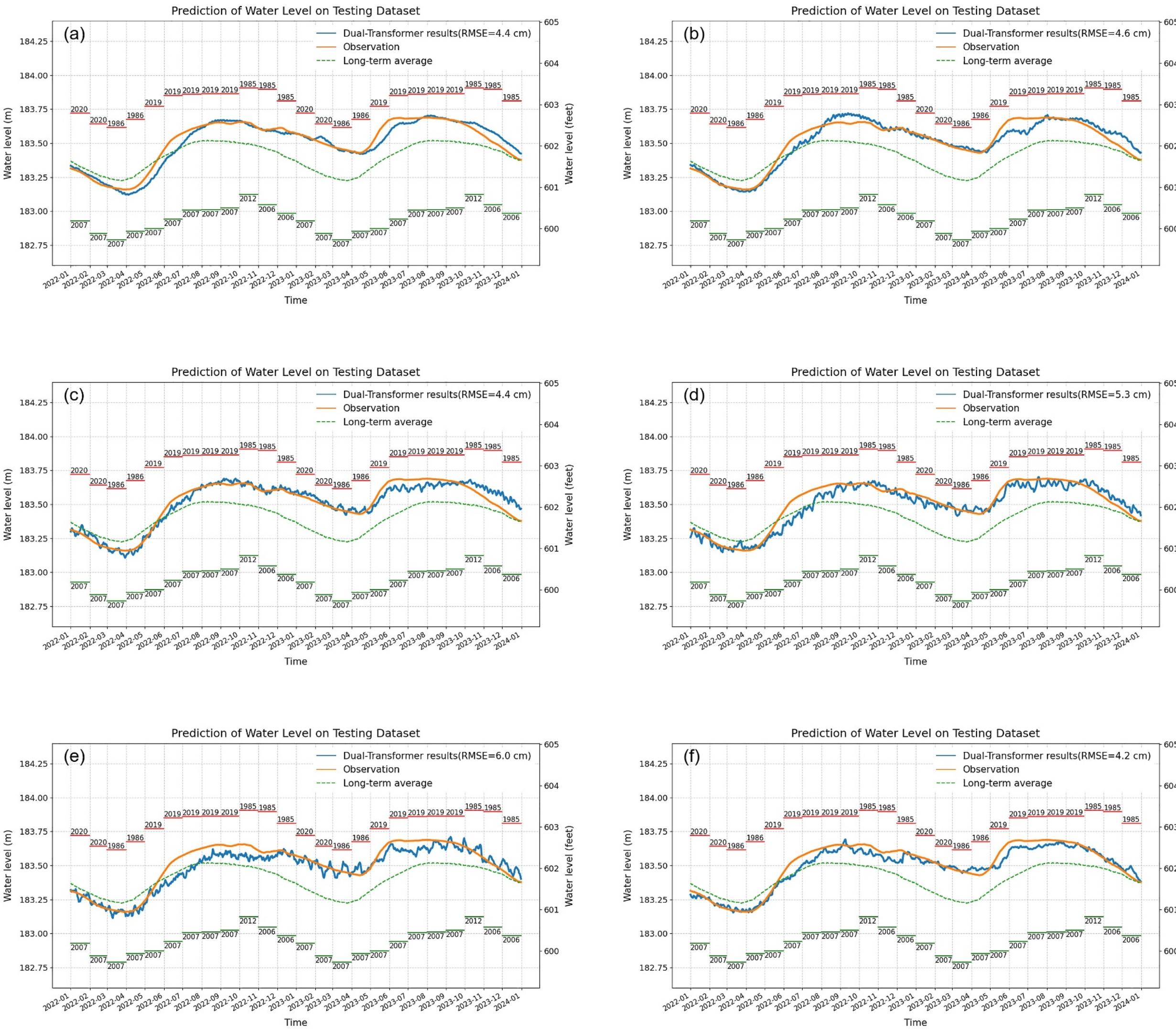

Figure 9. Model performance in predicting lake levels for Lake Superior with a lead time of 1, 2, 3, 4, 5, and 6 months (panels a–f, respectively), compared with observed values.

Prophet. The most intense corrections from the Critic are required at the lowest water levels. For example, red dots mark the 20 most significant adjustments made by the Critic to the Prophet's results corresponding to the 20 lowest water levels, respectively. This synergy clearly demonstrates the necessity and effectiveness of the Dual-Transformer in making seasonal lake level predictions.

4.3. Lake Level Prediction on Different Timescales

In addition to the Dual-Transformer dedicated to predicting lake levels 6 months into the future, we trained Dual-Transformers with identical architectures to predict lake levels at various time scales, including 1 week, 1 month, 2 months, 3 months, 4 months, and 5 months. As an example, Figure 9 shows the Dual-Transformer's performance in predicting lake levels for Lake Superior with lead times of 1–6 months. In all cases, the predictions closely match the target values.

Figure 10 compares the prediction errors for the Dual-Transformer with lead times ranging from 1 week to 6 months, showing RMSEs between 2.5 and 6 cm. The most accuracy results are achieved in the 1-week prediction task, which is understandable as water levels change minimally over such a short period. A "persistence model," which assumes an unchanged lake level over time, shows a significant error increase with a RMSE of 15–25 cm beyond a one-month outlook. The failure of the persistence model reinforces the highly dynamic nature of

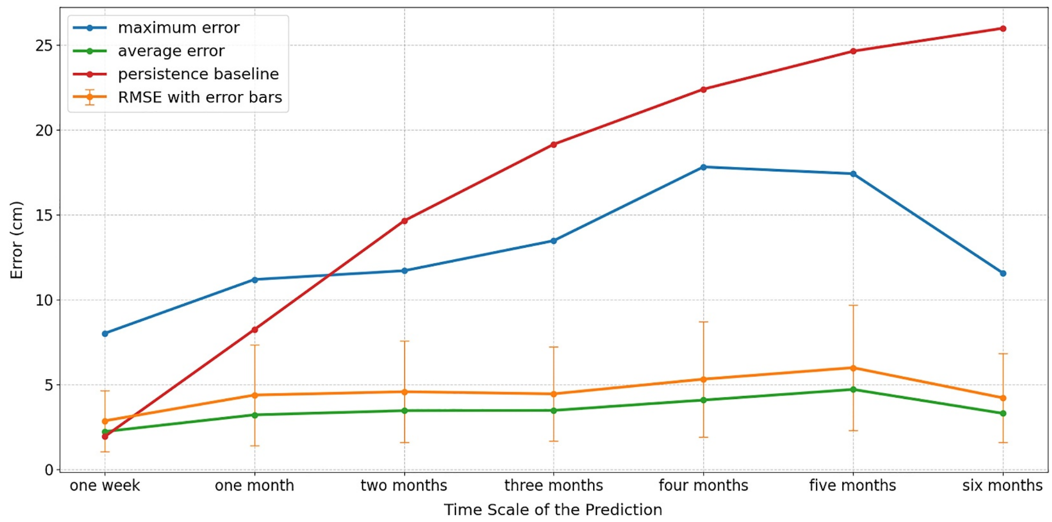

Figure 10. Prediction errors for Dual-Transformers with lead times of 1 week, 1 month, 2 months, 3 months, 4 months, 5 months, and 6 months. We also include "persistence model" as a baseline, which assumes a unchanged lake level over time.

lake level changes. Interestingly, the performance of the 6-month prediction model (RMSE = 4.2 cm) is better than that of the 4- and 5-month prediction models with RMSEs of ~6 cm. Non-linear performance trends across different lead times are common in predictive modeling, as model accuracy does not always decrease monotonically with increasing lead time. This phenomenon is influenced by the complex nature of environmental data and varying degrees of predictability at different time scales. Factors such as data periodicity, model architecture, feature relevance, and training dynamics can cause performance fluctuations across prediction horizons. Consequently, the slightly better performance of the 6-month model compared to the 4- and 5-month models is both plausible and expected within the context of complex geophysical forecasting. Still, this intriguing finding warrants further investigation, and a 6-month time scale may be the most effective choice for seasonal lake level predictions. By combining the predictions from the seven models, we can now predict 1-week, 1-month, 2-month, 3-month, 4-month, 5-month, and 6-month lake levels into the future (Figure 11).

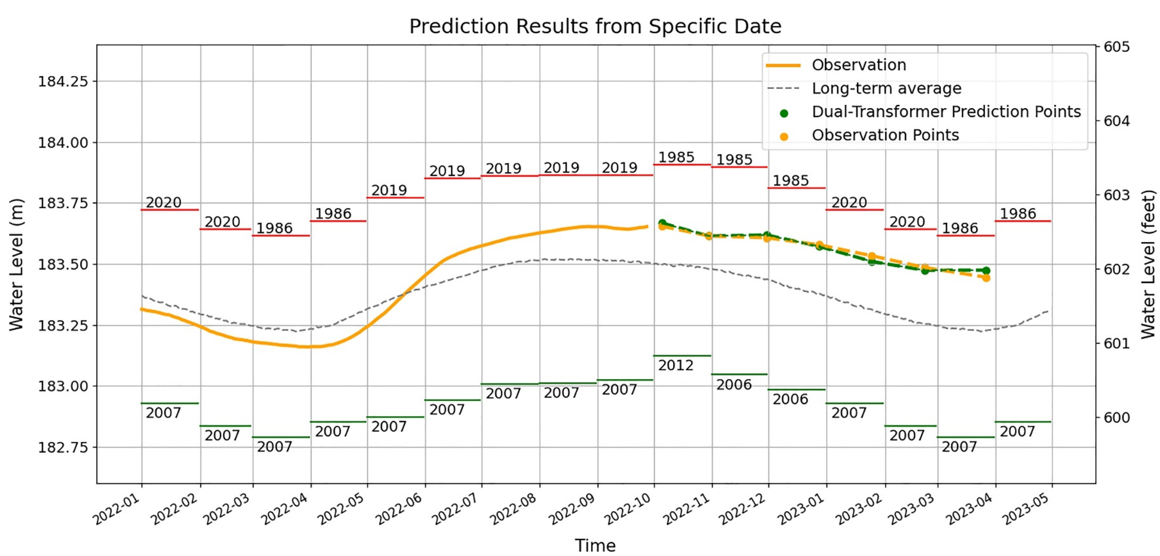

Figure 11. Combined predictions from the seven models, collectively predicting 1-week, 1-month, 2-month, 3-month, 4-month, and 5-month and 6-month lake levels into the future.

10.1029/2024JH000519

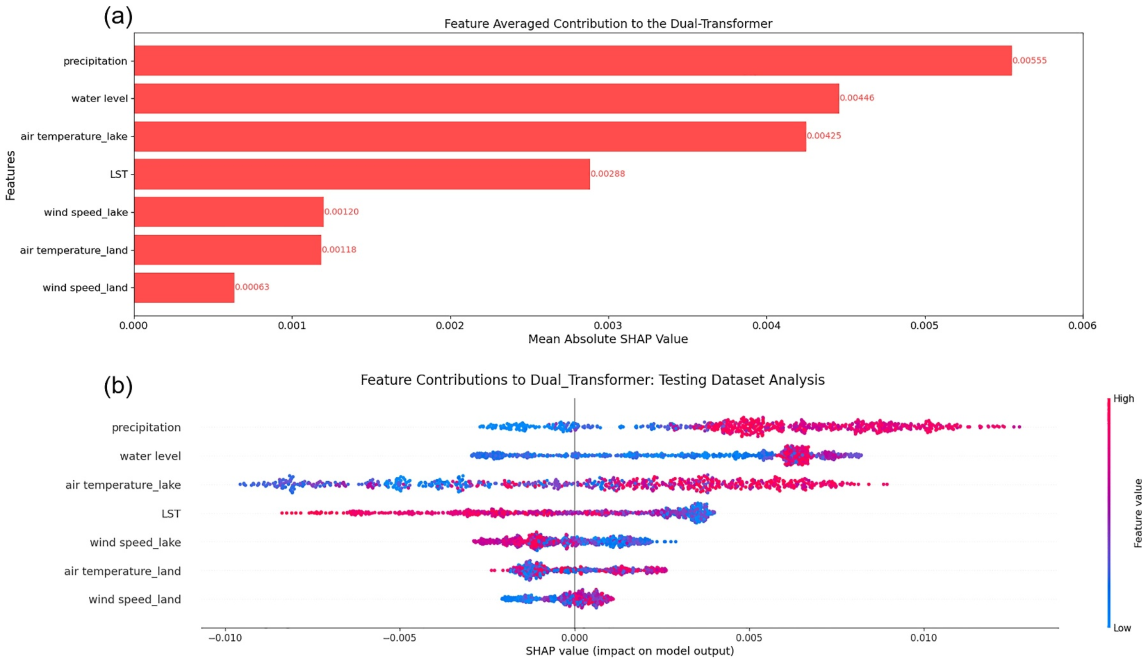

Figure 12. (a) Feature importance, (b) impacts of features on the prediction. In (a), the horizontal axis represents the absolute SHAP value, and a larger absolute SHAP value indicates a more significant influence of the feature on the prediction. In panel (b), the horizontal axis shows the SHAP values of the samples of each feature, where a positive SHAP value means a feature sample generates a higher-than-average water level prediction value and vice versa. A warm (cold) color indicates a feature sample has a high (low) value relative to its average.

4.4. Feature Importance Analysis

While deep learning models are often considered black boxes due to their complex nonlinear structures, it is important to understand and quantify the impact of each feature on the model's predictions. To that end, we apply SHapley Additive exPlanations (SHAP) analysis, a powerful tool for explainable artificial intelligence (XAI) in machine learning (Lundberg & Lee, 2017) to analyze feature contributions by computing Shapley values, which measure the marginal impact of each feature on the model's predictions.

Figure 12 presents the SHAP results for the trained Dual-Transformer model, highlighting three key findings. First, precipitation emerges as the most important feature for lake level prediction, aligning with our hydrological understanding. This is expected, as precipitation inherently encapsulates information for two key components of NBS: precipitation over the lake and river runoff, which is strongly influenced by precipitation over the watershed. Second, temperature-related features (air temperature and LST) exert a greater influence than wind-speed-related features. This is likely because air-lake temperature gradients provide crucial information about lake evaporation, a key component of NBS. Third, lake-related features have a greater influence on water levels than land-related features.

To further evaluate the contribution of different features to the prediction results, we conducted two ablation experiments to test three selected features: precipitation (the most influential feature), wind speed over land (the least important feature), and lake surface temperature (LST) (one of the temperature-related features).

The first method, feature substitution test, involves replacing the examined feature with its long-term average value and assessing the resulting performance change in the trained Dual-Transformer model. The prediction results on the testing data set are presented in Figure 13. Consistent with insights gained from the SHAP analysis, substituting precipitation causes a significant decline in prediction accuracy, with the predicted curve deviating most noticeably from the observations and leading to an increased RMSE of 9.4 cm. In contrast, replacing wind speed over land has the least impact on model performance, yielding an RMSE of 5.4 cm. The predicted lake level remain closely aligned with observations, though it is still less accurate than the original model, which achieved an RMSE of 4.2 cm.

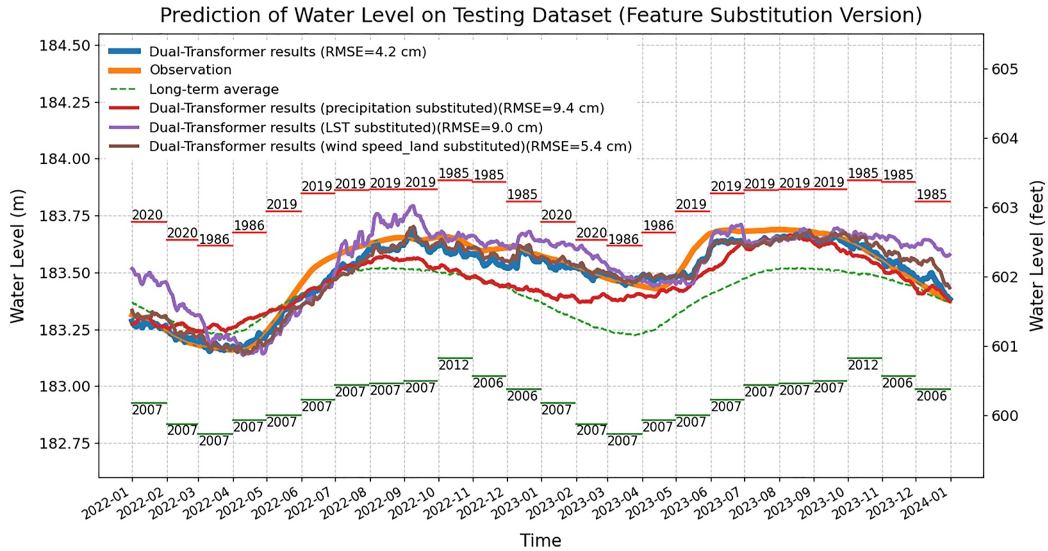

Figure 13. Impact of feature substitution on water level prediction performance. Predicted lake levels with substituted features are shown alongside observed values and the original model predictions.

The second method, feature removal test, involves removing the examined feature from the input data set and retraining the Dual-Transformer model to assess the model's performance. The prediction results on the testing data set are presented in Figure 14. Consistent with the findings from the first method, the worst performance occurs when precipitation was removed, resulting in a significantly increased RMSE of 12.0 cm. In contrast, removing wind speed over land has the least impact on model performance, yielding an RMSE of 5.4 cm.

The results of the feature substitution and removal tests are summarized in Table 2.

4.5. Expansion of the Dual-Transformer Framework to Lake Michigan-Huron and Lake Erie

The previous sections provide a detailed evaluation and analysis of the Dual-Transformer model's performance for Lake Superior, the largest of the Great Lakes. After demonstrating promising results, analyzing the contributions of different model components, and conducting a feature importance analysis to understand the predictive

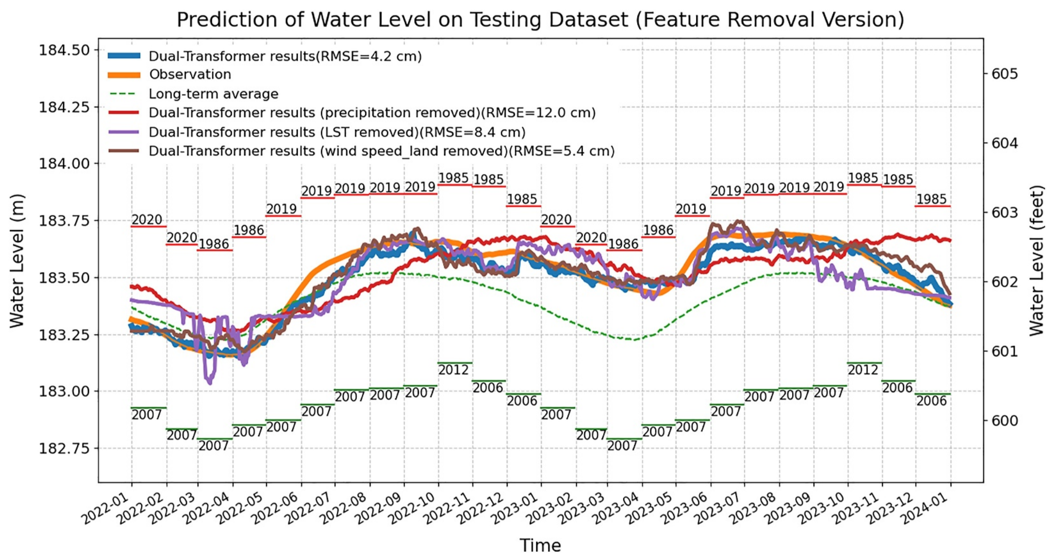

Figure 14. Impact of feature removal on water level prediction performance. Predicted lake levels with removed features are shown alongside observed values and the original model predictions.

Table 2
RMSE Comparison for Feature Substitution and Removal in Prediction Performance

	Original	Feature substitution test			Feature removal test		
Feature	-	Precipitation	LST	Wind speed_land	Precipitation	LST	Wind speed_land
Prediction RMSE (cm)	4.2	9.4	9.0	5.4	12.0	8.4	5.4

mechanisms of the Dual-Transformer, this section assesses whether the model can be effectively applied to other lakes within the Great Lakes system, evaluating the Dual-Transformer framework's robustness and adaptability.

Note that the Great Lakes system functions as a naturally interconnected hydrological network, where water flows from the headwater lake (Lake Superior) through a series of progressively lower lakes—Michigan-Huron, Erie, and Ontario—via connecting rivers and straits, ultimately discharging into the Atlantic Ocean through the St. Lawrence River. Lakes Michigan and Huron are hydraulically connected via the Straits of Mackinac, allowing for the free flow of water between them. As a result, the two lakes are typically treated as a single system for water level projections. This Lake Michigan-Huron convention is widely used by the Great Lakes hydrological community and government agencies.

Additionally, we have intentionally excluded Lake Ontario from our Dual-Transformer framework. This is because Lake Ontario's water levels are significantly influenced by human-made regulation plans, such as Plan 2014, managed by the International Lake Ontario-St. Lawrence River Board. This board regulates outflows from Lake Ontario through the Moses-Saunders Power Dam to balance various interests, including navigation, hydropower, and flood control. As a result, our framework—which is designed to model natural relationships between atmospheric conditions, lake thermal dynamics, and water levels—is not well-suited for Lake Ontario.

We first migrated our Dual-Transformer model to Lake Michigan-Huron and Lake Erie, using the same input features as in the Lake Superior experiment. This is referred to as a baseline configuration. In addition, since Lake Michigan-Huron and Lake Erie are progressively lower lakes, their lake levels are influenced by upper lake levels, incorporating lake level information from the upper lakes may improve prediction accuracy. For Lake Michigan-Huron, the upper lake is Lake Superior, while for Lake Erie, the upper lakes are Lake Michigan-Huron and Lake Superior. To test this hypothesis, we conducted additional experiments by including predicted lake levels from the upper lakes as additional input features. It should be noted that the upper lake levels information for both historical and future periods are predicted from our upper lake Dual-Transformer models, ensuring a rigorous and fair evaluation without incorporating observational lake level data to artificially enhance model performance.

For Lake Michigan-Huron model, we tested including Lake Superior's lake level as an additional input feature and the predicted Lake Superior's lake level by our Lake Superior Dual Transformer used by the Prophet and Critic. For Lake Erie, we evaluated two additional cases: (a) including only Lake Michigan-Huron's lake level predicted by our Lake Michigan-Huron Dual Transformer model, and (b) including the lake levels of both Lake Superior and Lake Michigan-Huron predicted by our Dual Transformer models, respectively. Table 3 provides detailed information on these five experiments.

Figure 15 shows the lake level prediction results for Lake Michigan-Huron, comparing scenarios with and without Lake Superior's lake level as an additional input feature. In the baseline case, excluding Lake Superior's lake level leads to larger deviations from observed lake levels in the testing data set. The largest biases occur during 05/2022–08/2022 and 08/2023–12/2023 (lower panel, Figure 15). In contrast, incorporating Lake Superior's lake level significantly improves model performance, reducing prediction error. Specifically, when Lake Superior's

Table 3
Comparison of Input Features Across Different Experimental Conditions

	Lake Michigan-Huron		Lake Erie		
Feature configuration	Baseline	Baseline + Lake Superior lake level	Baseline	Baseline + Lake Michigan-Huron lake level	Baseline + Lake Michigan-Huron and Superior lake level
Prediction RMSE (cm)	5.7	4.6	11.6	8.4	7.2

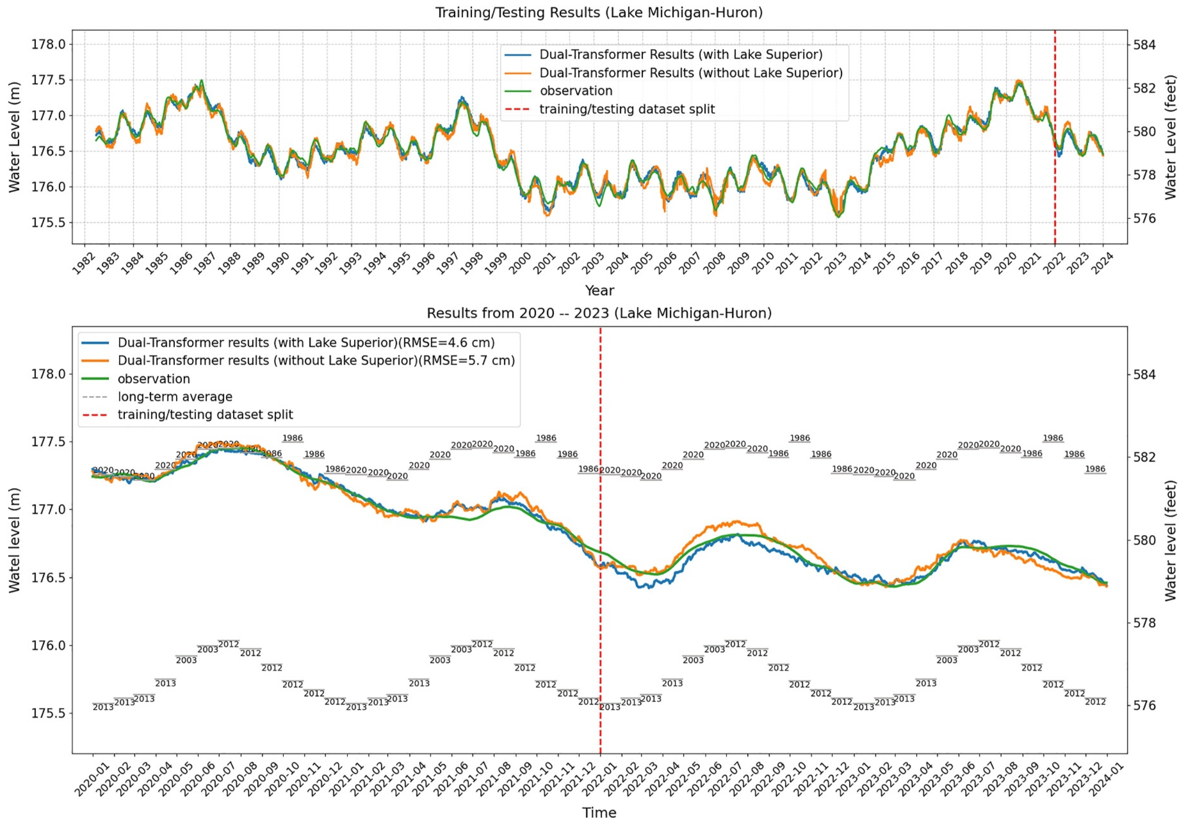

Figure 15. The performance of the Dual-Transformer with MLP for Lake Michigan-Huron on the training and testing data sets, separated by a vertical line.

lake level is included, the RMSE on the testing data set is 4.6 cm, with a correlation coefficient of 0.94. Excluding Lake Superior's lake level increases the RMSE to 5.7 cm and lowers the correlation coefficient to 0.90.

Figure 16 displays the lake level prediction results for Lake Erie under three different scenarios. Similar to the results for Lake Michigan-Huron, incorporating upper lake levels significantly improves model performance. The figure shows that the most accurate predictions occur when the lake levels of both Lake Superior and Lake Michigan-Huron are included as additional input features. In contrast, the worst performance occurs in the baseline case, in which neither upper lake's lake level is incorporated. Specifically, when both upper lakes' lake levels are included, the RMSE on the testing data set is 7.2 cm, with a correlation coefficient of 0.90. When only Lake Michigan-Huron's lake level is included, the RMSE increases to 8.4 cm, and the correlation coefficient slightly drops to 0.89. Finally, excluding both upper lakes' lake levels results in the highest RMSE of 11.6 cm and the lowest correlation coefficient of 0.86.

These results demonstrate that predicting lower lake levels using the Dual-Transformer framework in the Great Lakes system requires a sequential prediction approach rather than treating each lake individually. Specifically, the findings from Figure 16 highlight that the accuracy of Lake Erie's lake level predictions is significantly improved when the model first predicts and incorporates predicted lake levels from the upper lakes—Lake Superior and Lake Michigan-Huron. This approach aligns with the nature that the Great Lakes system functions as a hydrologically connected network, where lake levels in lower lakes are influenced by inflows from the upper lakes. Consequently, attempting to predict lower lake levels in isolation, without incorporating information from higher lakes, leads to significant errors and reduced model reliability in our model framework.

5. Summary and Conclusion

This study introduces a novel Dual-Transformer framework that integrates two modified Transformer models: the Prophet, responsible for predicting underlying trends, and the Critic, which refines these predictions by incorporating future atmospheric conditions. The final forecast is generated by an integrated MLP model that

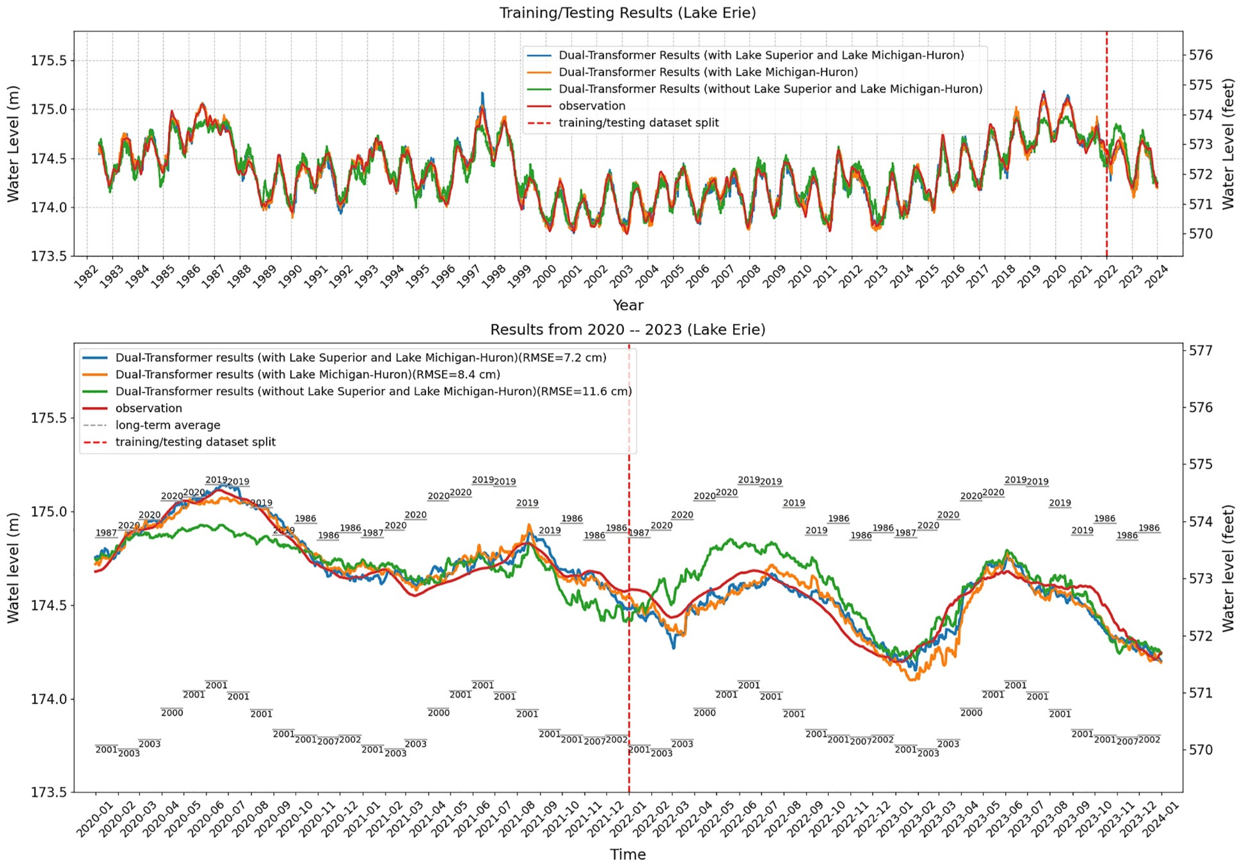

Figure 16. The performance of the Dual-Transformer with MLP for Lake Erie on the training and testing data sets, separated by a vertical line.

dynamically weights the outputs of both models. These three components of the Dual-Transformer are trained simultaneously, enabling coordinated optimization. The Dual-tranformer framework demonstrated exceptional performance and its superiority over other deep learning models under identical input formation.

To ensure practical applicability, the Dual-Transformer model utilizes only input features that are readily accessible from reanalysis data sets or operational model outputs. The selection of these features is grounded in a fundamental understanding of the hydrological processes that govern lake level changes. The chosen atmospheric and lake-related input features are key state variables that mechanistically regulate the NBS components of the lake's water budget, including over-lake precipitation, evaporation, and basin runoff, which collectively drive lake level changes. Unlike traditional mechanistic modeling, which requires a suite of atmospheric, hydrological, and hydrodynamic models to explicitly simulate NBS components, our approach bypasses this need. Instead, the Dual-Transformer model learns and extracts hidden relationships between input features and lake levels, enabling direct prediction without the intermediate step of explicitly resolving individual NBS components.

The key contribution of this study lies in the development of an innovative learning framework that achieves high accuracy in predicting lake levels with a lead time of up to six months—assuming the input features are relatively accurate, as demonstrated using CFSR reanalysis data. Since any model's performance relies on the accuracy of input features, any degradation in their quality will result in decreased prediction accuracy. In real-world applications, the input features for the Critic component of the Dual-Transformer, which takes future atmospheric information, would have to come from a forecast model rather than a reanalysis data set. As a result, the Dual-Transformer model may not achieve the same prediction accuracy demonstrated in this study. However, this discrepancy arises from the degradation of input feature quality rather than a limitation of the Dual-Transformer framework itself. Addressing this issue and improving atmospheric forecasting models is a crucial task for the atmospheric modeling community but falls beyond the scope of this study. Our work represents a step toward improving lake level prediction rather than a definitive solution to the challenge.

It should be highlighted that both mechanistic modeling and machine learning-based approaches play crucial yet complementary roles in Great Lakes water level simulation and prediction. Mechanistic models, which have been widely used in Great Lakes studies, explicitly resolve the physical processes governing lake level changes and incorporate the impact of regulation planning. These models provide crucial insights into the underlying hydrological dynamics and are essential for process-based understanding and scenario analysis. In contrast, our Dual-Transformer framework offers an alternative data-driven approach that learns hidden relationships between atmospheric and lake-related variables, enabling direct lake level prediction and serving as a complementary tool that has potential to improve prediction accuracy, provide rapid assessments, and integrate with traditional methods to advance water resource management and decision-making in the Great Lakes region.

Lastly, the Dual-Transformer model's computational efficiency is noteworthy. It operates six orders of magnitude faster than conventional mechanistic models used in our previous study, Kayastha et al. (2022), delivering predictions in less than one second on a standard personal computer. This efficiency demonstrates its potential to be further developed as a data-driven predictive framework in the goal of improving real-time end-to-end water management. The Dual-Transformer model should be continuously updated to adapt to evolving climate and weather conditions, ensuring robust performance under changing climatic conditions. Additionally, integrating routing and regulation models will be essential in future research efforts.

Data Availability Statement

The source codes of the Dual-Transformer, Single Transformer, and LSTM developed in this study are available at Chen and Xue (2025). Observation data sets used in this study can be retrieved from the following sources: Atmospheric data were obtained from the Climate Forecast System Reanalysis (CFSR), available from January 1979 to March 2011 (Saha et al., 2010), and its successor, the Climate Forecast System Version 2 (CFSv2) analysis product, which extends from 2011 onward (Saha et al., 2011). Lake surface temperature (LST) data were sourced from the Great Lakes Surface Environmental Analysis (GLSEA), provided by the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) (CoastWatch Great Lakes Node. n.d.). Since GLSEA data is only available from 1995 onward, LST data from 1981 to 1994 was obtained from a reconstructed data set using an LSTM model by Kayastha et al. (2023b). Lake water levels are retrieved from NOAA Water Levels (n.d.) and Fisheries and Oceans Canada. (n.d.).

Acknowledgments
This study is supported by the U.S. Department of Energy, Office of Science, under Award Number DE-SC0024446 as part of the Center for Climate-driven Hazard Adaptation, Resilience, and Mitigation (C-CHARM). This research is also supported by the U.S. National Science Foundation under Award Number 2438826. The statements, findings, conclusions, and recommendations expressed herein are those of the authors and do not necessarily reflect the views of the United States Government or any agency thereof. This is Contribution No. 126 of the Great Lakes Research Center at Michigan Technological University. We thank Dr. Chenfu Huang for processing the lake water level data and Ms. Congling Zhang for developing the initial version of the LSTM model.

References

Barzegar, R., Aalami, M. T., & Adamowski, J. (2021). Coupling a hybrid CNN-LSTM deep learning model with a boundary corrected maximal overlap discrete wavelet transform for multiscale lake water level forecasting. *Journal of Hydrology*, *598*, 126196. https://doi.org/10.1016/j.jhydrol.2021.126196

Bennington, V., Notaro, M., & Holman, K. D. (2014). Improving climate sensitivity of deep lakes within a regional climate model and its impact on simulated climate. *Journal of Climate*, *27*(8), 2886–2911. https://doi.org/10.1175/jcli-d-13-00110.1

Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*(1), 376–399. https://doi.org/10.1029/2018ms001472

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. https://doi.org/10.1029/2018gl078510

Chen, Y., & Xue, P. (2025). Trained deep learning models for "dual-transformer deep learning framework for seasonal forecasting of Great Lakes water levels" [Software]. *Journal of Geophysical Research - Machine Learning and Computation (2025) (1.0). Zenodo*. https://doi.org/10.5281/zenodo.15276228

CoastWatch Great Lakes Node. (n.d.). Sea surface temperature (SST) from Great Lakes Surface Environmental Analysis (GLSEA). In *Geodetic coordinate system (LAT, LON)* (pp. 1995–2023). NOAA. Retrieved from https://apps.glerl.noaa.gov/erddap/files/GLSEA_GCS/

Dai, Z. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv Preprint arXiv:1901.02860*.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv Preprint arXiv:1702.08608*.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv Preprint arXiv:2010.11929*.

Feng, X., Ma, G., Su, S.-F., Huang, C., Boswell, M., & Xue, P. (2020). A multi-layer perceptron approach for accelerated wave forecasting in Lake Michigan. *Ocean Engineering*, *211*, 107526. https://doi.org/10.1016/j.oceaneng.2020.107526

Fisheries and Oceans Canada. (n.d.). Canadian tides and water levels data archive. Retrieved from https://www.isdm-gdsi.gc.ca/isdm-gdsi/twl-mne/inventory-inventaire/list-liste-eng.asp?user=ISDM-GDSI&region=MEDS&tst=1

Gronewold, A., Do, H., Mei, Y., & Stow, C. (2021). A tug-of-war within the hydrologic cycle of a continental freshwater basin. *Geophysical Research Letters*, *48*(4), e2020GL090374. https://doi.org/10.1029/2020gl090374

Gronewold, A. D., & Stow, C. A. (2014). Water loss from the Great Lakes. *Science*, *343*(6175), 1084–1085. https://doi.org/10.1126/science.1249978

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, C., Anderson, E., Liu, Y., Ma, G., Mann, G., & Xue, P. (2021a). Evaluating essential processes and forecast requirements for meteotsunami-induced coastal flooding. *Natural Hazards*, *110*(3), 1–26. https://doi.org/10.1007/s11069-021-05007-x

Huang, C., Zhu, L., Ma, G., Meadows, G. A., & Xue, P. (2021b). Wave climate associated with changing water level and ice cover in Lake Michigan. *Frontiers in Marine Science*, *8*. https://doi.org/10.3389/fmars.2021.746916

Hunter, T. S., Clites, A. H., Campbell, K. B., & Gronewold, A. D. (2015). Development and application of a North American Great Lakes hydrometeorological database—Part I: Precipitation, evaporation, runoff, and air temperature. *Journal of Great Lakes Research*, *41*(1), 65–77. https://doi.org/10.1016/j.jglr.2014.12.006

Kayastha, M. B., Liu, T., Titze, D., Havens, T. C., Huang, C., & Xue, P. (2023a). Reconstructing 42 Years (1979–2020) of Great Lakes surface temperature through a deep learning approach. *Remote Sensing*, *15*(17), 4253. https://doi.org/10.3390/rs15174253

Kayastha, M. B., Liu, T., Titze, D., Havens, T. C., Huang, C., & Xue, P. (2023b). Reconstructing 42 Years (1979–2020) of Great Lakes surface temperature through a deep learning approach [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.7995786

Kayastha, M. B., Ye, X., Huang, C., & Xue, P. (2022). Future rise of the Great Lakes water levels under climate change. *Journal of Hydrology*, *612*, 128205. https://doi.org/10.1016/j.jhydrol.2022.128205

Kurt, O. (2024). Model-based prediction of water levels for the Great Lakes: A comparative analysis. *Earth Science Informatics*, *17*(4), 1–17. https://doi.org/10.1007/s12145-024-01341-3

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., & Yan, X. (2019). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in Neural Information Processing Systems*, *32*. https://doi.org/10.48550/arXiv.1907.00235

Liu, Y., Racah, E., Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., et al. (2016). Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv Preprint arXiv:1605.01156*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4768–4777.

Lv, Z., Zhang, S., Jin, J., Wu, Y., & Ek, M. B. (2019). Coupling of a physically based lake model into the climate forecast system to improve winter climate forecasts for the Great Lakes region. *Climate Dynamics*, *53*(9), 6503–6517. https://doi.org/10.1007/s00382-019-04939-2

MacKay, M., & Seglenieks, F. (2013). On the simulation of Laurentian Great Lakes water levels under projections of global climate change. *Climatic Change*, *117*(1), 55–67. https://doi.org/10.1007/s10584-012-0560-z

Melillo, J. M., Richmond, T., Yohe, G., & others. (2014). Climate change impacts in the United States. *Third National Climate Assessment*, *52*.

Millerd, F. (2005). The economic impact of climate change on Canadian commercial navigation on the Great Lake. *Canadian Water Resources Journal*, *30*(4), 269–280. https://doi.org/10.4296/cwrj3004269

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). ClimaX: A foundation model for weather and climate. *arXiv Preprint arXiv:2301.10343*.

NOAA Water Levels. (n.d.). *Tides and currents: Water levels*. NOAA. Retrieved from https://tidesandcurrents.noaa.gov/stations.html?type=Water+Levels

Notaro, M., Zhong, Y., Xue, P., Peters-Lidard, C., Cruz, C., Kemp, E., et al. (2021). Cold season performance of the NU-WRF regional climate model in the Great Lakes Region. *Journal of Hydrometeorology*, *22*(9), 2423–2454. https://doi.org/10.1175/jhm-d-21-0025.1

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., et al. (2022). Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv Preprint arXiv:2202.11214*.

Pringle, W. J., Huang, C., Xue, P., Wang, J., Sargsyan, K., Kayastha, M. B., et al. (2024). Coupled Lake-Atmosphere-Land Physics Uncertainties in a Great Lakes Regional Climate Model. *Authorea Preprints*.

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Rasp, S., & Thuerey, N. (2021). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, *13*(2), e2020MS002405. https://doi.org/10.1029/2020ms002405

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat, F. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., et al. (2010). NCEP climate forecast system reanalysis (CFSR) selected hourly time-series products, January 1979 to December 2010. *Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory*. https://doi.org/10.5065/D6513W89

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., et al. (2011). NCEP climate forecast system version 2 (CFSv2) selected hourly time-series products, January 2011 to December 2023. *Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory*. https://doi.org/10.5065/D6N877VB

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., et al. (2014). The NCEP climate forecast system version 2. *Journal of Climate*, *27*(6), 2185–2208. https://doi.org/10.1175/jcli-d-12-00823.1

Samek, W. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv Preprint arXiv:1708.08296*.

Scher, S. (2018). Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, *45*(22), 12–616. https://doi.org/10.1029/2018gl080704

Scher, S., & Messori, G. (2018). Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2830–2841. https://doi.org/10.1002/qj.3410

Stakhiv, E., & Stewart, B. (2010). Needs for climate information in support of decision-making in the water sector. *Procedia Environmental Sciences*, *1*, 102–119. https://doi.org/10.1016/j.proenv.2010.09.008

Trebitz, A. S., & Hoffman, J. C. (2015). Coastal wetland support of Great Lakes fisheries: Progress from concept to quantification. *Transactions of the American Fisheries Society*, *144*(2), 352–372. https://doi.org/10.1080/00028487.2014.982257

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, Q., & Wang, S. (2020). Machine learning-based water level prediction in Lake Erie. *Water*, *12*(10), 2654. https://doi.org/10.3390/w12102654

Xue, P., Huang, C., Zhong, Y., Notaro, M., Kayastha, M. B., Zhou, X., et al. (2024). Enhancing winter climate simulations of the Great Lakes: Insights from a new coupled lake-ice-atmosphere (CLIAv1) model on the importance of integrating 3D hydrodynamics with a regional climate model. *Geoscientific Model Development Discussions*, *2024*, 1–41.

Xue, P., Pal, J. S., Ye, X., Lenters, J. D., Huang, C., & Chu, P. Y. (2017). Improving the simulation of large lakes in regional climate modeling: Two-way lake–atmosphere coupling with a 3D hydrodynamic model of the Great Lakes. *Journal of Climate*, *30*(5), 1605–1627. https://doi.org/10.1175/jcli-d-16-0225.1

Xue, P., Schwab, D. J., & Hu, S. (2015). An investigation of the thermal response to meteorological forcing in a hydrodynamic model of Lake Superior. *Journal of Geophysical Research: Oceans*, *120*(7), 5233–5253. https://doi.org/10.1002/2015jc010740

Xue, P., Wagh, A., Ma, G., Wang, Y., Yang, Y., Liu, T., & Huang, C. (2022). Integrating deep learning and hydrodynamic modeling to improve the Great Lakes Forecast. *Remote Sensing*, *14*(11), 2640. https://doi.org/10.3390/rs14112640

Xue, P., Ye, X., Pal, J. S., Chu, P. Y., Kayastha, M. B., & Huang, C. (2022). Climate projections over the Great Lakes Region: Using two-way coupling of a regional climate model with a 3-D lake model. *Geoscientific Model Development Discussions*, *15*(11), 1–37. https://doi.org/10.5194/gmd-15-4425-2022

Ye, X., Chu, P. Y., Anderson, E. J., Huang, C., Lang, G. A., & Xue, P. (2020). Improved thermal structure simulation and optimized sampling strategy for Lake Erie using a data assimilative model. *Journal of Great Lakes Research*, *46*(1), 144–158. https://doi.org/10.1016/j.jglr.2019.10.018

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, *31*(7), 1235–1270. https://doi.org/10.1162/neco_a_01199

Yuan, J., Liang, J.-H., Chassignet, E. P., Zavala-Romero, O., Wan, X., & Cronin, M. F. (2024). The K-profile parameterization augmented by deep neural networks (KPP_DNN) in the general ocean turbulence model (GOTM). *Journal of Advances in Modeling Earth Systems*, *16*(9), e2024MS004405. https://doi.org/10.1029/2024ms004405