Journal of Hydrology: Regional Studies 59 (2025) 102339

Contents lists available at ScienceDirect

Journal of Hydrology: Regional Studies

journal homepage: www.elsevier.com/locate/ejrh

Enhancing the accuracy and generalizability of reference evapotranspiration forecasting in California using deep global learning

Arman Ahmadi [a,*], Andre Daccache [b], Minxue He [c], Peyman Namadi [c], Alireza Ghaderi Bafti [d], Prabhjot Sandhu [c], Zhaojun Bai [e], Richard L. Snyder [f], Tariq Kadir [c]

[a] *Department of Environmental Science, Policy and Management, University of California, Berkeley, Berkeley, CA, USA*
[b] *Department of Biological and Agricultural Engineering, University of California, Davis, Davis, CA, USA*
[c] *California Department of Water Resources, Sacramento, CA, USA*
[d] *Department of Ocean Engineering, University of Rhode Island, Kingston, RI, USA*
[e] *Department of Computer Science, University of California, Davis, Davis, CA, USA*
[f] *Department of Land, Air and Water Resources, University of California, Davis, CA 95616, USA*

ARTICLE INFO

Keywords:
Time series forecasting
Global learning
Deep learning
Reference evapotranspiration
Water management

ABSTRACT

Study region: This research focuses on the Central Valley of California, a climatically homogeneous region known for its significant agricultural productivity and reliance on extensive irrigation. Our study utilizes monthly reference evapotranspiration (ET_O) time series data from 55 standardized weather stations as part of the California Irrigation Management Information System (CIMIS).

Study focus: ET_O is a critical component of regional water cycles, indicating atmospheric water demand. This study evaluates the potential of deep learning (DL) models for ET_O forecasting, particularly emphasizing the efficacy of a global learning scheme compared to traditional local learning. Global learning involves training forecasting models on pooled data from multiple time series, tested over new instances. We compared the performance of statistical models and advanced DL models, demonstrating significant accuracy enhancements in global learning schemes. We also explored automatic hyperparameter optimization for these models to achieve state-of-the-art forecasting accuracy, yielding RMSE values below 10 mm/month for one-year-ahead forecasts on new, unseen stations.

New hydrological insight for the region: Applying global learning methodologies to DL models markedly improved forecasting performance, showcasing an ability to generalize findings to ungauged regions and even newly established weather stations. This suggests a promising avenue for enhancing water resource management efficiency in data-scarce areas. Our findings argue that such data-centric methodological shifts could play a critical role in better managing the irrigation demands of the Central Valley, thereby supporting sustainable water usage and agricultural productivity in the region.

* Corresponding author.
E-mail address: a.ahmadi@berkeley.edu (A. Ahmadi).

https://doi.org/10.1016/j.ejrh.2025.102339
Received 6 January 2025; Received in revised form 6 March 2025; Accepted 19 March 2025
Available online 25 March 2025

1. Introduction

Hydrometeorological time series forecasting, particularly at the monthly scale, is crucial in water resources planning and management (Milly et al., 2008; Li et al., 2017). Accurate forecasts of variables such as precipitation, streamflow, and evapotranspiration are fundamental for effective decision-making in various sectors, including agriculture, hydropower generation, reservoir operations, and environmental conservation (Gleick, 2003; Chen et al., 2024; Quinn et al., 2024). Monthly time series forecasts are critical as they align with many water management activities' operational and planning timescales, balancing shorter-term weather predictions and longer-term climate projections (Mehdizadeh et al., 2019; Le et al., 2024).

Reference evapotranspiration (ET_O) is the amount of water that vaporizes from a virtual 0.12 m tall, vegetated surface using empirical equations to estimate canopy and aerodynamic resistances in a modified Penman-Monteith equation (Allen et al., 2005, 2006). In reality, ETo is approximately equal to the evapotranspiration of a broad expanse of a healthy, well-watered, cool-season grass pasture close to 0.12 m height (Allen et al., 1998). Assuming the weather data are measured over a large expanse of well-watered pasture maintained at approximately 0.12 m height, ET_O is a meteorological variable driven by air temperature, humidity, solar radiation, and wind speed (Allen et al., 1998, 2005, 2006; Ahmadi et al., 2022). The standardized ET_O is used as a direct measure of atmospheric evaporative power or water demand, making it a significant component in regional water cycles. Like other hydrometeorological time series at monthly time scales, monthly ET_O time series are primarily seasonality-driven, with a trivial trend component. This study focuses on one year-ahead monthly ET_O forecasting as an instance of hydrometeorological time series forecasting useful for regional water resources planning and management.

The Central Valley of California is chosen as the case study for this research. The Central Valley is one of the world's most productive agricultural regions, providing over a third of the vegetables and two-thirds of the fruits and nuts consumed in the United States (Hanak et al., 2011). Due to its extensive agricultural activities and limited in-season precipitation, the region is primarily dependent on irrigation for crop production. Optimal irrigation planning is imperative for sustainable water resources and food production in the Central Valley, as excessive irrigation depletes valuable ground and surface water resources and exacerbates soil salinity issues, further threatening agricultural sustainability. Consequently, accurate forecasts of ET_O are crucial for efficient water management and irrigation practices in this region.

A considerable amount of research exists on using data-driven approaches to estimate reference and actual evapotranspiration (Zhang et al., 2020; Chen et al., 2020; Dong et al., 2022; Ahmadi et al., 2024; Bafti et al., 2024). Moreover, numerous studies have focused on forecasting reference (Torres et al., 2011; Lee et al., 2024; Granata et al., 2024) and actual (Talib et al., 2021; Granata and Di Nunno, 2021; Babaeian et al., 2022) evapotranspiration with data-driven methodologies. For further instance, Gocić et al. (2015) explored soft computing methods for monthly ET_O forecasting, while Karbasi et al. (2022) investigated a hybrid deep learning model for weekly ET_O forecasting. Ferreira and da Cunha (2020) utilized deep learning (DL) models for multi-step ahead daily ET_O forecasts, and Chia et al. (2022) employed deep neural networks for long-term forecasting of monthly mean ET_O. In a recent study, Ahmadi et al. (2023) analyzed the accuracy, complexity, and data efficiency of several statistical, machine learning (ML), and DL methods for monthly ET_O forecasting. The current study is a follow-up of this research, focusing on global learning and cutting-edge DL models for 12-month ahead ET_O forecasting. The forecasting models evaluated in this study include widely used statistical methods frequently employed in hydrometeorological time series forecasting and state-of-the-art deep learning architectures developed explicitly for time series forecasting. These architectures have demonstrated superior performance on benchmark datasets and in forecasting competitions (Lim and Zohren, 2021; Makridakis et al., 2022).

Forecasting evapotranspiration using ML and DL has seen various advancements, such as the development of ensemble DL architectures (Granata and Di Nunno, 2021), Multilayer Perceptron-Random Forest stacked models (Granata et al., 2024), and hybrid ML models (Lee et al., 2024). These studies focus on optimizing model architectures to improve forecasting accuracy. In contrast, our research introduces a novel training strategy known as *global learning*, which represents a shift from model-centric improvements to data-centric innovations. Here, we show how this approach enhances the generalizability and accuracy of DL models by training them on diverse datasets from multiple locations, thereby significantly reducing the risk of overfitting. Our study advances beyond incremental improvements by comprehensively evaluating global learning's effectiveness in hydrometeorological time series forecasting using weather station data.

Contrary to local learning, which trains a forecasting model on a single time series, global learning schemes use a pool of multiple time series from different sources as the training set (Bandara et al., 2020; Smyl, 2020; Salinas et al., 2020). Global learning is specifically helpful for hydrometeorological time series forecasting when several stations in a climatically homogeneous region measure the variable of interest. For example, this study uses the ET_O time series from 47 weather stations in the Central Valley as its training set and tests the performance of the globally learned DL models on another eight unseen stations that constitute the test set. By leveraging data from where it is available, global learning enables reliable forecasting in poorly gauged regions with similar climatic characteristics and recently established stations.

This research aims to quantitatively analyze the effects of global learning on the performance of deep forecasting models in monthly ET_O forecasting. The research objectives are: 1) to assess the forecasting accuracy of statistical and advanced DL models in local learning scheme, 2) to compare the DL models' performance in local versus global learning, and 3) to achieve state-of-the-art deep forecasting accuracy through automatic hyperparameter optimization. In addition to quantifying performance improvements, our study discusses how global learning can enhance the generalizability of deep forecasting models and mitigate their risk of overfitting. This generalizability is especially important for partially-gauged regions and newly established weather stations. We hypothesize that globally learned DL models outperform both locally learned DL models and the best-performing statistical models while automatically tuned deep forecasting models achieve the highest accuracy. We also hypothesize that globally learned models are more generalizable to unseen instances and less prone to overfitting.

2. Study area and dataset

The Central Valley of California is the case study of this research (Fig. 1). The Central Valley, relying on extensive irrigation, is one of the most agriculturally productive regions worldwide (Hanak, 2011). This study uses monthly ET_O time series at 55 locations in the Central Valley from the *California Irrigation Management Information System (CIMIS)* program. CIMIS comprises over 145 automated and standardized weather stations to aid irrigators and water managers in planning and decision-making. CIMIS employs the Penman-Monteith equation and a modified version of Penman's equation to calculate ET_O. Hourly weather data is used to determine hourly ET_O, which is then summed over 24 h (midnight to midnight local time) to estimate daily ET_O. The monthly ET_O values reported by the CIMIS portal are aggregates of these daily values in metric units (mm). Further details about CIMIS data and the Penman-Monteith equation can be found in Ahmadi et al. (2022).

According to the CIMIS website (accessed July 2023), 42 inactive and 38 active stations are in the Central Valley (https://cimis.water.ca.gov/). After reviewing the maintenance and data quality reports on the CIMIS website, we eliminated eight stations: three stations had poor data quality, two stations had bare reference surfaces, two stations had alfalfa reference surfaces, and one was a non-ET_O site. Of the remaining 72 stations, 17 were excluded due to insufficient time series data. Ultimately, 55 stations with more than six consecutive years of data remained. Monthly data from January 1986 to June 2023 is downloaded for all stations. Data from eight active stations (i.e., stations with data available until June 2023) are used as the test set, and the remaining 47 stations constitute the training set (Fig. 1). The zoning classification of Fig. 1 defines California's homogeneous ET_O zones. More information about these zones is found in Ahmadi et al. (2022).

As Fig. 1 shows, both training and test sets are widely spread over the Central Valley and uniformly distributed. To ensure the representativeness of the test set and its statistical similarity to the training set, we compared the ET_O distributions in these sets (Fig. 2). Fig. 2 suggests that the ET_O distributions in training and test sets are reasonably similar.

3. Methodology

3.1. Time series forecasting platform

We employed Darts, a Python library for time series manipulation and forecasting (Herzen et al., 2022). Darts encompasses various forecasting models, ranging from statistical methods to state-of-the-art deep neural network architectures specifically designed for time series forecasting. Readers are directed to Herzen et al. (2022) for further details on this library.

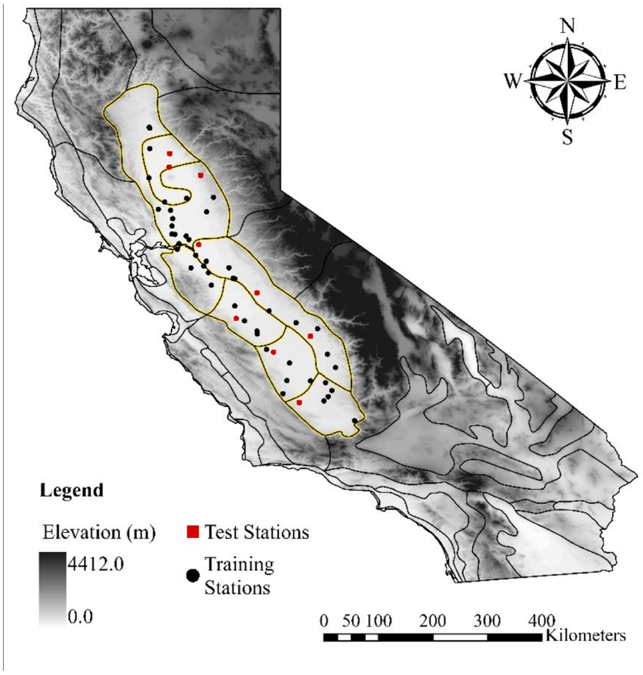


Fig. 1. The study area (the Central Valley of California, outlined by the yellow boundary) with CIMIS stations overlaid on a digital elevation model (DEM). The grayscale shading represents elevation (m), with darker regions indicating higher elevations. Black circles denote training stations, while red squares indicate test stations.

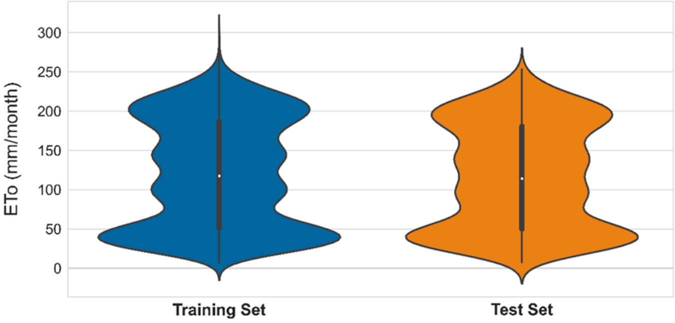


Fig. 2. Distribution of reference evapotranspiration in the test and training sets using violin plots. The shape of each distribution represents the density of data points, with inner box plots indicating the interquartile range and median value.

3.2. *Local vs. global learning*

Local learning refers to training a separate forecasting model for each individual time series. Therefore, in the local learning scheme of our study, a separate model is trained for each station in the test set. All the stations in the test set have data available through the end of June 2023. The last three years of data (i.e., July 2020 to June 2023) are used to test the models' performance, and models are trained over the remaining time steps (Fig. 3). The forecasting horizon for all models is 12 months (i.e., 12 time steps), and the stride is set to one month (Fig. 3) for consistent and smooth interval progression. Fig. 3 demonstrates how the model is continually updated and retrained as new data becomes available, always predicting 12 months ahead while moving forward in monthly increments through the 36-month (3-year) testing period.

In global learning, unlike local learning, a single forecasting model is trained using data from multiple time series. The global learning scheme of our study trains a single model using the data from all 47 stations in the training set (Fig. 1). Data from the eight stations in the test set are not introduced to this model in the training stage. Only after training is the model used to forecast the last three years of data for each of the eight stations in the test set. After the training stage, no further fine-tuning is performed on the globally learned models.

3.3. *Statistical forecasting models*

3.3.1. *Seasonal autoregressive integrated moving average (SARIMA)*

The ARIMA model is widely recognized in time series forecasting for its combination of autoregressive (AR), differencing (I), and

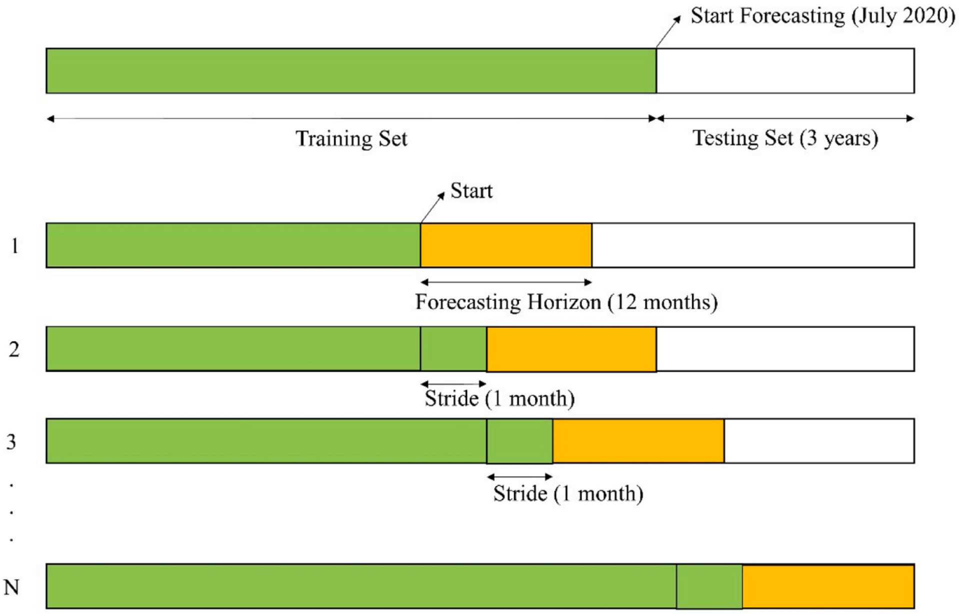


Fig. 3. Data splitting and forecasting time steps for local learning. The forecasting horizon is one year (i.e., 12 monthly time steps), with the last three years used as the test set.

moving average (MA) components (Contreras et al., 2003; Hyndman and Athanasopoulos, 2018). Differencing is employed to eliminate non-stationarity by computing the differences between consecutive observations. The AR component forecasts the variable of interest using a linear combination of its past values, while the MA component utilizes past forecast errors instead of the variable's past values in a regression-like manner. Although ARIMA is a non-seasonal model, it sets the foundation for the seasonal ARIMA (SARIMA) model, which integrates additional seasonal terms into the ARIMA framework.

In this study, we focus on the SARIMA model, which extends ARIMA by incorporating four additional hyperparameters: P, D, and Q, representing the seasonal order for the AR, differencing, and MA components, respectively, and m, which denotes the periodicity or number of time steps in a complete seasonal period. For monthly data, m is set to 12 (Hyndman and Athanasopoulos, 2018).

To optimize the hyperparameters of SARIMA, we used the *pmdarima* Python library. Each station in the test set had its hyperparameters optimized individually, after which these station-specific parameters were utilized to train the SARIMA models. The *AutoARIMA* function from the *pmdarima* library was used to identify the optimal set of parameters for the SARIMA model, resulting in a single fitted model. The maximum values for p, q, P, and Q were set to 5, and m was fixed at 12, aligning with our monthly data.

The Akaike Information Criterion (AIC) was employed to select the best model, with an alpha level of 0.05 for statistical significance. The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) unit root test was used to assess stationarity, and the Osborn-Chui-Smith-Birchenhall (OCSB) seasonal unit root test was performed for seasonal stationarity. Model parameters were optimized using the stepwise algorithm described by Hyndman and Khandakar (2008), and the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm with optional box constraints was used as the optimization algorithm.

3.3.2. *Holt-Winters' exponential smoothing*

Exponential smoothing, a statistical forecasting technique introduced in the late 1950s (Holt, 2004; Winters, 1960), utilizes weighted averages of past data points to predict future values. This univariate method assigns exponentially decreasing weights to older observations, giving more significance to more recent data (Hyndman and Athanasopoulos, 2018). Additional details about this method can be found in Kalekar (2004) and Hyndman and Athanasopoulos (2018). Our study applies an additive approach to both trend and seasonal components, with a seasonal period set to 12. Moreover, the trend component is damped for better accuracy.

3.3.3. *Theta method*

The Theta model, introduced by Assimakopoulos and Nikolopoulos (2000), is a univariate forecasting approach that modifies the local curvature of time series data by applying a coefficient known as "Theta" (a real number) to the second differences in the data. This method decomposes the original series into multiple lines, extrapolates each line using suitable forecasting models, and then combines the predictions to arrive at a final forecast. This study employs the 4Theta model, an enhanced variation of the original Theta method (Spiliotis et al., 2020). After conducting an exploratory analysis by testing different Theta values (i.e., a manual search), we selected Theta = 2 based on its empirical performance, which ensures a well-balanced decomposition of the time series and enhances forecast accuracy. The seasonal period is specified as 12, with a multiplicative type of seasonality. The Theta lines are integrated using an additive model, and the trend is modeled linearly. For additional details on the Theta and 4Theta models, readers should consult Assimakopoulos and Nikolopoulos (2000) and Spiliotis et al. (2020).

3.4. *Machine learning model (LightGBM)*

In this study, we utilize LightGBM as a machine learning forecasting tool. LightGBM, introduced by Ke et al. (2017) and developed by Microsoft as a free and open-source framework, offers an efficient implementation of the gradient boosting algorithm and optimizes memory usage. Gradient boosting is an ensemble technique combining decision tree models to form a more robust predictive model. The fitting process involves a gradient descent optimization algorithm that minimizes the loss gradient as the model parameters are adjusted. Within machine learning literature and competitions, gradient boosting and decision tree-based models have been shown to outperform other regression algorithms for tabular data (Shwartz-Ziv and Armon, 2022). Furthermore, various studies highlight that while gradient boosting algorithms perform similarly in accuracy and runtime, LightGBM often excels (Al Daoud, 2019). Consequently, this study adopts LightGBM as the primary machine learning model, utilizing 24 prior time steps. LightGBM forecasts the next 12 time steps with a Multi-Input Multi-Output (MIMO) strategy. For further details on the LightGBM methodology, readers can refer to Ke et al. (2017) and Al Daoud (2019).

3.5. *Deep learning models*

All DL models use a Multi-Input Multi-Output (MIMO) strategy. We performed manual hyperparameter tuning for all DL models in local learning. Nevertheless, as highlighted in the literature (Zhu et al., 2023; Zhu et al., 2024a and 2024b), precise hyperparameter tuning is crucial for achieving high performance in DL models. Consequently, we employed automatic hyperparameter optimization for selected DL architectures in global learning (see Section 3.6). DL models take 24 previous time steps as input and output the next 12 time steps. We used Google Colab GPU to train DL models.

3.5.1. *N-BEATS*

Neural basis expansion analysis for interpretable time series forecasting (N-BEATS), a deep neural network architecture introduced by Oreshkin et al. (2019), leverages backward and forward residual connections and a deeply stacked array of fully connected layers. Initially developed in 2019 to address univariate time series forecasting, N-BEATS is known for its rapid training capabilities and state-of-the-art performance across various datasets. For additional details on the N-BEATS architecture, readers are encouraged to consult Oreshkin et al. (2019).

In this study, we adopted the generic architecture described by Oreshkin et al. (2019). Specifically, we employed four stacks, each containing four blocks. Each block included four fully connected layers preceding the final backcast-forecast forking layer, with each layer comprising 16 neurons. The expansion coefficient dimension was set to five, and the rectified linear unit (ReLU) function was used as the activation function for the encoder/decoder intermediate layer. Our grid search indicated that N-BEATS performed optimally without dropout; thus, the dropout probability was set to zero. The model was trained over 100 epochs with a batch size of 32.

3.5.2. *Long short-term memory (LSTM)*

Long Short-Term Memory (LSTM), introduced by Hochreiter and Schmidhuber (1997), is a type of recurrent neural network (RNN). RNN models are well-suited for tackling problems involving sequential input data, such as time series. However, traditional RNN models often struggle with long-term dependencies, failing to retain information over extended sequences. LSTM architecture is designed to address this issue, making it a significant advantage for this model. For further details on LSTM, readers are referred to Hochreiter and Schmidhuber (1997) and Van Houdt et al. (2020). Our LSTM model features a single recurrent layer with 12 features in its hidden state. The dropout rate for this model is set to zero. The model is trained over 1000 epochs with a batch size of 8.

3.5.3. *Temporal convolutional network (TCN)*

While convolutional neural networks (CNNs) are often associated with raster data, they can also be adapted for sequential data with appropriate modifications. The temporal convolutional network (TCN), introduced by Bai et al. (2018), is a convolutional architecture designed explicitly for sequence modeling. In this study, we utilize a dilated TCN for forecasting. For more details on this model, readers are referred to Bai et al. (2018). Our model features a kernel size of 6 and 18 filters. The base of the exponent that determines the dilation at each level is set to 2. We applied weight normalization to the model and employed a dropout rate of 0.1. The model was trained over 1000 epochs with a batch size of 32.

3.5.4. *Transformer model*

The Transformer model, introduced by Vaswani et al. (2017), represents a state-of-the-art deep learning architecture. Unlike traditional architectures, the Transformer does not depend on recurrence or convolutions to produce its output; instead, it utilizes an encoder-decoder structure. Central to this architecture is the multi-head attention mechanism, which can simultaneously focus on different positions in the sequence, making the Transformer particularly suitable for time series forecasting. Additionally, its highly parallelizable nature makes it well-suited for training on GPUs. For further details on the Transformer model, readers are encouraged to consult Vaswani et al. (2017).

In our implementation, the transformer model's encoder and decoder inputs have 16 features, with one encoder layer and one decoder layer. The multi-head attention mechanism utilizes four heads. The feedforward network dimension is set to 128, and we use the ReLU activation function in the encoder/decoder intermediate layers. Based on grid search results, the dropout rate is set to 0.1. The model is trained over 1200 epochs with a batch size of 32.

3.5.5. *Temporal fusion transformer (TFT)*

The temporal fusion transformer (TFT), introduced by Lim et al. (2021), is a cutting-edge deep learning architecture designed for interpretable multi-horizon time series forecasting. This novel attention-based model integrates recurrent layers for local processing and a self-attention mechanism for capturing long-term dependencies. The TFT can learn temporal relationships at varying scales and includes specialized components for selecting relevant features. Readers are referred to Lim et al. (2021) for an in-depth understanding of this architecture.

In our study, we configured the TFT with a hidden state size of 16 and a hidden size of 8 for processing continuous variables. The architecture includes one layer for each LSTM encoder and decoder. The model employs four attention heads, with the multi-head attention query applied exclusively to the future (decoder) part. A gated residual network serves as the feedforward component. The training was conducted using PyTorch's mean squared error (MSE) loss function. The model was trained over 700 epochs with a batch size of 32.

3.5.6. *N-HiTS*

Neural hierarchical interpolation for time series forecasting (N-HiTS) is a novel DL model introduced by Challu et al. (2023). N-HiTS is like N-BEATS but aims to enhance performance while reducing computational costs by incorporating multi-rate sampling of inputs and multi-scale interpolation of outputs. By assembling its predictions sequentially and emphasizing components with different frequencies and scales, N-HiTS addresses the challenges of prediction volatility and computational complexity. Extensive experiments have demonstrated that N-HiTS significantly improves accuracy and efficiency, outperforming state-of-the-art Transformer models by reducing computation time dramatically. Detailed information about N-HiTS can be found in Challu et al. (2023).

In our N-HiTS architecture, we employed four stacks, each containing four blocks. Each block included four fully connected layers, with each layer comprising 32 neurons. The ReLU function was the activation function for the encoder/decoder intermediate layer. Similar to our N-BEATS model, the dropout probability was set to zero. The model was trained over 100 epochs with a batch size of 32.

3.5.7. N-Linear and D-Linear

N-Linear and D-Linear models are introduced by Zeng et al. (2023) as simple yet effective alternatives to Transformer-based models for long-term time series forecasting. Unlike Transformers, these "embarrassingly" simple models avoid the temporal information loss associated with self-attention mechanisms by employing a one-layer linear architecture. Despite their simplicity, extensive experiments show that N-Linear and D-Linear models outperform sophisticated Transformer models significantly in accuracy and efficiency. More details can be found in the original study introducing these models (Zeng et al., 2023). For our D-Linear model, the size of the kernel for the moving average is set to 24. We trained both N-Linear and D-Linear models over 100 epochs with a batch size of 32.

3.5.8. Time-series dense encoder (TiDE)

Time series dense encoder (TiDE) developed by Das et al. (2023) resembles Transformers but strives for enhanced performance and reduced computational cost by utilizing multilayer perceptron (MLP)-based encoder-decoders without the use of attention mechanisms. This approach maintains the simplicity and speed of linear models while effectively handling nonlinear dependencies, achieving near-optimal error rates for linear dynamical systems. Empirically, TiDE matches or surpasses previous methods on established long-term time-series forecasting benchmarks, operating faster than the best Transformer models. Detailed information about TiDE can be found in Das et al. (2023).

Our TiDE architecture consists of one residual block in the encoder layer and one in the decoder layer, while the width of the layers in the residual blocks is set to 128. The width of the layers in the past and the future covariate projection residual blocks is 4. The output of the decoder layer is 16-dimensional. The temporal decoder layers have a width of 32 units. The dropout probability is 0.1. The model is trained over 100 epochs with a batch size of 32.

3.6. Automatic hyperparameter optimization

We employed Optuna, a hyperparameter optimization framework for automatic hyperparameter tuning of the best-performing DL models in global learning (Akiba et al., 2019). Optuna allows users to construct the hyperparameter search space dynamically by offering a define-by-run API. Optuna has a user-friendly setup, and it provides efficient sampling and pruning algorithms for customization. Detailed information about Optuna and its algorithm can be found in Akiba et al. (2019). We used the Davis CIMIS station from the training set to optimize hyperparameters using Optuna, where the last three years of its data are used as the validation set and

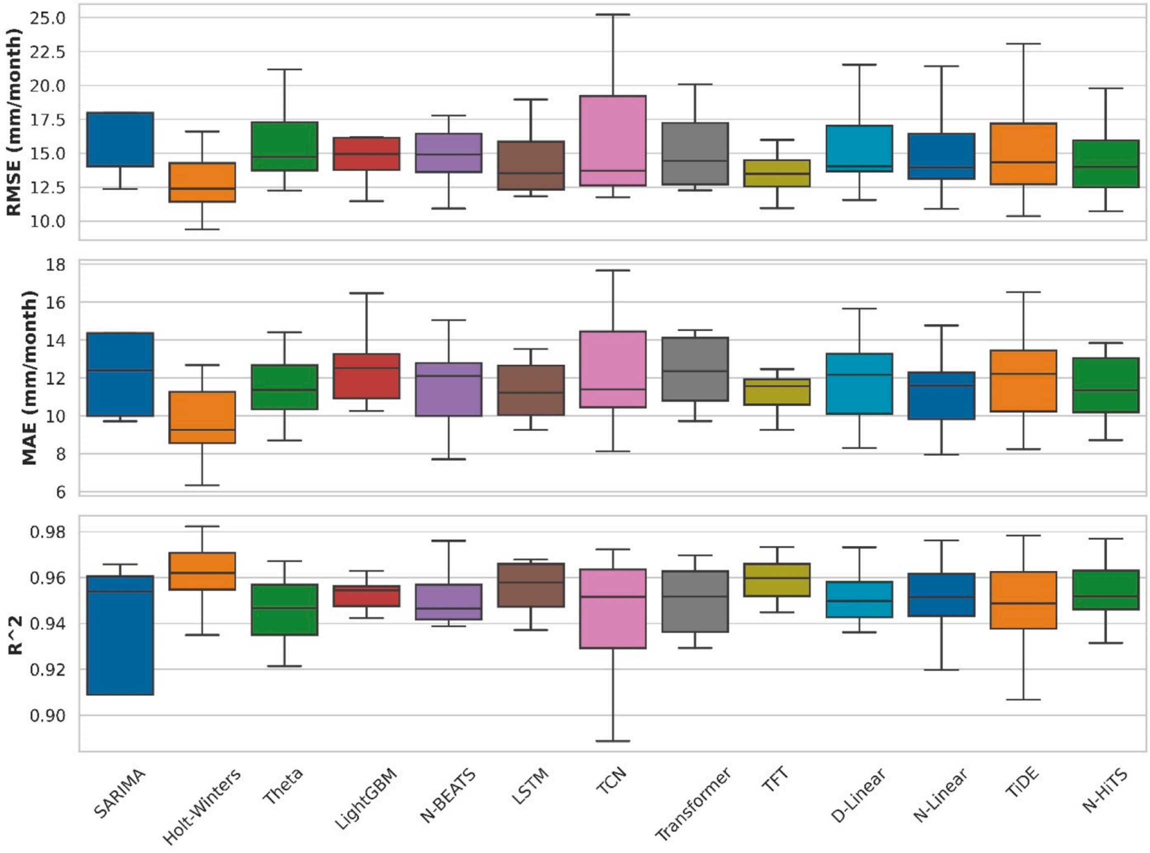


Fig. 4. Results of local learning on the test set for statistical, machine learning, and deep learning models.

the remaining as the train set.

4. Results and discussion

4.1. Local learning

Fig. 4 shows the performance of locally learned models on the test set. Holt-Winters is the most accurate forecasting model based on RMSE, MAE, and R2 scores. This finding aligns with Ahmadi et al. (2023) analysis of a similar dataset, where the Holt-Winters model's accuracy was comparable with the best-performing DL forecasting models.

As Fig. 4 suggests, DL models' performance in local learning falls behind simpler statistical models. This might be because DL models have more complex algorithms involving hundreds or even thousands of trainable parameters (readers are referred to Ahmadi et al. 2023 to learn more about the number of trainable parameters in DL and statistical models). We need extensive datasets to train these big models effectively, which is hard to provide in the case of local learning.

Another reason for the superiority of simpler models is the inherent simplicity of monthly ET_O and, more broadly, monthly hydrometeorological time series. These time series are usually dominated by clear seasonal patterns due to predictable climatic cycles with few nonlinear or complex interactions (Wilks, 2011). The Holt-Winters model excels in this context as it is explicitly designed to handle periodic data, incorporating specific components for modeling seasonality, trend, and level (Holt, 2004; Winters, 1960). Holt-Winters is less prone to overfitting as a simpler model, making it more effective for the straightforward patterns prevalent in monthly ET_O time series. In contrast, advanced DL models like N-BEATS, TiDE, and Transformers have high capacity and can capture complex patterns, but they are prone to overfitting when applied to these relatively simple patterns (Goodfellow et al., 2016).

4.2. Global learning

As Fig. 5 shows, global learning significantly enhances the performance of N-BEATS, TiDE, and N-HiTS. The performance of LSTM, D-Linear, and N-Linear stays about the same for local and global learning schemes, while the performance of TCN, Transformer, and TFT improves meaningfully in global learning. As mentioned earlier, the DL models require vast amounts of data to be trained effectively. Contrary to local learning, where models are trained over a single time series, in global learning, DL models learn from

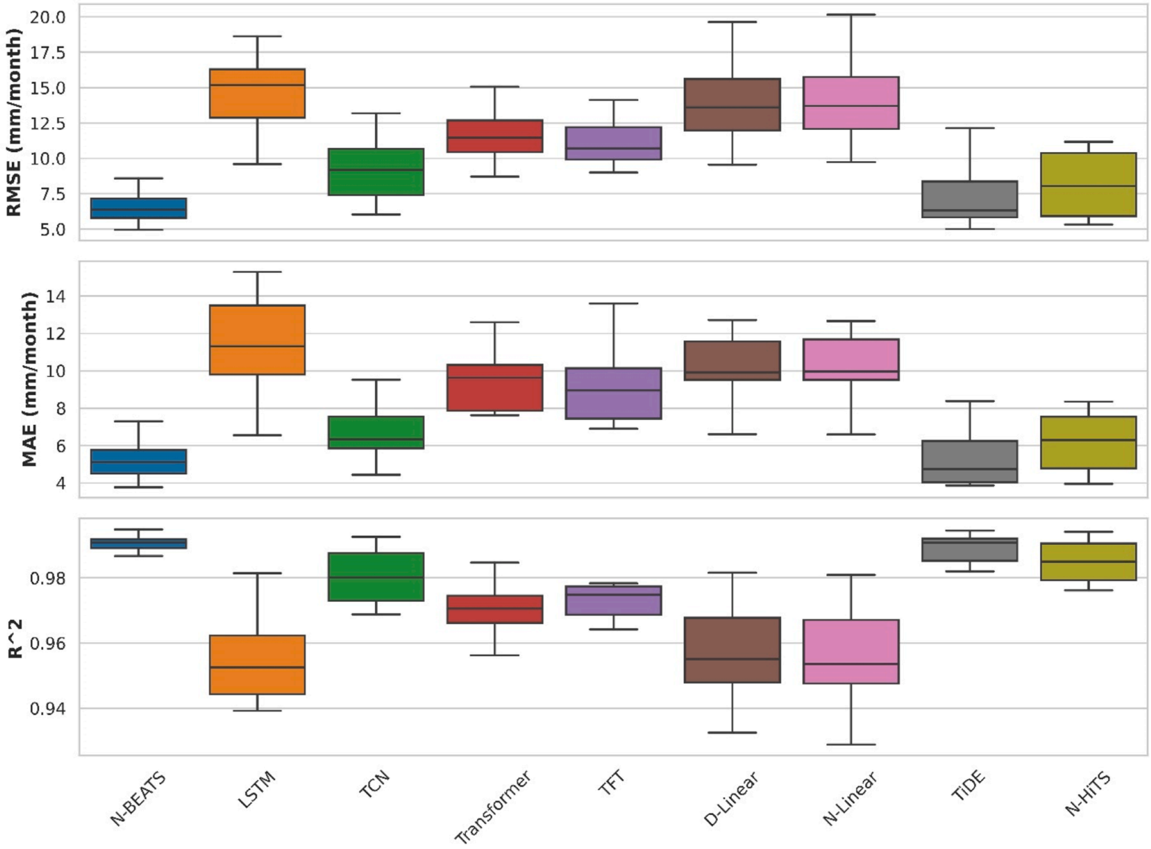


Fig. 5. Results of global learning on the test set for deep learning models.

multiple time series, which, as Fig. 5 suggests, results in higher forecasting accuracy.

N-BEATS, TiDE, and N-HiTS have high-capacity architectures capable of capturing complex patterns in data (Oreshkin et al., 2019; Das et al., 2023; Challu et al., 2023). When employed in global learning, these models can capture and generalize broader patterns across different time series in the dataset, boosting performance. Conversely, LSTM is designed to capture long-range dependencies in sequential data and might be sensitive to the length and type of sequences, which can vary between local and global learning schemes (Hochreiter and Schmidhuber, 1997). Furthermore, the linear architectures of D-Linear and N-Linear models may already effectively capture linear relationships within the local context, offering limited gains from additional global patterns (Zeng et al., 2023). These simpler models may thus experience a performance plateau, as they have less room for improvement when introducing more training data and adopting global schemes.

4.3. *Automatic hyperparameter optimization*

We employed Optuna for automatic hyperparameter tuning of the best-performing DL models: N-BEATS, TiDE, and N-HiTS. Fig. 6 demonstrates that while Optuna improved the performance of N-BEATS and N-HiTS, the TiDE model performed better with manually selected hyperparameters. This discrepancy arises from differences in model sensitivity to hyperparameter settings. TiDE, being an MLP-based encoder-decoder, has a relatively stable optimization landscape, making it less dependent on extensive hyperparameter search. Consequently, manual tuning provided more reliable performance in this case.

It should be noted that Fig. 6 also compares the best-performing DL models in the global learning scheme with Holt-Winters, which had the best accuracy in local learning. As can be seen, when trained globally, DL models outperform Holt-Winters by a significant margin. Automatic hyperparameter tuning enhances the performance of two of three models, making the difference between their and Holt-Winter's overall accuracies even larger.

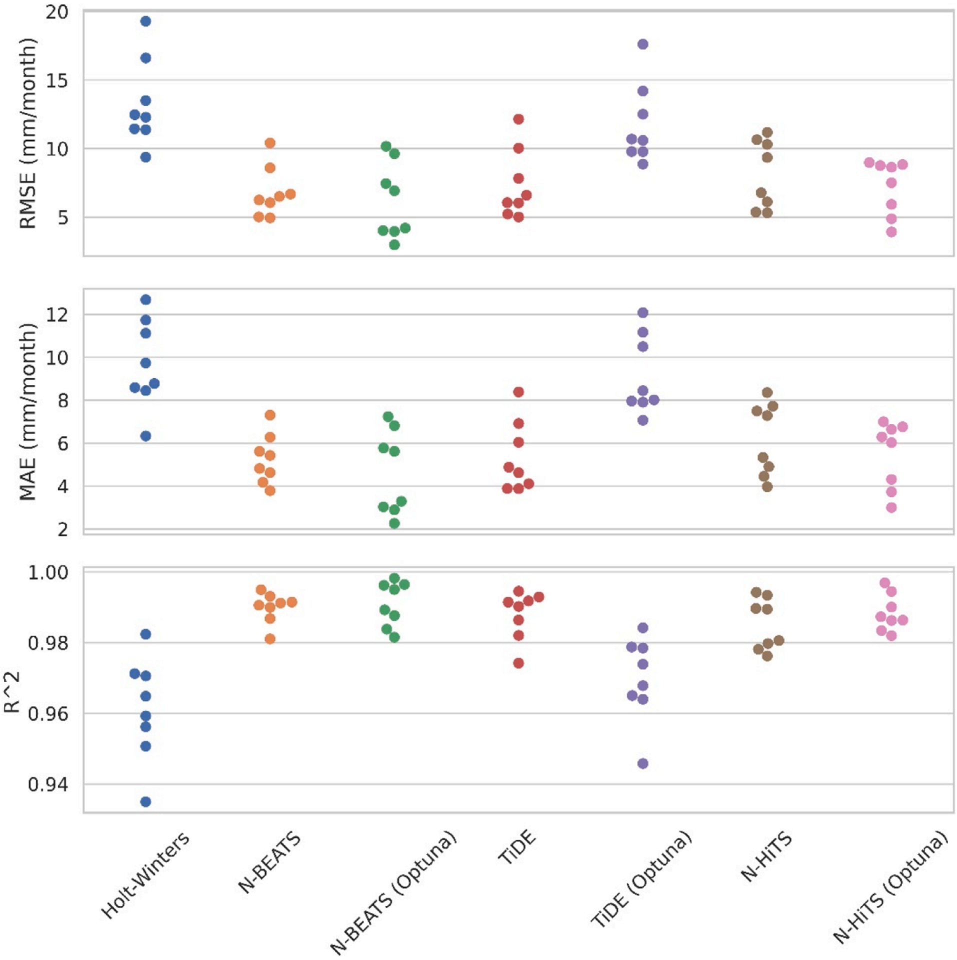


Fig. 6. Results of best-performing globally learned deep learning models on the test set with and without automatic hyperparameter tuning by Optuna in comparison to the Holt-Winters, as the best-performing statistical model.

4.4. Local versus global learning

Fig. 7 compares the forecasting accuracy in local versus global learning. We used the Durham CIMIS station to depict how these learning schemes affect the forecasted time series. Durham CIMIS station is one of the stations in the test set with data available from 1986 to 2023. As Fig. 7-a illustrates, TFT -the best-performing DL model in local learning- is very similar to Holt-Winters, which is a statistical forecasting model with a much simpler algorithm and lower computational cost. Conversely, when we use the global learning scheme, DL forecasting models generate accurate forecasts far closer to the observed values (Fig. 7-b and -c). In other words, we can infer that global learning unleashes the true potential of advanced deep forecasting models such as N-BEATS and N-HiTS. As mentioned earlier, N-BEATS and N-HiTS have high-capacity architecture and can model complex nonlinear patterns effectively. However, as our findings suggest, the local learning scheme cannot provide them with enough data. Along with their accuracy benefits, some DL models are computationally expensive and require strong processing units such as GPUs for efficient applications. However, as shown by Ahmadi et al. (2023), although N-BEATS architecture has more trainable parameters than LSTM, TCN, Transformer, and TFT, it is considerably less computationally expensive than these models (readers are referred to Table 1 of Ahmadi et al. 2023 for detailed information about the computational cost of forecasting models).

In addition to improving accuracy, using global learning for DL models can enhance their generalizability. Trained over a pooled dataset of hydrometeorological time series with characteristics like the target, these globally learned models can work reliably on unseen instances. This is especially important in hydrometeorological time series in ungauged locations and newly established measurement stations. Deep global forecasting models can leverage large data availabilities from adjacent locations or even stations from other regions with matching climatic characteristics. A related observation was made in research on actual evapotranspiration estimation (Bafti et al., 2024), where capturing spatial dependencies of adjacent cells using attention-based U-Net models improved accuracy across multiple locations. It is important to note that the high forecasting accuracy of the globally-learned DL models, as shown in Fig. 7-b and -c, is not due to overfitting. These models were not trained on data from the Durham CIMIS station. Thus, they had no prior exposure to this station's data during training, eliminating the possibility of overfitting to this specific dataset. This holds true for all the stations in the test set, further ruling out the possibility of overfitting in our results.

Monthly hydrometeorological time series are usually seasonality-driven; therefore, global learning can open the door for mixing different variables in the training set, making the models more nuanced to complex patterns. Future studies can apply global learning to various hydrometeorological time series and climates. As larger training datasets often result in more accurate DL models, we hypothesize that including more time series from different hydrometeorological variables leads to better performance. We additionally hypothesize that incorporating a broader array of hydrometeorological variables into the training set could enhance the generalizability of deep forecasting models by capturing additional dependencies that influence evapotranspiration and other hydrological processes.

5. Summary and conclusion

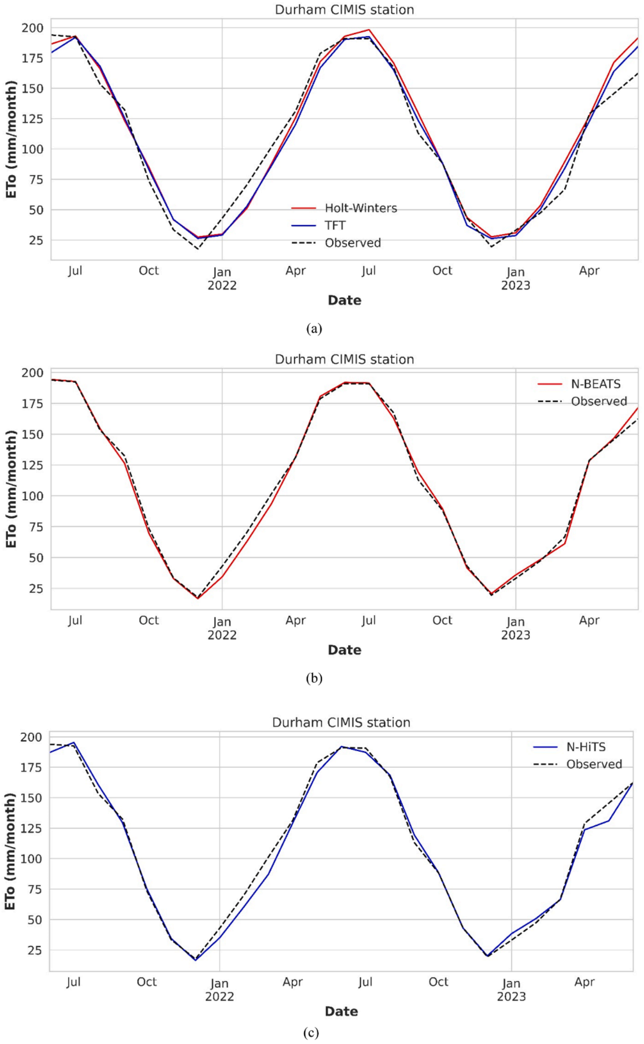
This study explored how global learning affects the forecasting accuracy and generalizability of deep learning models. Our findings revealed that global learning significantly enhances the accuracy of deep learning models for reference evapotranspiration forecasting in the Central Valley of California. Global learning could unleash the potential of high-capacity deep architectures by providing them with more training data. We also demonstrated how global learning enhances the generalizability of deep forecasting models and makes them less prone to overfitting. Deep global learning can leverage large data availabilities in gauged regions and facilitate reliable hydrometeorological time series forecasting in ungauged locations and recently established stations with short data histories.

While this study demonstrates the advantages of global learning for hydrometeorological forecasting, some limitations should be considered. First, the approach relies on the availability of sufficient and diverse training data, which may not always be feasible. Second, deep learning models require substantial computational resources, particularly when training on large datasets. Future research should explore methods to integrate physics-based constraints and domain-specific knowledge into deep learning architectures, potentially enhancing their interpretability and reliability. Moreover, future studies can test the performance improvement of global learning for other seasonality-driven hydrometeorological time series and different climates and regions.

Our study's broader impacts are significant for both research and practical water management applications. By demonstrating the efficacy of global learning, we introduce a data-centric approach that enhances hydrometeorological forecasting, which is especially beneficial in regions with limited data. This methodology can specifically help newly established weather stations lacking long-term datasets, aiding water managers in optimizing resource allocations and decision-making to promote sustainable practices. Furthermore, our work sets the stage for future research exploring global learning in diverse hydrological regions and climates, addressing data scarcity challenges, and enhancing decision-making in partially-gauged and developing areas.

CRediT authorship contribution statement

Namadi Peyman: Writing – review & editing, Conceptualization. **He Minxue:** Writing – review & editing, Conceptualization. **Daccache Andre:** Writing – review & editing, Supervision, Conceptualization. **Ahmadi Arman:** Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Kadir Tariq:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization. **Snyder Richard L.:** Writing – review & editing, Supervision, Conceptualization. **Bai Zhaojun:** Writing – review & editing, Funding acquisition, Conceptualization. **Sandhu Prabhjot:** Conceptualization. **Bafti Alireza Ghaderi:** Writing – review & editing.



(a)

(b)

(c)

Fig. 7. Accuracy in global vs. local learning in forecasting reference evapotranspiration in Durham CIMIS station. Best statistical and deep learning local learning models (a), N-BEATS global learning with Optuna (b), N-HiTS global learning with Optuna (c).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the California Department of Water Resources and the University of California, Davis grant number 4600014165-01.

Data availability

All data used in this manuscript is publicly available. Links are provided in the manuscript.

References

Ahmadi, A., Daccache, A., Sadegh, M., Snyder, R.L., 2023. Statistical and deep learning models for reference evapotranspiration time series forecasting: a comparison of accuracy, complexity, and data efficiency. Comput. Electron. Agric. 215, 108424.

Ahmadi, A., Daccache, A., Snyder, R.L., Suvočarev, K., 2022. Meteorological driving forces of reference evapotranspiration and their trends in California. Sci. Total Environ. 849, 157823.

Ahmadi, A., Kazemi, M.H., Daccache, A., Snyder, R.L., 2024. SolarET: a generalizable machine learning approach to estimate reference evapotranspiration from solar radiation. Agric. Water Manag. 295, 108779.

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M., 2019, July. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).

Al Daoud, E., 2019. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. Int. J. Comput. Inf. Eng. 13 (1), 6–10.

Allen, R.G., Pereira, L.S., Raes, D. and Smith, M., 1998. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. *Fao, Rome*, *300*(9), p.D05109.

Allen, R.G., Pruitt, W.O., Wright, J.L., Howell, T.A., Ventura, F., Snyder, R.L., Itenfisu, D., Steduto, P., Berengena, J., Baselga Yrisarry, J., Smith, M., Pereira, L.S., Raes, D., Perrier, A., Alves, I., Walter, I., Elliott, R., 2006. A recommendation on standardized surface resistance for hourly calculation of reference ETo by the FAO56 penman-monteith method. Agric. Water Man. 81, 1–22.

Allen, R.G., Walter, I.A., Elliott, R.L., Howell, T.A., Itenfisu, D., Jensen, M.E., Snyder, R.L., 2005. The ASCE standardized reference evapotranspiration equation. Am. Soc. Civ. Eng. Reston, Virginia, 192p.

Assimakopoulos, V., Nikolopoulos, K., 2000. The theta model: a decomposition approach to forecasting. Int. J. Forecast. 16 (4), 521–530.

Babaeian, E., Paheding, S., Siddique, N., Devabhaktuni, V.K., Tuller, M., 2022. Short-and mid-term forecasts of actual evapotranspiration with deep learning. J. Hydrol. 612, 128078.

Bafti, A., Ahmadi, A., Abbasi, A., Kamangir, H., Jamali, S., Hashemi, H., 2024. Automated actual evapotranspiration estimation: hybrid model of a novel attention-based U-Net and metaheuristic optimization algorithms. Atmos. Res. 297, 107107. https://doi.org/10.1016/j.atmosres.2023.107107.

Bai, S., Kolter, J.Z. and Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv: 1803.01271*.

Bandara, K., Bergmeir, C., Smyl, S., 2020. Forecasting across time series databases using recurrent neural networks on groups of similar series: a clustering approach. Expert Syst. Appl. 140, 112896.

Challu, C., Olivares, K.G., Oreshkin, B.N., Ramirez, F.G., Canseco, M.M. and Dubrawski, A., 2023, June. Nhits: Neural hierarchical interpolation for time series forecasting. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 6, pp. 6989-6997).

Chen, S., Feng, Y., Mao, Q., Li, H., Zhao, Y., Liu, J., Wang, H., Ma, D., 2024. Improving the accuracy of flood forecasting for Northeast China by the correction of global forecast rainfall based on deep learning. J. Hydrol., 131733

Chen, Z., Zhu, Z., Jiang, H., Sun, S., 2020. Estimating daily reference evapotranspiration based on limited meteorological data using deep learning and classical machine learning methods. J. Hydrol. 591, 125286.

Chia, M.Y., Huang, Y.F., Koo, C.H., Ng, J.L., Ahmed, A.N., El-Shafie, A., 2022. Long-term forecasting of monthly mean reference evapotranspiration using deep neural network: a comparison of training strategies and approaches. Appl. Soft Comput. 126, 109221.

Contreras, J., Espinola, R., Nogales, F.J., Conejo, A.J., 2003. ARIMA models to predict next-day electricity prices. IEEE Trans. Power Syst. 18 (3), 1014–1020.

Das, A., Kong, W., Leach, A., Mathur, S., Sen, R. and Yu, R., 2023. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*.

Dong, J., Zhu, Y., Jia, X., Han, X., Qiao, J., Bai, C., Tang, X., 2022. Nation-scale reference evapotranspiration estimation by using deep learning and classical machine learning models in China. J. Hydrol. 604, 127207.

Ferreira, L.B., da Cunha, F.F., 2020. Multi-step ahead forecasting of daily reference evapotranspiration using deep learning. Comput. Electron. Agric. 178, 105728.

Gleick, P.H., 2003. Water use. Annu. Rev. Environ. Resour. 28 (1), 275–314.

Gocić, M., Motamedi, S., Shamshirband, S., Petković, D., Ch, S., Hashim, R., Arif, M., 2015. Soft computing approaches for forecasting reference evapotranspiration. Comput. Electron. Agric. 113, 164–173.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.

Granata, F., Di Nunno, F., 2021. Forecasting evapotranspiration in different climates using ensembles of recurrent neural networks. Agric. Water Manag. 255, 107040.

Granata, F., Di Nunno, F., de Marinis, G., 2024. Advanced evapotranspiration forecasting in Central Italy: stacked MLP-RF algorithm and correlated nystrom views with feature selection strategies. Comput. Electron. Agric. 220, 108887.

Hanak, E., 2011. Managing California's water: From conflict to reconciliation. Public Policy Instit. of CA.

Herzen, J., Lässig, F., Piazzetta, S.G., Neuer, T., Tafti, L., Raille, G., Van Pottelbergh, T., Pasieka, M., Skrodzki, A., Huguenin, N., Dumonal, M., 2022. Darts: user-friendly modern machine learning for time series. J. Mach. Learn. Res. 23 (124), 1–6.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Holt, C.C., 2004. Forecasting seasonals and trends by exponentially weighted moving averages. Int. J. Forecast. 20 (1), 5–10.

Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting: principles and practice. OTexts.

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. J. Stat. Softw. 27, 1–22.

Kalekar, P.S., 2004. Time series forecasting using holt-winters exponential smoothing. Kanwal Rekhi Sch. Inf. Technol. 4329008 (13), 1–13.

Karbasi, M., Jamei, M., Ali, M., Malik, A., Yaseen, Z.M., 2022. Forecasting weekly reference evapotranspiration using auto encoder decoder bidirectional LSTM model hybridized with a boruta-catboost input optimizer. Comput. Electron. Agric. 198, 107121.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30.

Le, M.H., Kim, H., Do, H.X., Beling, P.A., Lakshmi, V., 2024. A framework on utilizing of publicly availability stream gauges datasets and deep learning in estimating monthly basin-scale runoff in ungauged regions. Adv. Water Resour. 188, 104694.

Lee, J., Bateni, S.M., Jun, C., Heggy, E., Jamei, M., Kim, D., Ghafouri, H.R., Deenik, J.L., 2024. Hybrid machine learning system based on multivariate data decomposition and feature selection for improved multitemporal evapotranspiration forecasting. Eng. Appl. Artif. Intell. 135, 108744.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., Di, Z., 2017. A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. Wiley Interdiscip. Rev. Water 4 (6), e1246.

Lim, B., Arık, S.Ö., Loeff, N., Pfister, T., 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int. J. Forecast. 37 (4), 1748–1764.

Lim, B., Zohren, S., 2021. Time-series forecasting with deep learning: a survey. Philos. Trans. R. Soc. A 379 (2194), 20200209.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2022. M5 accuracy competition: results, findings, and conclusions. Int. J. Forecast. 38 (4), 1346–1364.

Mehdizadeh, S., Fathian, F., Safari, M.J.S., Adamowski, J.F., 2019. Comparative assessment of time series and artificial intelligence models to estimate monthly streamflow: a local and external data analysis approach. J. Hydrol. 579, 124225.

Milly, P.C., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., Stouffer, R.J., 2008. Stationarity is dead: whither water management? Science 319 (5863), 573–574.

Oreshkin, B.N., Carpov, D., Chapados, N. and Bengio, Y., 2019. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv: 1905.10437*.

Quinn, J.D., Reed, P.M., Giuliani, M., Castelletti, A., 2024. Average domination: A new multi-objective value metric applied to assess the benefits of forecasts in reservoir operations under different flood design levels. Adv. Water Resour. 185, 104638.

Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T., 2020. DeepAR: probabilistic forecasting with autoregressive recurrent networks. Int. J. Forecast. 36 (3), 1181–1191.

Shwartz-Ziv, R., Armon, A., 2022. Tabular data: deep learning is not all you need. Inf. Fusion 81, 84–90.

Smyl, S., 2020. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. Int. J. Forecast. 36 (1), 75–85.

Spiliotis, E., Assimakopoulos, V., Makridakis, S., 2020. Generalizing the theta method for automatic forecasting. Eur. J. Oper. Res. 284 (2), 550–558.

Talib, A., Desai, A.R., Huang, J., Griffis, T.J., Reed, D.E., Chen, J., 2021. Evaluation of prediction and forecasting models for evapotranspiration of agricultural lands in the Midwest US. J. Hydrol. 600, 126579.

Torres, A.F., Walker, W.R., McKee, M., 2011. Forecasting daily potential evapotranspiration using machine learning and limited climatic data. Agric. Water Manag. 98 (4), 553–562.

Van Houdt, G., Mosquera, C., Nápoles, G., 2020. A review on the long short-term memory model. Artif. Intell. Rev. 53 (8), 5929–5955.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wilks, D.S., 2011. Statistical methods in the atmospheric sciences. Academic press.

Winters, P.R., 1960. Forecasting sales by exponentially weighted moving averages. Manag. Sci. 6 (3), 324–342.

Zeng, A., Chen, M., Zhang, L. and Xu, Q., 2023, June. Are transformers effective for time series forecasting?. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 9, pp. 11121-11128).

Zhang, Y., Zhao, Z., Zheng, J., 2020. CatBoost: a new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. J. Hydrol. 588, 125087.

Zhu, S., Di Nunno, F., Ptak, M., Sojka, M., Granata, F., 2023. A novel optimized model based on NARX networks for predicting thermal anomalies in Polish lakes during heatwaves, with special reference to the 2018 heatwave. Sci. Total Environ. 905, 167121.

Zhu, S., Di Nunno, F., Sun, J., Sojka, M., Ptak, M., Granata, F., 2024a. An optimized NARX-based model for predicting thermal dynamics and heatwaves in rivers. Sci. Total Environ. 926, 171954.

Zhu, S., Shinohara, R., Matsuzaki, S.I.S., Kohzu, A., Watanabe, M., Nakagawa, M., Di Nunno, F., Sun, J., Zhou, Q., Granata, F., 2024b. Diel temperature patterns unveiled: high-frequency monitoring and deep learning in Lake Kasumigaura. Ecol. Indic. 169, 112958.