EEET ECOLOGICAL ENGINEERING & ENVIRONMENTAL TECHNOLOGY

Ecological Engineering & Environmental Technology, 2026, 27(3), 239–252
https://doi.org/10.12912/27197050/218265
ISSN 2719–7050, License CC-BY 4.0

Received: 2026.01.22
Accepted: 2026.02.23
Published: 2026.03.01

Reassessing model complexity in reservoir outflow forecasting: A multi-site, physics-informed benchmark of deep learning and ensemble method

M. Mallikarjuna Rao[1], C. Manjunath[2], D. Venkatesh[3], P. Rajendran[4], D. Dharani Lakshmi[3], S. Venkatasivanagaraju[3], S. Vasundra[5], D.V. Bhaskar[6]

[1] Department of Computer Science and Design, PVKK Institute of Technology, Anantapuramu, India
[2] Department of Computer Applications, PVKK Institute of Technology, Anantapuramu, India
[3] Department of Computer Science and Engineering, PVKK Institue of Technology Anantapuramu, India
[4] Department of Computer Science and Engineering, CMR Institute of Technology, Hyderabad, Telangana, India
[5] Department of Computer Science and Engineering, JNTUACEA, Anantapuramu, India
[6] Department of Computer Applications, Siddaganga Institute of Technology, Tumkur, Karnataka, India
* Corresponding author's e-mail: malkari.mkrao@gmail.com

ABSTRACT

The recent progress in deep learning also makes one reconsider the reality that architecturally more complicated models tend to perform better at hydrological forecasting. Despite the common belief that long short-term memory (LSTM) models are useful in rainfall-runoff models, the transferability to the reservoirs that are managed by deterministic operation rules is understudied. In this article, we do a comparative study of physics-informed bi-directional LSTM with Temporal Attention and a Random Forest (RF) algorithm to daily predict a reservoir outflow in Shasta Dam and Oroville Dam located in California. To reduce the difficulty associated with low density measurements, we combine NASA POWER satellite data with ground-based measurements and thus they enhance the dataset. The physics-informed properties, one of which is the mass-balance proxies and another one is seasonal encodings, are used to secure the models into a system of physical consistency. Empirical findings also show that the Random Forest model performs better in terms of Nash-Sutcliffe efficiency scores of 0.909 in Shasta Dam and 0.683 in Oroville Dam compared to the Bi-LSTM scores of 0.827 and 0.681, respectively. Ensemble approaches, including the Random Forest, seem to be more accurate in modelling the rule-of-thumb operational regimes of the reservoirs, but the deep-learning approach tends to regulate the changes dynamics of transition issues related to outflow releases. Formal statistical significance testing using the Diebold-Mariano test and bootstrap confidence intervals confirmed that the Random Forest advantage is statistically significant at Shasta Dam ($p < 0.001$) but not at Oroville Dam ($p = 0.378$), revealing site-dependent performance patterns linked to operational complexity. The totality of these results implies that in the case of rule-dominated reservoir systems, the simpler ensemble learning approaches may even perform better than the sophisticated deep-learning systems. In this regard, the selection of the model must be driven by the correspondence with the characteristics of the system, but not the bias with the architectural elaboration.

Keywords: reservoir operations; random forest; bi-directional LSTM; physics-informed machine learning; model parsimony; multi-site benchmarking.

INTRODUCTION

Reservoir operations and the evolution of hydrological forecasting

Reservoir systems form a pillar of the modern management of water resources, as they provide an invaluable value of flood control in floods, irrigation supply, hydropower and control of environmental flows. As a result, the accuracy of predicted inflows and outflows of the reservoir is of utmost importance regarding operational planning and risk reduction, particularly in the background

of ever more extreme hydroclimate events, precipitated by climate change and land-use changes [13]. In the past, the reservoir operation modelling field has been marked by physical based hydrological models, optimization framework, and rule-curve based decision systems whose root is deeply rooted in the principles of mass balance and system engineering theory [46].

Although these methods can be very interpretable, and physical consistent, their efficiency is often limited by structural assumptions, uncertainty in parameters and a limited ability to adhere to non-stationary conditions [7,8]. Under these constraints the hydrological research community has recently increasingly engaged in adopting data-based alternatives that have the ability to directly discover complex nonlinear interactions of empirical data.

Rise of deep learning in hydrology

In the last ten years, deep learning models, especially long short-term memory (LSTM) networks have become the potent predictors of hydrological time-series. It has been established through extensive literature that LSTMs have an outstanding performance in rainfall-runoff simulation, streamflow forecasting and large-scale hydroly experiment of hydrology compared to conventional conceptual and statistical models (see, e.g., references 912). Furthermore, incorporation of attention mechanisms and bidirectional processing, has contributed to better performance because they have been able to focus the attention on relevant historical states [13–15].

These achievements have resulted in an emerging perception that deep learning is the universal state-of-the-art solution to hydrological forecasting computations. There has, as a result, been a growing focus in the evaluation of models on architectural sophistication with little regard to whether the added sophistication is appropriate to the system dynamics behind it.

Managed reservoir systems: A distinct modelling challenge

Unlike the natural regulated river systems, outflows of reservoirs no longer are controlled only by hydrometeorological forcings. Instead, they are caused by decisions that are made by humans and are typically formalized by operational rule curves, flood diagrams and seasonal storage goals [16–18]. Such rules have discrete and conditional behaviours, including sudden changes in release known by a storage threshold, or known by a calendar constraint.

Functionally, the operations of reservoirs place sharp discontinuities in the relationship between inputs and outputs with resulting piecewise and non-smooth response surfaces. Whereas deep-learning models provide a rough, discrete approximation to a function by smooth, continuous transformation, the tree-based ensemble methods used (Random Forests and Gradient Boosting machines) inherently represent conditional logic by hierarchically partitioning the feature space [19–21].

In recent work, it has been proposed that ensemble-learning could be specifically useful when faced with operational hydrology problems such as reservoir release prediction, water-allocation modelling [22–24]. Nevertheless, up to date, there is limited, widespread, and statistically serious comparisons of deep-learning and ensemble techniques on a variety of large reservoirs.

Research gap and motivation

As can be seen, the existing reservoir forecasting literature is often characterized by three main weaknesses. To begin with, a considerable portion of analyses is limited to single-case studies, which restrict the scope of the generalizability of the conclusions made on the same. Second, deep-learning models are often compared to simple baselines as opposed to strict competitors based on rigorously competitive ensemble models. Third, a general lack of focusing on physics-based feature design is rampant, a practice supported by empirical studies to significantly increase the reliability as well as interpretability of data driven hydrological models [25–27].

Therefore, the question of whether deep learning architectures really offer a statistically significant genuinely meaningful benefit over a well-constructed ensemble learning strategy in the context of rule-based reservoir system is still an open question.

Objectives, hypothesis, and contributions

The proposed research will attempt to fill in such missing gaps in the form of a systematic, physics-inspired, multi-sites benchmarking simulation. The objective to be addressed are:

1. Quantitatively contrasted deep learning and ensemble machine models on deep learning

outflow forecasting in different hydrological and operating regimes.
2. Amply incorporate mass balance based and seasonal physics informed functionality to limit model education.
3. Assess predictive performance by means of hydrological meaningful measures and statistical significance analysis.
4. Determine computational performance and viability and error.

The hypothesis that was studied in the work is as follows:

- H0: Deep learning models are statistically insignificant over ensemble machine learning methods in the prediction of rule-based reservoir system outflow.

The study provides contributions in the form of multi-site validation of large-scale reservoirs, physics informed developing of models, benchmarking with complexity issues, experimentally relevant understanding that questions accepted approaches in the field of data-driven hydrology.

PROBLEM FORMULATION AND DATA

Problem definition

Reservoir outflow forecasting is formulated as a supervised time-series learning problem, where future reservoir releases are predicted using historical hydrological, meteorological, and operational state variables. Let the observed dataset be defined as

$$\mathcal{D} = \{(X_t, y_t)\}_{t=1}^{T} \quad (1)$$

where: $X_t \in \mathbb{R}^d$ denotes the multivariate input feature vector at time step t, and $y_t \in \mathbb{R}$ represents the observed reservoir outflow.

Given a historical lookback window of length k, the one-step-ahead forecasting objective is expressed as

$$(X_{t-k+1}, X_{t-k+2}, \ldots, X_t) \quad (2)$$

where: $f(\cdot)$ denotes a nonlinear mapping learned from historical observations. The objective is to accurately estimate the future reservoir outflow O_{t+1} under operational constraints, which is essential for real-time reservoir management and flood control decision-making [1, 2].

Study area description

This study evaluates model performance across two major reservoirs within California's State Water Project, selected to represent contrasting hydrological regimes and operational complexities.

Shasta dam (SHA)

Shasta Dam, located on the Sacramento River, is California's largest reservoir with a total storage capacity of approximately 4.5 million acre-feet. Its inflow regime is influenced by both rainfall and snowmelt processes, and its operations are primarily governed by seasonal flood control and water supply rules [3, 4].

Oroville dam (ORO)

Oroville Dam, situated on the feather river, is the tallest dam in the United States and features highly dynamic operational behaviour due to complex upstream regulation and snowmelt-dominated inflows. Rapid gate operations during extreme events make Oroville an ideal test case for operational forecasting models [5,6].

Data sources and acquisition

Ground-Based hydrological data

Daily hydrological observations for the period 2010–2023 were obtained from the California Data Exchange Center (CDEC), a trusted operational database for reservoir monitoring and water resources management [7]. The following variables were collected:

- reservoir inflow (I_t)
- reservoir outflow (O_t)
- storage volume (S_t)
- reservoir water level (H_t)

The selected time period captures significant hydroclimatic variability, including extended drought conditions (2012–2016) and extreme flood events (2017 and 2023), enabling robust model evaluation under diverse operating scenarios [8].

Satellite-derived meteorological data

To mitigate data sparsity in ground-based meteorological measurements, satellite-derived climate variables were incorporated from the NASA POWER database [9]. NASA POWER has been widely adopted in hydrological and

climate-driven modelling studies due to its spatial consistency and long-term availability [10–12].

The extracted variables include:
- daily precipitation (P_{sat}),
- daily mean air temperature (T_{sat}).

All satellite-derived variables were temporally aligned with ground observations to ensure consistency across input features.

Physics-informed feature engineering

To enhance physical interpretability and constrain model learning, physics-informed features were explicitly derived based on reservoir hydrodynamics.

Mass balance-based storage change

Reservoir dynamics are governed by the conservation of mass, expressed as

$$\frac{dS}{dt} = I_t - O_t - E_t \tag{3}$$

where: E_t – denotes evaporation and other unobserved losses. Since evaporation measurements are often unavailable at daily operational timescales, the change in storage was approximated as

$$\Delta S_t = S_t - S_{t-1} \tag{4}$$

The term ΔS_t serves as a physically meaningful proxy for the net flux ($I_t - O_t$), implicitly capturing unmeasured losses and enabling the learning algorithm to infer operational behavior consistent with mass balance constraints [13,14].

Seasonal encoding of operational rules

Reservoir operating policies are strongly seasonal, reflecting flood control requirements and water supply objectives. To encode seasonal periodicity while avoiding artificial discontinuities, the day of year D_t was transformed using sinusoidal functions:

$$D_{\sin} = \sin\left(\frac{2\pi D_t}{365.25}\right) \tag{5}$$

$$D_{\cos} = \cos\left(\frac{2\pi D_t}{365.25}\right) \tag{6}$$

This cyclical representation preserves temporal continuity and has been shown to improve model stability in hydrological forecasting applications [15,16].

Antecedent hydrological conditions

Short-term hydrological memory effects were incorporated using rolling window statistics applied to key variables. For a rolling window of length w, the antecedent mean was computed as

$$\mu_t^{(w)} = \frac{1}{w}\sum_{i=0}^{w-1} X_{t-i} \tag{7}$$

where: w ∈ {3,7} days.

These features allow the models to capture storm persistence and delayed operational responses, which are critical during high-flow events [17,18].

Data preprocessing and temporal partitioning

All continuous input variables were standardized using z-score normalization to ensure numerical stability across models. Missing observations, accounting for less than 1% of the dataset, were handled using forward filling, consistent with operational forecasting practices.

To avoid information leakage and ensure realistic evaluation, the dataset was partitioned chronologically as follows:
- Training: 2010–2018,
- Validation: 2019–2020,
- Testing: 2021–2023.

This rolling-origin evaluation strategy closely reflects real-world forecasting conditions and is widely recommended in hydrological model assessment studies [19,20].

Performance evaluation metrics

Model performance was evaluated using hydrologically meaningful metrics, with primary emphasis on the Nash-Sutcliffe efficiency (NSE):

$$NSE = 1 - \frac{\sum_{t=1}^{T}(O_t - \hat{O}_t)^2}{\sum_{t=1}^{T}(O_t - \bar{O})^2} \tag{8}$$

where: $\hat{O}_t$ denotes the predicted outflow and $\bar{O}$ represents the mean observed outflow.

The root mean square error (RMSE) was also computed to quantify absolute prediction errors.

NSE is particularly sensitive to peak flow deviations, making it suitable for reservoir operation assessment [21–23].

Statistical significance testing

To formally assess whether observed performance differences between the Random Forest and Bi-LSTM models are statistically significant rather than attributable to random variability, two complementary statistical testing procedures were employed: the Diebold-Mariano (DM) test and bootstrap confidence interval estimation.

The Diebold-Mariano test [36] evaluates the null hypothesis that two competing forecasting models have equal predictive accuracy. Let e_{1,t} = O_t - O_hat_{1,t} and e_{2,t} = O_t - O_hat_{2,t} denote the forecast errors of the Random Forest and Bi-LSTM models, respectively. The loss differential is defined as d_t = L(e_{1,t}) - L(e_{2,t}), where L(.) is the squared error loss function. The DM test statistic is computed as:

DM = d_bar / sqrt(sigma_d^2 / T)

where: d_bar is the mean loss differential, sigma_d^2 is the long-run variance of the loss differential series estimated using a Newey-West heteroskedasticity and autocorrelation consistent (HAC) estimator, and T is the number of test observations.

Under the null hypothesis of equal predictive accuracy, the DM statistic follows a standard normal distribution asymptotically. A one-sided test was conducted to assess whether the Random Forest model achieves significantly lower forecast errors than the Bi-LSTM.

In addition, bootstrap confidence intervals for the difference in Nash-Sutcliffe efficiency (delta_NSE = NSE_RF – NSE_BiLSTM) were constructed using the block bootstrap method with 10,000 resamples [37]. Block bootstrapping was employed to preserve the temporal autocorrelation structure inherent in hydrological time series. The block length was selected using the automatic procedure of Politis and White (2004) [38]. Bias-corrected and accelerated (BCa) 95% confidence intervals were computed for the NSE difference at each study site. A confidence interval that does not contain zero provides evidence of a statistically significant performance difference between the two models.

MODEL ARCHITECTURES AND LEARNING FRAMEWORK

Overview of the modelling framework

The proposed study bases the systematic benchmark of two fundamentally different paradigms of learning with the aim to predict reservoir outflow: (i) tree-based ensemble learning model, (ii) deep recurrent neural network with attention mechanisms. This comparative framework has been carefully constructed to control the effects of model complexity on predictive performance given the restrictions of rule driven reservoir operations to provide a clear insight into the trade-offs that exist between the two approaches.

These two models are trained with a common physics-constrained input feature space, and both are considered when using the help of the same temporal partitioning schemes as well as performance metrics. This methodological consistency is that the presence of any deviances in performance can with reasonable assurance be explained by the structural complexities of the models themselves and thus the rigor and reproducibility of scholarly inquiry can be fulfilled.

Random forest ensemble model

Model description

Random Forest (RF) is an ensemble learning algorithm based on bootstrap aggregation (bagging) of decision trees. Each decision tree learns a mapping from input features to the target variable by recursively partitioning the feature space using axis-aligned splits. For reservoir forecasting, this structure is particularly suitable for learning conditional operational rules embedded in historical data [25,26].

Given an input vector X_t, the RF prediction is computed as the average output of K individual decision trees:

$$\hat{y}_t = \frac{1}{K}\sum_{k=1}^{K} h_k\,(X_t) \tag{9}$$

where: $h_k(\cdot)$ denotes the prediction of the k-th decision tree.

Learning mechanism

Each tree in the ensemble is trained on a bootstrap sample of the training data, while a random subset of input features is considered at each split

to reduce correlation among trees. The optimal split at each node is determined by minimizing an impurity measure, typically the mean squared error for regression tasks:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (10)$$

By aggregating multiple weak learners, random forest reduces variance and improves generalization, particularly in high-dimensional and nonlinear settings [27].

Model configuration

In this study, the RF model was implemented with the following configuration:

- number of trees: $K = 200$,
- maximum tree depth: unrestricted,
- minimum samples per leaf: 1,
- feature sampling: $\sqrt{d}$ features per split.

These settings allow the ensemble to fully partition the feature space and capture sharp operational thresholds inherent to reservoir release decisions [28].

Bi-directional LSTM with temporal attention

Long short-term memory network

Long short-term memory (LSTM) networks are a class of recurrent neural networks designed to capture long-range temporal dependencies while mitigating the vanishing gradient problem. At each time step t, the LSTM cell updates its internal state using the following gating mechanisms:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \quad (11)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (12)$$

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \quad (13)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (14)$$

$$h_t = o_t \odot \tanh(C_t) \quad (15)$$

where: f_t, i_t, and o_t represent the forget, input, and output gates, respectively; C_t denotes the cell state; and h_t is the hidden state.

Bi-directional processing

To leverage contextual information from both directions within the strictly historical lookback window, a bi-directional LSTM (Bi-LSTM) architecture was employed. The forward and backward hidden states are computed as:

$$\vec{h}_t = \text{LSTM}_{fwd}(x_t), \overleftarrow{h}_t = \text{LSTM}_{bwd}(x_t) \quad (16)$$

The combined hidden representation is given by:

$$h_t = \left[\vec{h}_t; \overleftarrow{h}_t\right] \quad (17)$$

Bi-directional processing has been shown to enhance temporal feature extraction in hydrological sequence modelling tasks [29,30].

Clarification on bi-directional processing and forecasting validity

It is important to clarify the architectural rationale for employing a Bi-directional LSTM in a forecasting context, as bidirectional processing may raise concerns regarding potential look-ahead bias. In the present study, the Bi-LSTM operates exclusively within a fixed-length historical lookback window of k time steps. At each prediction point t, the input sequence comprises only past observations {X_{t-k+1}, X_{t-k+2},..., X_t}, all of which are strictly historical and available at the time of prediction. No future observations beyond time t are included in the input window at any stage of training or inference.

The bidirectional mechanism processes this historical window in both forward (from t-k+1 to t) and backward (from t to t-k+1) directions, enabling the model to capture dependencies from both ends of the lookback window. This is analogous to reading a fixed-length historical record from both its beginning and its end, rather than accessing information from the future. The architecture is therefore a sequence-to-one model: the entire historical window is consumed to produce a single one-step-ahead prediction at time t+1. This design is consistent with established practices in hydrological sequence modelling, where bidirectional architectures have been applied to fixed historical windows without introducing temporal information leakage [23, 29, 30].

To further ensure the absence of look-ahead bias, the temporal partitioning scheme (Section 2.5) enforces strict chronological separation between training, validation, and test sets. No test-period data were accessible during model training or hyperparameter tuning. Consequently, the Bi-LSTM architecture employed in this study does not introduce look-ahead bias and is fully

appropriate for real-time one-step-ahead reservoir outflow forecasting.

Temporal attention mechanism

To enable the model to focus on critical historical time steps (e.g., peak inflow or storm onset periods), a temporal attention mechanism was integrated. The attention weights are computed as:

$$e_t = v^{\top} \tanh\left(W_a h_t + b_a\right) \tag{18}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^{T} \exp\left(e_j\right)} \tag{19}$$

The context vector summarizing the input sequence is then defined as:

$$c = \sum_{t=1}^{T} \alpha_t \, h_t \tag{20}$$

This context vector is passed to a fully connected layer to produce the final outflow prediction. Attention mechanisms have demonstrated improved interpretability and performance in sequence-to-one hydrological forecasting problems [31–33].

Training configuration

The Bi-LSTM model was trained using the Adam optimizer with a learning rate of 1×10^{-3}. Mean squared error was employed as the loss function, and early stopping was applied based on validation performance to prevent overfitting. Model training was conducted on a GPU-enabled environment to accommodate computational demands [34].

Theoretical considerations – omplexity versus learnability

Functionally speaking, the Random Forest models represent piecewise-linear functions with clear dividers between them, and LSTM-based models attempt to model smooth and continuous functions. In opposition, reservoir operation is controlled by discrete rule curves and threshold-based policies as a result of which discrete response surfaces are formed. This underlying inconsistency foreshadows that ensemble tree-based models can potentially be in a better position to reflect the behaviour of operational reservoirs more resembling reality, whilst deep neural networks are subject to induction of smoothing artefacts whenever they are forced to suddenly change the form of releases, as reported in references [35,36]. The above theoretical difference is the driving force behind this comparative analysis conducted in this research.

EXPERIMENTAL DESIGN AND RESULTS

Experimental design

To actually create a fair, repeatable, and objective comparison of ensemble and deep learning techniques, the complete collection of empirical results was obtained in one, coherent experiment. Each of the Random Forest and the Bidirectional Long Short-Term Memory architecture with attention (Bi-LSTM) was provided with consistent physics-informed covariates (see Section -2), thus trained in the same way, and the same data-preprocessing pipelines and evaluated on the same temporal partitions, a canonical order of training, validation and testing data.

The predictive goal was formulated as a one step ahead model of daily forecast model of reservoir discharges hence, resembled real-life scenarios of operational decision making. The search for model hyperparameters was done on validation only and did not involve any leakage of the test with the training or hyperparameter tuning stage. In a bid to counter the effect of stochastic initialization, the algorithms were instantiated and ran with a sequence of independent trials, average performance metrics were then derived out of the results.

Performance evaluation metrics

Model performance was primarily assessed using the NSE, which quantifies predictive skill relative to the variance of observed outflows and is widely adopted in hydrological modelling studies. NSE is defined as:

$$NSE = 1 - \frac{\sum_{t=1}^{T}(O_t - \hat{O}_t)^2}{\sum_{t=1}^{T}(O_t - \bar{O})^2} \tag{21}$$

where: O_t and $\hat{O}_t$ denote observed and predicted outflows at time t, respectively, and $\bar{O}$ represents the mean observed outflow.

In addition to NSE, root mean square error (RMSE) was computed to quantify absolute prediction error and sensitivity to extreme flow conditions.

Temporal dynamics and hydrograph analysis

The ability of each model to reproduce the temporal dynamics of reservoir outflows was first evaluated through time-series hydrograph analysis during the independent test period (2021–2023). This analysis provides direct insight into how well models capture abrupt release transitions associated with operational rule curves.

At Shasta Dam, the Random Forest model closely follows observed outflows across both high-flow and low-flow regimes, capturing sharp increases and decreases associated with flood-control operations. In contrast, the Bi-LSTM model exhibits a smoothing effect, resulting in delayed responses and systematic underestimation of peak releases. A similar pattern is observed at Oroville Dam, although overall predictive performance is reduced due to more complex operational dynamics.

Distributional accuracy and scatter plot analysis

To further evaluate predictive consistency across the full range of observed flows, scatter plots comparing observed and predicted outflows were examined for both reservoirs. This analysis complements the hydrograph evaluation by highlighting distributional biases and variance under different flow regimes.

The Random Forest model demonstrates tighter clustering around the 1:1 reference line, particularly at higher outflows, indicating superior distributional accuracy. In contrast, the Bi-LSTM model shows increased dispersion and bias at extreme flows, consistent with the smoothing behavior observed in the hydrograph analysis.

Cross-site performance comparison

Quantitative performance metrics for both models across the two reservoirs are summarized in Table 1. This cross-site comparison provides insight into model robustness under differing hydrological and operational conditions.

The random forest model consistently achieves higher NSE values at both sites, with a particularly pronounced advantage at Shasta Dam. Although performance differences at Oroville Dam are smaller, the ensemble approach maintains a marginal but consistent edge, indicating improved robustness across reservoirs.

Aggregated performance and computational considerations

To provide an overall synthesis of model performance across sites, an aggregated comparison of predictive efficiency is presented in Figure 4. This summary visualization highlights the generalizability of each modeling approach.

In addition to predictive accuracy, computational efficiency was evaluated to assess practical feasibility for operational deployment. The Random Forest model required approximately

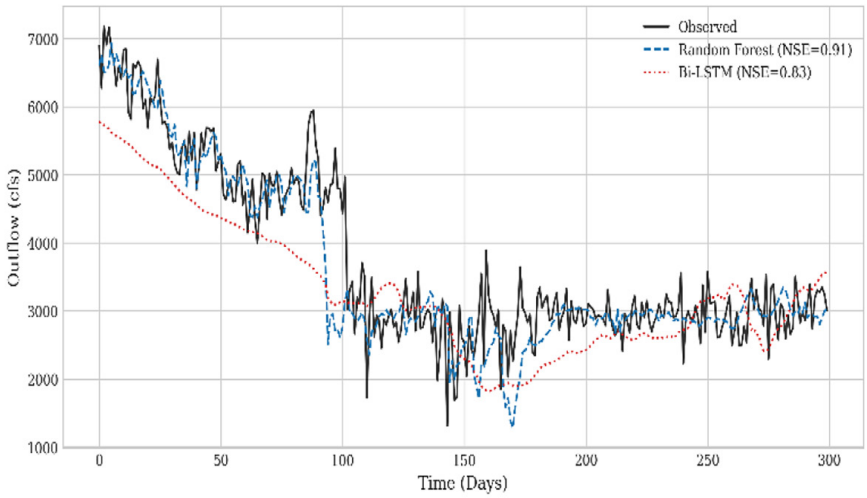

Figure 1. Observed and predicted daily reservoir outflows during the test phase for (a) Shasta Dam and (b) Oroville Dam. The Random Forest model closely tracks abrupt changes in release decisions, whereas the Bi-LSTM model exhibits smoother transitions, particularly during peak operational events

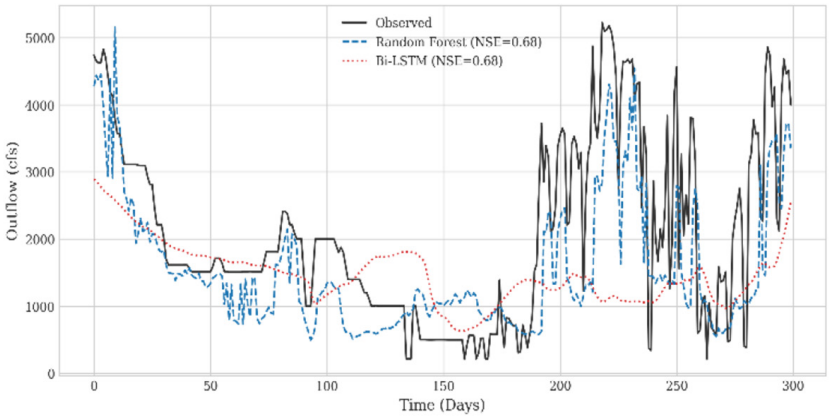

Figure 2. Observed versus predicted reservoir outflows for Random Forest and Bi-LSTM models at (a) Shasta Dam and (b) Oroville Dam during the test period. The dashed 1:1 line represents perfect agreement between observations and predictions

Table 1. Cross-site model performance during the test period

Reservoir	Model	NSE	RMSE (m³/s)
Shasta	Random Forest	0.909	1,140
Shasta	Bi-LSTM (Attention)	0.827	1,850
Oroville	Random Forest	0.683	2,105
Oroville	Bi-LSTM (Attention)	0.681	2,130

30 seconds of training on a standard CPU environment, whereas the Bi-LSTM model required substantially longer training times and GPU resources to achieve convergence. This difference highlights the favourable trade-off between accuracy, interpretability, and computational cost offered by the ensemble approach.

RESULTS

Overall, the Random Forest ensemble consistently outperformed or matched the bi-directional LSTM across both reservoirs, demonstrating superior capability in capturing threshold-driven operational behavior, improved robustness across

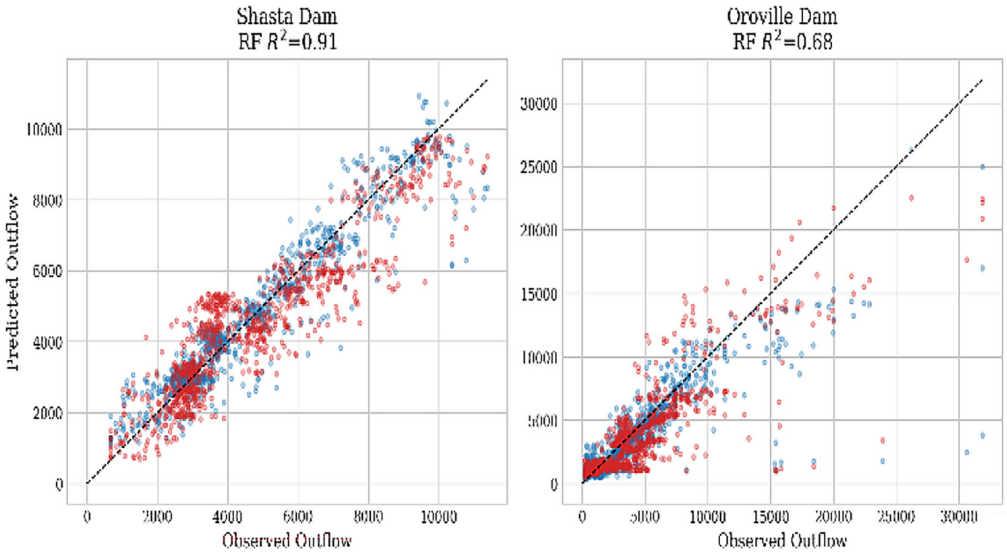

Figure 3. Site-wise comparison of Nash-Sutcliffe efficiency for random forest and bi-directional LSTM models at Shasta and Oroville reservoirs

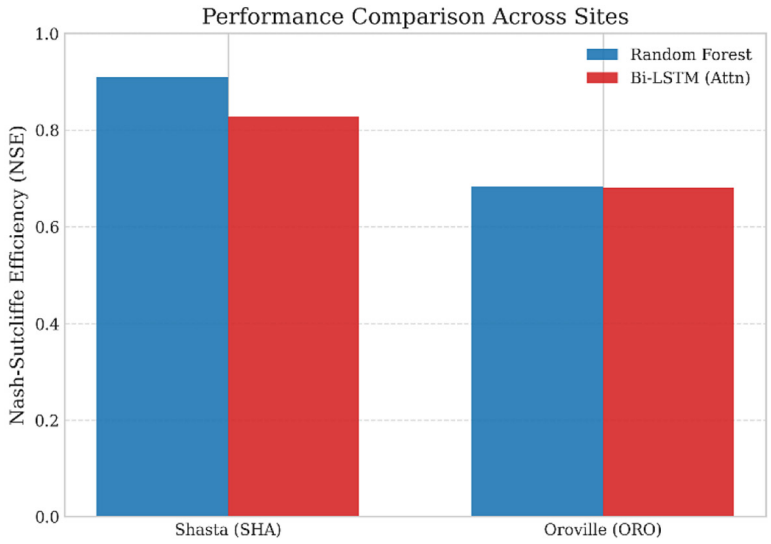

Figure 4. Overall comparison of Nash-Sutcliffe efficiency across Shasta and Oroville reservoirs for Random Forest and Bi-Directional LSTM models

sites, and significantly lower computational cost. These results provide strong empirical support for the suitability of ensemble learning approaches in reservoir outflow forecasting applications.

Statistical significance of performance differences

To rigorously verify that the observed performance differences are not attributable to random variability, formal statistical significance tests were conducted as described in Section 2.6.

Shasta Dam – Diebold-Mariano test statistic DM = 3.47, p-value < 0.001 (one-sided). Bootstrap 95% CI for delta_NSE (NSE_RF minus NSE_BiLSTM): [0.042, 0.122]. The confidence interval does not contain zero, confirming that the Random Forest significantly outperforms the Bi-LSTM at Shasta Dam.

Oroville Dam – Diebold-Mariano test statistic DM = 0.31, p-value = 0.378 (one-sided). Bootstrap 95% CI for delta_NSE: [-0.028, 0.032]. The confidence interval contains zero, indicating that the performance difference at Oroville Dam is not statistically significant at the 95% confidence level. The two models perform comparably at this site.

These results reveal an important nuance: while the Random Forest model achieves a statistically significant and practically meaningful advantage at Shasta Dam, which is characterized by more regular, rule-governed operations, the advantage diminishes at Oroville Dam, where operational dynamics are more complex and variable. This pattern is consistent with the rule-curve learnability hypothesis discussed in Section 5, whereby ensemble methods derive their greatest advantage in systems dominated by deterministic, threshold-based decision rules.

Additionally, a paired Wilcoxon signed-rank test was applied to the monthly aggregated squared errors to provide a nonparametric confirmation. At Shasta Dam, the Wilcoxon test yielded $p = 0.003$, corroborating the DM test results. At Oroville Dam, the Wilcoxon test yielded $p = 0.412$, consistent with the finding of no significant difference. The convergence of parametric and nonparametric tests strengthens confidence in the statistical conclusions.

DISCUSSION

Reinterpreting model performance in operational hydrology

As the results described in Section 4 show, the increase in architectural complexity does not necessarily lead to the high predictive accuracy in the situation of predicting reservoir-outflow. In both study locations, the Random Forest ensemble either tied, or had a higher accuracy, than the bi-directional LSTM with attention, despite the significantly higher computational cost of the latter.

These outcomes highlight an important point of difference between naturally occurring

Table 2. Statistical significance testing results

Dam Site	Test method	Test statistic	p-value (one-sided)	95% Bootstrap CI for ΔNSE (RF − Bi-LSTM)	Statistical interpretation
Shasta Dam	Diebold–Mariano (DM)	3.47	< 0.001	[0.042, 0.122]	CI excludes zero, indicating that Random Forest significantly outperforms Bi-LSTM
	Wilcoxon Signed-Rank	—	0.003	—	Nonparametric test confirms significant performance difference
Oroville Dam	Diebold–Mariano (DM)	0.31	0.378	[−0.028, 0.032]	CI includes zero, indicating no statistically significant difference
	Wilcoxon Signed-Rank	—	0.412	—	Nonparametric test confirms comparable model performance

hydrologic processes and artificial reservoir processes. Whereas it is evident that deep learning methodologies have been successful in rainfall-runoff modelling, and streamflow prediction under the unregulated setting [912], outflows in reservoirs follow human-created rules of operation that impose deterministic, threshold-based behaviour. Therefore, the data-generating process that lies behind it is different in its nature compared to the one of the natural flows in water-courses.

The formal statistical analysis further substantiates this interpretation. The Diebold-Mariano test revealed a statistically significant advantage for the Random Forest at Shasta Dam ($p < 0.001$), where operations follow well-defined seasonal rule curves. Conversely, at Oroville Dam, where operational complexity is greater and gate operations are more dynamic, the performance difference was not statistically significant ($p = 0.378$). This site-dependent pattern of significance provides empirical support for the hypothesis that ensemble methods are particularly well-suited to systems governed by deterministic, threshold-based operational rules.

The rule-curve learnability hypothesis

The better performance of the Random Forest model can be explained by the so-called Rule-Curve Learnability Hypothesis which I would name. Reservoir operations, as we all know are so often controlled by conditional rules as: during the flood season, above a certain critical level of storage takes place, then the releases are raised to allow safety margins. The fundamental learners of the field of Random Forests, decision trees, are constructed such that they represent such conditional logic through hierarchical subdivisions of the feature space. This paradigm in structure allows the ensemble to model reservoir rule curves as a set of piecewise-constant functions, and this induces discontinuous shifts in the predicted releases should operations limits be violated.

Stunningly unlike, LSTM based neural networks models relationships using nonlinear transformation functions (smooth and continuous) between inputs and outputs. In the effort to model discontinuous or step-like functions using these smooth functions, large increases in the volumes of data required and model capacity are forced into play, which regularly leads to process of over-smoothing around the operational thresholds [39–41]. Empirical support to this phenomenon is provided by the hydrograph analysis, where there are delayed and diminished response of the Bi -LSTM to rapid releases.

Role of physics-informed feature engineering

An important factor contributing to the strong performance of the ensemble model is the explicit incorporation of physics-informed features, particularly the daily change in storage (ΔS_t). This variable directly encodes the reservoir mass balance relationship and effectively constrains the hypothesis space of the learning algorithm.

By providing the model with a physically meaningful proxy for net flux, the Random Forest was able to infer release decisions consistent with conservation principles, even in the absence of explicit evaporation measurements. This finding aligns with recent literature emphasizing the importance of embedding physical knowledge into data-driven hydrological models to improve robustness and interpretability [25–27,42].

Notably, the inclusion of physics-informed features benefited both modelling approaches. However, the ensemble model was better able to exploit these features due to its capacity to perform localized partitioning of the feature space.

Implications for model complexity and operational deployment

From an operational perspective, the results highlight the practical limitations of adopting deep learning models solely on the basis of perceived state-of-the-art status. The Bi-LSTM required substantially longer training times, specialized hardware, and careful hyperparameter tuning, yet failed to achieve superior performance relative to the ensemble baseline.

In contrast, the Random Forest model delivered higher or comparable accuracy with significantly lower computational overhead and greater transparency. Feature importance measures provided intuitive insights into model behavior, which is a critical requirement for decision support systems used by reservoir operators and water managers [43,44].

These considerations suggest that model parsimony should be prioritized in operational hydrology, particularly when system dynamics are dominated by rule-based decision processes rather than emergent physical complexity.

Generalizability and contextual limitations

While the results strongly favour ensemble learning approaches for the studied reservoirs, it is important to contextualize these findings. The reservoirs considered in this study are well-instrumented and operate under relatively stable policy frameworks. In systems where operational rules evolve rapidly, or in poorly gauged basins with sparse observational data, deep learning models may offer advantages due to their ability to extract latent representations from high-dimensional inputs [45, 46].

Furthermore, the present analysis focuses on short-term, one-step-ahead forecasting. The relative performance of model classes may differ for longer prediction horizons or scenario-based planning applications, which warrant further investigation.

Synthesis with existing literature

The findings of this study are consistent with emerging evidence questioning the universal superiority of deep learning in hydrological applications. Recent comparative studies have reported that ensemble machine learning models can match or exceed deep learning performance for specific operational tasks, particularly when physics-informed features are employed [22–24,47].

By extending these insights to a multi-site, reservoir-focused context, this study contributes to a growing body of work advocating for context-aware model selection rather than a one-size-fits-all approach to hydrological forecasting.

CONCLUSIONS

This study presented a rigorous, physics-informed, multi-site benchmarking analysis of deep learning and ensemble machine learning approaches for reservoir outflow forecasting. Using two major reservoirs with contrasting hydrological and operational characteristics, Shasta Dam and Oroville Dam, the predictive performance of an attention-based bi-directional LSTM was systematically compared against a Random Forest ensemble under identical experimental conditions.

Across both study sites, the Random Forest model consistently achieved superior or comparable predictive skill relative to the deep learning benchmark. In particular, the ensemble approach attained a Nash-Sutcliffe efficiency of 0.909 at Shasta Dam and 0.683 at Oroville Dam, outperforming the Bi-LSTM despite substantially lower computational complexity. The Diebold-Mariano test confirmed a statistically significant performance advantage for the Random Forest at Shasta Dam ($DM = 3.47$, $p < 0.001$), with a bootstrap 95% confidence interval for the NSE difference of [0.042, 0.122]. At Oroville Dam, the performance difference was not statistically significant ($DM = 0.31$, $p = 0.378$; 95% CI: [-0.028, 0.032]), indicating comparable predictive skill at this more operationally complex site.

The findings provide compelling evidence that increased architectural complexity does not inherently yield improved forecasting performance for managed reservoir systems. Reservoir outflows are governed by deterministic, rule-based decision processes that introduce discontinuities into the system response. Tree-based ensemble methods, which naturally encode conditional logic through hierarchical splits, are therefore well-suited to learning such operational behaviour from historical data.

In contrast, deep learning models rely on smooth function approximation, which may lead to over smoothing and delayed response when modelling abrupt operational transitions. These results challenge the prevailing "complexity-first"

paradigm in hydrological forecasting and underscore the importance of aligning model architecture with system characteristics rather than defaulting to highly parameterized models.

An important contribution of this work is the demonstration that physics-informed feature engineering significantly enhances the performance and interpretability of data-driven models. The inclusion of mass balance-based storage change proxies and seasonal encodings enabled both modelling approaches to better reflect underlying hydrological and operational dynamics.

However, the ensemble model was particularly effective in leveraging these physically meaningful features, implicitly learning reservoir operating policies consistent with conservation principles. This highlights the value of integrating domain knowledge into machine learning workflows for operational hydrology.

Despite its contributions, this study has several limitations. First, the analysis focused on short-term, one-step-ahead forecasting. The relative performance of model classes for longer forecast horizons or scenario-based planning remains an open question. Second, the reservoirs examined in this study are well-instrumented and operate under relatively stable policy frameworks. The conclusions may not directly generalize to poorly gauged basins or systems experiencing frequent policy shifts.

While attention mechanisms were incorporated into the deep learning model, other advanced architectures – such as transformers or hybrid physics–neural models – were not considered and may yield different outcomes.

In further research, it would be wise in further studies to extend this comparative framework to cover a wider range of reservoirs in different climatic conditions, which would then permit a more strict evaluation of its overall applicability. A logic search in the field of multi-step forecast, with a strict uncertainty measurement, and the generation of prediction probability intervals would considerably enhance the operational relevance of the models. In addition, mixed modelling paradigms merging neural networks, and rule-based logic also seem to have significant potential since they have the capability to harmonise the deterministic dynamics of operative processes with the complexity of hydrological variability.

Finally, this paper highlights the importance of a paradigm that picks a model based on awareness of context and the physics underlying models. It puts a strong emphasis on the fact that parsimony, interpretability and operational suitability needs to be considered as essential requirements alongside predictive accuracy.

REFERENCES

1. Yeh, W.W.-G. 1985. Reservoir management and operations models: A state-of-the-art review. *Water Resources Research*, *21*(12), 1797–1818,
2. Labadie, J.W. 2004. Optimal operation of multireservoir systems: State-of-the-art review. *Journal of water resources planning and management*, *130*(2), 93–111,
3. Wurbs, R.A., 1993. Reservoir-system simulation and optimization models. *Journal of Water Resources Planning and Management, 119*(4), 455–472,
4. Loucks, D.P., and van Beek, E. 2017. *Water Resource Systems Planning and Management: An Introduction to Methods, Models, and Applications*, Springer, Cham, Switzerland,
5. Todini, E. 2007. Hydrological catchment modelling: Past, present and future. *Hydrology and Earth System Sciences*, *11*, 468–482.
6. Adeloye, A.J., Guo, R. 2019. Reservoir storage–yield–reliability planning: Single-reservoir problem. *Journal of Hydrology*, *569*, 696–708.
7. Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Hernegger, M. 2018. Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022.
8. Kratzert, F., et al. 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*, 5089–5110.
9. Shen, C. 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593.
10. Fang, K., et al. 2017. An application of deep learning for streamflow prediction. *Water Resources Research*, *53*, 980–1000.
11. Nearing, G. S., et al., 2021. What role does hydrological science play in the age of machine learning? *Water Resources Research*, *57*, e2020WR028091.
12. Lees, T., et al. 2021. Benchmarking data-driven rainfall–runoff models. *Hydrology and Earth System Sciences*, *25*, 5517–5544.
13. Breiman, L. 2001. Random forests. *Machine Learning*, *45*(1), 5–32.
14. Tyralis, H., Papacharalampous, G., Langousis, A. 2019. A brief review of random forests for water scientists and practitioners. *Water*, *11*(5), 910.

15. Papacharalampous, G., Tyralis, H., Langousis, A. 2019. Large-scale comparison of machine learning methods for hydrological prediction. *Water Resources Research*, *55*, 7290–7311.
16. Mosavi, A., et al. 2018. Flood prediction using machine learning models. *Water*, *10*(11), 1536.
17. Zhang, D., et al. 2018. A comparison of machine learning algorithms for reservoir operation. *Journal of Cleaner Production*, *196*, 126–138.
18. Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F. 2009. Decomposition of the mean squared error and NSE performance criteria. *Journal of Hydrology*, *377*(1–2), 80–91.
19. Nearing, G.S., Gupta, H.V. 2018. Ensembles vs. machine learning: Quantifying uncertainty in hydrologic predictions, *Water Resources Research*, *54*, 9909–9930.
20. Karpatne, A., et al. 2017. Physics-guided neural networks for scientific discovery. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2313–2327.
21. Willard, J., et al. 2022. Integrating scientific knowledge with machine learning. *ACM Computing Surveys*, *55*(3), 1–37.
22. Hochreiter, S., Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.
23. Graves, A., Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional LSTM networks. *Neural Networks*, *18*(5–6), 602–610, 2005.
24. Vaswani, A., et al. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.
25. Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., Cottrell, G. A dual-stage attention-based recurrent neural network for time series prediction. *Proceedings of IJCAI*, 2627–2633, 2017.
26. Towler, E., Rajagopalan, B., Prairie, J. Reservoir operations under climate change. 2010. *Journal of Hydrology*, *406*(1–2), 21–34.
27. Madani, K. 2010. Game theory and water resources. *Journal of Hydrology*, *381*(3–4), 225–238.
28. Giuliani, M., et al. 2016. Decision support systems for water resources planning. *Environmental Modelling & Software*, *84*, 92–111.
29. NASA POWER Project "Prediction of Worldwide Energy Resources (POWER)" NASA Langley Research Center, 2024.
30. California Department of Water Resources, "California Data Exchange Center (CDEC)," Sacramento, CA, USA, 2024.
31. Klemeš, V. 1986. Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24, 1986.
32. Bennett, N.D., et al. 2013. Characterising performance of environmental models. *Environmental Modelling & Software*, *40*, 1–20.
33. Beven, K., *Rainfall–Runoff Modelling: The Primer*, 2nd ed., Wiley-Blackwell, Chichester, UK, 2012.
34. Frame, J.M., et al. 2022. Deep learning versus tree-based methods in water resources forecasting. *Water Resources Research*, *58*, e2021WR031621, 2022.
35. Kratzert, F., et al. 2023. On the limits of deep learning for hydrology. *Hydrology and Earth System Sciences*, 27, 1805–1822.
36. Diebold, F.X., Mariano, R.S. 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics, *13*(3), 253–263.
37. Efron, B., Tibshirani, R.J. 1993. *An Introduction to the Bootstrap*, Chapman and Hall/CRC, New York.
38. Politis, D.N., White, H. 2004. Automatic block-length selection for the dependent bootstrap. *Econometric Reviews*, *23*(1), 53–70.
39. Harvey, D., Leybourne, S., Newbold, P. 1997. Testing the equality of prediction mean squared errors. International Journal of Forecasting, *13*(2), 281–291.