Diagnostic assessment of reservoir release policies using LSTM across the continental U.S.

Matthew Chen[1], **Jonathan D. Herman**[1]

[1]Department of Civil & Environmental Engineering, University of California, Davis

Key Points:

- LSTMs can model reservoir releases with physically justifiable storage dynamics in internal cell states without using observed storage input.
- Model accuracy is only weakly related to the accuracy of the learned storage representation, but is strongly related to the degree of regulation.
- Large sampled pooled training does not generalize well to out-of-sample reservoirs, but finetuning improves accuracy for reservoirs with scarce data.

Corresponding author: Matthew Chen, `mtwchen@ucdavis.edu`

Abstract

The influence of reservoirs on the water cycle introduces significant uncertainty for hydrologic prediction. The representation of reservoirs in hydrologic models ideally must be accurate, interpretable, and transferable across sites. Recent studies have highlighted the potential for data-driven methods, including long short-term memory (LSTM) networks, to accurately capture reservoir releases. However, the performance of LSTM models of reservoir releases has not yet been diagnosed on a large-sample dataset to understand whether their accuracy is physically justified. This study evaluates the ability of LSTMs to represent reservoir release policies across the continental U.S., leveraging the recently developed ResOpsUS dataset. In particular, we focus on four key challenges to the development and application of LSTMs for this purpose: architecture selection, mass conservation, nonstationarity in time, and large-sample pooled training. We find that in many cases the LSTM succeeds in encoding physically interpretable storage dynamics in its internal states. However, the model accuracy is only weakly related to the strength of the learned storage representation; instead, it is log-linearly related to the degree of regulation. In addition, LSTMs struggle to generalize in time, where distributional shifts in operating conditions may cause unstable accuracy, and in space, where large sample pooled training fails to improve performance. This study contributes to the growing literature on interpreting deep learning models of human-hydrologic systems.

1 Introduction

Reservoirs are critical infrastructure that balance human and environmental needs such as flood control (Boulange et al., 2021), water supply (Biemans et al., 2011), hydroelectricity, and environmental flows (Adams et al., 2017; Yin et al., 2014). However, because reservoir releases are managed to consider complex tradeoffs between multiple competing operating objectives, they are fundamentally dependent on human decisions and cannot be modeled as a physical hydrologic process (Longyang & Zeng, 2023; Yang et al., 2016). The role of reservoirs in altering surface flows is widely recognized (Nilsson et al., 2005; Zhou et al., 2016; Galelli et al., 2025), and human intervention in the water cycle introduces significant uncertainty for hydrologic prediction (Thompson et al., 2013). Further, reservoirs are often represented simplistically in hydrologic models, which may not be able to capture realistic operating rules (Pokhrel et al., 2016). For example, Dang et al. (2020) found that the misrepresentation of dams results in erroneous parameterizations of hydrologic models. Further, Hodgkins et al. (2024) found that when reservoir storage was neglected in continental-scale hydrologic models, their errors increase nonlinearly with reservoir storage.

The ideal approach to represent reservoir releases in hydrologic models would be to incorporate the control policies directly as they are defined by operating agencies. However, these rule curves are not always well-documented, and do not generalize across basins. It is also widely recognized that true release decisions, which require undocumented operator judgment to adapt to current conditions, constraints, and competing objectives, often deviate from these rules (Oliveira & Loucks, 1997). Given these challenges, alternative models of reservoir release policies have been developed in main categories: generic control policies, optimization methods, and data-driven policies. Generic control policies have low data requirements and are highly transferable, but may not accurately reproduce observed flows at fine temporal resolutions (Haddeland et al., 2006; Hanasaki et al., 2006). On the other hand, optimization methods seek to find optimal releases based on one or more operating objectives (Turner & Galelli, 2016). These methods can provide valuable decision support, although in simulation the predefined objectives and simplifying assumptions often fail to capture the complexity of real-world operating conditions (Giuliani et al., 2021). Finally, data driven methods infer operating policies directly from historical inflow, storage, and release records, such as calibrating the parameters of a generic policy or other functional form against observed data (Tefs et al., 2021; Turner

et al., 2020; Yassin et al., 2019; Zhao et al., 2016). For example, Turner et al. (2020) found that such data driven policies are more accurate than generic policies where key parameters are set uniformly across reservoirs.

Within this category of data-driven control policies, several recent studies have highlighted the potential for fully empirical machine learning methods to accurately capture reservoir releases and outperform other data-driven methods (Coerver et al., 2018; Dong et al., 2023; Ehsani et al., 2016; Gangrade et al., 2022; Longyang & Zeng, 2023; Yang et al., 2016). Data-driven methods are supported by recent unprecedented high resolution reservoir datasets on a national scale (Hou et al., 2022; Steyaert et al., 2022), providing an opportunity to develop and analyze data driven reservoir models on large samples, much like the CAMELS dataset has done for rainfall-runoff modeling (Addor et al., 2017). In particular, the widespread success of LSTMs in rainfall-runoff modeling (Kratzert et al., 2018) holds promise for modeling reservoir releases: a model with explicit accumulation of one observable state should be able to capture release decisions, provided that the inputs align with relevant decision-making processes.

While LSTMs have been applied to models of reservoir releases, these studies have not investigated the key benefit of LSTMs for this purpose — the accumulation of storage. Because the LSTM maintains memory cell states through time (Hochreiter & Schmidhuber, 1997), it can in principle integrate past inflows and outflows to infer a latent storage state, analogous to how rule-based models enforce mass balance. This makes reservoir modeling distinctly different than hydrologic modeling, because the accumulated state is easily measured. However, many studies using LSTMs (e.g., (Longyang & Zeng, 2023; Tran et al., 2025; Fan et al., 2023; Gangrade et al., 2022)) use observed storage directly as an input variable, rather than testing whether the network can infer it through accumulation. When observed inflow and storage are both provided, the model is implicitly provided with recently observed outflow via mass balance. This potentially allows the model to exploit the high persistence and autocorrelation with previous observations rather than learning the true data generating process (i.e. rules of human decision making). Data leakage in this way may artificially inflate the accuracy of LSTMs and reduce generalization ability especially in cases where observed storage is not available but is simulated or estimated (e.g. when coupling with hydrologic models).

Here we develop a diagnostic framework for LSTMs to understand their ability to internally accumulate storage states and respect mass conservation, and to generalize over time and space. This includes several steps with direct parallels in rainfall-runoff modeling: architecture selection (Fuente et al., 2024), analysis of cell states (Lees et al., 2022; Kratzert, Herrnegger, et al., 2019), conservation of mass (Hoedt et al., 2021), and large-sample pooled training (Kratzert et al., 2024). Training a data driven model on a standardized large sample of reservoirs captures a diverse range of operating conditions and strategies, and potentially enables the extrapolation of policies to data-scarce regions (Turner et al., 2021). However, it is not clear whether reservoir release policies, once trained, can generalize across basins in the same way as physical hydrologic processes. Finally, modeling reservoir releases raises the additional challenge of nonstationarity, as the operator preferences and the structure of the policy itself may change during the period of record (Mason et al., 2018).

This study contributes a large sample diagnostic framework for LSTM-driven reservoir releases that addresses the key challenges of physical interpretability and generalization in time and space. For the former, we ask if the LSTM learns to capture latent representations of mass balance and storage dynamics for the reservoir operations problem. Our approach is to correlate unseen observed storage data with LSTM cell states, first individually and then as a linear model fitted to the cell states. For generalization in time, we provide a method to monitor performance in time through moving windows so that unstable performance due to nonstationary policies can be quickly detected. Finally, for generalization in space, we test if performance can be improved through large

sample pooling, as well as extrapolation to data scarce regions through finetuning and transfer learning.

2 Methods

2.1 Long Short-Term Memory Networks

Here we give a brief introduction to the Long Short-Term Memory (LSTM) architecture (Hochreiter & Schmidhuber, 1997). The LSTM model addresses the problem of unstable gradients in training recurrent neural networks by conserving long term information using memory cells managed by several gating mechanisms, which control the flow of information through element-wise matrix multiplication with gate values ranging between 0 and 1. These allow the model to learn temporal relationships and long-term dependencies. Further, the ability of the LSTM to dynamically accumulate information makes it a well-suited candidate to model dynamical systems (Jordan et al., 2021; Kratzert, Klotz, et al., 2019; Wang, 2017) such as reservoir control. In a LSTM, every timestep t has a hidden state $h^{(t)}$ and a memory cell state $c^{(t)}$. The cell states store and maintain long term information, where the information from the cell state can be released into the hidden state where it can be used for prediction. This flow of information is managed by the output gate. As new inputs arrive, the model can also save and remove information from the cell state, which are managed by the input gate and forget gate, respectively. For example, in the reservoir control problem, storage states can be modeled by memory cells, where mass accumulation is managed by the input and forget gates, and release decisions can then be modeled based on the accumulated storage and day of the year, as managed by the output gate. Note that a LSTM architecture does not conserve mass unless explicitly tailored to do so (Hoedt et al., 2021).

The gate values at each timestep depend on the previous hidden state $h^{(t-1)}$ and the new input $x^{(t)}$ while the logistic function $\sigma : \mathbb{R} \rightarrow [0, 1]$ enforces that the gates values fall between 0 and 1. The forget gate, $f^{(t)}$, parameterized by the weight matrices W_f, U_f, and b_f, controls what information is retained versus forgotten from the previous cell state (Eq. 1).

$$f^{(t)} = \sigma\left(W_f h^{(t-1)} + U_f x^{(t)} + b_f\right) \tag{1}$$

The input gate, $i^{(t)}$, controls the information flow from the new input into the cell state. This gate is parameterized by the weight matrices W_i, U_i, and b_i (Eq. 2).

$$i^{(t)} = \sigma\left(W_i h^{(t-1)} + U_i x^{(t)} + b_i\right) \tag{2}$$

Finally, the output gate, $o^{(t)}$, controls information flow from the cell state to the hidden state to make a prediction at the current timestep. This gate is parameterized by the weight matrices W_o, U_o, and b_o (Eq. 3).

$$o^{(t)} = \sigma\left(W_o h^{(t-1)} + U_o x^{(t)} + b_o\right) \tag{3}$$

After the gate values are computed, a candidate cell state update $\tilde{c}^{(t)}$ is computed from the previous hidden state and data input from the current timestep using a hyperbolic tangent activation function (Eq. 4).

$$\tilde{c}^{(t)} = \tanh\left(W_c h^{(t-1)} + U_c x^{(t)} + b_c\right) \tag{4}$$

The cell state is then updated based on the values of the forget and input gates (Eq. 5).

$$c^{(t)} = f^{(t)} \odot c^{(t-1)} + i^{(t)} \odot \tilde{c}^{(t)} \tag{5}$$

Finally, the hidden state is computed based on the value of the output gate, which is used to derive the final prediction (Eq. 6).

$$h^{(t)} = o^{(t)} \odot \tanh\left(c^{(t)}\right) \tag{6}$$

In the supplement (Figure S1), the operations from Equations 1-6 are presented as a computational graph.

2.2 Data Processing

Reservoir inflow, storage, and release data are drawn from two sources: ResOpsUS (Steyaert et al., 2022), which covers the majority of large reservoirs in the continental U.S. over the period 1980-2020 on a daily timestep; and longer records from the U.S. Bureau of Reclamation RISE system (1940s-present) to support more detailed modeling of specific reservoirs in the Western U.S. We filter the ResOpsUS data to only the reservoirs with records that are at least 90% complete over the period 1980-2020, which leaves 116 reservoirs. We analyze 4 additional reservoirs from the U.S. Bureau of Reclamation to support longer records (see Figure S2 map in the supplement for all sites considered in the study). The reservoir data records including inflow and release timeseries are split into training, validation, and testing portions. The training portion, representing the first 60% of the available timeseries, is used directly for model training. The validation portion, representing the next 20% of the available record, is used for hyperparameter tuning, model selection, and early stopping. Early stopping prevents overfitting by interrupting the training process based on the validation data as a proxy for out-of-sample performance (Li et al., 2019). The validation set provides some measure of out-of-sample performance, especially if the validation set is not overutilized in the modeling process (i.e. overfitting to validation or "data leaking"). Finally, the testing portion, representing the last 20% of the available data record, is used solely for the estimation of out-of-sample performance. The testing data is untouched throughout the model building process; however, it is also the furthest away from the training set in time. This may be a challenge if the reservoir operating policy has changed, a challenge we investigate later in the study.

Prior to modeling, the data is linearly transformed to be zero mean and unit standard deviation, based on statistics from the training set. In early experiments, data normalization (scaling between 0 and 1) was also tested but provided little benefit over standardization. Additionally, while the majority of records are verified to be complete, missing values are imputed using the training mean. We also split the data into batches of 3 years (the batch sequence length was chosen between 0.5 and 5 years in preliminary testing). Using longer sequences allow the model to capture longer term dependencies but may incur difficulty due to vanishing or exploding gradients in the training process where gradient magnitudes may become exponentially large or small over many steps of backpropagation through time.

2.3 Model Selection, Hyperparameter Tuning, and Benchmarks

While model selection experiments are conducted on the ResOPS dataset to compare model architectures and evaluate LSTM performance against benchmark models over many reservoirs (see Section 3.1), we conduct hyperparameter tuning on an individual reservoir (Shasta Reservoir, California) to select optimal hyperparameters using exhaustive grid search. Specifically, we select the model with the optimal validation loss

over a predefined grid of hyperparameters, averaging over 5 different random seeds to account for stochasticity in the optimization algorithm. We tune the number of LSTM layers (1 or 2), the size of the LSTM hidden layer (between 5 and 50), the hidden size of the feed-forward network (between 5 and 50), and the dropout regularization probability (0.3, 0.5, or 0.7). Shasta reservoir was chosen for its long data record and its representative degree of regulation (discussed in Section 3.3). It is computationally infeasible to tune hyperparameters using grid search for all 116 reservoirs in the large sample dataset individually, so the LSTM hyperparameters selected here are applied throughout the study. While we recognize that optimal hyperparameters for Shasta Reservoir may not be optimal for a different reservoir, tuning to a specific reservoir will provide a general idea of the architecture necessary to capture reservoir operations. That is, we assume that reservoir operating policies across the continental US are somewhat similar in terms of their complexity, even if their operating purposes and typical release patterns differ.

For model selection we consider four main LSTM architectures along with several other machine learning benchmarks (Figure 1). Since we are interested in the ability of the LSTM to learn to conserve mass and learn reservoir storages implicitly in its cell states, the modeling task is to predict reservoir releases based only on two inputs, the inflow and the day of the year. By doing so, we assume that storage is primarily driven by inflow and outflow, and that this dominates other external effects such as evaporation and seepage. Model 1 is a standard "vanilla" LSTM model where data is first processed by an LSTM cell and then headed by a single layer feed-forward neural network to provide additional non-linear flexibility in learning the operating policy. Model 2 is similar to Model 1, but adopts an autoregressive structure in which the previous output is concatenated as an additional input for the current prediction.

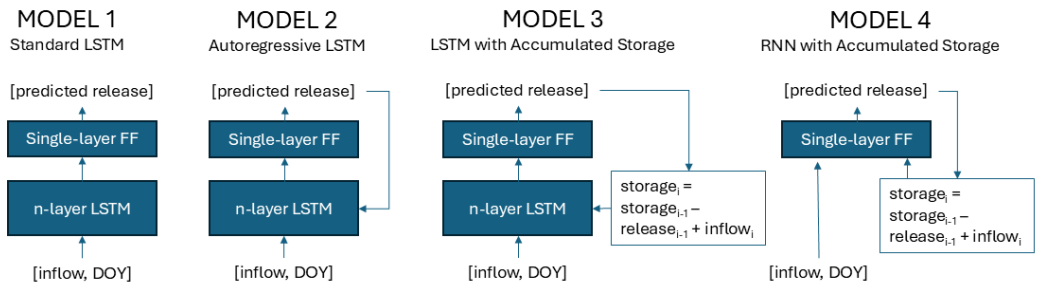

Figure 1. LSTM model architectures. "FF": feed-forward neural network; "DOY": day of the year.

For Model 3 and Model 4, we propose two alternative architectures that accumulate implied storage states internally. Specifically, Model 3 explicitly models the mass balance for implied storages as an additional internal state and concatenates them as input so that mass conservation is learned without relying on the LSTM gating mechanisms. Model 4 is similar to Model 3, except without the LSTM gating mechanisms, which results in a model resembling a mass-accumulating recurrent neural network (RNN). However, Model 4 is unable to capture longer-range dependencies through learned cell states beyond storage, unlike in Model 3. All four models are trained using the squared error loss function. These LSTM experiments were conducted in the Pytorch deep learning library (Paszke et al., 2019). In training, model parameters were optimized using the Adam algorithm, a first-order stochastic gradient descent algorithm with momentum and strong empirical performance (Kingma & Ba, 2015).

We consider autoregressive linear and random forest models with 5 lags as benchmarks for other machine learning architectures against the LSTM. Specifically, we model reservoir releases as a function of the current day of the year and inflow, as well as the previous 5 inflow values. We also consider benchmarks where the current observed storage is another data input, since the linear and random forest models cannot learn to preserve information over time unlike the LSTM architecture (these are denoted linear-S and random forest-S, respectively). This modeling problem is represented by Equation 7, where $\hat{y}^{(t)}$ is the predicted target release at time t, and x_{in}, x_{sto}, x_{DOY} are the inflow, storage, and day of year features, respectively. These benchmark models, trained with squared-error loss, are implemented in the open-source scikit-learn library (Pedregosa et al., 2011).

$$\hat{y}^{(t)} = f\left(x_{in}^{(t)}, x_{in}^{(t-1)}, x_{in}^{(t-2)}, x_{in}^{(t-3)}, x_{in}^{(t-4)}, x_{in}^{(t-5)}, x_{sto}^{(t)}, x_{DOY}^{(t)}\right) \tag{7}$$

Since we are interested in the LSTM learning storage information implicitly, it is useful to also add a benchmark where observed storage is explicitly provided to the LSTM model. As such we also compare a version of Model 1 (denoted as Model 1-S) that is provided inflow, storage and the day of the year rather than only inflow and day of the year.

Finally, as a benchmark against a rule-based model, we adapt a simplified reservoir release policy which includes an exponential hedging release rule (Eq. 8) and a flood control rule (Eq. 9) as used in the larger context of climate adaptation planning in Steinschneider et al. (2023); Chen and Herman (2024). $R(t), S(t)$ are releases and implied storages for time t, respectively, and $R_m(t), S_m(t)$ denote the median observed release and storage (in the training data) for that day of the year. The flood control rule applies if the day of the year is in between $[x_1, x_2]$, and storage exceeds a threshold $S(t) > x_4 S_m(t)$. x_0, x_1, x_2, x_3, x_4 are fitted parameters resulting in 5 parameters per reservoir. We match the LSTM training methodology by fitting the policy parameters using the training dataset, using early stopping on the validation dataset to prevent overfitting, and providing a final estimate for out-of-sample performance on the test dataset. This ensures that the results are directly comparable with the LSTM and other machine learning benchmarks. Additional details about the model formulation can be found in Steinschneider et al. (2023).

$$\frac{R(t)}{R_m(t)} = \left(\frac{S(t)}{S_m(t)}\right)^{x_0} \tag{8}$$

$$R(t)' = R(t) + x_3(S(t) - x_4 S_m(t)) \tag{9}$$

Table 1 summarizes R^2 scores for Models 1-4 on Shasta Reservoir after hyperparameter tuning, as well as the linear, random forest, and rule-based benchmarks, and variants of models where observed storage is a provided input.

We are interested in comparing validation scores so that the test data is completely withheld from the model building process; the test data can later be used for further analysis of model behavior (such as the behavior of cell states) and provide a final estimate for out-of-sample performance. Between Models 1-4, Model 1 (validation $R^2 = 0.69$) and Model 2 (validation $R^2 = 0.70$) both perform reasonably well in validation, however, Model 1 is more parsimonious and efficient to train. After running the hyperparameter tuning process with exhaustive grid search on Shasta Reservoir, we find the following optimal configuration for Model 1: 1 LSTM layer, 30 LSTM hidden units, 15 feed-forward hidden units, and a dropout probability of 0.3. Figure 2 provides a visualization of the hyperparameter tuning results. Note that smaller architectures, particularly with 5 LSTM or feed-forward hidden units, and higher dropout (0.7) are associated with higher validation error; tuning results are more uniform beyond these cases.

Table 1. Train, validation, and test R^2 scores for Models 1-4 and benchmark models. Models with observed storage as input are denoted with -S.

Model	Storage Inputted	Train (1944-1991)	Validation (1992-2007)	Test (2008-2022)
LSTM Model 1	No	0.73	0.69	0.43
LSTM Model 2	No	0.74	0.70	0.44
LSTM Model 3	No	0.59	0.60	0.52
LSTM Model 4	No	0.60	0.67	0.41
LSTM Model 1-S	Yes	0.82	0.75	0.64
Linear	No	0.4	0.4	0.14
Random Forest	No	0.64	0.57	0.37
Linear-S	Yes	0.42	0.43	0.21
Random Forest-S	Yes	0.69	0.64	0.56
Rule Based Model	No	0.69	0.68	0.66

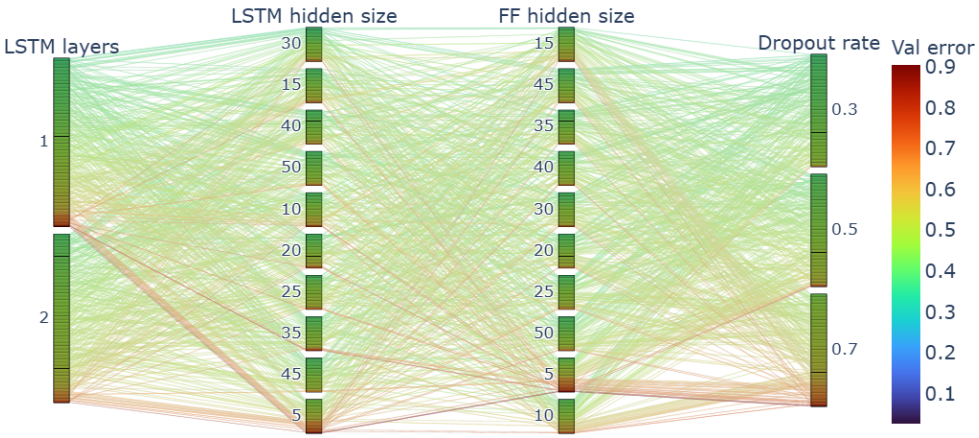

Figure 2. Hyperparameter tuning of LSTM Model 1 on Shasta Reservoir.

2.4 Model Diagnostics

Beyond model selection, our goal is to diagnose the behavior and performance of LSTM reservoir models, specifically, assessing their physical interpretability and performance over time, and testing if learned policies are generalizable in both space and time. For large sample experiments, we leverage reservoir inflow, storage, and release records from the ResOps dataset (Steyaert et al., 2022), and for additional experiments that benefit from longer history we utilize records from the U.S. Bureau of Reclamation. For each reservoir in the selected dataset, we conduct data processing as described above, selecting 60% of the available record for training, 20% for validation, and the last 20% for testing.

2.4.1 Analysis of Cell States

We investigate whether a LSTM model of reservoir releases can learn storage representations in its memory cell states without being given storage data explicitly, with the larger goal of determining if such models are physically interpretable. This is inspired by the success of LSTM in capturing hydrologic states such as snowpack in its memory cells (Kratzert, Herrnegger, et al., 2019). Similar to Kratzert, Herrnegger, et al. (2019), we study the correlation coefficients between individual memory cell states compared to observed storage with the purpose of uncovering if the LSTM learns to internally represent storage in an interpretable way; the ideal would be a model that aligns with physical understanding of reservoir storage and release decisions.

Further, the LSTM may learn a distributed representation of storage across multiple of its cell states. To address this, we employ a linear probe (Liu et al., 2019; Alain & Bengio, 2018), i.e., a supervised ordinary least squares model fitted between frozen cell states and observed storage on training data to test if storage can be represented by a linear combination of cell states. While useful latent representations may be encoded nonlinearly, more complex probes (e.g. multi-layer perceptrons) are not only less interpretable, but also introduce uncertainty in which model, the probe itself or the original model, learns the useful relationship. If internal states are a sufficient statistic of the input features, i.e. providing a lossless encoding, a nonlinear probe with sufficient capacity could in principle learn any task as a function of the original features (Hewitt & Liang, 2019). As such, we argue that linear probes are simple enough that if storage can be linearly extracted from cell states on unseen data, we can conclude that the LSTM itself learns mass balance in a physically meaningful and accessible way. This interpretation of linear probes has also been accepted in the broader machine learning literature (Belinkov, 2021), and similar methodology has previously been applied to analyze hydrological states (Lees et al., 2022), though not yet for reservoir operations. Here, similar to the analysis of individual states, we study the correlation coefficients between observed storage and the linear probe on unseen (test) data, as well as their relationship with accuracy and the degree of regulation.

2.4.2 Model Performance vs. Degree of Regulation

Previous studies have shown that LSTM rainfall-runoff models perform worse on managed basins, particularly those with higher degrees of regulation (Ouyang et al., 2021). Consequently, we hypothesize that this result extends to reservoir models directly, i.e. higher degrees of regulation in a reservoir adversely affects performance. Specifically, we compare performance from the large sample of individually trained reservoirs against the log ratio of mean inflow to maximum storage as a proxy for capacity. This ratio represents the inverse of the degree of regulation. We then compute Pearson's correlation coefficient between the LSTM R^2 performance and the log mean-inflow max-storage ratio. Statistical inference is done using randomization testing and Monte Carlo resampling, i.e. via permutation test, to determine the p-value against the null hypothesis of

no correlation. As a robustness check against outliers, we also include the 95% confidence interval of Thiel-Sen slopes (Sen, 1968).

2.4.3 Model Performance Over Time

While overfitting can lead to a downward shift between training set and out-of-sample performance, reservoir policies themselves may also change over time. Any difference between the out-of-sample and test distributions, i.e. distributional shift in operating conditions, may cause a declining trend in performance. To understand this problem, we first train a new "initial" LSTM model on the first 30 years and validate on the next 10 years, and then analyze R^2 performance on rolling and sliding 20-year windows to capture how performance changes over time. Notably, this experiment is challenged by limited record lengths: the length of the initial training window is chosen so that the model can learn a reasonable representation of the operating policy while the moving window size is chosen to balance signal and noise. For this reason, we select two example reservoirs for this experiment leveraging longer records available from the U.S. Bureau of Reclamation (Shasta and Folsom, both in California) with records of 70-80 years.

2.4.4 Large Sample Pooled Training

Finally, we test a top-down modeling approach, that is, learning a general model by training on all available data. To assess the ability of a simultaneously trained LSTM reservoir policy to generalize, we randomly select 70% of ResOps reservoirs (where at least 90% of the data record is available), pool and train on them simultaneously, and test out-of-sample performance using the remaining 30% of reservoirs. This is not to be confused with data splitting in time, where we train, validate, and test on the same reservoir. Here, out-of-sample testing is done on held out reservoirs, not held out time. Recognizing that reservoirs may be operated differently, we include the main use from the GRanD dataset (Lehner et al., 2011) and four categories of degree of regulation (low, medium, high, and very high) based on the 25, 50, and 75th percentiles of the log mean-inflow max-storage ratio as additional features.

We also compare the test performance after fine-tuning the pooled model on individual reservoirs. Finetuning in this context refers to calibrating a pre-trained model to a specific reservoir by running additional training iterations from data unique to the reservoir of interest. This is related to the concept of transfer learning in the machine learning literature in which a pre-trained model trained on a large dataset can be adapted to improve performance for a potentially different task on a smaller dataset (Tan et al., 2018). The idea of "knowledge transfer" has shown to be successful in a variety of domains including image recognition (Iorga & Neagoe, 2019) and natural language processing (Ruder et al., 2019). In this case, we can train and validate (on a 75% training and 25% validation split, respectively) using 5-30 year subsets of the complete data record for the held-out reservoirs as the finetuning data, and finally test using the last 20% of the complete record so that results between finetuning, individual training, and the pooled training model are comparable. Performance is also compared to an individual model trained only on the finetuning subset with the same train/validation split, which represents individual training in a data-scarce scenario. Note that validation scores here are not directly comparable since they vary in length depending on the amount of fine-tuning data used while the test set is the same.

3 Results

3.1 Large Sample Model Performance

Figure 3 summarizes performance over the ResOPS dataset for LSTM Models 1-4 as well as the benchmarks listed in Table 1. Comparing median scores in validation

for the purpose of LSTM model selection, Model 1 (median val R^2=0.618) outperforms Model 2 (median val R^2=0.587), Model 3 (median val R^2=0.511), Model 4 (median R^2=0.368), as well as the rule-based model (median val R^2=0.504). Additionally, Model 1 is the most accurate model in validation for 62 out of all 116 sites (53%). This solidifies our model selection choice of Model 1 for the large sample experiments moving forward. We also note that the deep learning architectures where implied storages were recursive inputs (Model 3 and 4) did not improve performance compared to a purely statistical version (Model 1).

In final testing, the ranking between Model 1-4 median scores remains the same as the training and validation set. Additionally, Model 1 (median test R^2=0.567) still outperforms the rule based model (median test R^2=0.433), as well as the other architectures even with observed storage inputs: linear-S (median test R^2=0.333) and random forest-S (median test R^2=0.555). Model 1 is still the most accurate model for the majority of sites as well for 63 out of 116 reservoirs (54%). This highlights the advantage of LSTM to learn non-linear temporal relationships compared other model architectures.

Unsurprisingly, Model 1 trained with observed storage (Model 1-S) has the highest performance (median val R^2=0.713, median test R^2=0.658). However, as discussed in the introduction, previous observed storages implicitly encode previous observed outflows which can artificially inflate accuracy. When Model 1-S is tested with simulated storage instead of observed storage (olive), the persistence from recent observed releases is no longer present and performance collapses (median train R^2=0.07, val R^2=0.03, test R^2=0.04). This is the situation expected when coupling reservoir policies with hydrologic models, as the decision would depend on the modeled reservoir storage rather than the historical observation.

These aggregate performance results for the machine learning and rule based models are generally consistent with the initial model selection finding for Shasta reservoir in Table 1, though the ranking may be different for individual reservoirs. For example, the rule-based model performed much better in testing compared to the LSTM in the individual case of Shasta reservoir. This requires foreknowledge that the fixed rule structure of the model is a good fit for the reservoir in question. If it is not, performance will likely suffer, whereas the fully empirical LSTM has enough degrees of freedom to adapt to any operating policy leading to its stronger performance overall. The aggregate rankings are also consistent between the train, validation, and test periods. We also observe large spreads of R^2 scores for each model, with high and low extreme values across reservoirs. For example, the interquartile range for Model 1 in validation is 0.357 with a maximum score of 0.978 and a minimum score of -1.656.

Across all tested models, we observe severe declines in performance in the test period compared to the validation and training periods. Models 1, 2, 3, and 4 show declines of 0.08, 0.07, 0.08, and 0.05 in median R^2 between the training and test periods. The same is true for the rule based model, which had a decline of 0.12 in median R^2. Declines for individual reservoirs can be far greater than the median decline over all reservoirs - for example, Model 1 trained for Shasta reservoir (Table 1) has a decline of 0.3 in R^2 between the training and testing periods. This could be due to either poor generalization (i.e. overfitting), policy changes over time, or a combination. We explore the issue of performance over time further in Section 3.5.

Figure 4 shows R^2 with respect to geographic location, and we find no apparent spatial patterns. Climate and hydrologic factors alone do not appear to be a strong indicator of model performance.

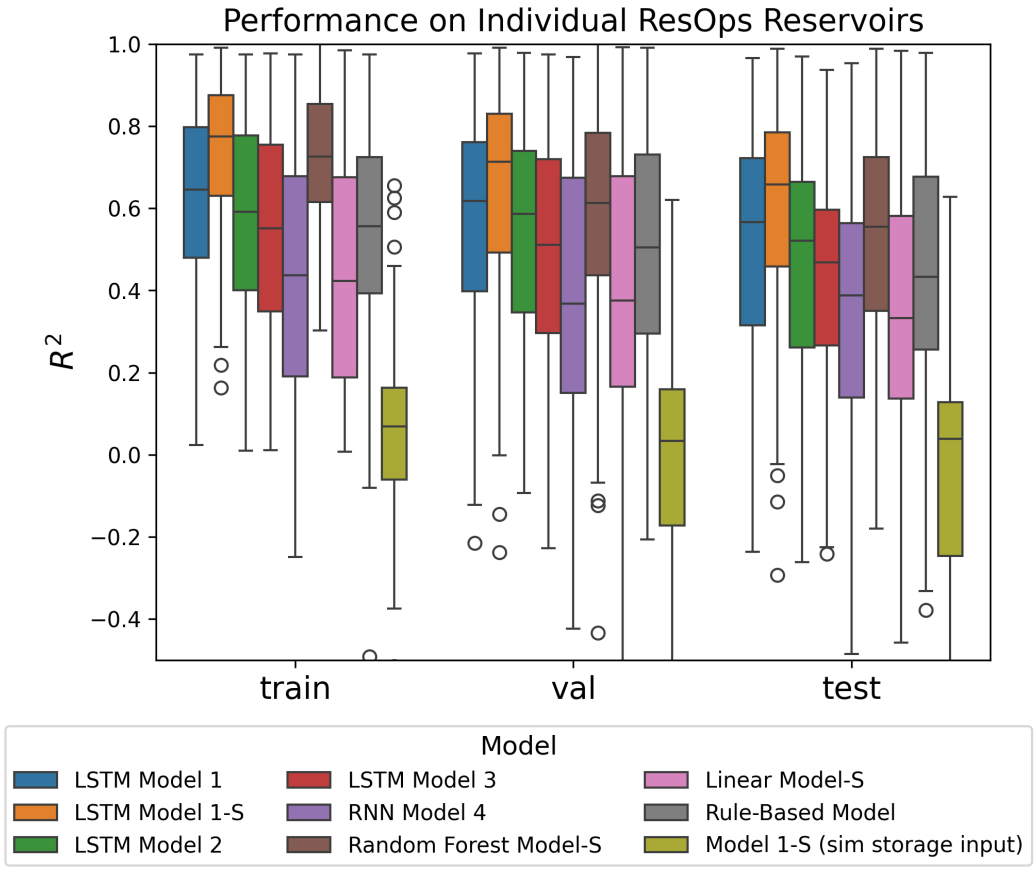

Figure 3. Comparison of individual model performance over ResOPS reservoirs. LSTM policies and benchmarks are described in Section 2.3.

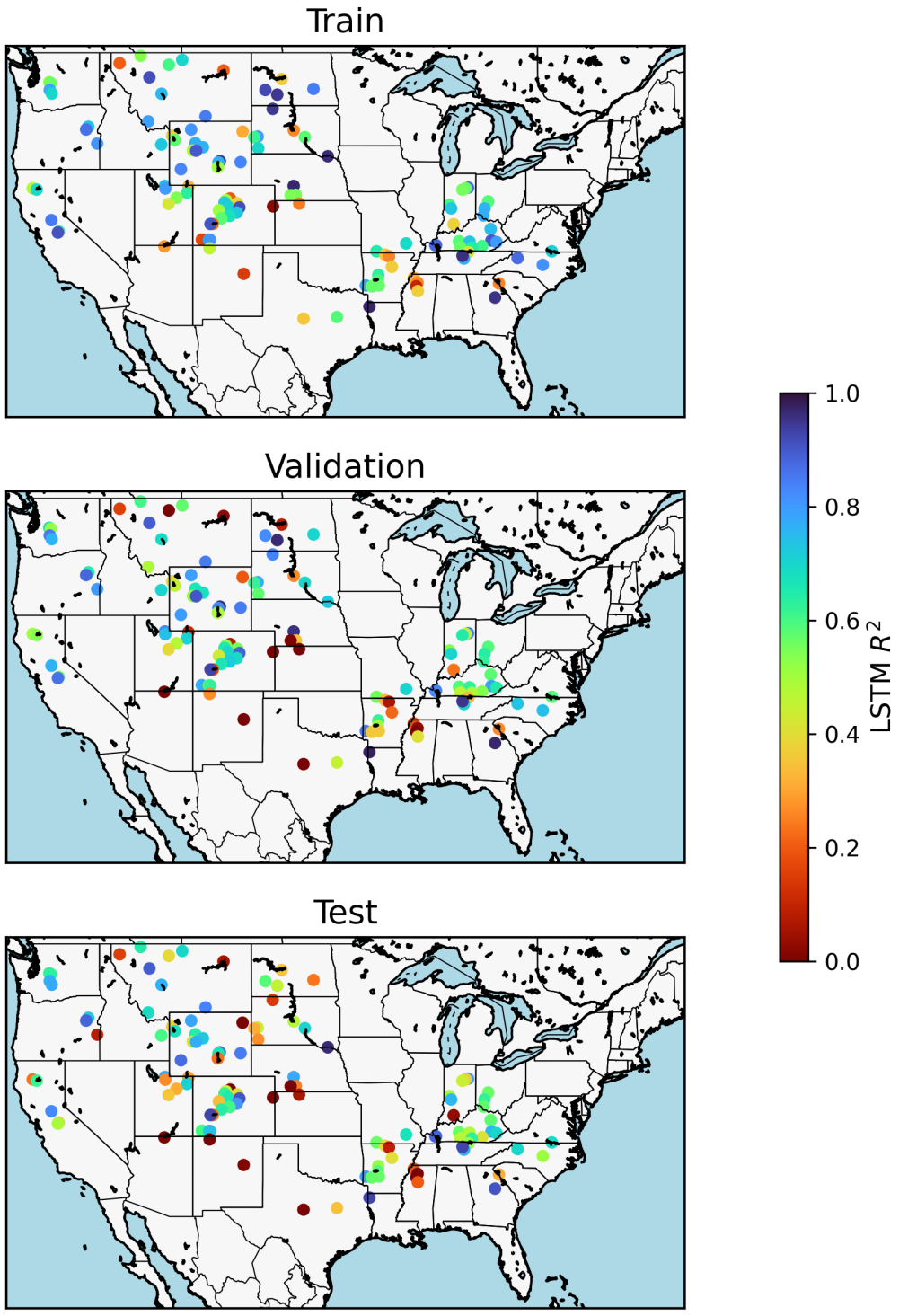

Figure 4. Map of train, validation, and test R^2 scores for individually trained LSTMs

3.2 Analysis of Cell States and Observed Storage

We compare the memory cell states of Model 1 trained on ResOPS reservoirs with observed storage to see if the model learns physically interpretable states internally. Figure 5a provides the distribution of the maximum absolute correlation coefficient between the cell states and observed storage on test data for each model. The distribution is roughly symmetric, with a median correlation coefficient of 0.531 and a standard deviation of 0.223. For the linear probes, the median correlation coefficient is 0.67 with a standard deviation of 0.237. This shows that for many sites, the LSTM successfully learns an accurate storage representation and is making release decisions aligned with physical understanding by incorporating mass balance in its latent feature embeddings. Additionally, we find that the learned storage state is distributed across multiple cell states and can be linearly extracted, as demonstrated through the success of the linear probes to encode storage as opposed to individual cell states.

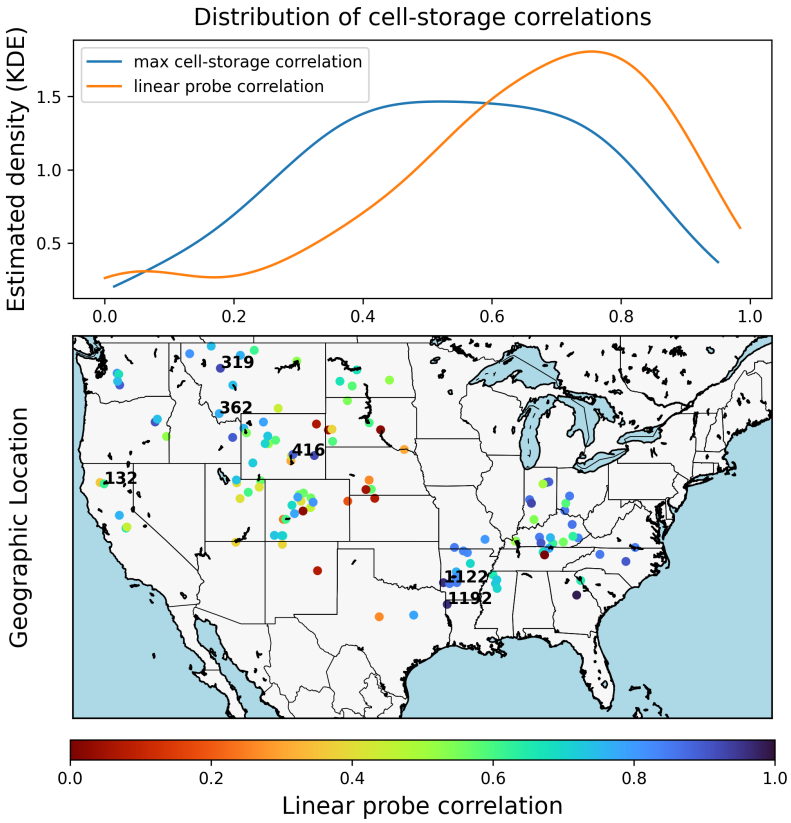

Figure 5. (a) Distribution of cell state-storage correlations over ResOPS reservoirs, using individual cells and a linear probe distributed across all cells; (b) map of linear probe correlation between cell states and observed storage.

Despite a high median linear probe correlation, the spread is large, suggesting that storage representations should be evaluated on a case by case basis. There also exists a geographic pattern in Figure 5b, where the southeast region appears to have stronger linear probe correlations. This can partly be explained by overlap with the reservoir's main use as identified by the GRaND dataset (see Supplemental Figure S3). Flood control reservoirs have the highest median linear probe correlation (0.77), followed by navigation (0.74), hydroelectricity (0.72), irrigation (0.62), water supply (0.56), and recreation (0.40). This result is unsurprising: flood control storages are driven by flood peaks

and depend on shorter term accumulations, which are easier to model, while storage accumulation for irrigation and water supply occur on longer timescales and require modeling longer-range dependencies. However, we note that real world reservoir operations incorporate additional complexity by balancing multiple competing objectives beyond a single main use.

Although statistically significant via permutation testing, we find in Figure 6a that the linear probe correlation is a weak predictor of model performance (r=0.36). That is, there are models with strong storage representations but poor performance, and models with strong performance but poor storage representation. We also find that the highest performing reservoirs with the lowest linear probe correlations also tend to have the lowest degrees of regulation as measured by the log inverse ratio of max storage and mean inflow. In this quadrant of the plot, learning storage is not necessary to produce accurate releases. However, looking at the relationship as a whole, we find no significant evidence that the linear probe correlation is related to the degree of regulation in Figure 6b.

Figure 7 shows the timeseries of observed storages and the corresponding linear probe in the testing period for 6 specific sites: 1122 (Gillham), 898 (Glendo), 1755 (Barren River Lake), 1042 (Norfork), 895 (Gavins Point), and 601 (Navajo). These sites are chosen from different quadrants of Figure 6a and are also marked in Figure 5b. Both the linear probe and storages in these plots have been scaled. In the best case scenario, both storage representation and model accuracy are high. For example, the Gillham linear probe (r=0.96) accurately captures the storage dynamics driven primarily by short term accumulations of flood peak storage, and the accuracy is also high (test $R^2 = 0.79$). Similarly, the Glendo linear probe (r=0.93) successfully captures the annual seasonality of storage and the test $R^2 = 0.85$. However, a good storage representation does not guaranty good accuracy. For example, Barren River Lake's linear probe captures storage dynamics well (r=0.90) both for flood peak storage and annual seasonality, albeit performance is poor (test $R^2 = 0.47$). Norfork is a starker example where the linear probe correlation is 0.82 but the test R^2 is only 0.09. In this case, we can deduce that the error in release predictions occurs between the mapping from storage to release, and not from inflow/outflow to storage. Conversely, good model accuracy can also be achieved with poor storage representation. For example, Gavins Point achieves a test $R^2 = 0.97$ but the linear probe poorly captures storage dynamics (r=0.31). This is especially true for models with low degrees of regulation, since releases are more sensitive to inflows directly and may not require storage accumulation to make accurate predictions. Finally, in the worst case, poor performance is accompanied by poor storage representation, which is the case for Navajo (linear probe r=0.38, model test $R^2 = -0.06$). In summary, we find that LSTMs are capable of representing storage dynamics including flood peak accumulation and annual seasonality, but the relationship between linear probe accuracy and release accuracy varies widely across sites.

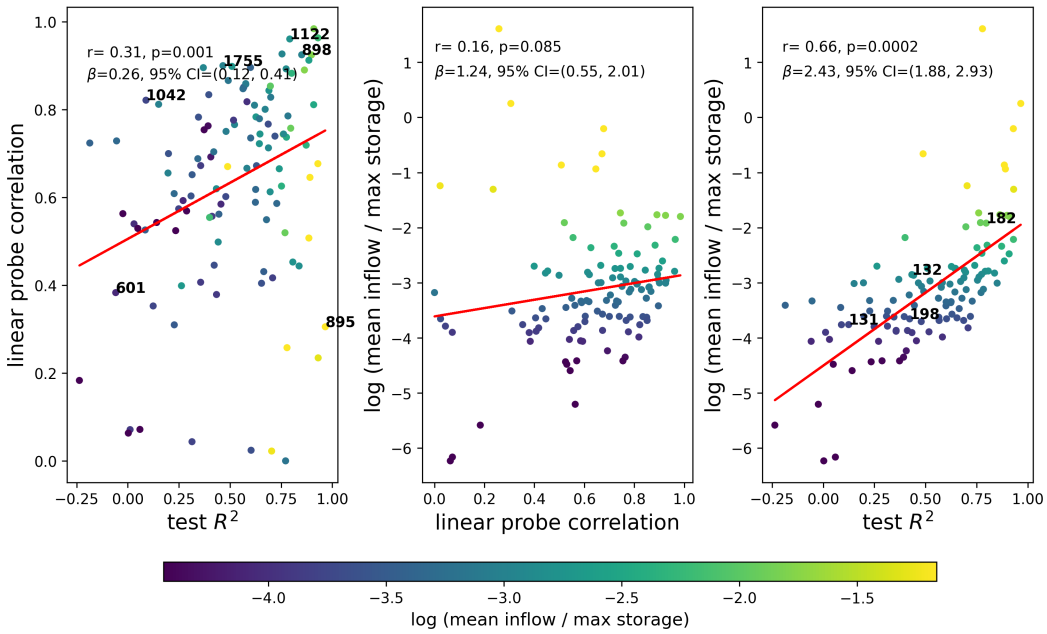

Figure 6. Scatterplots between (a) test R^2 score and linear probe correlation with observed storage, (b) linear probe correlation and the inverse degree of regulation (log mean-inflow max-storage ratio), and (c) test R^2 score and the inverse degree of regulation

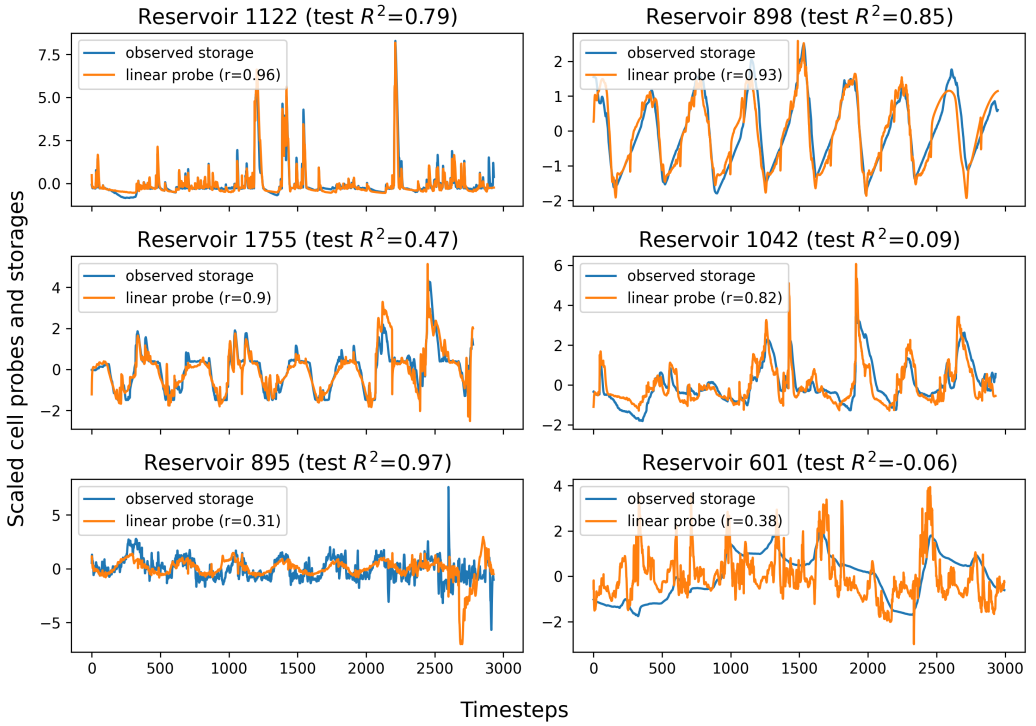

Figure 7. Timeseries of linear probes and observed storage in the test period for selected sites.

3.3 Degree of Regulation and Model Performance

We aim to further understand the model performance in the context of the degree of regulation. A lower degree of regulation may indicate shorter lag times between inflow and release, i.e., release predictions are more directly sensitive to inflow, leading to better model accuracy. Figure 6c shows R^2 performance against the degree of regulation (inverted) and observed storage. We find that the Pearson correlation between R^2 score and the degree of regulation is strong (r=0.66). Randomization inference shows that the correlation coefficients are significant at the 0.05 level, rejecting the null hypothesis of no correlation. The 95% CI of Thiel-Sen slopes (β) also does not contain zero, demonstrating that the association is robust to outliers. If the residence time of water is shorter (lower degree of regulation) and outflow is more directly sensitive to inflow, the model does not need to rely on storage as an intermediate state which may depend on longer range depedencies and adds an additional layer of modeling uncertainty, and therefore increases accuracy in general.

Four sites, Folsom (182), Shasta (132), Trinity (131), and New Melones (198), are selected from Figure 6c to represent different degrees of regulation. These reservoirs have longer inflow records from the U.S. Bureau of Reclamation dating back to the construction of the reservoir, providing several additional decades prior to the ResOpsUS dataset that can be used to analyze performance behavior. Folsom represents a low degree of regulation, while Trinity and New Melones have higher degrees of regulation. Shasta reservoir falls in between. Figure 8 plots the predicted and observed releases for these four selected reservoirs using LSTM Model 1.

Both Shasta and Folsom capture peak releases reasonably well, but Shasta is more prone to false-positive peaks, which are especially prevalent in the test period compared to training or validation. Specifically, the test period has a lower frequency of release peaks compared to the training or validation periods, which indicates that the operating conditions (hydrology and/or operating policy) have shifted since the training period, leading to poor performance. In contrast to Shasta, Folsom has a lower degree of regulation so the outflows are more sensitive to inflow directly. As previously discussed, predictions for low DOR reservoirs are less reliant on long range dependencies which corresponds to higher model performance. Consistent with Figure 6c, New Melones and Trinity reservoirs have much lower performance corresponding to their high degree of regulation. While both learn reasonable seasonal releases, the models have largely ignored peak releases, especially for Trinity reservoir. These results confirm the finding that the degree of regulation adversely affects model performance.

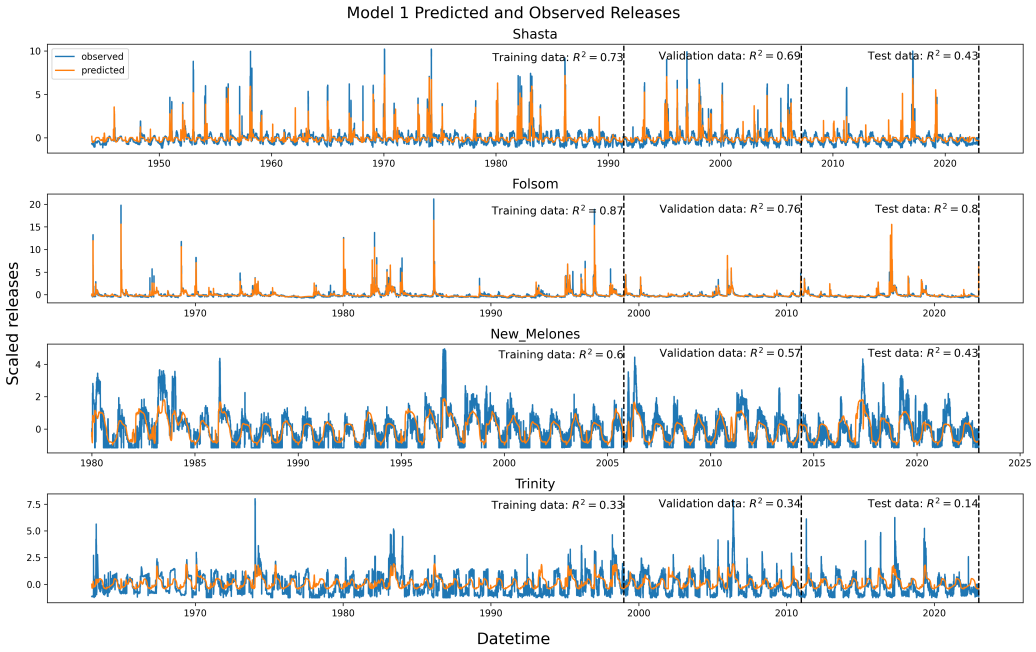

Figure 8. Timeseries plots for predicted and observed releases for Shasta, Folsom, New Melones, and Trinity Reservoirs using Model 1.

3.4 Performance Over Time

The drop in performance between the train and test period may be explained partly by changes in the reservoir operating policy during that time. We investigate this question using an "initial model" trained on the first 30 years of record and validated on the next 10 years for two reservoirs with longer records (USBR), reasonable performance, and differing degrees of regulation, specifically, Folsom and Shasta.

Figure 9 shows performance in 20-year rolling and sliding windows for these selected reservoirs. Folsom shows an initial drop in performance apparent in the rolling windows, but then stabilizes. This behavior is expected with some degree of overfitting, but in this case performance does not continue to decline. In contrast, performance for Shasta continues to decline and does not stabilize. This is also consistent with declining train, validation, and test R^2 for Shasta in Table 1, and overall in Figure 3. This supports the hypothesis that changing operating policies and/or climatic conditions are resulting in declining performance. Further, this behavior can be exacerbated by the initial overfitting, that is, the model becomes overly attuned to operations in the training period which makes performance sensitive to even slight changes. This result suggests that performance should be monitored in time, and if performance drifts the model may require re-training.

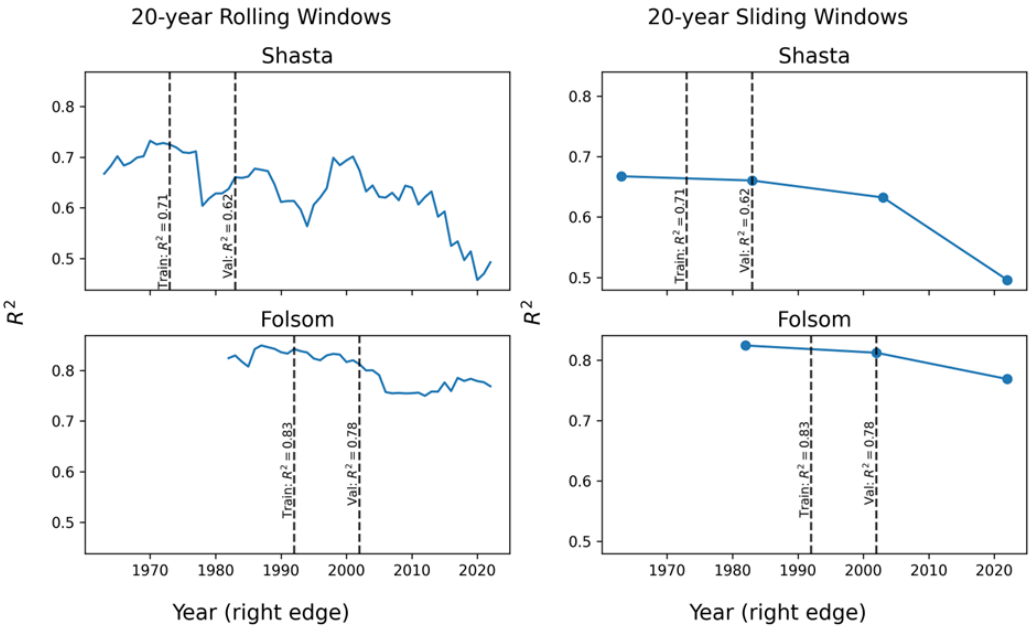

Figure 9. 20-year rolling and sliding R^2 performance for initial models trained for Shasta and Folsom reservoirs

Shasta reservoir has a higher degree of regulation and may be more sensitive to changes in policy, while Folsom has a lower degree of regulation so that releases that are more sensitive to inflows directly. We hypothesize a lower degree of regulation can make the model less prone to performance drift. Using the large ResOps sample, this conjecture is supported by a weak but statistically significant negative correlation (r=-0.24, p=0.013) between the difference in train and test R^2 and the log mean-inflow-max-storage ratio (See Supplemental Figure S4). Overall, we find that declines in performance over time due to changes in operating conditions are location specific and are only weakly explained by degree of regulation of each site. We contribute an approach to analyze the problem of performance drift using moving windows, provided that there is a long enough data record to train an initial model and track performance changes over time.

3.5 Pooled Training and Finetuning

After training models to reservoirs individually, we answer the question of whether stronger results can be achieved by training on a pool of reservoirs, and if pooling is generalizable to out of sample sites (OOS). Figure 10 compares R^2 scores on the last 20% of record for OOS reservoirs, comparing individually trained models, the pooled model, as well as finetuning the pooled model with a 5-30 year horizon of data. Recall that the training and validation periods for each fine-tuning process do not align, although we can compare performance on the same testing period. The pooled model (median score of 0.363) performs significantly worse than training individually (median score of 0.577). This result confirms that given the feature space (that is, across standardized inflows, the main operating use, and the degree of regulation), we are unable to find a strong reservoir policy that generalizes across reservoirs. This result suggests that the generalization ability of LSTMs observed in rainfall-runoff modeling may not extend to models of reservoir release policies, as these tend to be location-specific.

In finetuning or transfer learning, the pretrained (pooled) model undergoes site specific training. This leverages knowledge from training on other reservoirs in the pool and can be especially useful when site specific data is scarce. In Figure 10, we also compare finetuning performance against individually trained models using the same data horizon. We find that finetuning generally improves performance compared to individual training using the same data horizon; however, this effect shrinks for longer horizons and more available training data. For example, the median R^2 for finetuning on a 5 year horizon is 0.40 while the median R^2 for the individually trained model on a 5 year horizon is 0.32. For a 30 year horizon, the median R^2 score for both finetuning and individual training is 0.57. This suggests that while reservoir operations do not generalize across sites with pooled training, finetuning improves performance especially for data scarce scenarios.

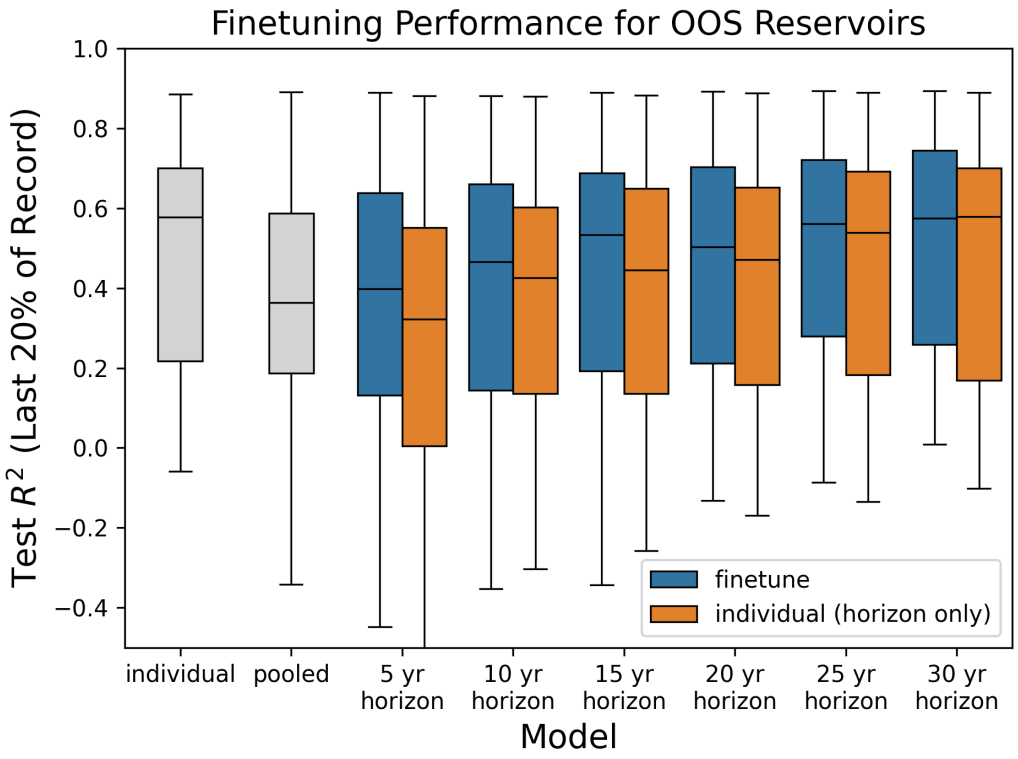

Figure 10. R^2 performance scores on the last 20% of available record for out-of-sample (OOS) reservoirs (n = 35) for individual and pooled Model 1 and finetuning on 5-30 years data

4 Discussion and Conclusion

We find that across a large sample of reservoirs, LSTMs show the ability to model reservoir release policies as well as or better than other data-driven methods. In addition, for many cases, the LSTM's internal dynamics encode a latent state consistent with reservoir storage even without using observed storage as an input. This implies that the LSTMs have learned to use mass balance to integrate past inflows and outflows to predict releases in a physically meaningful way. We demonstrate that these learned storage states can be linearly decoded from cell states in unseen data via ordinary least squares. This suggests that previous findings where LSTM rainfall-runoff models learn to represent hydrologic states (Kratzert, Herrnegger, et al., 2019; Lees et al., 2022) are reproducible for the reservoir operations problem. However, we find that the relationship be-

tween cell-storage correlation and model accuracy is weak although statistically significant. That is, for many reservoirs with strong performance, and especially for those with low degrees of regulation, there does not necessarily exist an interpretable storage representation in their cell states. In this case, future work should focus on enforcing physical constraints directly in the model process, either through alternative architectures (Fuente et al., 2024; Hoedt et al., 2021), or by penalizing physical violations in custom loss functions (Zheng et al., 2022). Although including observed storage directly as an LSTM input is an alternative and improves model performance, this is impractical for many applications requiring projections beyond a single timestep, for example, climate scenario projections. We find that performance collapses when an LSTM release model is trained with observed storage but is instead run with simulated storage. While true observed storage inputs can allow the model to exploit persistence in recent observed releases, the same cannot be done for simulated implied storage.

LSTM accuracy does not appear to be related to location, and by extension hydroclimatic factors, alone. This is consistent with feature importance analysis from previous studies which found that hydrologic variables such as precipitation and temperature are much less influential compared to inflow and storage for release predictions (Fan et al., 2023). Instead, we find that performance is strongly log-linearly related to the degree of regulation. For reservoirs with lower degrees of regulation, the release decision is more directly related to inflow and does not depend on longer-timescale accumulation, and consequently, accuracy is stronger for these models. These results are consistent with Ouyang et al. (2021) which found that LSTM runoff models perform worse on managed basins with higher degrees of regulation since these models need also to correctly account for the impact of reservoir operations on downstream discharge (Dang et al., 2020).

We also find that model performance may in some cases be unstable through time. While a decline in performance between the training and test periods is expected due to overfitting, performance may continue to decline due to distributional shift. That is, the learned operating policy under the training period may not apply to the future where operator preferences have evolved to adapt to changing hydroclimatic conditions (Mason et al., 2018). Although difficult to resolve due to the long data records needed to support training, we retroactively monitor performance over moving windows after training an initial model. We find that declines in performance are largely site specific, although they are weakly related to the degree of regulation. Looking forward, performance should be continuously monitored to account for accuracy drifts due to shifts in regulation or hydrology, and trigger retraining if necessary.

LSTM release policies may also struggle to generalize in space. Pooled training does not show an improvement on out-of-sample reservoirs, especially in the absence of finetuning. This suggests that the transferability of trained LSTMs for rainfall-runoff modeling, where additional hydrologic diversity is an asset to model performance, does not hold universally for modeling reservoir release policies (Kratzert et al., 2024). However, finetuning can help increase performance in data-scarce scenarios, although accuracy remains low. Future work can address regional pooling in contrast to continental scale pooling as done in this study; the performance of regional pooling compared to individual training remains an open question. This may be especially promising for basins that are jointly operated, since here we assume independent operations.

This study contributes a diagnostic framework for LSTM reservoir release models focusing on their physical interpretability and generalization ability in space and time. LSTMs remain a promising fully empirical and data-driven method to efficiently embed human decision making in large scale hydrological models, owing in part to the strength of the artificial neural network as a universal approximator (Hornik et al., 1989) and the increasing availability of reservoir operational data (Steyaert et al., 2022; Hou et al., 2022). However, we highlight key difficulties to LSTM modeling for this purpose—in particular, challenges in storage representation and maintaining stable performance—and pro-

vide a method to evaluate these challenges before and during coupling with larger models to ensure both accuracy and physical realism.

Data Availability Statement

All code corresponding to methods and figure generation can be found in the public repository: https://github.com/Matt2371/DL-reservoir-modeling, which is permanently archived at (Chen, 2025).

Acknowledgments
This work was supported by the U.S. National Science Foundation Graduate Research Fellowship Program, as well as NSF Grants 2041826 and 2205239. All conclusions are those of the authors.

References

Adams, L. E., Lund, J. R., Moyle, P. B., Quiñones, R. M., Herman, J. D., & O'Rear, T. A. (2017, 9). Environmental hedging: A theory and method for reconciling reservoir operations for downstream ecology and water supply. *Water Resources Research*, *53*, 7816-7831. doi: 10.1002/2016WR020128

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017, 10). The camels data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*, 5293-5313. Retrieved from `https://hess.copernicus.org/articles/21/5293/2017/` doi: 10.5194/hess-21-5293-2017

Alain, G., & Bengio, Y. (2018, 11). Understanding intermediate layers using linear classifier probes. Retrieved from `http://arxiv.org/abs/1610.01644`

Belinkov, Y. (2021, 9). Probing classifiers: Promises, shortcomings, and advances. Retrieved from `http://arxiv.org/abs/2102.12452`

Biemans, H., Haddeland, I., Kabat, P., Ludwig, F., Hutjes, R. W., Heinke, J., ... Gerten, D. (2011). Impact of reservoirs on river discharge and irrigation water supply during the 20th century. *Water Resources Research*, *47*. doi: 10.1029/2009WR008929

Boulange, J., Hanasaki, N., Yamazaki, D., & Pokhrel, Y. (2021, 12). Role of dams in reducing global flood exposure under climate change. *Nature Communications*, *12*. doi: 10.1038/s41467-020-20704-0

Chen, M. (2025). *Matt2371/dl-reservoir-modeling: first submission (v1.0.0).* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.17861203` doi: 10.5281/zenodo.17861203

Chen, M., & Herman, J. D. (2024, 4). Detection time for nonstationary reservoir system performance driven by climate and land-use change. *Journal of Water Resources Planning and Management*, *150*. doi: 10.1061/jwrmd5.wreng-6184

Coerver, H. M., Rutten, M. M., & Giesen, N. C. V. D. (2018). Deduction of reservoir operating rules for application in global hydrological models. *Hydrology and Earth System Sciences*, *22*, 831-851. doi: 10.5194/hess-22-831-2018

Dang, T. D., Chowdhury, A. F. K., & Galelli, S. (2020, 1). On the representation of water reservoir storage and operations in large-scale hydrological models: Implications on model parameterization and climate change impact assessments. *Hydrology and Earth System Sciences*, *24*, 397-416. doi: 10.5194/hess-24-397-2020

Dong, N., Guan, W., Cao, J., Zou, Y., Yang, M., Wei, J., ... Wang, H. (2023, 4). A hybrid hydrologic modelling framework with data-driven and conceptual reservoir operation schemes for reservoir impact assessment and predictions. *Journal of Hydrology*, *619*. doi: 10.1016/j.jhydrol.2023.129246

Ehsani, N., Fekete, B. M., Vörösmarty, C. J., & Tessler, Z. D. (2016, 4). A neural network based general reservoir operation scheme. *Stochastic Environmental Research and Risk Assessment*, *30*, 1151-1166. doi: 10.1007/s00477-015-1147-9

Fan, M., Zhang, L., Liu, S., Yang, T., & Lu, D. (2023, 3). Investigation of hydrometeorological influences on reservoir releases using explainable machine learning methods. *Frontiers in Water*, *5*. Retrieved from https://www.frontiersin.org/articles/10.3389/frwa.2023.1112970/full doi: 10.3389/frwa.2023.1112970

Fuente, L. A. D. L., Ehsani, M. R., Gupta, H. V., & Condon, L. E. (2024, 2). Toward interpretable lstm-based modeling of hydrological systems. *Hydrology and Earth System Sciences*, *28*, 945-971. doi: 10.5194/hess-28-945-2024

Galelli, S., Turner, S. W. D., Pokhrel, Y., Ng., J. Y., Castelletti, A., Bierkens, M. F. P., ... Biemans, H. (2025, 7). Advancing the representation of human actions in large-scale hydrological models: Challenges and future research directions. *Water Resources Research*, *61*. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039486 doi: 10.1029/2024WR039486

Gangrade, S., Lu, D., Kao, S. C., & Painter, S. L. (2022, 12). Machine learning assisted reservoir operation model for long-term water management simulation. *Journal of the American Water Resources Association*, *58*, 1592-1603. doi: 10.1111/1752-1688.13060

Giuliani, M., Lamontagne, J. R., Reed, P. M., & Castelletti, A. (2021, 12). *A state-of-the-art review of optimal reservoir control for managing conflicting demands in a changing world* (Vol. 57). John Wiley and Sons Inc. doi: 10.1029/2021WR029927

Haddeland, I., Skaugen, T., & Lettenmaier, D. P. (2006, 4). Anthropogenic impacts on continental surface water fluxes. *Geophysical Research Letters*, *33*. doi: 10.1029/2006GL026047

Hanasaki, N., Kanae, S., & Oki, T. (2006, 7). A reservoir operation scheme for global river routing models. *Journal of Hydrology*, *327*, 22-41. doi: 10.1016/j.jhydrol.2005.11.011

Hewitt, J., & Liang, P. (2019, 9). Designing and interpreting probes with control tasks. Retrieved from http://arxiv.org/abs/1909.03368

Hochreiter, S., & Schmidhuber, J. U. (1997). Long short-term memory. *Neural Computation*, *9*, 1735-1780. Retrieved from http://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf?casa_token=Styd-71DQioAAAAA:QvJW2dBxd-5ihwSumWqKKmT6VaWieXAj1b5KxTSL4OM1OO2or0Ybd4NVUKQraCEmRfAQ doi: https://doi.org/10.1162/neco.1997.9.8.1735

Hodgkins, G. A., Over, T. M., Dudley, R. W., Russell, A. M., & LaFontaine, J. H. (2024, 2). The consequences of neglecting reservoir storage in national-scale hydrologic models: An appraisal of key streamflow statistics. *Journal of the American Water Resources Association*, *60*, 110-131. doi: 10.1111/1752-1688.13161

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., ... Klambauer, G. (2021, 1). Mc-lstm: Mass-conserving lstm. *Proceedings of the 38th International Conference on Machine Learning*, *139*, 4275-4286. Retrieved from http://arxiv.org/abs/2101.05186

Hornik, K., Stinchcombe, M., & White, H. (1989, 1). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359-366. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/0893608089900208 doi: 10.1016/0893-6080(89)90020-8

Hou, J., Dijk, A. I. V., Beck, H. E., Renzullo, L. J., & Wada, Y. (2022, 7). Remotely sensed reservoir water storage dynamics (1984-2015) and the influence

of climate variability and management at a global scale. *Hydrology and Earth System Sciences*, *26*, 3785-3803. doi: 10.5194/hess-26-3785-2022

Iorga, C., & Neagoe, V.-E. (2019, 6). A deep cnn approach with transfer learning for image recognition. In *2019 11th international conference on electronics, computers and artificial intelligence (ecai)* (p. 1-6). IEEE. doi: 10.1109/ECAI46879.2019.9042173

Jordan, I. D., Sokół, P. A., & Park, I. M. (2021, 7). Gated recurrent units viewed through the lens of continuous time dynamical systems. *Frontiers in Computational Neuroscience*, *15*. doi: 10.3389/fncom.2021.678158

Kingma, D. P., & Ba, J. (2015, 12). Adam: A method for stochastic optimization. In *Proceedings of the 3rd international conference on learning representations (iclr 2015)* (p. 1-15). Retrieved from `http://arxiv.org/abs/1412.6980`

Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). Hess opinions: Never train an lstm on a single basin. *Hydrol. Earth Syst. Sci. Discuss. [preprint]*, 1-19. Retrieved from `https://doi.org/10.5194/hess-2023-275` doi: 10.5194/hess-2023-275

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., & Klambauer, G. (2019). Neuralhydrology – interpreting lstms in hydrology. In *Explainable ai: Interpreting, explaining and visualizing deep learning* (Vol. 11700, p. 347-362). doi: 10.1007/978-3-030-28954-6_19

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018, 11). Rainfall-runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, *22*, 6005-6022. doi: 10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019, 12). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*, 5089-5110. doi: 10.5194/hess-23-5089-2019

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., Bruijn, J. D., ... Dadson, S. J. (2022). Hydrological concept formation inside long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, *26*, 3079-3101. doi: 10.5194/hess-26-3079-2022

Lehner, B., Liermann, C. R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., ... Wisser, D. (2011, 11). High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. *Frontiers in Ecology and the Environment*, *9*, 494-502. Retrieved from `https://esajournals.onlinelibrary.wiley.com/doi/10.1890/100125` doi: 10.1890/100125

Li, M., Soltanolkotabi, M., & Oymak, S. (2019, 3). Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *Proceedings of machine learning research* (Vol. 108, p. 4313-4324). Retrieved from `https://proceedings.mlr.press/v108/li20j.html`

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019, 4). Linguistic knowledge and transferability of contextual representations. Retrieved from `http://arxiv.org/abs/1903.08855`

Longyang, Q., & Zeng, R. (2023, 6). A hierarchical temporal scale framework for data-driven reservoir release modeling. *Water Resources Research*, *59*. doi: 10.1029/2022WR033922

Mason, E., Giuliani, M., Castelletti, A., & Amigoni, F. (2018, 4). Identifying and modeling dynamic preference evolution in multipurpose water resources systems. *Water Resources Research*, *54*, 3162-3175. doi: 10.1002/2017WR021431

Nilsson, C., Catherine, Reidy, A., Dynesius, M., & Revenga, C. (2005). Fragmentation and flow regulation of the world's large river systems. *Science*, *308*, 405-408. Retrieved from `www.sciencemag.orgSCIENCEVOL30815APRIL2005`

Oliveira, R., & Loucks, D. P. (1997). Operating rules for multireservoir systems. *Water Resources Research*, *33*, 839-852. doi: 10.1029/96WR03745

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., & Shen, C. (2021, 8). Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy. *Journal of Hydrology*, *599*. doi: 10.1016/j.jhydrol.2021.126455

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019, 12). Pytorch: An imperative style, high-performance deep learning library. In *33rd conference on neural information processing systems (neurips 2019)* (p. 1-12). Retrieved from `http://arxiv.org/abs/1912.01703`

Pokhrel, Y. N., Hanasaki, N., Wada, Y., & Kim, H. (2016, 7). *Recent progresses in incorporating human land–water management into global land surface models toward their integration into earth system models* (Vol. 3). John Wiley and Sons Inc. doi: 10.1002/wat2.1150

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the north* (p. 15-18). Association for Computational Linguistics. doi: 10.18653/v1/N19-5004

Sen, P. K. (1968, 12). Estimates of the regression coefficient based on kendall's tau. *Journal of the American Statistical Association*, *63*, 1379-1389. Retrieved from `http://www.tandfonline.com/doi/abs/10.1080/01621459.1968.10480934` doi: 10.1080/01621459.1968.10480934

Steinschneider, S., Herman, J. D., Kucharski, J., Abellera, M., & Ruggiero, P. (2023, 1). Uncertainty decomposition to understand the influence of water systems model error in climate vulnerability assessments. *Water Resources Research*, *59*. doi: 10.1029/2022WR032349

Steyaert, J. C., Condon, L. E., Turner, S. W., & Voisin, N. (2022, 2). Resopsus, a dataset of historical reservoir operations in the contiguous united states. *Scientific Data*, *9*, 34. doi: 10.1038/s41597-022-01134-7

Tefs, A. A., Stadnyk, T. A., Koenig, K. A., Déry, S. J., MacDonald, M. K., Slota, P., ... Hamilton, M. (2021, 7). Simulating river regulation and reservoir performance in a continental-scale hydrologic model. *Environmental Modelling and Software*, *141*. doi: 10.1016/j.envsoft.2021.105025

Thompson, S. E., Sivapalan, M., Harman, C. J., Srinivasan, V., Hipsey, M. R., Reed, P., ... Blöschl, G. (2013, 12). Developing predictive insight into changing water systems: Use-inspired hydrologic science for the anthropocene. *Hydrology and Earth System Sciences*, *17*, 5013-5039. doi: 10.5194/hess-17-5013-2013

Tran, H., Zhou, T., Tan, Z., Fang, Y., & Leung, L. R. (2025, 11). Improving the prediction of daily reservoir releases over the conus using conditioned lstm. *Journal of Hydrology*, *661*. doi: 10.1016/j.jhydrol.2025.133750

Turner, S. W., Doering, K., & Voisin, N. (2020, 10). Data-driven reservoir simulation in a large-scale hydrological and water resource model. *Water Resources Research*, *56*. doi: 10.1029/2020WR027902

Turner, S. W., & Galelli, S. (2016, 2). Water supply sensitivity to climate change: An r package for implementing reservoir storage analysis in global and regional impact studies. *Environmental Modelling and Software*, *76*, 13-19. doi: 10.1016/j.envsoft.2015.11.007

Turner, S. W., Steyaert, J. C., Condon, L., & Voisin, N. (2021, 12). Water storage and release policies for all large reservoirs of conterminous united states. *Journal of Hydrology*, *603*. doi: 10.1016/j.jhydrol.2021.126843

Wang, Y. (2017, 5). A new concept using lstm neural networks for dynamic system identification. In *2017 american control conference (acc)* (p. 5324-5329). IEEE. doi: 10.23919/ACC.2017.7963782

Yang, T., Gao, X., Sorooshian, S., & Li, X. (2016, 3). Simulating california reservoir operation using the classification and regression-tree algorithm combined with

a shuffled cross-validation scheme. *Water Resources Research*, *52*, 1626-1651. doi: 10.1002/2015WR017394

Yassin, F., Razavi, S., Elshamy, M., Davison, B., Sapriza-Azuri, G., & Wheater, H. (2019, 9). Representation and improved parameterization of reservoir operation in hydrological and land-surface models. *Hydrology and Earth System Sciences*, *23*, 3735-3764. doi: 10.5194/hess-23-3735-2019

Yin, X. A., Yang, Z. F., Petts, G. E., & Kondolf, G. M. (2014, 5). A reservoir operating method for riverine ecosystem protection, reservoir sedimentation control and water supply. *Journal of Hydrology*, *512*, 379-387. doi: 10.1016/j.jhydrol.2014.02.037

Zhao, G., Gao, H., Naz, B. S., Kao, S. C., & Voisin, N. (2016, 12). Integrating a reservoir regulation scheme into a spatially distributed hydrological model. *Advances in Water Resources*, *98*, 16-31. doi: 10.1016/j.advwatres.2016.10.014

Zheng, Y., Liu, P., Cheng, L., Xie, K., Lou, W., Li, X., ... Zhang, W. (2022, 4). Extracting operation behaviors of cascade reservoirs using physics-guided long-short term memory networks. *Journal of Hydrology: Regional Studies*, *40*, 101034. Retrieved from `https://linkinghub.elsevier.com/retrieve/pii/S2214581822000477` doi: 10.1016/j.ejrh.2022.101034

Zhou, T., Nijssen, B., & Lettenmaier, D. P. (2016). The contribution of reservoirs to global land surface water storage variations*. *J. Hydrometeor.*, *17*, 309-325. Retrieved from `http://dx.doi.org/10.1175/JHM-D-15-` doi: 10.1175/JHM-D-15

Supplementary Materials

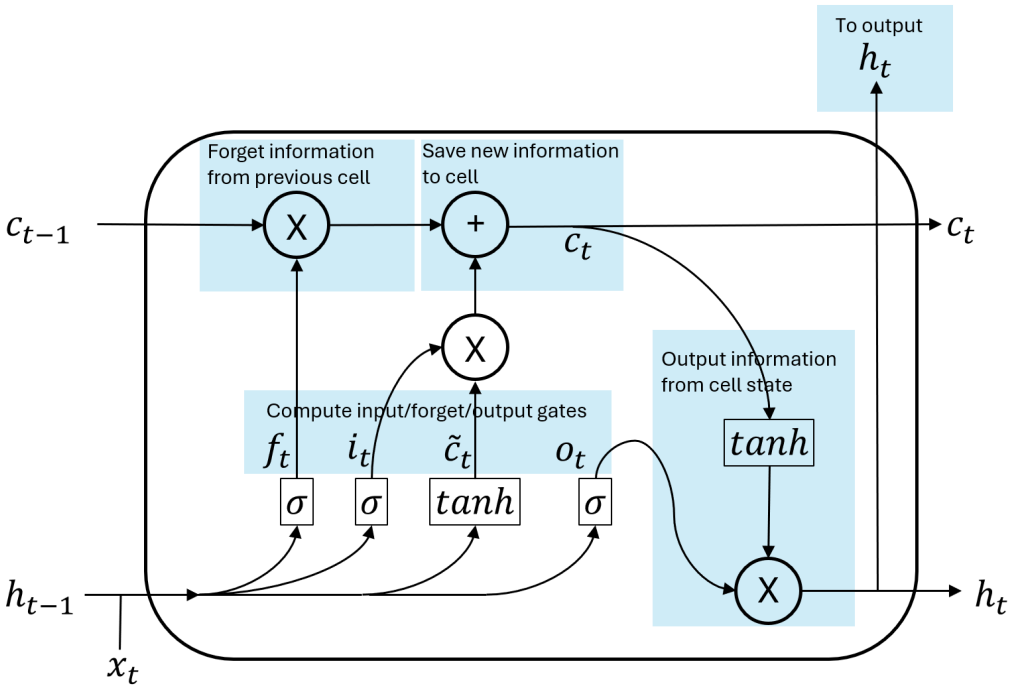

Figure S1. The LSTM architecture represented as a computational graph.

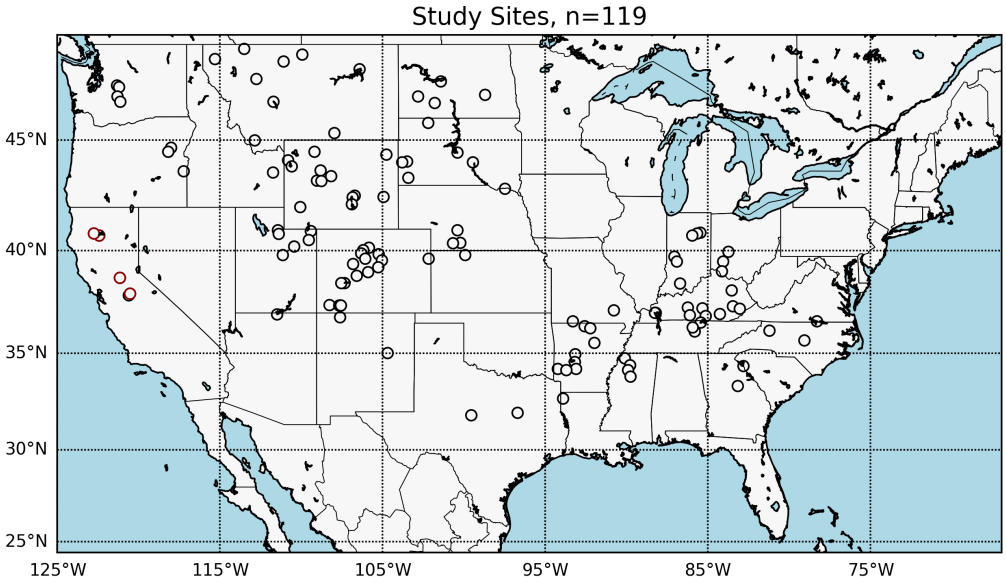

Figure S2. Map of 119 study sites across the continental US. Reservoirs from the U.S. Bureau of Reclamation are highlighted in red, otherwise data is from the ResOpsUS dataset.

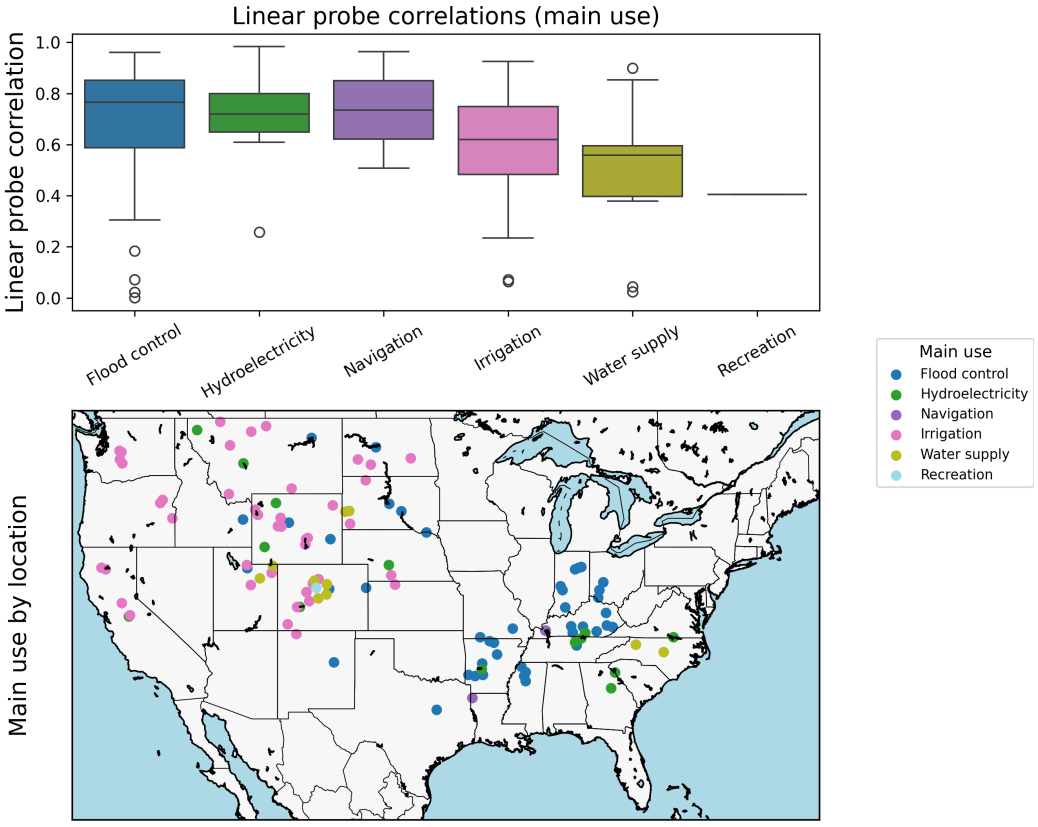

Figure S3. Linear probe correlations with observed storage in testing grouped by main use

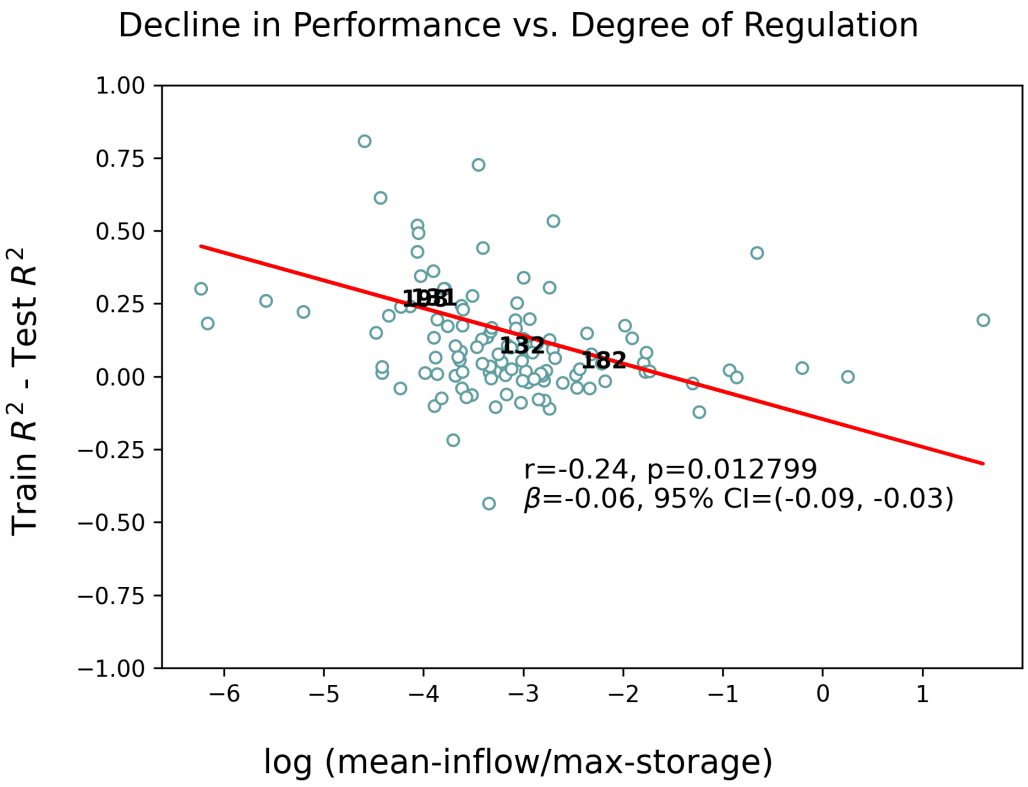

Figure S4. Decline in R^2 in training and test versus log mean-inflow-max-storage ratio.