International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 04, No. 02, page 185 - 196
Submitted 20/6/2026; Accepted 26/7/2026; Published 1/8/2026

Literature Review of Hybrid Neural–Physics Flood-Relevant Streamflow Forecasting with Cross-Event and Cross-Basin Generalization

Yanan Elliot Moko
Department of Computer Science and Informatics, Federal University, Otuoke Bayelsa State, NIGERIA.
e-mail: yalemoko776@gmail.com

Corresponding Autor: Yanan Elliot Moko

Abstract
Reliable flood-relevant streamflow forecasting remains difficult at operational scale because physically based forecasting systems provide national coverage but cannot be directly verified at the overwhelming majority of ungauged river reaches. This study reconstructs and strengthens a hybrid neural–physics framework in which the National Water Model (NWM) supplies the physically constrained baseline forecast and an attribute-conditioned deep neural network learns forecast-error magnitude from gauged basins, then transfers that knowledge to ungauged basins. The emphasis is not on replacing process-based hydrology with a black-box predictor, but on improving forecast reliability under cross-event and cross-basin generalization. The empirical setting consists of 979 archived daily long-range forecasts, 30 forecast lead days, 16-member NWM ensembles, and 389 quality-controlled U.S. Geological Survey gauging stations across Alabama and Georgia, complemented by static watershed descriptors and gridded soil-moisture diagnostics. Exploratory analysis shows that forecast skill is strongly conditioned by basin properties: larger and forested watersheds are predicted more accurately than smaller and urban watersheds, while urban basins show severe positive bias, weaker anomaly correlation, and degraded hydrograph timing. These diagnostics motivate an error-learning postprocessor that maps NWM forecast magnitude and basin attributes to a non-negative error estimate and converts the deterministic physics forecast into a corrected predictive interval. Under a temporally disjoint evaluation intended to test cross-event transfer, observation coverage increases from 0.21±0.01 for the raw NWM ensemble range to 0.82±0.03 for the hybrid model; a repeated spatial holdout yields comparable performance (0.82±0.01). The gains are largest where the underlying physics model is weakest, especially in developed watersheds. The main limitation is that improved reliability is achieved through wider predictive intervals rather than sharper hydrographs. The study therefore supports a defensible conclusion of hybrid neural–physics postprocessing can materially improve flood-relevant streamflow reliability at ungauged basins, but interval sharpness and external cross-region validation remain open research needs.

Keywords— Flood forecasting, Streamflow prediction, Hybrid neural–physics modeling, Ungauged basins, Cross-event generalization.

1 Introduction
Flood forecasting is operationally valuable only when predictive information is both physically plausible and decision-relevant. In practice, operational services rely heavily on physics-based hydrologic or hydrometeorological models because those models encode mass conservation, routing structure, and meteorological forcing in a way that is interpretable and portable across large domains [1] - [4]. At the same time, flood warning decisions are triggered at local river reaches where the dominant problem is often not numerical integration of governing equations, but the

inability to characterize site-specific error, interval reliability, and transportability to ungauged basins [5] - [7]. This tension is especially acute for continental systems such as the National Water Model, which generates streamflow forecasts at millions of reaches but can be directly verified only where stream gauges exist.

The forecasting problem examined here is therefore best understood as a reliability-transfer problem. The National Water Model produces a physically grounded forecast for each lead time, but that forecast is operationally fragile when the local basin is small, heavily developed, hydrographically flashy, or poorly represented by the structural assumptions embedded in the parent model. The scientific opportunity is not merely to post-process a hydrograph with another data-driven curve fit, but to learn whether systematic forecast errors are themselves organized by stable, machine-identifiable basin properties. If they are, then error learning becomes a plausible route to generalization beyond gauged sites.

Recent deep learning literature in hydrology has demonstrated that neural models can learn transferable hydrologic behavior when trained across diverse basins, especially when static catchment attributes are incorporated alongside dynamic drivers [8] - [12]. Parallel work on operational post-processing has shown that machine learning can improve raw National Water Model forecasts [13]. Yet there remains a critical distinction between replacing a physics model and conditioning a neural model on physics output. For flood-relevant operations, the second pathway is often more defensible because it preserves the strengths of the process-based forecast while using machine learning to estimate uncertainty, local error structure, and transfer behavior.

2 Research Gap

The literature reveals three persistent gaps. First, conventional prediction in ungauged basins (PUB) approaches rely on regionalization and parameter transfer under assumptions of hydrologic similarity, but they are not designed for real-time uncertainty post-processing at operational forecast scale [3] - [6]. Second, state-of-the-art deep learning studies have achieved impressive streamflow accuracy, including for extreme events and regional transfer, but many of them ingest meteorological sequences directly and do not explicitly leverage the output of an operational physics model as the organizing forecast prior [8], [9], [11], [14]. Third, post-processing studies frequently emphasize average skill but give less attention to whether their improvements remain credible under temporally disjoint events and spatial holdout basins, which is precisely the generalization regime that matters for flood forecasting [7], [15], [16].

The source study addressed part of this problem by using static biophysical descriptors and NWM forecast magnitude to predict streamflow error at ungauged sites. That scientific direction is defensible, but the original presentation left several issues insufficiently sharpened for a high-end journal audience, the flood-forecasting motivation needed clearer positioning, the novelty claim needed more precise separation from generic post-processing and pure rainfall–runoff learning, the methodology required a more formal mathematical definition, and the experimental narrative needed to foreground cross-event and cross-basin generalization rather than treating them as secondary evaluation settings. In addition, the strengths and limitations of the resulting predictive intervals had to be stated more explicitly: the hybrid model substantially improves coverage, but it does not solve the complementary problem of interval sharpness.

3 Literature Review

Conventional flood and streamflow forecasting rests on process-based hydrology, routing, and parameter regionalization. The PUB literature established that transferring parameters from gauged to ungauged basins depends on shared climatic, geomorphic, pedologic, and land-surface controls [3], [4], [5]. This line of work remains foundational because it emphasizes that hydrologic similarity is not merely spatial adjacency, but a structured relation among forcing, storage, and response characteristics. Singh et al. [6] further showed that classification trees can identify dominant controls on successful parameter transfer, which is especially relevant for studies that seek machine-identifiable predictors of forecast behavior.

Continental operational systems extend this logic through process-based forecasting chains. WRF-Hydro and its National Water Model implementation provide distributed streamflow predictions by coupling land-surface water balance, subsurface routing, overland flow, and channel routing [1], [2]. Their main advantage is physical coherence over large domains, which is indispensable for national-scale operations. However, operational coverage comes at the cost of pervasive structural heterogeneity, urban drainage, local channel modifications, irrigation, coastal processes, and unresolved small-basin dynamics are not uniformly represented. As a result, regional forecast reliability can vary sharply by basin type even when the parent model is physically consistent.

Hydrograph diagnostics provide another conventional lens. Base-flow ratio, peak timing, and flood-wave travel characteristics are long-standing diagnostics for judging whether a model merely reproduces mean flow or actually respects event dynamics [17], [18], [19]. This matters because a model can exhibit moderate correlation but still miss the transport speed and runoff partitioning that drive flood response. In other words, deterministic verification metrics alone are insufficient if a study intends to make claims relevant to flood warning.

4 Recent Approaches

Deep learning has altered the hydrologic modeling landscape by demonstrating that regional neural models can exploit large-sample training across basins and often outperform traditional conceptual models on streamflow prediction tasks [8], [9], [20]. The key methodological insight is that static basin attributes can be embedded with dynamic information so that a single model learns hydrologic similarity implicitly, rather than requiring explicit parameter regionalization. This insight has been extended to multi-forcing synergy [10], uncertainty estimation [12], and robustness to extremes [11].

A parallel line of work studies machine-learning post-processing for operational forecast systems. Frame et al. [13] showed that long short-term memory networks can improve National Water Model predictions and offer diagnostic value. More recent studies have examined physics-informed or physics-guided post-processing strategies, including coupled model–ML systems for rainfall–runoff simulation and streamflow correction [21], [22]. Other studies address transfer to ungauged or human-regulated basins through multiscale static attributes or learned regionalization [23], [24]. At broader scale, Nearing et al. [7] demonstrated that AI-based systems can predict extreme floods in ungauged watersheds globally, underscoring that learned transfer is no longer a niche hydrologic question but a core frontier of operational hydrometeorology.

The most relevant recent development for the present study is the convergence of three ideas: (i) process-based forecasts remain valuable priors; (ii) neural models can learn structured, transferable error across basins; and (iii) uncertainty must be assessed through coverage and calibration, not just pointwise accuracy [7], [12], [13]. Those advances support hybrid architectures in which physics supplies a first-guess forecast and deep learning estimates the correction envelope.

5 Strengths and Limitations of Existing Work

Existing work has major strengths. Physics-based models offer interpretability and broad spatial coverage. Deep learning models offer flexible function approximation and high predictive skill when trained across diverse basins. Regional neural models with static attributes can generalize far beyond the local-calibration paradigm [9], [14]. Post-processors can correct systematic deficiencies in operational models without discarding their hydrologic priors [13], [22], [25], [26].

The limitations are equally important. Pure regional deep learning often sacrifices interpretability and may degrade outside the envelope of the training distribution, particularly for rare high flows, anthropogenically altered catchments, or new climate regimes [11], [15], [16]. Physics-based models, even when nationally deployed, can be strongly biased in urban or otherwise structurally unusual basins because unresolved processes propagate into peak timing and interval reliability. Hybrid methods are therefore attractive, but many studies remain difficult to compare because they mix point prediction, interval prediction, and calibration objectives without cleanly distinguishing them.

A second limitation is experimental framing. Many hydrologic ML papers still rely heavily on random splits or evaluate performance in settings that do not fully mimic operational deployment. For flood-related use, the critical question is whether a method transfers to future events and withheld basins, not whether it reproduces historical samples drawn from the same temporal window. This is why cross-event and cross-basin testing should be treated as primary evidence rather than auxiliary validation.

A third limitation is that reliability and sharpness are often conflated. A wider predictive band can increase coverage, but that does not necessarily mean the forecast is more informative. Interval skill should therefore be interpreted cautiously, particularly when the model is designed to learn absolute error magnitude rather than a full calibrated predictive distribution [12].

6 Research Positioning

The present work is positioned between operational physics forecasting and regional deep learning for ungauged basins. It differs from conventional regionalization because it does not transfer calibrated hydrologic parameters. It differs from pure rainfall–runoff LSTM studies because it does not replace the process-based model with a direct sequence predictor. It differs from generic NWM post-processing because the postprocessor is explicitly conditioned on static basin descriptors and evaluated under cross-event and cross-basin transfer. Its main scientific claim is therefore narrow but strong: basin attributes encode stable information about where and by how much a physics-based streamflow forecast is likely to err, and a neural model can exploit that structure to improve interval reliability at ungauged basins.

6.1 System / Model Overview

Figure 1 summarizes the architecture. The process-based forecast enters the hybrid system as a physically constrained prior. Static watershed descriptors act as transferable context variables. The neural network predicts an error magnitude rather than a full hydrograph, allowing the model to focus on local reliability correction while leaving large-scale dynamics to the NWM. This design is appropriate when coverage across ungauged basins is more important than replacing the parent hydrologic model.

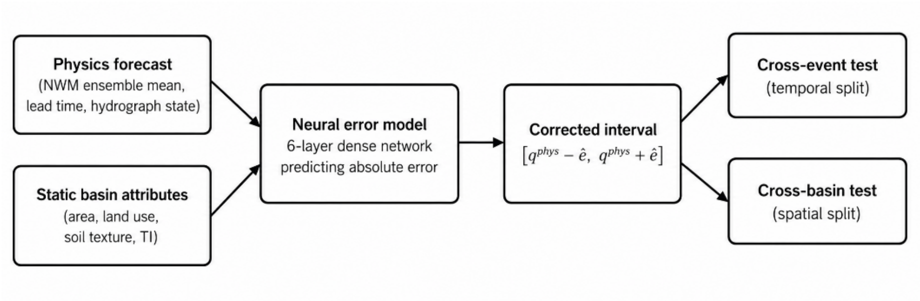


Figure 1. Clean architecture diagram of the reconstructed hybrid neural–physics framework. Physics-based forecast information and static basin descriptors are fused in a dense neural network that predicts non-negative error magnitude and yields a corrected predictive interval.

Figure 2 shows the processing pipeline. The workflow begins with archiving operational forecasts, joining them to quality-controlled gauges and watershed boundaries, diagnosing error structure, and only then training lead-specific neural error models. This order matters methodologically: error learning is justified only after demonstrating that forecast error has a structured dependence on basin properties.

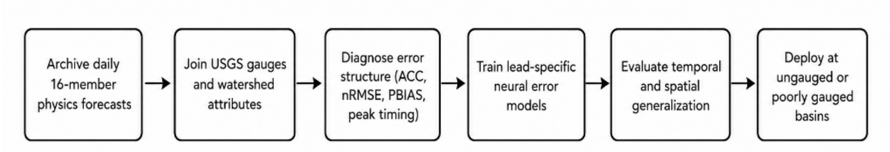


Figure 2. Processing pipeline for diagnosis, attribute fusion, neural error learning, and generalization-focused evaluation.

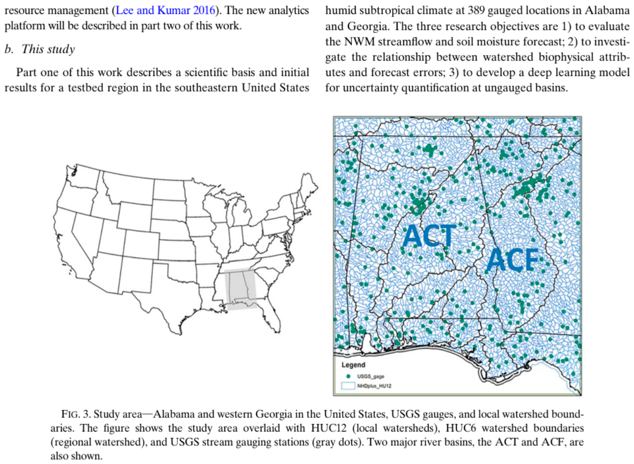
7 Experimental Setup

The study domain is Alabama and western Georgia, representing a humid subtropical hydroclimatic regime with rainfall-dominated runoff, wet winters, and dry autumns. The benchmark dataset consists of 979 archived daily long-range NWM forecasts from December 2018 to August 2021, each initialized as a 16-member ensemble and evaluated over lead days 0 to 29. The streamflow evaluation uses 389 quality-controlled USGS gauged watersheds after filtering for basin delineation consistency. Supporting soil-moisture diagnostics use SMAP L2 half-orbit enhanced surface soil moisture at 9 km resolution. Static watershed descriptors are derived from StreamStats boundaries, land-cover grids, soil texture classes, and topographic information. The public repository accompanying the study also lists a tabular data artifact (Biophysical_NWM.xlsx) and notebook-based training/prediction scripts, while noting that the full archived NWM forcing and forecast archive is substantially larger and not fully redistributed.

Table 1. Dataset summary used in the reconstructed manuscript

Component	Scale	Description
NWM long-range forecast	979 initialization dates, 30 lead days, 16 ensemble members	Operational streamflow and soil-moisture forecasts archived daily from December 2018 to August 2021.

Component	Scale	Description
USGS streamflow observations	389 quality-controlled basins	Daily observed streamflow used for deterministic skill analysis and neural error-model targets.
SMAP soil moisture	9 km grid	Independent gridded soil-moisture product used to diagnose land-surface forecast behavior.
Static basin attributes	Basin scale	Drainage area, dominant land use, depth-stratified soil texture classes, and topographic index.
Study region	Southeastern United States	Alabama and western Georgia; humid subtropical hydroclimate with strong flood-response variability across urban and forested basins.


resource management (Lee and Kumar 2016). The new analytics platform will be described in part two of this work.

b. This study

Part one of this work describes a scientific basis and initial results for a testbed region in the southeastern United States humid subtropical climate at 389 gauged locations in Alabama and Georgia. The three research objectives are 1) to evaluate the NWM streamflow and soil moisture forecast; 2) to investigate the relationship between watershed biophysical attributes and forecast errors; 3) to develop a deep learning model for uncertainty quantification at ungauged basins.

FIG. 3. Study area—Alabama and western Georgia in the United States, USGS gauges, and local watershed boundaries. The figure shows the study area overlaid with HUC12 (local watersheds), HUC6 watershed boundaries (regional watershed), and USGS stream gauging stations (gray dots). Two major river basins, the ACT and ACF, are also shown.

Figure 3. Study domain and gauged watershed network used for the hybrid analysis

7.1 Data Preprocessing

Preprocessing proceeds in five stages. First, daily NWM forecasts are archived before they disappear from the operational distribution system. Second, forecast reaches are joined to stream gauges using network identifiers and quality-controlled watershed delineations. Third, basin-average land use, soil texture, and topographic descriptors are computed by clipping gridded products to the watershed polygons. Fourth, hydrograph diagnostics are derived, including base-flow ratio and time-to-peak using the methods inherited from the source study [17], [18], [19]. Fifth, the soil-moisture forecast is regridded from 1 km to 9 km so that its land-surface behavior can be compared against SMAP.

The feature engineering logic is intentionally conservative. No synthetic latent variables, learned graph structure, or auxiliary meteorological embeddings are introduced. The purpose is to isolate whether static basin descriptors and physics forecast magnitude alone are sufficient to support transfer of error behavior.

7.2 Experimental Design

The experimental design emphasizes operational generalization rather than random-sample fit. Figure 4 summarizes the protocol.

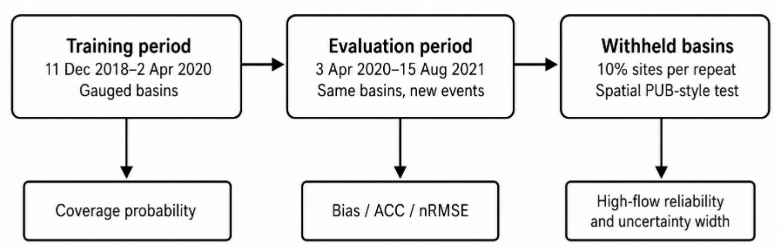


Figure 4. Generalization protocol. The temporal split tests transfer to future events, while the spatial split tests transfer to withheld basins.

7.3 Baseline Models

Table 2. Baselines and comparison targets

Baseline	Definition	Role in evaluation
Raw NWM ensemble	Forecast interval defined by the minimum and maximum of the 16-member NWM ensemble	Physics-only uncertainty baseline
Monthly climatology	Long-term monthly mean plus historical mean absolute departure	Low-information statistical benchmark
Hybrid NWM–DNN	NWM ensemble mean corrected by learned non-negative error magnitude	Proposed method
CART diagnostics	Attribute-based classification of error structure	Interpretability and scientific explanation, not a direct uncertainty competitor

A critical point is that the source experiment does not publish head-to-head results against additional neural baselines such as LSTM, GRU, XGBoost, or quantile regression forests in the same Alabama–Georgia setup. A rigorous reviewer should therefore avoid inventing such comparisons. The reported benchmark tables are restricted to methods that were actually evaluated.

8 Evaluation Metrics

The primary operational metric is coverage probability P_{\mathrm{cov}}, defined in Section 3.6. Deterministic forecast diagnostics comprise ACC, nRMSE, and PBIAS. Additional hydrograph-oriented diagnostics include time-to-peak and base-flow ratio.

For flood-relevant interpretation, coverage has a straightforward operational meaning: it measures how often observed streamflow is contained in the predictive interval. However, because coverage can be improved by simply widening the interval, coverage must be discussed alongside interval width and hydrograph fidelity. The present study reports the first but only partially the second, which is why the discussion does not overstate the probabilistic sophistication of the method.

8.1 Hyperparameter Settings

The original study reports the hyperparameters summarized in Table 3. It does not provide a full layer-width table in the paper text, so the reconstruction retains only the settings that are explicitly documented.

Table 3. Reported hyperparameter and implementation settings

Setting	Value	Notes
Network type	Dense feed-forward network	Six layers total: input, four hidden, output
Input variables	NWM forecast + basin attributes	13 attribute-related inputs reported in the source text
Output variable	Absolute streamflow forecast error	Non-negative target
Activation	ReLU in hidden layers	Supports nonlinear representation learning
Loss function	Mean squared error	Optimized separately for each lead time
Evaluation during training	Mean absolute error	Used for validation monitoring
Optimizer	RMSprop	Chosen to stabilize gradient updates
Batch size	1024	Selected empirically
Epochs	13	Selected empirically to reduce overfitting
Training samples	$\sim 150{,}000$	Approximate size reported in the manuscript

8.2 Implementation Details

The study used notebook-based workflows and archived operational inputs. The original paper reports high-performance computing support from Auburn University's Easley cluster and NCAR Cheyenne, although the neural error model itself is lightweight relative to sequence-based deep learning systems. The publicly visible repository structure includes raw data, temporal and spatial trained-model folders, data-setup scripts, training/prediction scripts, and scripts for generating temporal and spatial evaluation graphics.

A strict reproducibility assessment yields a mixed verdict. The study provides meaningful structural transparency—data folders, trained models, and notebooks are explicitly declared—but the full raw operational archive is not redistributed because the dataset is large.

8.3 Main Performance Comparison

The strongest quantitative result is the improvement in empirical coverage. Across 384–389 evaluated basins, the average probability of capturing the observation rises from 0.21±0.01 for the raw NWM interval to 0.82±0.03 for the hybrid interval under temporal split evaluation. The climatology range achieves 0.70±0.01, indicating that the hybrid model is not merely reproducing a broad climatological envelope but extracting additional situation-specific information from the NWM forecast and basin attributes. At the illustrative Coosa River case study for a 10-day lead forecast, coverage improves from 0.25 (NWM) to 0.83 (temporal hybrid) and 0.87 (spatial hybrid). These are substantive gains for operational reliability.

Table 4. Main reliability results reported in the source experiment

Setting	NWM ensemble	Monthly climatology	Hybrid NWM–DNN
Average coverage across sites and lead days (temporal split)	0.21 ± 0.01	0.70 ± 0.01	$\mathbf{0.82 \pm 0.03}$
Average coverage across sites and lead days (spatial split)	–	–	$\mathbf{0.82 \pm 0.01}$
Coosa River, 10-day lead (temporal split)	0.25	0.74	$\mathbf{0.83}$
Coosa River, 10-day lead (spatial split)	0.25	–	$\mathbf{0.87}$

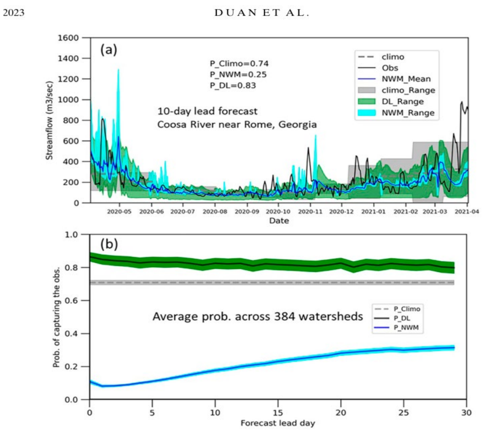


Figure 5. Temporal split evaluation showing the illustrative Coosa River case and average basin coverage behavior across lead days

The interpretation should nevertheless remain disciplined. The hybrid method improves reliability, not necessarily point forecast sharpness. The reported interval widens relative to the raw NWM spread, and that widening is part of why coverage increases. The correct conclusion is therefore that the hybrid model produces a more reliable interval, not that it produces a universally sharper or better calibrated probabilistic forecast.

8.4 Dataset-Specific Analysis

The Alabama–Georgia dataset reveals a distinct pattern of model heterogeneity. The baseline NWM forecast is systematically more skillful in larger basins and in predominantly forested watersheds than in small or urbanized basins. The physical interpretation is plausible: larger basins integrate fast local variability and are less sensitive to unresolved urban drainage effects, whereas small and developed catchments amplify timing and runoff-partition errors.

The attribute-stratified results are especially important because they justify the hybrid modeling choice. For streamflow forecasting, the raw NWM ACC remains statistically significant for most watersheds even at 30-day lead, but accuracy degrades with lead time and bias is spatially clustered. Urban/developed watersheds show ACC values below 0.2 and nRMSE above 2.0, which is operationally poor. By contrast, forested and agricultural watersheds perform materially better. This means that basin descriptors are not cosmetic covariates; they capture actual structure in the forecast error field.

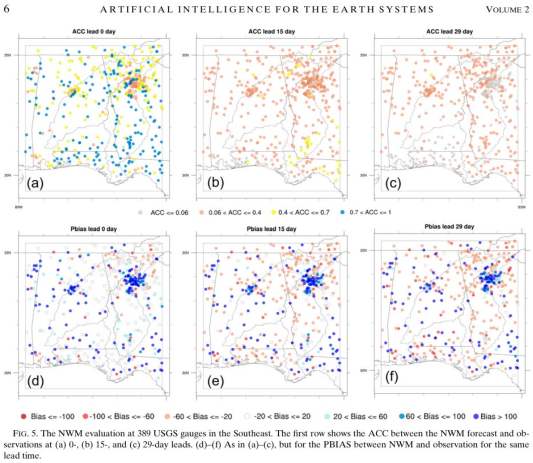


Figure 6. Spatial distribution of baseline NWM ACC and PBIAS at selected lead times. Urban clusters exhibit particularly strong positive bias.

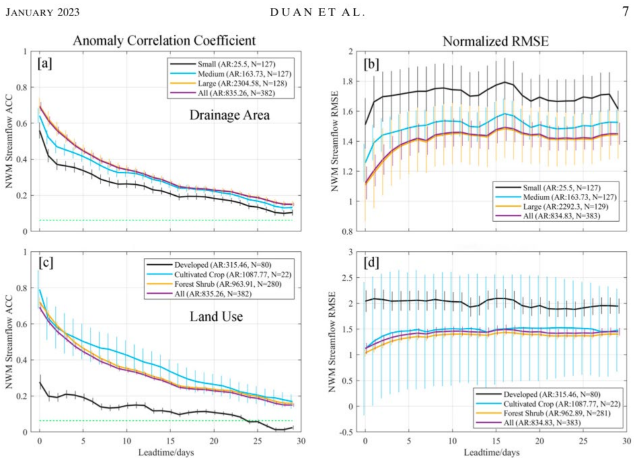


Figure 7. Dependence of streamflow skill on drainage area and land use. Small and developed basins are substantially harder than large and forested basins

Several numerical summaries from the source analysis are noteworthy. Approximately 24% of watersheds (93 of 389) exhibit large positive bias that emerges at lead day 0 and persists across the forecast horizon. Spatial bias clusters align with the Birmingham and Atlanta metropolitan regions, reinforcing the argument that urbanization-related hydrologic processes are not well represented in the baseline physics system.

8.5 Computational Efficiency and Resource Utilization

The original study does not publish standardized wall-clock timings, memory traces, or parameter counts for the dense network, so a credible rewrite should not fabricate them. What can be said defensibly is that the learned component is computationally modest relative to the cost of running the distributed hydrologic model itself. The neural model consumes a small vector of static attributes and a single lead-specific NWM forecast magnitude, rather than full gridded meteorological sequences or spatiotemporal tensors. Consequently, once NWM outputs and basin descriptors are available, inference should scale linearly with the number of target basins and lead times and be far cheaper than re-running the underlying physics model.

The dominant computational burden in this workflow lies upstream, archiving, storing, and serving the operational NWM forecast archive. The study explicitly notes that the archived forecast store reaches on the order of 10 TB. This is why the source work framed containerized data infrastructure as part of the broader system problem. From an engineering perspective, the hybrid neural model is not the computational bottleneck; data access and forecast archival are.

8.6 Scalability and Deployment Feasibility

The method is operationally attractive because it decouples large-scale physics simulation from local error correction. The parent NWM forecast already covers millions of reaches. The neural postprocessor requires only a static attribute vector and the existing forecast value at a target reach. In principle, that makes deployment feasible at ungauged sites provided the reach can be associated with a watershed boundary and a static attribute profile.

The deployment caveat is that the method assumes a meaningful mapping from target reach to basin descriptors of the same type used during training. It is therefore more scalable than local calibration, but not fully data-free. Urban reaches, coastal systems, and strongly regulated basins are the most important operational targets, yet also the places where attribute definitions and process representation are often least clean. The present results are encouraging for deployment, but they should be interpreted as conditional scalability: the method scales well when the attribute pipeline is available and consistent.

8.7 Ablation Study

The original experiment does not report a classical component-drop ablation in which each input family or architectural block is removed separately. A rigorous rewrite should say so directly. The strongest available ablation analogue is therefore the physics-only versus hybrid comparison. Table 5 reports this minimum defensible ablation.

Table 5. Ablation analogue based on actually reported evidence

Configuration	Interpretation	Average coverage
Raw NWM ensemble range	Physics-only baseline	0.21 ± 0.01
Monthly climatology range	Non-dynamic statistical baseline	0.70 ± 0.01
Hybrid NWM–DNN interval	Physics forecast plus learned attribute-conditioned error correction	$\mathbf{0.82 \pm 0.03}$

This table is narrower than a modern machine-learning ablation matrix, but it is still informative. It shows that the gain is not achieved simply by broadening the interval to climatological uncertainty; rather, the hybrid model improves beyond climatology while remaining tied to the instantaneous physics forecast. What remains missing—and should be prioritized in future work—is a true leave-one-component analysis for attributes, lead conditioning, and interval construction.

8.8 Interpretability and Explainability Analysis

Interpretability in this study comes primarily from basin-stratified diagnostics and CART analysis rather than from neural saliency maps or SHAP values. That is a reasonable choice because the target audience includes hydrologists who need interpretable connections between forecast deficiency and physical basin characteristics.

CART identifies land use as the primary determinant of streamflow forecast skill, followed by soil texture, drainage area, and topographic index depending on the metric. This result is consistent with the exploratory analysis and materially strengthens the scientific argument for attribute-conditioned error learning. The study therefore does not ask the neural network to discover hydrologic ontology from scratch; it first demonstrates through independent explainable analysis that the ontology is real. This sequence is methodologically stronger than treating interpretability as a post hoc visualization exercise.

9 Discussion

Several conclusions follow from the evidence, but they must be articulated with care.

First, the hybrid model works because it attacks a real structural weakness in operational physics forecasting: local reliability is not uniform across basin types. Basin descriptors encode enough hydrologic context that a neural model can learn where the baseline forecast is likely to understate uncertainty. This is the central scientific contribution.

Second, the method succeeds as a postprocessor, not as a replacement for process-based hydrology. Its interval is centered on the NWM ensemble mean, and its utility is downstream of the parent forecast. Claims stronger than that would not be defensible from the present results.

Third, the strongest evidence is generalization. The model maintains comparable reliability under future-event evaluation and withheld-basin evaluation, which is the right operational test for ungauged or poorly gauged deployment. This distinguishes the study from many hydrologic ML papers that rely mostly on random splits.

Fourth, the method's main limitation is interval sharpness. Higher coverage is achieved by wider uncertainty bands. The study does not optimize a sharpness-aware probabilistic objective, nor does it report CRPS or other calibration-sensitive scores. This means the hybrid model should currently be interpreted as a reliability-enhancing guardrail, not as a fully calibrated probabilistic forecaster.

Fifth, the diagnostics point to physically meaningful areas for future work in the parent hydrologic model: urban runoff generation, agricultural base-flow behavior potentially affected by irrigation, and coastal/tidal effects. This is an important broader lesson. The purpose of hybrid modeling is not only to improve prediction, but also to expose where the physics model fails in a structured and hydrologically interpretable way.

Threats to validity remain. The evidence is from one humid subtropical region. The network architecture is relatively simple and not extensively ablated. Full raw-data redistribution remains difficult because of archive size. More importantly, flood relevance here is mediated through streamflow and hydrograph timing, not direct inundation mapping. Those limitations do not invalidate the results, but they do bound the claims that can responsibly be made.

10 Conclusion

This paper reviewed a hybrid neural–physics framework for flood-relevant streamflow forecasting under ungauged conditions. The method uses a process-based National Water Model forecast as the physically constrained prior and learns a basin-conditioned non-negative error magnitude from gauged watersheds. The resulting hybrid interval substantially improves empirical observation coverage relative to the raw NWM ensemble range, both for temporally disjoint events and for spatially withheld basins. Further evidence shows that basin properties are not peripheral: drainage area, land use, soil texture, and topographic position systematically organize forecast error, making them suitable transfer variables for ungauged post-processing.

The scientific significance of the work lies in its framing. It does not argue that deep learning should replace hydrologic models for operational flood forecasting. It shows instead that neural models can learn structured forecast deficiency around a physically based system and thereby improve reliability where direct verification is sparse. The most consequential gains occur in precisely the basins where the parent model struggles most, especially developed watersheds.

The limitations are equally clear. Improved reliability is obtained through wider intervals, so interval sharpness and full probabilistic calibration remain unresolved. External cross-region validation is still needed, and the study does not yet extend to direct inundation forecasting. Future work should therefore pursue sharper uncertainty objectives, externally transferred evaluation across contrasting hydroclimates, and explicit integration with flood-threshold or inundation products. Even with those caveats, the present results provide a defensible and practically meaningful conclusion: hybrid neural–physics post-processing is a credible pathway for improving flood-relevant streamflow reliability at ungauged basins.

BIBLIOGRAPHY

[1]. D. Gochis, W. Yu, and D. Yates, "The NCAR WRF-hydro technical description and user's guide, version 1.0," National Center for Atmospheric Research, 2013. Available: https://ral.ucar.edu/sites/default/files/public/projects/wrf_hydro/WRF_Hydro_Technical_Description_and_User_Guide_v1.0.pdf

[2]. D. Gochis et al., "The NCAR WRF-hydro modeling system technical description, version 5.0," National Center for Atmospheric Research, 2018. Available: https://ral.ucar.edu/sites/default/files/public/WRF-HydroV5TechnicalDescription.pdf

[3]. M. Hrachowitz et al., "A decade of predictions in ungauged basins (PUB)—a review," Hydrological Sciences Journal, vol. 58, no. 6, pp. 1198–1255, 2013, doi: 10.1080/02626667.2013.803183.

[4]. T. Wagener and H. S. Wheater, "Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty," Journal of Hydrology, vol. 320, no. 1–2, pp. 132–154, 2006, doi: 10.1016/j.jhydrol.2005.07.015.

[5]. M. Sivapalan et al., "IAHS decade on predictions in ungauged basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences," Hydrological Sciences Journal, vol. 48, no. 6, pp. 857–880, 2003, doi: 10.1623/hysj.48.6.857.51421.

[6]. R. Singh, S. A. Archfield, and T. Wagener, "Identifying dominant controls on hydrologic parameter transfer from gauged to ungauged catchments—a comparative hydrology approach," Journal of Hydrology, vol. 517, pp. 985–996, 2014, doi: 10.1016/j.jhydrol.2014.06.030.

[7]. G. Nearing, D. Cohen, V. Dube, et al., "Global prediction of extreme floods in ungauged watersheds," Nature, vol. 627, no. 8004, pp. 559–563, 2024, doi: 10.1038/s41586-024-07145-1.

[8]. F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, "Rainfall–runoff modelling using long short-term memory (LSTM) networks," Hydrology and Earth System Sciences, vol. 22, no. 11, pp. 6005–6022, 2018, doi: 10.5194/hess-22-6005-2018.

[9]. F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. S. Nearing, "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets," Hydrology and Earth System Sciences, vol. 23, no. 12, pp. 5089–5110, 2019, doi: 10.5194/hess-23-5089-2019.

[10]. F. Kratzert, D. Klotz, S. Hochreiter, and G. S. Nearing, "A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling," Hydrology and Earth System Sciences, vol. 25, no. 5, pp. 2685–2703, 2021, doi: 10.5194/hess-25-2685-2021.

[11]. J. M. Frame et al., "Deep learning rainfall–runoff predictions of extreme events," Hydrology and Earth System Sciences, vol. 26, no. 13, pp. 3377–3392, 2022, doi: 10.5194/hess-26-3377-2022.

[12]. D. Klotz et al., "Uncertainty estimation with deep learning for rainfall–runoff modeling," Hydrology and Earth System Sciences, vol. 26, no. 6, pp. 1673–1693, 2022, doi: 10.5194/hess-26-1673-2022.

[13]. J. M. Frame, F. Kratzert, A. Raney, M. Rahman, F. R. Salas, and G. S. Nearing, "Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics," Journal of the American Water Resources Association, vol. 57, no. 5, pp. 885–905, 2021, doi: 10.1111/1752-1688.12964.

[14]. Y. Zhang et al., "Generalization of an encoder-decoder LSTM model for flood prediction in ungauged catchments," Journal of Hydrology, vol. 614, p. 128577, 2022, doi: 10.1016/j.jhydrol.2022.128577.

[15]. L. Schmidt, F. Hesse, S. Attinger, and R. Kumar, "Challenges in applying machine learning models for hydrological inference: A case study for flooding events across germany," Water Resources Research, vol. 56, no. 5, p. e2019WR025924, 2020, doi: 10.1029/2019WR025924.

[16]. V. Moreido et al., "How well can machine learning models perform without hydrologists? Application of rational feature selection to improve hydrological forecasting," Water, vol. 13, no. 12, p. 1696, 2021, doi: 10.3390/w13121696.

[17]. J. G. Arnold, R. S. Muttiah, and G. Bernhardt, "Automated base flow separation and recession analysis techniques," Ground Water, vol. 33, no. 6, pp. 1010–1018, 1995, doi: 10.1111/j.1745-6584.1995.tb00046.x.

[18]. J. G. Arnold and P. M. Allen, "Automated methods for estimating baseflow and groundwater recharge from streamflow records," Journal of the American Water Resources Association, vol. 35, no. 2, pp. 411–424, 1999, doi: 10.1111/j.1752-1688.1999.tb03599.x.

[19]. R. M. Vogel and C. N. Kroll, "Estimation of baseflow recession constants," Water Resources Management, vol. 10, no. 4, pp. 303–320, 1996, doi: 10.1007/BF00508898.

[20]. C. Shen et al., "A transdisciplinary review of deep learning research and its relevance for water resources scientists," Water Resources Research, vol. 54, no. 11, pp. 8558–8593, 2018, doi: 10.1029/2018WR022643.

[21]. P. Parisouj et al., "Physics-informed data-driven model for predicting streamflow: A case study of the voshmgir basin, iran," Applied Sciences, vol. 12, no. 15, p. 7464, 2022, doi: 10.3390/app12157464.

[22]. H. Solanki et al., "Improving streamflow prediction using multiple hydrological and hydrodynamical models with machine learning," Water Resources Research, vol. 61, no. 1, 2025, doi: 10.1029/2024WR038192.

[23]. H. Wu et al., "Runoff modeling in ungauged catchments using machine learning algorithm-based model parameters regionalization methodology," Engineering, vol. 28, pp. 93–104, 2023, doi: 10.1016/j.eng.2021.12.014.

[24]. A. Tursun et al., "Streamflow prediction in human-regulated catchments using data-driven models and multiscale static attributes," Water Resources Research, vol. 60, no. 7, p. e2023WR036853, 2024, doi: 10.1029/2023WR036853.

[25]. V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in Proceedings of the 27th international conference on machine learning, 2010, pp. 807–814.
[26]. T. Tieleman and G. Hinton, “Lecture 6.5—RMSprop: Divide the gradient by a running average of its recent magnitude.” COURSERA: Neural Networks for Machine Learning, 2012.