JAWRA Journal of the American Water Resources Association

AMERICAN WATER RESOURCES ASSOCIATION JAWRA JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

TECHNICAL NOTE OPEN ACCESS

Machine Learning for a Heterogeneous Water Modeling Framework

Jonathan M. Frame[1] | Ryoko Araki[2] | Soelem Aafnan Bhuiyan[3] | Tadd Bindas[4] | Jeremy Rapp[5] | Lauren Bolotin[2,6] | Emily Deardorff[2,6] | Qiyue Liu[7] | Francisco Haces-Garcia[8] | Mochi Liao[9] | Nels Frazier[6,10] | Fred L. Ogden[10]

[1]University of Alabama, Tuscaloosa, Alabama, USA | [2]San Diego State University, San Diego, California, USA | [3]George Mason University, Fairfax, Virginia, USA | [4]Pennsylvania State University, College Station, Pennsylvania, USA | [5]Michigan State University, East Lansing, Michigan, USA | [6]Lynker, Fort Collins, Colorado, USA | [7]University of Illinois Urbana-Champaign, Urbana, Illinois, USA | [8]University of Houston, Houston, Texas, USA | [9]Duke University, Durham, North Carolina, USA | [10]NOAA/NWS Office of Water Prediction, Tuscaloosa, Alabama, USA

Correspondence: Jonathan M. Frame (jmframe@crimson.ua.edu)

Received: 25 May 2024 | **Revised:** 18 January 2025 | **Accepted:** 27 January 2025

Funding: This project was conducted as part of the National Oceanic & Atmospheric Administration (NOAA) National Water Center Innovators Program, Summer Institute. The Summer Institute was administered by the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) and the University of Alabama under the Cooperative Institute for Research to Operations in Hydrology (CIROH; NOAA Cooperative Agreement NA22NWS4320003). This research was partially supported by NOAA cooperative agreement (grant number NA19NES4320002).

Keywords: decision support systems | hydrologic modeling | model ensemble | National Water Model | open-source community model) | operational modeling | physics-informed machine learning

ABSTRACT

This technical note describes recent efforts to integrate machine learning (ML) models, specifically long short-term memory (LSTM) networks and differentiable parameter learning conceptual hydrological models (δ conceptual models), into the next-generation water resources modeling framework (Nextgen) to enhance future versions of the U.S. National Water Model (NWM). We address three specific methodology gaps of this new modeling framework: (1) assess model performance across many ungauged catchments, (2) diagnostic-based model selection, and (3) regionalization based on catchment attributes. We demonstrate that an LSTM trained on CAMELS catchments can make large-scale predictions with Nextgen across the New England region and match the average flow duration curve observed by stream gauges for streamflow with low exceedance probability (high flows), but diverges from the mean in high exceedance probability (low flows). We demonstrate improvements in peak flow predictions when using δ conceptual model, but results also suggest that performance increases may come at a cost of accurately representing hydrologic states within the conceptual model. We propose a novel approach using ML to predict the most performant mosaic modeling approach and demonstrate improved distributions of efficiency scores throughout the large sample of basins. Our findings advocate for the future development of ML capabilities within Nextgen for advancing operational hydrological modeling.

1 | Introduction

The operational U.S. National Water Model (NWM) is based on WRFHydro (Cosgrove et al. 2024), originally deployed in 2016. Since then, tremendous progress has occurred within the machine learning (ML) community in developing and operationalizing robust and reliable ML modeling approaches for hydrologic forecasting (Nearing et al. 2020; Nearing et al. 2024). In recent years, the NOAA Summer Innovator's Program, which provides a testbed for continental scale modeling (Cosgrove et al. 2024), has explored ML capabilities for future versions of the NWM (Deardorff et al. 2022; Wang et al. 2023). Johnson et al. (2023) called for heterogeneous model formulations and diagnostic model selection for large-domain hydrological modeling. NOAA's Office of Water Prediction initiated the development of the standards based, model agnostic, next-generation water resources modeling framework

https://doi.org/10.1111/1752-1688.70000

(Nextgen; Ogden et al. 2021; https://github.com/NOAA-OWP/ngen, accessed January 2025), for the purposes of coupling spatially varied models and supporting community development for continental-scale modeling (Cosgrove et al. 2024). Participating federal agencies anticipate that standards based frameworks will facilitate research contributions to water prediction, including work by students (Ogden et al. 2021; Araki et al. 2024). The purpose of this technical note is to review and synthesize student-led contributions of ML to Nextgen, primarily through the NOAA Summer Innovator's Program (Summer Institute) (SI; Bales and Flowers 2022; Bales and Flowers 2021; Bales 2019). Research themes at the SI in 2021, 2022, and 2023 were specifically designed to address the above-mentioned issues. This note concludes with recommendations for future ML development for Nextgen, which we hope bridges the gap between the fast-paced ML research and the need for robust analysis for operational water modeling.

In this technical note, we deploy ML and physics-informed ML models for streamflow prediction in Nextgen. Among the various ML models, long short-term memory (LSTM) network models for hydrological forecasting excel in terms of streamflow prediction skill and accuracy (Nearing et al. 2024), and have been included as a key modeling approach in the Nextgen framework (Frame 2022). Differentiable parameter learning conceptual hydrological models (δ conceptual models) utilize neural networks to generate static, or time-varying, parameters for process-based equations, tuning the NN through the operations of the physical equations, and has been proposed as a path for unifying ML and conceptual modeling (Shen et al. 2023). δ conceptual models have been shown to perform well in ungauged regions (Feng et al. 2023), addressing the challenge of parameter regionalization. This method works for different conceptual models of hydrological processes (Espinoza et al. 2024), and has been used for hydraulic routing (Bindas et al. 2024).

Given the hundreds of thousands (~800,000) of catchments within the CONUS scale Nextgen hydrofabric, one of the most pressing challenges is identifying the region-specific most performant modeling approach (Peckham et al. 2023). ML provides an ideal toolkit for dealing with the model selection problem, specifically high dimensional classification (Jehn et al. 2020). Representative characteristics and attributes of hydrologic catchments are needed to generalize ML and δ models, as well as identifying the region-specific model approach. Knoben et al. (2024) describe the model selection problem as determining which models provide close-to-best performance across gauged basins using KGE uncertainty, finding that just four models can cover 95% of basins, but their approach does not address ungauged basins, whereas we use machine learning to predict the most appropriate ensemble of models at locations without gauged data based on catchment attributes.

To summarize and graphically outline this technical note, Figure 1 shows (right) evaluating distributed catchments with a downstream gauge, (left) model selection based on ensemble model weights, (center right) δ conceptual modeling, and (center left) role of static catchment attributes for regionalization to ungauged basins.

2 | Methods

2.1 | Data and Models

2.1.1 | CAMELS Catchments

The experiments presented in this note utilize the Catchment Attributes and Meteorological dataset for Large Sample Studies (CAMELS; Newman et al. 2015; Addor et al. 2017) for model training, calibration, and validation. The US National Center

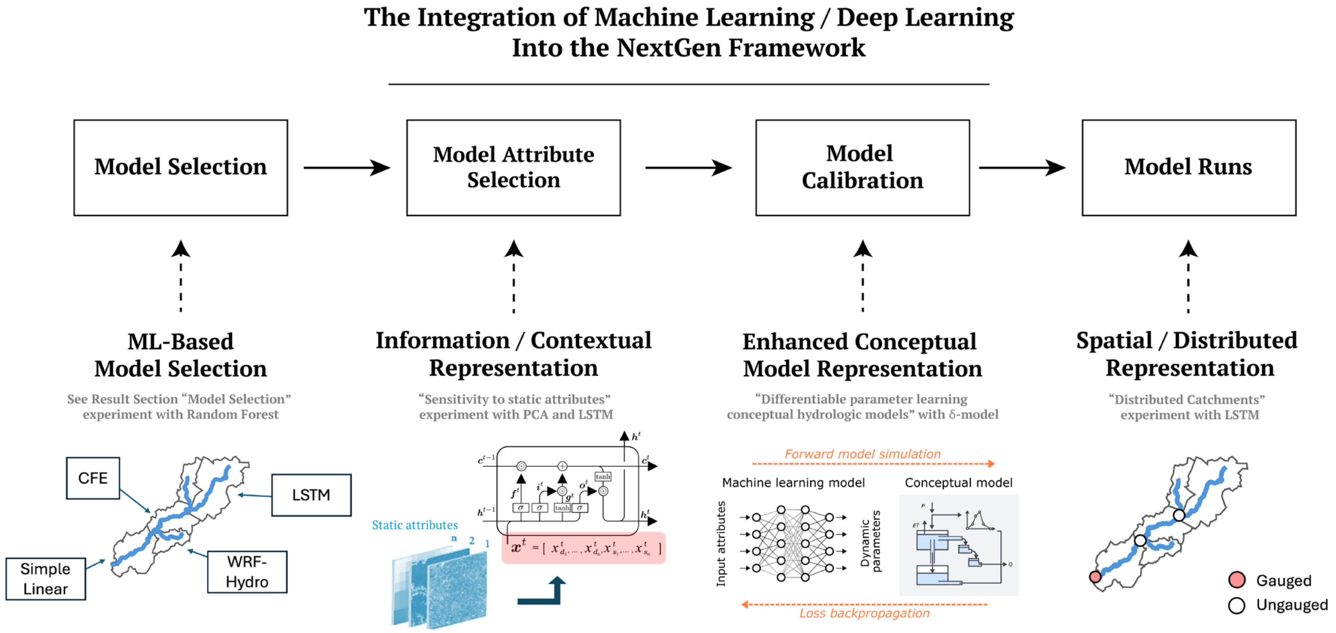
The Integration of Machine Learning / Deep Learning
Into the NextGen Framework
Model Selection
Model Attribute Selection
Model Calibration
Model Runs
ML-Based Model Selection
See Result Section "Model Selection" experiment with Random Forest
Information / Contextual Representation
"Sensitivity to static attributes" experiment with PCA and LSTM
Enhanced Conceptual Model Representation
"Differentiable parameter learning conceptual hydrologic models" with δ-model
Spatial / Distributed Representation
"Distributed Catchments" experiment with LSTM
CFE
LSTM
Simple Linear
WRF-Hydro
Static attributes
n
2
1
c^{t-1}
h^{t-1}
f^t
i^t
g^t
o^t
σ
tanh
c^t
h^t
$x^t = [\, x^t_{d_1}, \ldots, x^t_{d_n}, x^t_{s_1}, \ldots, x^t_{s_n} \,]$
Forward model simulation
Machine learning model
Conceptual model
Input attributes
Dynamic parameters
Loss backpropagation
Gauged
Ungauged

FIGURE 1 | Graphical project flowchart and summary. This figure shows four distinct, but interwoven project components including: (left) selecting the right model/s for each individual catchment, (left center) analysis of the attributes best suited for regionalization of models, (right center) utilizing differential conceptual modeling to improve model performance and overcome the burden of calibration/regionalization, and finally (right) distributed ungauged catchments contributing to a downstream gauge, highlighting the need to assess model performance in ungauged catchments.

for Atmospheric Research curated these data (NCAR; https://ral.ucar.edu/solutions/products/camels, accessed March 2020). CAMELS data include corresponding daily streamflow records from US Geological Survey (USGS) gauges, and meteorological forcing data. We use hourly streamflow records compiled by Gauch et al. (2021).

2.1.2 | National Water Model (NWM)

We include streamflow output from the 25-year (1993–2019) retrospective simulation by the NWM version 2.1 (https://docs.opendata.aws/nwm-archive/readme.html, accessed February 2024). NWM versions 3.1 and below apply Noah-MP (Niu et al. 2011) on a 1 km grid, and adds surface and lateral shallow subsurface routing on a 250 m grid, 1-D hydrologic channel routing, and a conceptual groundwater model to provide hydrological forecasts for the entire contiguous U.S., Hawaii, Puerto Rico, and portions of Alaska. A detailed background on NWM is provided by Cosgrove et al. (2024).

2.1.3 | Nextgen Hydrofabric (Hydrofabric)

This hydrofabric is used in the “Distributed catchments” experiment described below. The hydrofabric is a user-defined realization of the hydrologic features of the landscape. NOAA and the USGS have jointly developed a set of tools called “hyRefactor” that allow a user to parse the national reference hydrofabric, which was based on the NHD+ dataset. These tools allow the user to specify catchments and flow paths, a desired catchment size distribution, and stream length between adjacent channel junctions. The hydrofabric consists of four essential components as defined in the OGC WaterML 2.0, Part 3 Hydrologic Features (HY_Features) conceptual data model: catchment divides, flow paths, water bodies, and the network topology defined at connectors among the other three elements. Static attributes essential for runoff and routing modeling can be generated for each catchment as ancillary data. The hydrofabric is defined using software and data from open, reproducible hydrological analyses (Johnson 2022). The Nextgen NWM hydrofabric realizes catchments at a scale of 3–15 km^2, and imposes minimum channel reach lengths of 500 m.

2.1.4 | HydroAtlas

We use HydroAtlas static attributes (Linke et al. 2019) to test the sensitivity of large-sample ML and δ conceptual modeling for effective regionalization. There are 280 HydroAtlas attributes, which include physical, biological, and climatological basin characteristics. We processed these to correspond to the hydrofabric basins. Then, for model training, we averaged hydrofabric basin attribute values to represent the CAMELS basins.

2.1.5 | Long Short-Term Memory Network (LSTM)

LSTM is a form of recurrent neural network (RNN) distinguished by its ability to capture and maintain information about the state and dynamics of a system over extended periods. LSTMs incorporate a gated mechanism, including input, output, and forget gates, which enhance the network's ability to learn from long-term dependencies by mitigating the vanishing gradient problem encountered in other RNN architectures. This LSTM operates on a 1-h time step and is one of the initial models compatible with Nextgen (Frame 2022, Chapter 5).

2.1.6 | Conceptual Functional Equivalent (CFE) to the WRF-Hydro-Based National Water Model

The NOAA CFE model (Ogden et al. 2024) applies appropriate conceptualizations to replace the computationally expensive 250 m routing grid use in WRF-Hydro while retaining similar water balance features, providing a simplified version of the NWM (https://github.com/NOAA-OWP/cfe, accessed February 2024). The CFE model applies the same rainfall partitioning scheme and groundwater nonlinear reservoir, while simplifying vadose zone modeling. The model applies a geomorphological instantaneous unit hydrograph for surface flow routing, and a Nash Cascade for lateral subsurface flow routing from catchments. CFE eliminates the need for the NWM's detailed spatial discretization, offering a computationally efficient approach. CFE is one of the initial models compatible with Nextgen (Ogden et al. 2024; Cunha et al. 2021).

2.2 | Model Framework Constraints

The problem we are addressing is identifying the region-specific most performant multi-model ensemble approach for large-scale hydrologic forecasting. This task requires that Nextgen models conform to specific requirements for compatibility.

2.2.1 | Basic Model Interface (BMI)

Nextgen uses BMI as a common model code architecture and component-based abstraction enabling the integration of plug-and-play models in a heterogeneous modeling framework (Peckham et al. 2013; Hutton et al. 2020).

2.3 | Experimental Design

2.3.1 | Model Performance

We evaluate streamflow predictions against stream gauges using four metrics of performance. Nash–Sutcliffe Efficiency (NSE) measures the relative magnitude of the residual variance compared to the variance of observations with values ranging from $-\infty$ to one and values above zero indicate predictions more accurate than the mean observed data (Nash and Sutcliffe 1970). Normalized NSE (NNSE) adjusts NSE by normalizing the data between zero and one, making it suitable for comparing datasets across different scales. Kling–Gupta Efficiency (KGE) combines correlation, variability, and bias into one statistic to provide a holistic measure of model accuracy (Gupta et al. 2009). For NSE, NNSE, and KGE, a score of one represents optimal performance. The flow duration curve (FDC) describes the frequency and magnitude of flows, comparing the percentage of time-specific flow rates are exceeded.

2.3.2 | Distributed Catchments

The USGS stream gauge network across the United States has a median catchment area of 167 km^2. Therefore, the models in Nextgen will be deployed on catchments smaller than the gauged watersheds of which they are part. We demonstrate that an LSTM trained on the gauged catchment areas can be applied to catchments at the finer spatial resolution of the Nextgen hydrofabric.

We applied the LSTM to the HUC 01 (New England) region. Streamflow in HUC 01 results from snowmelt, rainfall, and rain-on-snow events. Groundwater contributions to streamflow are significant in this region, which is typified as postglacial landscape with thin, coarse soils overlying bedrock, and broad meadows infilled with fine sediments and organics. The hydrofabric for this region includes about 10,000 catchments with a combined drainage area of 191,020 km^2. The HUC 01 region includes 26 CAMELS basins. We compared the FDC from each hydrofabric catchment with the FDC from stream gauges corresponding to CAMELS basins.

2.3.3 | Multi-Model Ensemble Versus Best-Predicted Model (Model Selection)

The problem of selecting the most performant model for each catchment is challenging. A few automation algorithms exist to optimize model structure and parameters simultaneously (Spieler et al. 2020; Chlumsky et al. 2021). Our approach uses a random forest regressor (RFR) to predict the most performant multi-model ensemble of streamflow predictions. We evaluate our approach on basins that are not included in the training set of the RFR, assuring suitability for ungauged catchments.

To identify the most performant multi-model ensemble approach in each catchment, our framework trains RFR on a large sample of catchments to predict model performance based on catchment attributes. We trained/calibrated four models using data from 495 CAMELS basins across CONUS using. The diverse set of candidate models included: two existing BMI-compliant models, CFE (conceptual) and LSTM (deep learning), the NWM v2.1 (physics-inspired), and a simple linear (regression) model predicting streamflow directly from precipitation. These models are used to demonstrate the identification of the most performant multi-model combinations.

The normalized Nash–Sutcliffe (NNSE) from each model simulation and CAMELS catchment attributes (Newman et al. 2015; Addor et al. 2017) were used to train the RFR, to predict the comparative model performance and thus advise model selection. A weight is placed on the model prediction of a basin based on the *predicted* NNSE value of each model, for any particular basin. The resulting multi-model ensemble streamflow prediction for each basin (Q^*) is the weighted sum:

$$Q^* = \sum_{i=1}^{n} \left(\frac{NNSE_i}{\sum_{j=1}^{n} \left(NNSE_j \right)} * Q_i \right)$$

of the ensemble member's (i) streamflow (Q_i) values. We included a threshold for including a candidate model that the predicted NNSE must be within 0.2 of the highest predicted NNSE, otherwise it is dropped from the ensemble.

After the models are configured within the decision support system, an estimate of model performance for any basin can be obtained using only catchment attributes. We expect to provide users of Nextgen with a reproducible workflow for applying the most performant multi-model ensemble for any given hydrofabric catchment using only widely available static attributes, which will ensure its applicability across CONUS.

2.3.4 | Differentiable Parameter Learning Conceptual Hydrologic Models (δ Conceptual Models)

We implemented the CFE model using a differentiable (δ) platform (PyTorch) to enable gradient tracking through the model. Through gradient descent, the conceptual states of CFE trained a neural network (NN) to learn parameters, which in our case are time-varying. This integration of NN, and differentiable CFE, is termed δ-CFE, following the experiments demonstrated by Tsai et al. (2021) and Feng et al. (2023). The predicted, dynamic, parameters are: saturated vertical hydraulic conductivity (K_s) [m s^{-1}] and the primary groundwater flux coefficient (C_{gw}) [h^{-1}]. By using differentiable modeling, we can optimize streamflow output based on internal dynamics of the δ-CFE, allowing for efficient learning, and prediction, of the system state and mass balance transition at every 1-h time step. We tested our δCFE against a standard CFE to ensure our δCFE makes identical predictions to the soon-to-be operational version. We also tested to ensure that δCFE recovers synthetic parameters, verifying the embedded NN within δCFE can learn known parameters from a digital twin hydrograph. Our model fully complies with BMI standards, keeping its direct applicability to Nextgen.

2.3.5 | Sensitivity to Static Attributes

The need for static attributes as model inputs to represent true basin characteristics, for regionalization and ungauged predictions, cannot be understated. Yet, no comprehensive sensitivity test of static attributes has been presented in the literature. This is true for individual model performance for LSTM and δ conceptual models, and for model selection criteria. Deardorff (2022) demonstrated that slight differences in calculations of static attributes can severely degrade LSTM performance. Using a spatially complete static attribute dataset (HydroAtlas) eliminates the need to reproduce static attribute calculations, which decreases the potential for errors in the model inputs.

To explore added skill provided from static attributes (i.e., NOAH-MP parameters) supplied to non-process-based models readily available within the current NWM framework, we train an ensemble of LSTMs and review their performance. Secondly, we proxy more complicated subsampling of relevant hydrological and contextual states via geospatial centroid sampling of HydroAtlas attributes overlapping NWM hydrofabric generations. To this end, we generate hydrofabric features covering the spatial extents of the CAMELS basins, a common benchmarking extent used for evaluating LSTM streamflow models. Next, we expand the number of static attributes available in each hydrofabric feature by sampling the 281 hydro-environmental attributes in the HydroAtlas dataset using each hydrofabric feature catchment spatial extent. We perform a sensitivity analysis of these sets of attributes on LSTM performance to determine if there is a need for more hydro-environmental contextual information in the Nextgen framework. We aim to inform the decision of what attributes to include in the operational Nextgen hydrofabric. As the hydrofabric has a finer spatial discretization than CAMELS, our results will help determine the role of attribute heterogeneity on model performance.

3 | Results and Discussion

3.1 | Distributed Catchments

Figure 2 shows the average flow duration curve (FDC) for the example simulation period across 1000 randomly chosen HUC 01 sub-catchments over the 3-year simulation period, as compared to the observed flows from CAMELS catchments within the region. The average of predicted streamflows match up well with the higher flows (lower percent exceedance), but tend to slightly overestimate the lowest flows (highest percent exceedance). It is known that both ML and δ models struggle to predict low flows (Feng et al. 2023). In this particular case, however, we are predicting runoff on catchments much smaller (3–15 km^2) than the training set of CAMELS basins with areas up to 20,000 km^2. The minimum and maximum FDC of the HUC 01 LSTM Nextgen have a much greater variation than the observed CAMELS FDC. This is to be expected, given the large number of Nextgen hydrofabric catchments.

The simulated catchment responses come from watersheds with areas that are much smaller than those contributing to the gauging stations on which the LSTM was trained. This demonstrates that the LSTM can make effective runoff predictions at each catchment within the hydrofabric, even when it is trained on larger, gauged, watersheds. The conceptual and process-based modules within Nextgen require calibration, which will include some sort of regionalization strategy to apply to ungauged basins. The LSTM does not need to be calibrated in the same manner, and no regionalization is required, as LSTM has been shown to make predictions in ungauged basins with median NSE of 0.69 (Kratzert et al. 2019).

In the Nextgen framework, a routing scheme is required to combine the runoff through a network of smaller catchments, effectively creating a spatially distributed LSTM model. There is a lingering question of whether or not the LSTM model should be trained in a manner that includes this routing scheme. A similar approach was taken by Yu et al. (2024) and Yang et al. (2024), further demonstrating the suitability of LSTM with routing for a spatially distributed simulation; however, these implementations operate on a daily time step, whereas our results are simulated with an hourly timestep. Vrugt (2024) advocates for distribution-based calibration/training models and evaluation. Training the LSTM on the flow duration curve itself, rather than a squared error metric, could be more suitable for this type of operational environment, including this distribution-based evaluation.

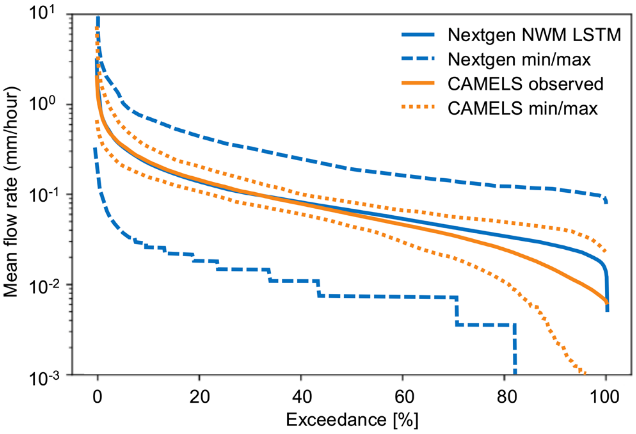

FIGURE 2 | Average, maximum, and minimum flow duration curves for HUC 01, New England, from streamflow observation and model predictions using the LSTM module within the Next Generation U.S. National Water Model (Nextgen).

3.2 | Model Selection

Figure 3 illustrates the distributions of model performance in the form of the cumulative distribution (top) and box plots (bottom) of the performance metric NNSE for the LSTM (predictions in ungauged basins), CFE, NWM v. 2.1, and a linear (regression) model. The LSTM results are from basins withheld from the training set, while the other three models were calibrated/trained on the individual basin being evaluated. Also included in the figure are performance distributions of the most performing model of the four (referred to as the best actual model; BAM), the best model predicted by the RF (BPM), and the multi-model ensemble with weights predicted by the RF (i.e., weighted ensemble; WE). These results are meant to demonstrate the efficacy of model selection, not to demonstrate the superiority of any one particular model.

For catchments with NNSE values above approximately 0.6, the performance of the WE predictions closely matches that of the best actual model. Interestingly, the multi-model ensemble streamflow prediction remains the most effective approach for catchments with NNSE values below 0.6, suggesting its robustness across a range of hydrologic conditions. Notably, the multi-model ensemble generally outperforms the LSTM model, which stands as the most effective single model in the study.

Employing machine learning (ML) to determine the optimal weights for a multi-model ensemble streamflow prediction could ease the calibration process for hydrologic models across diverse regions and catchments. However, this approach necessitates running all candidate models with predicted ensemble weights significantly greater than zero across each appropriate catchment, which might offset the computational efficiencies gained, especially in scenarios where computationally streamlined models like CFE could be directly applied. This trade-off highlights a crucial consideration in the balance between achieving

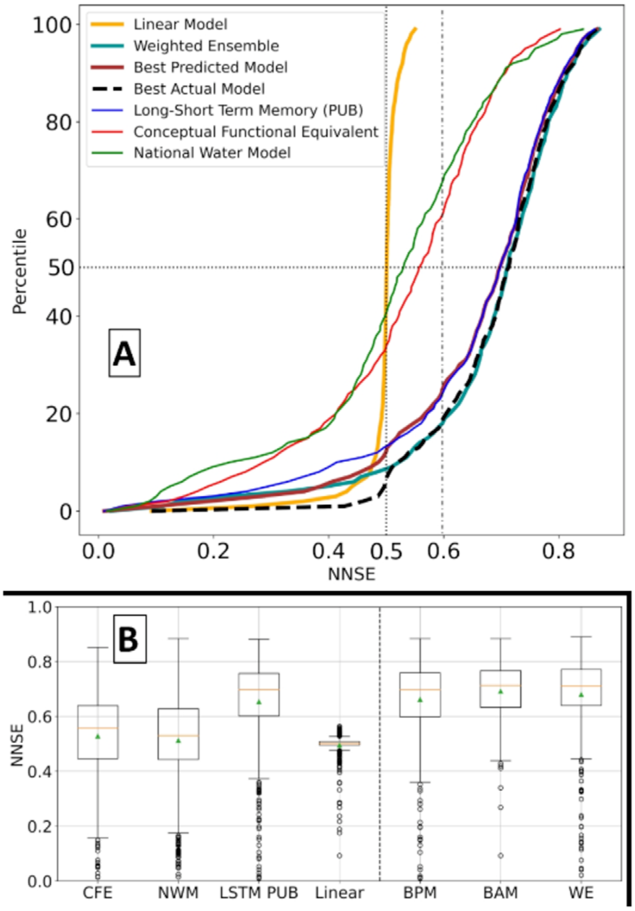

FIGURE 3 | Cumulative distribution function (top, A) and boxplot (bottom, B) of all individual models, the best actual model for each watershed, the best predicted model for each watershed by the random forest regressor (RFR), and ensemble models using weights derived from RFR model performance predictions.

optimal model performance and maintaining computational efficiency in large-scale hydrologic modeling efforts.

3.3 | Differentiable Parameter Learning Conceptual Hydrologic Models

Figure 4 presents a hydrograph from the δCFE model (middle), along with associated precipitation and temperature data (top), and the time series for dynamic parameters C_{gw} (the second from the bottom) and K_s (bottom). The hydrograph demonstrates that δCFE closely matches the observed peak runoff events and shows a sharper recession curve compared to its calibrated counterpart, indicating a distinct response in the δ model's simulation of runoff processes.

The δCFE matching of observed peaks is the result of the dynamic parameters adjusting to release more water than usual when a stormflow-generating precipitation event occurs. The sharper recession curve following peak events suggests an accelerated drainage effect caused by the model's depleted state values. Similar behavior has been observed in the δ version of another conceptual model called the Simple Hydrologic Model (Espinoza et al. 2024). While this feature allows δCFE to align closely with observed peaks, it may result in an underestimation of the antecedent storage state for the subsequent rainfall events. Song et al. (2024) proposed a capillary rise term to allow more vertical movement of water to fill in these depleted states as a potential solution that improves low-flow predictions. The predicted time-varying K_s is smaller than the original value, and the time-varying C_{gw} is larger. The effect of greater equifinality is potentially much larger with time-varying parameters. Despite the advantages of employing differentiable parameter learning for enhanced model performance, there is a risk that such δ conceptual models may misrepresent physical states and processes of hydrologic systems, yet be interpreted as physically representative of the hydrological system.

3.4 | Sensitivity to Static Attributes

Figure 5 depicts the variation in model performance, measured by the Nash–Sutcliffe Efficiency (NSE), in predicting hydrologic conditions for ungauged basins under increasing number of static attributes. The color gradient from blue to green in the figure indicates an increasing count of static attributes sampled from the HydroAtlas, revealing how these attributes impact model accuracy. Here, we incrementally added static attributes from a principal component analysis (PCA) of the HydroAtlas attributes. The best median NSE score of 0.62 is made with 10 PCA attributes. That median NSE score is notably lower than the LSTM PUB predictions (0.69) published by Kratzert et al. (2019). This discrepancy may come from the different training and test splits, the model hyperparameters, or the selected test basins. Our purpose here is to assess the relative performance given the number of static attributes and the source of data.

To compare against the PCA attributes, we selected any 10 random attributes for our model. We found that selecting any 10 random HydroAtlas attributes results in median, 75th, and 25th percentile NSE values of 0.61, 0.71, and 0.43, respectively. The optional number of HydroAtlas PCA attributes (10) results in median, 75th, and 25th percentile NSE values of 0.62, 0.72, and 0.43, respectively. This indicates that the PCA preprocessing of static attributes contributes nothing of value for model predictability. However, this analysis does help us understand the requirements of static attributes for model performance. Model accuracy improves as the number of HydroAtlas attributes increases to 10; beyond this point, adding more attributes lowers performance for this particular LSTM figuration. It is possible that increasing the number of static attributes degrades the performance because the LSTM does not have sufficient complexity in the parameter space to learn from each variable, or because there is not enough diversity in the training data to warrant the number of static attributes.

Static attributes from other sources were also tested. The analysis shows that calibrated parameters from the NOAH-MP model fail to provide useful information for LSTM predictions.

4 | Conclusion and Recommendations

The standards based, model-agnostic Nextgen is structured to integrate ML predictions. Incorporating LSTM and δ conceptual models within its diverse modeling environment

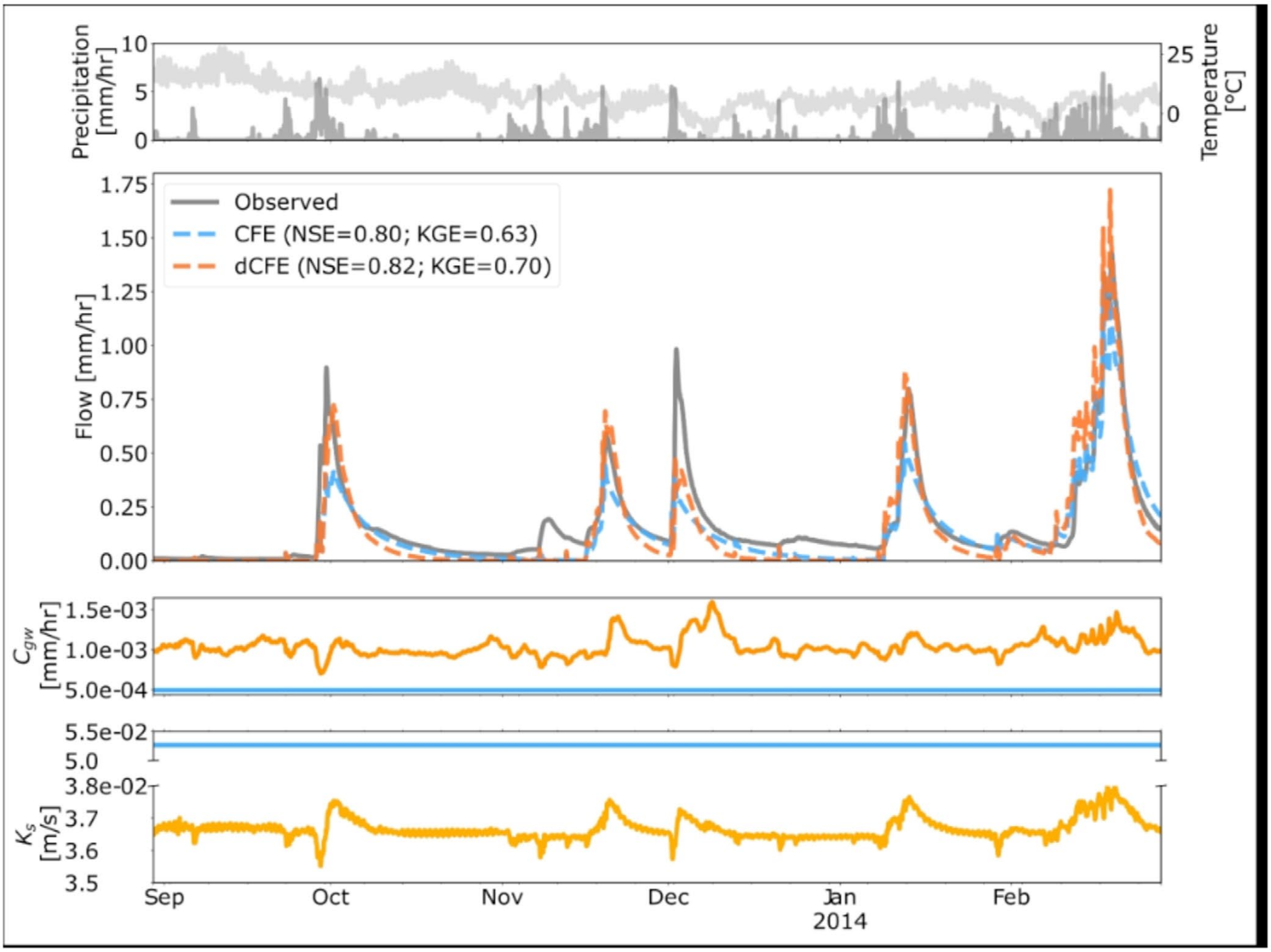

FIGURE 4 | Time series from δCFE for a Nehalem river, OR (USGS gauge 13,401,000), including the input forcings (top), hydrographs (middle), and dynamic parameters used in the simulation (bottom two).

showcases the framework's flexibility and underscores the potential for ML-based forecasting improvements to the operational NWM. The complexities of hydrologic variability and model selection will benefit from the implementation of ML in the framework.

We recommend the continued development of LSTM as the benchmark model for hydrologic predictions within Nextgen, due to their proven adaptability and accuracy in streamflow forecasting, particularly in ungauged basins. Though the LSTM predictions tend to overestimate low-flow conditions, this is likely a result of the mismatch of the training catchment scale and the simulation catchment scale, which may be addressed with distributed-catchment training. Concurrently, the continued development of δ models should be pursued for their unique strength in regionalization and enhancing model performance across diverse hydrologic conditions. However, with δ models, it is crucial to examine potential degradation of hydrologic states that may arise from overly focusing on streamflow prediction accuracy. To address this, we recommend simultaneous development of pure deep-learning models capable of predicting a broader array of hydrologic variables including surface storage and subterranean states. These models could offer a more holistic view of watershed dynamics, potentially mitigating the risk of oversimplification and ensuring a more comprehensive understanding and modeling of hydrologic processes. Looking forward, a reconceptualization of hydrologic processes to better utilize time-varying parameters may be a path to identifying nonlinear, hysteretic and scale-dependent hydrologic process representations, finding the Holy Grail of scientific hydrology (Beven 2006).

The results shown in the Model Selection section suggest that the problem of finding the most performant model in a particular basin can be approached using a heterogeneous model ensemble. While our results are consistent with previous studies that struggled to identify an “optimal” model for each individual basin with machine learning (e.g., Nearing et al. 2024), we showed that machine learning can offer effective strategies to combine multiple models, even in ungauged basins.

Immediate future progress should expand δCFE to include more time-varying parameters and performance assessment in ungauged basins compared to standard regionalization methods. Such advances could be coupled with the model selection framework, which would inform the need for regionalization strategies, and directly compare pure deep learning, pure process-based models, and somewhat in-between δ conceptual

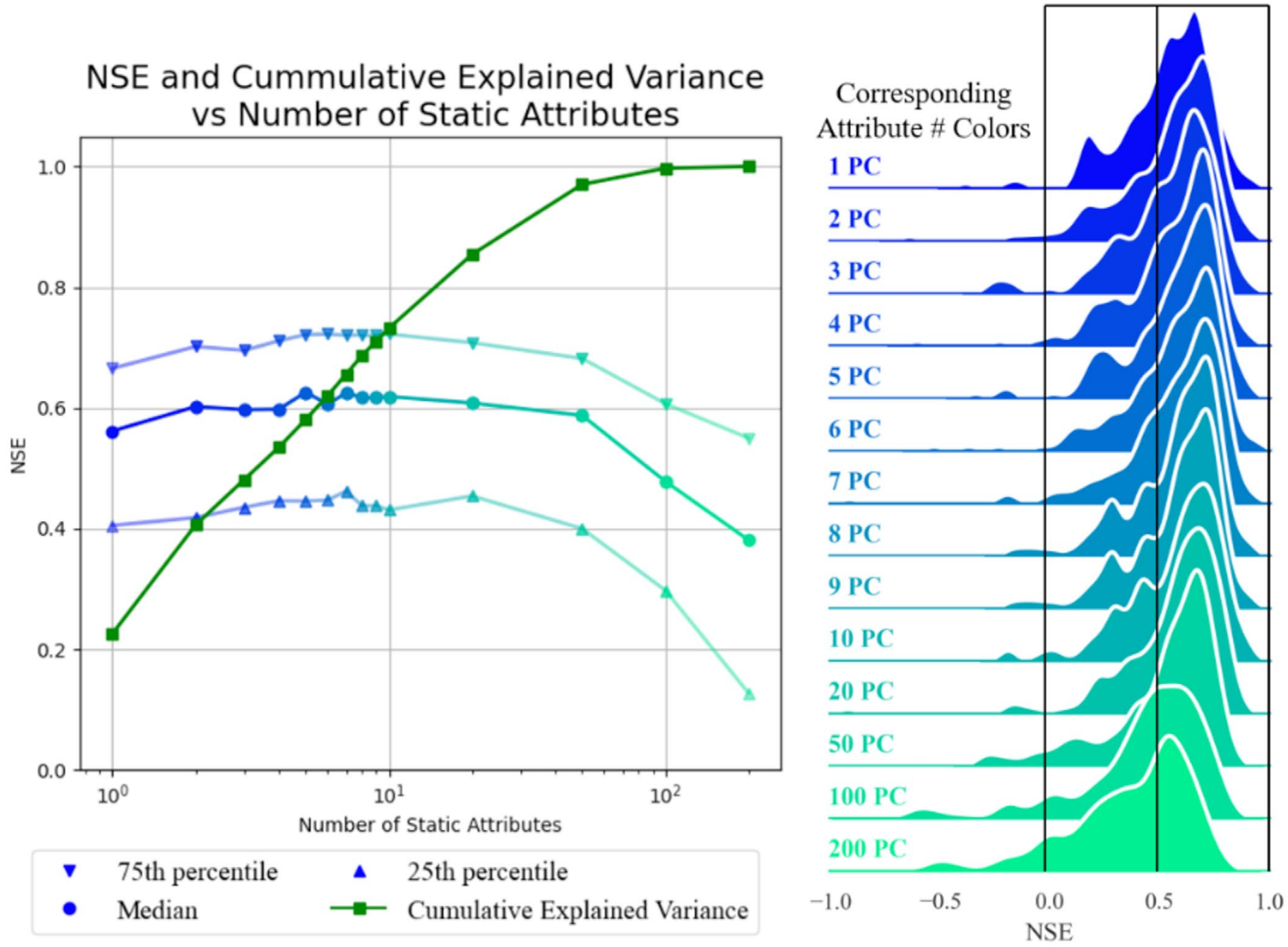

FIGURE 5 | Distributions of LSTM model performance across 131 "ungauged" basins, with different sets of static catchment attributes. Left: Model performance given the number of static catchment attributes. Right: Distributions of performance for different static attribute sets.

models. A continued attribute analysis would inform and refine the criteria for model selection, ensuring that the choice of models is not only based on current capabilities but also on a nuanced understanding of how different catchment attributes influence model accuracy and reliability.

A potential limitation within the current Nextgen software architecture, in terms of ML, is its inability to train ML models (it can only make inferences with previously trained models). As it stands, ML models destined for Nextgen must be trained using external tools, with their weights saved and then imported into Nextgen. This external training requirement introduces an additional step that could potentially slow down the iterative process of model development and operationalization.

This technical note is timely given the current surge in research and development activities for Nextgen, where machine learning (ML) is becoming a focal point of interest. We underscore that ML integration into hydrologic modeling is not a novel pursuit within our community; rather, it is a path we have collectively been navigating. We aim to ensure that our findings and methodologies contribute meaningfully to the transition from research to operational advancements. By sharing our experiences and insights, we hope to guide and enrich the ongoing efforts to harness ML's full potential in improving hydrologic forecasting and model selection within Nextgen, and more broadly research to operations in hydrology (Burian et al. 2023).

Author Contributions

Jonathan M. Frame: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – original draft, writing – review and editing. **Ryoko Araki:** conceptualization, data curation, formal analysis, investigation, software, visualization, writing – review and editing. **Soelem Aafnan Bhuiyan:** conceptualization, data curation, formal analysis, software, writing – review and editing. **Tadd Bindas:** conceptualization, data curation, formal analysis, investigation, methodology, software, validation, writing – review and editing. **Jeremy Rapp:** conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – review and editing. **Lauren Bolotin:** conceptualization, formal analysis, investigation, writing – review and editing. **Emily Deardorff:** conceptualization, formal analysis, investigation, writing – review and editing. **Qiyue Liu:** conceptualization, data curation, formal analysis, investigation, software, writing – review and editing. **Francisco Haces-Garcia:** conceptualization, data curation, formal analysis, investigation, software, visualization, writing – review and editing. **Mochi Liao:** conceptualization, formal analysis, investigation, software, writing – review and editing. **Nels Frazier:** data curation, software, visualization. **Fred L. Ogden:** conceptualization, formal analysis, investigation, methodology, project administration, resources, supervision, writing – review and editing.

Acknowledgments

This project was conducted as part of the National Oceanic & Atmospheric Administration (NOAA) National Water Center Innovators Program, Summer Institute. The Summer Institute was administered by the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) and the University of Alabama under the Cooperative Institute for Research to Operations in Hydrology (CIROH; NOAA Cooperative Agreement NA22NWS4320003). This research was partially supported by NOAA cooperative agreement (grant number NA19NES4320002). The research results presented herein represent the personal opinion of the authors, and not the official policy of the U.S. Department of Commerce or the National Oceanic and Atmospheric Administration.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All data and code used in this paper are publicly available in the following locations: U.S. National Water Model: https://docs.opendata.aws/nwm-archive/readme.html; HydroAtlas: https://www.hydrosheds.org/hydroatlas; CAMELS data: https://ral.ucar.edu/solutions/products/camels; LSTM for Nextgen: https://github.com/NOAA-OWP/lstm; CFE: https://github.com/NWC-CUAHSI-Summer-Institute/cfe_py; dCFE: https://github.com/NWC-CUAHSI-Summer-Institute/dCFE; Static attribute PCA: https://github.com/NWC-CUAHSI-Summer-Institute/attribute_pca; Model selection: https://github.com/NWC-CUAHSI-Summer-Institute/model-selection.

References

Addor, N., A. J. Newman, N. Mizukami, and M. P. Clark. 2017. "The CAMELS Data Set: Catchment Attributes and Meteorology for Large-Sample Studies." *Hydrology and Earth System Sciences* 21, no. 10: 5293–5313. https://doi.org/10.5194/hess-21-5293-2017.

Araki, R., F. L. Ogden, and H. K. McMillan. 2024. "Testing Performance Measures to Reproduce Streamflow and Soil Moisture Seasonality in the Conceptual-Functional Equivalent (CFE) Model." In Review for *Journal of the American Water Resources Association*.

Bales, J. 2019. "Featured Collection Introduction: National Water Model II." *Journal of the American Water Resources Association* 55, no. 4: 938–939. https://doi.org/10.1111/1752-1688.12786.

Bales, J., and T. Flowers. 2021. "Featured Collection Introduction: National Water Model III." *Journal of the American Water Resources Association* 57, no. 2: 205–208. https://doi.org/10.1111/1752-1688.12913.

Bales, J., and T. Flowers. 2022. "Featured Collection Introduction: National Water Model IV." *Journal of the American Water Resources Association* 58, no. 1: 1–4. https://doi.org/10.1111/1752-1688.13001.

Beven, K. 2006. "Searching for the Holy Grail of Scientific Hydrology: Qt=(S, R, Δt)A as Closure." *Hydrology and Earth System Sciences* 10: 609–618. https://doi.org/10.5194/hess-10-609-2006.

Bindas, T., W. P. Tsai, J. Liu, et al. 2024. "Improving River Routing Using a Differentiable Muskingum-Cunge Model and Physics-Informed Machine Learning." *Water Resources Research* 60, no. 1: e2023WR035337. https://doi.org/10.1029/2023WR035337.

Burian, S. J., T. M. Graziano, E. P. Clark, et al. 2023. "CIROH Implements a Research + Operations Paradigm at the Intersection of Hydrologic, Computer, Data, and Decision Sciences." AGU Fall Meeting 2023, San Francisco, CA, 11-15 December 2023, Hydrology / Next-Generation Water Resources Modeling: Collaborative Development at the Intersection of Domain, Computer, Data, and Decision Sciences I, Poster No. 179, id. H31V-179.

Chlumsky, R., J. Mai, J. R. Craig, and B. A. Tolson. 2021. "Simultaneous Calibration of Hydrologic Model Structure and Parameters Using a Blended Model." *Water Resources Research* 57, no. 5: e2020WR029229. https://doi.org/10.1029/2020WR029229.

Cosgrove, B., D. Gochis, T. Flowers, et al. 2024. "NOAA's National Water Model: Advancing Operational Hydrology Through Continental-Scale Modeling." *JAWRA Journal of the American Water Resources Association* 60: 247–272. https://doi.org/10.1111/1752-1688.13184.

Cunha, L., K. Jennings, A. Wood, et al. 2021. "Next Generation National Water Model: Strategy and Preliminary Performance of Initial Model Formulations." Paper presented at the AGU Fall Meeting 2021, New Orleans, LA, December 13–17, AGU Fall Meeting Abstracts, id. H54G-04.

Deardorff, E. 2022. "Benchmarking the Performance of Deep Learning Streamflow Models and Finetuning Methods Against the National Water Model in Small, Semi-Arid Watersheds Using Hydrologic Signatures." https://digitalcollections.sdsu.edu/do/f657d252-6989-44ff-9561-dac40d55106b&sa=D&source=docs&ust=1710913696050628&usg=AOvVaw0AobMPdEhvc1pBA-uLUKk_.

Deardorff, E., A. Modaresi, J. Bales, and T. Flowers. 2022. "National Water Center Innovators Program Summer Institute Report 2022."

Espinoza, E. A., R. Loritz, M. Á. Chaves, N. Bäuerle, and U. Ehret. 2024. "To Bucket or Not to Bucket? Analyzing the Performance and Interpretability of Hybrid Hydrological Models With Dynamic Parameterization." *Hydrology and Earth System Sciences* 28, no. 12: 2705–2719. https://doi.org/10.5194/hess-28-2705-2024.

Feng, D., H. Beck, K. Lawson, and C. Shen. 2023. "The Suitability of Differentiable, Physics-Informed Machine Learning Hydrologic Models for Ungauged Regions and Climate Change Impact Assessment." *Hydrology and Earth System Sciences* 27: 2357–2373. https://doi.org/10.5194/hess-27-2357-2023.

Frame, J. M. 2022. *Deep Learning for Operational Streamflow Forecasts: A Long Short-Term Memory Network Rainfall-Runoff Module for the U.S. National Water Model*. ProQuest Dissertations Publishing.

Gauch, M., F. Kratzert, D. Klotz, G. Nearing, J. Lin, and S. Hochreiter. 2021. "Rainfall-Runoff Prediction at Multiple Timescales With a Single Long Short-Term Memory Network." *Hydrology and Earth System Sciences* 25, no. 4: 2045–2062. https://doi.org/10.5194/hess-25-2045-2021.

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez. 2009. "Decomposition of the Mean Squared Error and Nse Performance Criteria: Implications for Improving Hydrological Modelling." *Journal of Hydrology* 377, no. 1–2: 80–91.

Hutton, E. W. H., M. D. Piper, and G. E. Tucker. 2020. "The Basic Model Interface 2.0: A Standard Interface for Coupling Numerical Models in the Geosciences." *Journal of Open Source Software* 5, no. 51: 2317. https://doi.org/10.21105/joss.02317.

Jehn, F. U., K. Bestian, L. Breuer, P. Kraft, and T. Houska. 2020. "Using Hydrological and Climatic Catchment Clusters to Explore Drivers of Catchment Behavior." *Hydrology and Earth System Sciences* 24, no. 3: 1081–1100. https://doi.org/10.5194/hess-24-1081-2020.

Johnson, J. M. 2022. "National Hydrologic Geospatial Fabric (hydrofabric) for the Next Generation (Nextgen) Hydrologic Modeling Framework." HydroShare. http://www.hydroshare.org/resource/129787b468aa4d55ace7b124ed27dbde.

Johnson, J. M., S. Fang, A. Sankarasubramanian, et al. 2023. "Comprehensive Analysis of the NOAA National Water Model: A Call for Heterogeneous Formulations and Diagnostic Model Selection." *Journal* *of Geophysical Research: Atmospheres* 128, no. 24: e2023JD038534. https://doi.org/10.1029/2023JD038534.

Knoben, W. J. M., A. Raman, G. J. Gründemann, et al. 2024. "Technical Note: How Many Models Do We Need to Simulate Hydrologic Processes Across Large Geographical Domains?" *Hydrology and Earth System Sciences Discussions*. https://doi.org/10.5194/hess-2024-279.

Kratzert, F., D. Klotz, M. Herrnegger, A. K. Sampson, S. Hochreiter, and G. S. Nearing. 2019. "Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning." *Water Resources Research* 55, no. 12: 2019WR026065. https://doi.org/10.1029/2019WR026065.

Linke, S., B. Lehner, C. Ouellet Dallaire, et al. 2019. "Global Hydro-Environmental Sub-Basin and River Reach Characteristics at High Spatial Resolution." *Scientific Data* 6: 283. https://doi.org/10.1038/s41597-019-0300-6.

Nash, J. E., and J. V. Sutcliffe. 1970. "River Flow Forecasting Through Conceptual Models Part I—A Discussion of Principles." *Journal of Hydrology* 10, no. 3: 282–290.

Nearing, G., D. Cohen, V. Dube, et al. 2024. "Global Prediction of Extreme Floods in Ungauged Watersheds." *Nature* 627, no. 8004: 559–563. https://doi.org/10.1038/s41586-024-07145-1.

Nearing, G. S., F. Kratzert, A. K. Sampson, et al. 2020. "What Role Does Hydrological Science Play in the Age of Machine Learning?" *Water Resources Research* 57: e2020WR028091. https://doi.org/10.1029/2020wr028091.

Newman, A. J., M. P. Clark, K. Sampson, et al. 2015. "Development of a Large-Sample Watershed-Scale Hydrometeorological Data Set for the Contiguous USA: Data Set Characteristics and Assessment of Regional Variability in Hydrologic Model Performance." *Hydrology and Earth System Sciences* 19, no. 1: 209–223. https://doi.org/10.5194/hess-19-209-2015.

Niu, G.-Y., Z.-L. Yang, K. E. Mitchell, et al. 2011. "The Community Noah Land Surface Model With Multiparameterization Options (Noah-MP): 1. Model Description and Evaluation With Local-Scale Measurements." *Journal of Geophysical Research* 116, no. D12: 1381–1419. https://doi.org/10.1029/2010jd015139.

Ogden, F. L., A. B. Avant, R. B. Bartel, et al. 2021. "The Next Generation Water Resources Modeling Framework: Open Source, Standards Based, Community Accessible, Model Interoperability for Large Scale Water Prediction." American Geophysical Union, Fall Meeting, H43D-01. American Geophysical Union.

Ogden, F. L., X. Feng, A. J. Khattak, and P. LaFollette. 2024. "The Conceptual Functional Equivalent to the WRF-Hydro Based NOAA/NWS National Water Model (Version 3.1 and Earlier)." NOAA Technical report 57.

Peckham, S. D., E. W. Hutton, and B. Norris. 2013. "A Component-Based Approach to Integrated Modeling in the Geosciences: The Design of CSDMS." *Computers & Geosciences* 53: 3–12. https://doi.org/10.1016/j.cageo.2012.04.002.

Peckham, S. D., K. Jennings, J. L. Garrett, et al. 2023. "Developing a River Basin Repository for the Next Generation Water Resources Modeling Framework." Poster Presented at the American Geophysical Union Fall Meeting, Poster Hall A-C - South (Exhibition Level, South, MC), December 13, 2023, 08:30–12:50. Abstract H31V-1801.

Shen, C., A. P. Appling, P. Gentine, et al. 2023. "Differentiable Modelling to Unify Machine Learning and Physical Models for Geosciences." *Nature Reviews Earth and Environment* 4: 552–567. https://doi.org/10.1038/s43017-023-00450-9.

Song, Y., W. J. M. Knoben, M. P. Clark, et al. 2024. "When Ancient Numerical Demons Meet Physics-Informed Machine Learning: Adjoint-Based Gradients for Implicit Differentiable Modeling." *Hydrology and Earth System Sciences* 28: 3051–3077. https://doi.org/10.5194/hess-28-3051-2024.

Spieler, D., J. Mai, J. R. Craig, B. A. Tolson, and N. Schütze. 2020. "Automatic Model Structure Identification for Conceptual Hydrologic Models." *Water Resources Research* 56, no. 9: e2019WR027009. https://doi.org/10.1029/2019WR027009.

Tsai, W.-P., D. Feng, M. Pan, et al. 2021. "From Calibration to Parameter Learning: Harnessing the Scaling Effects of Big Data in Geoscientific Modeling." *Nature Communications* 12, no. 1: 5988. https://doi.org/10.1038/s41467-021-26107-z.

Vrugt, J. 2024. "Distribution-Based Model Evaluation and Diagnostics: Elicitability, Propriety, and Scoring Rules for Hydrograph Functionals." *Water Resources Research* 91. e2023WR036710. https://doi.org/10.1029/2023WR036710.

Wang, M., E. Hamidi, and D. Mccay. 2023. "National Water Center Innovators Program Summer Institute Report 2023." https://www.cuahsi.org/uploads/library/doc/SI2023_Report.pdf.

Yang, Y., D. Feng, H. Beck, et al. 2024. "Global Daily Discharge Estimation Based on Grid-Scale Long Short-Term Memory (LSTM) Model and River Routing." Presented at the AGU Fall Meeting 2024, Washington, DC.

Yu, Q., B. A. Tolson, H. Shen, M. Han, J. Mai, and J. Lin. 2024. "Enhancing Long Short-Term Memory (LSTM)-Based Streamflow Prediction With a Spatially Distributed Approach." *Hydrology and Earth System Sciences* 28, no. 9: 2107–2122. https://doi.org/10.5194/hess-28-2107-2024.