

Document Version

Final published version

Licence

CC BY

Citation (APA)

Matin, N., Herath, H. M. V. V., Saha, A., Marshall, L., & Babovic, V. (2026). Hydrologically constrained genetic programming for interpretable rainfall–runoff model discovery: A process–informed machine learning approach. *Journal of Hydrology*, 680, Article 136493. <https://doi.org/10.1016/j.jhydrol.2026.136493>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



Hydrologically constrained genetic programming for interpretable rainfall–runoff model discovery: A process–informed machine learning approach

Naila Matin^{a,*}, Herath Mudiyanse Viraj Vidura Herath^b, Abhishek Saha^{c,d}, Lucy Marshall^b, Vladan Babovic^a

^a Department of Civil and Environmental Engineering, National University of Singapore, 117576, Singapore

^b School of Civil Engineering, Faculty of Engineering, The University of Sydney, Sydney, New South Wales, Australia

^c Hydroinformatics Institute, 118256, Singapore

^d Delft Institute of Applied Mathematics, Delft University of Technology, Delft, the Netherlands

HIGHLIGHTS

- HydroXGP discovers interpretable rainfall–runoff model structures from data.
- Genetic programming evolves models from process-based building blocks.
- Model discovery is constrained through mass-conserving simulation.
- Learned structures transfer beyond the basin used for training.
- Strong out-of-sample skill relative to established hydrological benchmark models.

ARTICLE INFO

This manuscript was handled by Yuefei Huang, Editor-in-Chief, with the assistance of Yonghong Hao, Associate Editor

Keywords:

Genetic programming
Rainfall–runoff modelling
Process-informed machine learning
Conceptual hydrological models
Model structure discovery

ABSTRACT

Advances in machine learning have improved rainfall–runoff prediction, but integrating these capabilities with interpretable, process-based model structures remains an open challenge. This study seeks a middle ground between conceptual hydrological modelling and modern data-driven prediction by automating model structure discovery while retaining interpretability and hydrological consistency. We introduce HydroXGP (Hydrological Model Explorer using Genetic Programming) to reduce reliance on legacy model choices and subjective a priori structural assumptions by learning model structure directly from data during optimization. The framework evolves models from a modular, process-based component library adapted from the open-source MARRMoT toolbox, making the search process-informed through reusable flux and routing components and hydrologically constrained by enforcing mass balance during simulation. GP jointly optimizes model structure and parameters, while each discovered model remains explicit and interpretable. Evaluation on the CAMELS-US dataset showed that a proof-of-concept test in a snow-dominated basin achieved strong performance over the out-of-sample test period (NSE = 0.849), and, when transferred to a hydrologically similar basin, the learned structure retained useful predictive skill (NSE = 0.724). Across 18 basins (one representative basin per U.S. hydrologic region), median NSE reached 0.702, generally outperforming the mean performance of an ensemble of conceptual benchmark models while remaining competitive with some regional LSTM baselines. The findings support a modelling perspective in which structural hypotheses can be revisited and adapted as new data, diagnostics, or environmental signals become available, enabling basin-specific configurations that evolve in response to changing information rather than relying on a single fixed template.

* Corresponding author.

Email address: nailamatin@nus.edu.sg (N. Matin).

<https://doi.org/10.1016/j.jhydrol.2026.136493>

Received 8 April 2026; Received in revised form 21 August 2026; Accepted 24 September 2026

Available online 25 September 2026

0022-1694/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Hydrological models serve as the primary tools for quantifying catchment dynamics and supporting water resources management. Over time, a wide range of model types have been developed, from parsimonious conceptual formulations to complex physically based distributed systems (Clark et al., 2017; Horton et al., 2022). This proliferation is often interpreted as evidence that no single structure is universally appropriate, as the ‘uniqueness of place’ requires models to be tailored to specific landscape characteristics, dominant processes, and modelling objectives (Beven, 2020). Consequently, every modelling application requires the selection of a specific model structure, a decision that fundamentally influences the simulation of states and fluxes, and ultimately, the conclusions drawn from the study (Mendoza et al., 2015). Ideally, this choice should be guided by the adequacy of a model’s physical representation for the landscape and research question at hand (Fenicia et al., 2014). However, in practice, the choice is frequently driven by ‘legacy’ factors, such as familiarity, code availability, and institutional preference, rather than by rigorous testing of structural suitability for the unique hydrological conditions of a given basin (Addor and Melsen, 2019).

The fact that a model selected based on legacy can be calibrated to reproduce historical streamflow data does not guarantee that it correctly characterizes the underlying physical processes of the catchment. Indeed, large-sample experiments demonstrate that structurally distinct models can achieve similar predictive skill (often referred to as structural equifinality) while simulating markedly different internal flux dynamics (Khatami et al., 2019; Knoben et al., 2020). Without rigorous assessment of structural adequacy, such models may therefore yield the right results for the wrong reasons (Kirchner, 2006; Gupta et al., 2012), compromising their reliability under future or changing environmental conditions. This recognition has motivated calls to adopt the method of multiple working hypotheses, emphasizing systematic comparison of alternative process representations rather than reliance on inherited structures (Clark et al., 2011). Advancing hydrological understanding therefore requires not only parameter estimation but also principled exploration of different model structures. To support such exploration, the community has developed flexible and modular modelling frameworks. For example, FUSE (Clark et al., 2008) recombines components from established models to diagnose structural errors, while SUPERFLEX (Fenicia et al., 2011; Kavetski and Fenicia, 2011) provides a generic language of building blocks for custom model construction. Building on these ideas, MARRMoT (Knoben et al., 2019a) provides standardized libraries of conceptual models and process equations for systematic comparison of alternative structures. Raven (Craig et al., 2020), DECIPHER (Coxon et al., 2019), and RoGeR (Schwemmler et al., 2024) likewise support flexible configurations across process representations and spatial scales. In parallel, the growth of large-sample datasets has created new opportunities to test structural hypotheses across climatic and physiographic diversity, balancing site-specific insight with broader generalization (Gupta et al., 2014).

Yet flexibility alone does not resolve subjectivity. Even within modular frameworks, model assembly typically requires expert judgement regarding which storages, flux formulations, and routing schemes to include (Kiraz et al., 2023). Consequently, while flexible frameworks expand the hypothesis space, navigating that space remains largely manual and influenced by modeller experience or institutional precedent. To overcome these barriers, recent research has pivoted toward automated model structure identification. For example, AMSI (Spieler et al., 2020; Spieler and Schütze, 2024) treats structural decisions as optimization variables and searches efficiently for high-performing configurations, while sparse regression approaches such as SINDy have been adapted to discover governing equations and latent states directly from rainfall–runoff data (Dantzer and Kerkez, 2024). More recently, INDRA (Eythorsson and Clark, 2025) has explored AI-assisted hydrological workflows using multi-agent large language models to guide model conceptualization and configuration. These advances reduce reliance on

manual trial-and-error, but they still leave a large combinatorial space of plausible structures to be explored efficiently and evaluated robustly.

At the same time, deep learning has opened a different route to high-skill rainfall–runoff prediction and reshaped discussions about how predictive skill relates to process representation and robustness. Large-sample Long Short-Term Memory (LSTM) models trained across many basins have achieved strong predictive performance, including in ungauged settings when combined with catchment attributes (Kratzert et al., 2019, 2024). Evaluations under hydrological extremes highlight both their promise and the trade-offs introduced by architectural constraints such as mass conservation (Frame et al., 2022). Despite these empirical gains, such models are still often regarded as “black-box” approaches because they learn an input–output relationship from data, while the internal transformations and fitted weights are not straightforward to interpret as hydrological processes. In standard LSTM rainfall–runoff models, this also means that the model does not explicitly represent named storages and fluxes or enforce conservation laws in the way traditional hydrological models do (Reichstein et al., 2019). Related perspectives argue that machine learning can strengthen process-based modelling through automation and decision support rather than replacing physically grounded simulators (Clark et al., 2025). Even so, their empirical generalization has prompted alternative interpretations. Nearing et al. (2021) suggest that successful transfer to ungauged basins is consistent with learning hydrologically relevant structure from large-sample data. However, it remains uncertain whether these internal states reflect stable cause-effect mechanisms or mainly exploit correlations that may shift under change (De la Fuente et al., 2024; Bayati et al., 2026).

A growing family of physics-guided hybrid approaches attempts to bridge process-based reasoning and modern machine learning by encoding hydrological hypotheses in trainable, differentiable systems. For example, Feng et al. (2022) proposed δ -models, which embed neural networks within a conceptual rainfall–runoff backbone such as HBV to learn time-varying parameters that adapt to changing conditions. He et al. (2024) introduced deep process learning, which constrains neural networks using physically motivated priors to improve realism and identifiability. Across these hybrid strategies, differentiable programming enables gradient-based optimization of high-dimensional parameterizations within a physically grounded computational graph (Shen et al., 2023). However, the underlying conceptual structure is generally fixed a priori, leaving exploration of alternative structures largely outside the learning loop.

While differentiable hybrid approaches can improve parameter learning and enforce physical constraints within a predefined model structure, they rarely provide mechanisms to explore alternative structures systematically. Addressing this challenge requires methods capable of optimizing discrete structural decisions, such as component inclusion, connectivity, and flux-law selection, together with continuous parameters. Genetic programming (GP) (Koza, 1992) provides one such approach, using evolutionary search to construct symbolic expressions and model architectures directly from data. In hydroinformatics, evolutionary algorithms have long been used for automated model calibration (Babovic and Wu, 1994), and GP has been valued for its interpretability and for producing explicit mathematical relationships between inputs and outputs (Babovic and Keijzer, 2002). It has also been used to test structural hypotheses about dominant hydrological processes (Selle and Muttil, 2011; Chadalawada et al., 2017) and to incorporate domain-specific concepts, such as reservoirs and routing functions, directly into learning frameworks (Havlíček et al., 2013). More recent studies have moved towards hydrologically informed GP by constraining the search space with building blocks drawn from conceptual modelling frameworks, thus inducing rainfall–runoff models with recognizable process components (Chadalawada et al., 2020; Herath et al., 2021a,b). These efforts demonstrate feasibility, but they also highlight some continuing needs. For example, richer and extensible libraries of process components and flux functions are needed so discovery is not confined to a small set of templates, along with finer control over how catchment

response is generated across multiple processes and timescales. We also need numerically robust simulation of GP-assembled models so that selection pressure reflects structural adequacy rather than solver artefacts. Furthermore, many existing methods do not provide readily usable code or standardized toolkits, which limits broader adoption, reproducibility, and transparency. Finally, systematic benchmarking against established process-based conceptual models and modern deep learning baselines is needed to characterize the trade-offs between accuracy, robustness, and hydrological adequacy.

In this study, we introduce HydroXGP (Hydrological Model eXplorer using Genetic Programming), a process-informed, hydrologically constrained genetic-programming framework for discovering basin-specific conceptual rainfall–runoff model structures within a modular architecture. The overarching goal is to establish a practical middle ground between flexible conceptual modelling and modern machine learning by automating structure discovery while keeping candidate hypotheses interpretable and grounded in hydrologic process components. Specifically, the framework is process-informed in that candidate models are assembled from reusable hydrologic components drawn from established conceptual modelling practice, and hydrologically constrained in that each candidate is evaluated in a simulator that tracks water balance and updates storages based on fluxes between components. Admissible model structures are further restricted by a master architecture that permits only hydrologically plausible connections and compositions of the model building blocks. Because every candidate must be runnable as a time-stepping model, the framework favours numerically stable, feasible candidates, while those that cannot be simulated reliably are penalized during selection. Within this constrained search space, GP jointly optimizes discrete structural decisions and continuous parameters, allowing model topology to adapt to catchment behaviour. The ‘physics’ enforced by HydroXGP therefore refers to hydrologic consistency—mass conservation, storage–flux relationships, and connectivity constraints—together with numerically stable, feasible state evolution under time stepping. Discovered model structures are represented explicitly, making them amenable to inspection, reuse, and extension. The framework is evaluated through three targeted experiments using the large-sample Catchment Attributes and Meteorology for Large-Sample Studies (CAMELS) dataset (Newman et al., 2015), designed to address three research questions: (1) Feasibility: can process-informed GP discover parsimonious and mass-conserving, hydrologically consistent structures that achieve strong out-of-sample performance; (2) Transferability: to what extent can discovered structures be transferred between hydrologically similar basins, and can transfer reduce search effort under constrained computational budgets; and (3) Robustness: how do automatically induced structures perform across contrasting hydroclimatic settings relative to established conceptual and deep learning benchmarks. To promote transparency, reusability, and community use, HydroXGP is released as an open-source, readily extensible Python toolkit via HydroShare (<http://www.hydroshare.org/resource/49b2089113724baf9faa5bb45246db2f>). By situating evolutionary structure discovery within a modular conceptual modelling ecosystem, this work aims to reconcile predictive performance with interpretable, process-grounded modelling and to support more systematic testing of structural hypotheses in hydrology.

2. Methods

The proposed framework, HydroXGP, operationalizes process-informed structure discovery by assembling hydrological components under a master architecture and evaluating each candidate through mass-conserving simulation. Candidate models are encoded as executable configuration graphs rather than hard-coded implementations, keeping discovered structures explicit and interpretable. Methodologically, the framework contributes: (i) extensible model specification via JSON-based configurations; (ii) stable and scalable numerical simulation of connected graphs using sparse-matrix representations;

and (iii) extensive evaluation across a broad library of process-based equations and hydrological components. The following sections describe the component library and representation, the GP-based optimization procedure, and the numerical simulation used to evaluate candidate models.

2.1. Model building blocks and representation

The process-component library used in this study is adapted from the Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) (Knoben et al., 2019a; Trotter et al., 2022; Knoben and Trotter, 2022), an open-access and standardized model library that has substantially advanced systematic structural testing in hydrology. By providing consistent implementations of a large set of established conceptual models (47 in the current version) and representing their constituent flux equations and routing kernels as reusable building blocks, MARRMoT enables transparent and rigorous model comparison. Its careful separation of model equations from the numerical solution further facilitates the reuse of equations for automated structure exploration, making it an ideal foundation for the present work. Accordingly, the structural search space in this study is defined by a curated subset of MARRMoT components; the selected building blocks and their occurrences across the source model structures are listed in Appendix Table A.1, while Fig. 3 provides a conceptual overview of the hydrological process space in which these components can be connected.

Hydrologically, water enters the framework as precipitation, which may first be partitioned into rainfall and snowfall according to air temperature. When the snow module is active, snowfall accumulates in a snow storage and is subsequently released as snowmelt according to a temperature-dependent melt formulation. Rainfall may then pass through an interception storage, with throughfall entering the upper soil-moisture storage or generating infiltration-excess runoff. Within the soil-storage system, different process equations can be selected to represent evapotranspiration limitation, saturation-excess runoff, interflow, baseflow-type drainage, and percolation. Water may be transferred from the upper soil storage to a slower conceptual storage through drainage or percolation, while upward redistribution may occur through capillary-rise or groundwater return fluxes. The resulting runoff components may subsequently pass through unit-hydrograph operators and routing storages before contributing to outlet discharge. Importantly, the master architecture represents the full set of process pathways and structural components available to HydroXGP rather than a single rainfall–runoff model. Each discovered model therefore activates only a subset of these components and pathways and, where alternatives are available, selects the corresponding process equations listed in Table A.1. Related automated structure-identification studies have similarly constructed their search spaces using different selections and arrangements of hydrological building blocks (Chadalawada et al., 2020; Herath et al., 2021a; Spieler et al., 2020); the master architecture used here therefore represents one possible design.

Computationally, each candidate model is stored as a directed, machine-readable JSON configuration graph (Fig. 2(b) shows a simplified form). The nodes correspond to either storages or auxiliary dummy elements: real storages define the solver state vector, while dummy nodes implement algebraic operators for flux splitting and unit-hydrograph routing without introducing additional storage states. The edges represent flux definitions and specify a source node, a destination via a *target* field (another storage or an external sink such as outlet discharge Q or actual evapotranspiration E_a), and a process-equation identifier that resolves to a callable function in the component library. Optional modules are toggled via activation flags; when a module is deactivated, pruning removes it and redirects any incoming flux targets to its predefined bypass destination to preserve downstream connectivity. A simplified schematic of the overall pipeline (GP individual → model configuration → stepwise simulation) is shown in Fig. 2.

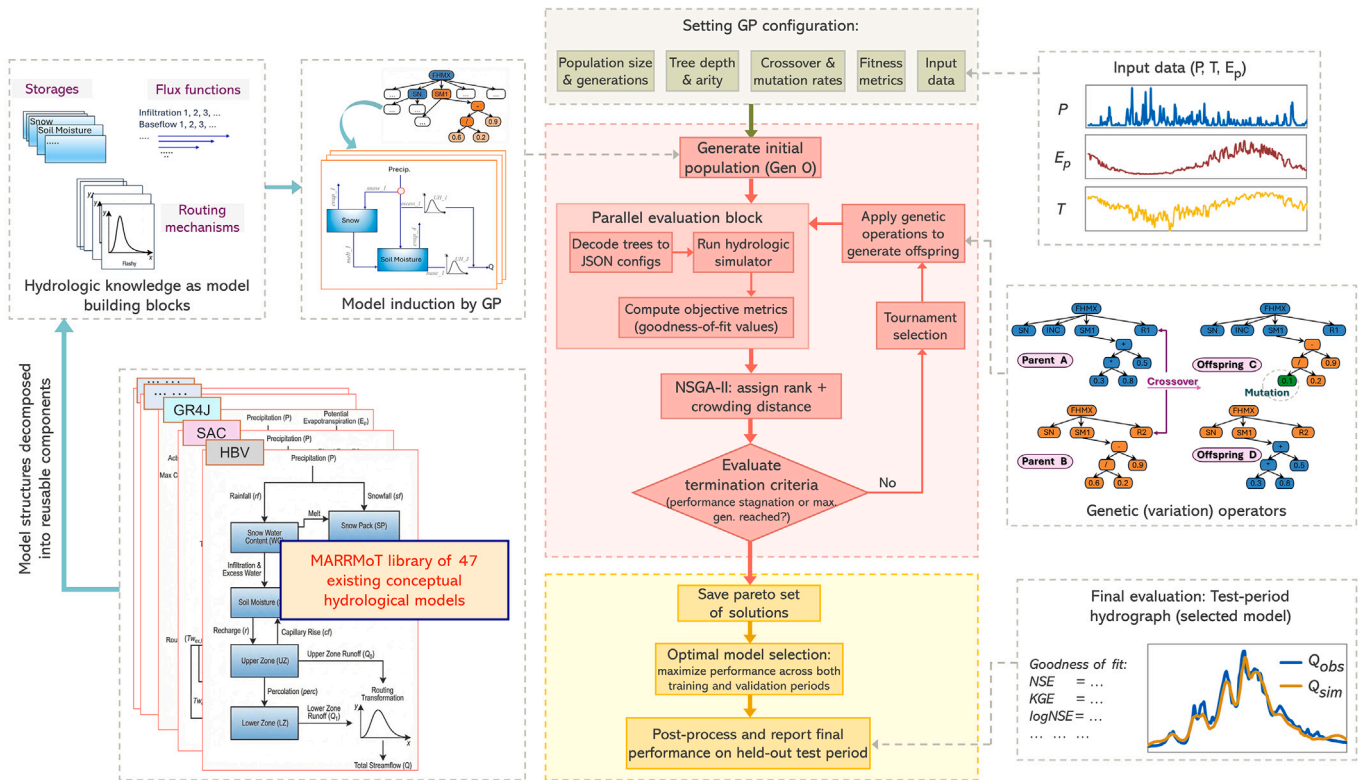


Fig. 1. Overview of HydroXGP workflow for conceptual hydrological model induction. Model structures are assembled from decomposed, reusable process components in the MARRMoT repository and evolved using genetic programming. Candidate models are evaluated using multi-objective performance metrics and ranked with NSGA-II; the resulting Pareto set is used to select a final model for evaluation on an independent test period.

The present study uses a non-exhaustive subset of MARRMoT flux and routing components. Because the master architecture excludes some structural patterns present across the full set of MARRMoT models, not all models are exactly reproducible within this search space. For example, the current architecture includes vertical transfers between the upper soil store and deeper storages and delayed contributions to streamflow within a closed catchment water balance, whereas more specialized MARRMoT representations—such as two-way water exchange between an aquifer and the river channel or between shallow and deep aquifers (MODHYDROLOG, m36; Chiew and McMahon, 1994) and groundwater gains from or losses to neighbouring catchments (GR4J, m07; Perrin et al., 2003)—are not included in the current implementation. Nevertheless, the framework is designed to be readily extensible with minimal effort and without modifying the simulator core: new flux equations or routing schemes can be added to the component library, and structural extensions (e.g., new storage types or connection patterns) can be introduced by editing the master architecture and associated bypass rules.

2.2. Optimization and model induction with GP

The optimization follows the evolutionary loop shown in Fig. 1. Each candidate model is encoded as a GP individual that specifies which process components are active, which alternative formulations are selected, and the associated parameter values. In our implementation, an individual consists of 66 arithmetic expression trees (one subtree per encoded entry). Each tree uses the function set $\{+, -, \times, /\}$ and terminals in $[0, 1]$, with tree depth limited to a user-defined maximum to prevent unbounded expression growth. The value 66 was not chosen a priori, but is the number of configurable decisions remaining after the selected subset of MARRMoT-derived components and wiring patterns was embedded in the master configuration used in this study (Table A.1; Fig. 3).

These components and connections were selected because they could be combined within a common executable architecture without making the graph representation overly complex. Once this master architecture is fixed, each candidate model is specified by 66 configurable entries, each representing one of three decision types: binary flags (active/inactive), categorical options (discrete choices), or continuous parameters (scaled to predefined bounds adopted from the original MARRMoT formulations). The normalized entries are first mapped to their model values according to the type of each entry (Table A.3), and the mapped values then instantiate an executable model configuration that is simulated as described in Section 2.3. This fixed-length encoding allows GP to evolve structure and parameters together, while controlling complexity by requiring all choices to map to the configurable entries defined by the master configuration. The master architecture constrains allowable structural connections and process orderings, preventing hydrologically implausible configurations outside the predefined architecture. Together, these constraints keep the evolutionary search feasible and support interpretation of the discovered models.

The search begins with an initial population of size N (here $N = 500$), generated to span diverse combinations of discrete structural decisions (e.g., module activations and alternative process/routing choices) and continuous parameters within predefined bounds. In each evolutionary generation, all candidates are evaluated by executing the corresponding model over the training period, after discarding a prescribed warm-up window (here a 365-day spin-up), and computing objective scores. Fitness evaluation is parallelized across CPU workers to reduce wall-clock time (i.e., elapsed real time) when assessing large populations over many generations. Candidates that fail during instantiation or simulation (e.g., solver non-convergence) are treated as infeasible and assigned strongly penalized, dominated objective values such that they are consistently disfavoured during selection. This biases the evolutionary search toward numerically feasible candidate models.

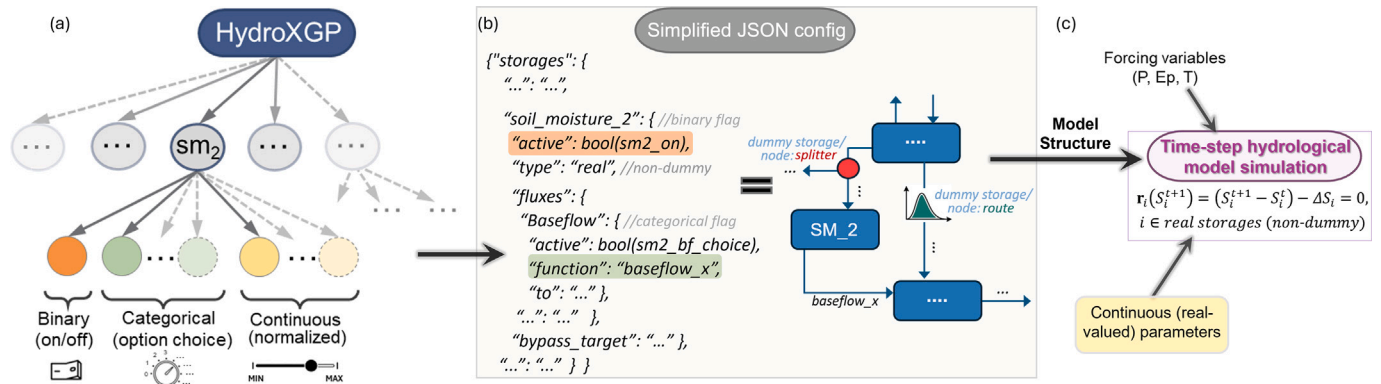


Fig. 2. From GP candidate to executable hydrological model in HydroXGP. (a) a GP individual is evaluated into a fixed-length normalized vector that maps to structural flags and parameter values. (b) these values instantiate an executable configuration graph within the master JSON configuration (architecture shown schematically in Fig. 3) via activation switches, process/routing selections, and bypass redirection of inactive components. (c) the assembled model is simulated stepwise under forcings (P , E_p , T) and the candidate parameter set; residuals/Jacobian are formed over real storages only, excluding dummy split/routing operators.

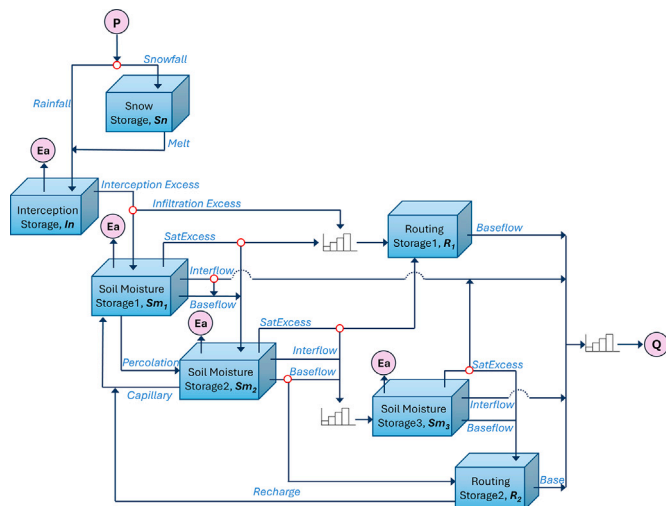


Fig. 3. Conceptual hydrological process space available to HydroXGP for rainfall-runoff model discovery, corresponding to the directed master configuration graph stored as JSON. Blue boxes denote real storages (dynamic states), red junction markers indicate dummy nodes used for flux splitting, and routing symbols denote unit-hydrograph routing operations applied to the corresponding fluxes. Arrows indicate potential process pathways within the master architecture; individual discovered models retain only the storages, fluxes, process formulations, and routing components selected through GP search. Process labels follow the terminology and functional roles used in the corresponding MARRMoT source models; Appendix Table A.1 lists the options available for each component. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The workflow supports both single-objective and multi-objective optimization; here we adopt a multi-objective formulation using three complementary discharge-based metrics: (i) Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970) to characterize overall hydrograph fit with sensitivity to high flows, (ii) Kling–Gupta efficiency (KGE; Gupta et al., 2009) to balance correlation, variability, and bias components, and (iii) log-transformed NSE (logNSE) to increase sensitivity to low-flow behaviour and recession dynamics (omitted for basins with frequent near-zero flows). All objectives are expressed in maximization form, with full definitions provided in Appendix Table B.1. While NSE remains one of the most widely used performance metrics in hydrologic modelling today (Melsen et al., 2025), calibration

based primarily on peak-flow variance can under-emphasize recessions and low-flow dynamics (Pushpalatha et al., 2012). Pareto-based multi-objective selection can constrain structural uncertainty and highlight balanced model structures that remain hydrologically consistent under contrasting conditions (Saavedra et al., 2022). This motivated us to adopt a multi-objective approach, as such approaches can encourage the evolution of candidate models that are robust across flow regimes (Chadalawada and Babovic, 2019).

Selection is performed using the non-dominated sorting genetic algorithm II (NSGA-II; Deb et al., 2002). After evaluation, NSGA-II assigns each candidate a Pareto rank (non-dominance level) via non-dominated sorting, and computes crowding distance within ranks to preserve diversity along the trade-off surface. Parents are selected using tournament selection, and offspring are generated by applying crossover and mutation to the candidate encoding: crossover recombines parental information, while mutation introduces random perturbations to promote exploration (for a more detailed overview of fundamental GP mechanics and operators, see Babovic and Keijzer, 2000). Variation operators respect parameter bounds and admissibility constraints, and complexity controls are applied to prevent excessive growth and keep evaluation cost tractable. Elitism is enforced through NSGA-II environmental selection on the combined parent-offspring population to form the next generation. The loop repeats for up to $G_{\max}$ generations (here $G_{\max} = 200$), with early stopping enabled if performance no longer improves over an independent validation period.

Upon termination, the outcome is a Pareto set of non-dominated candidate models representing different objective trade-offs. Here, “non-dominated” means that no model is strictly better than another across all objectives simultaneously; this set corresponds to the first Pareto front (Pareto rank 1, or Front 0 in 0-indexed implementations). To select a single model for reporting and subsequent analysis, a validation-based procedure is used. Model evolution is performed on the training period, while every n generations (here, $n = 1$), eligible candidate models are evaluated on an independent validation period using $S_{\text{val}} = w \cdot \text{NSE} + (1 - w) \cdot \bar{O}$, where w is a configurable weighting parameter and $\bar{O}$ denotes the mean of the remaining objectives (KGE and logNSE in this study; alternative or additional objectives can be specified within the framework). This score does not affect the Pareto-based evolutionary search, but is used for early stopping and for identifying the best-validation generation, from which the final model is selected by maximizing the mean of training and validation NSE, favouring solutions that perform consistently across both periods and generalize beyond the calibration window. Because a multi-objective search does not yield a uniquely preferred solution, this final selection step necessarily involves some modelling preference. A sensitivity analysis using $w = 0.5, 0.8$, and 1.0

revealed limited dependence on the weighting scheme: $w = 0.5$ and $w = 0.8$ selected identical models in structure, topology, and parameter values, while NSE-only weighting ($w = 1.0$) produced only minor topological differences; test-period NSE varied by less than 0.01 across the sensitivity analysis. However, the weighting involves user discretion rather than representing a uniquely correct choice and may be adjusted when a particular application places greater emphasis on specific aspects of streamflow behaviour (Wagener et al., 2026; Spieler and Stadnyk, 2026). A more systematic assessment of how alternative objective combinations and weighting schemes influence the selected model structures would require a dedicated investigation; within the scope of the present study, we adopted $w = 0.8$, assigning greater weight to NSE because it remains a widely recognized efficiency metric in hydrological modelling and provides direct comparability with the benchmark studies considered here, while retaining a non-zero contribution from the complementary objectives to discourage selection based on NSE alone. To account for run-to-run variability inherent in evolutionary search, the optimization is repeated with different random seeds. The held-out test period is not used during the model training phase and is instead used for out-of-sample evaluation of the completed models.

2.3. Simulation and graph-based sparsity

For each candidate produced by the evolutionary search, model evaluation requires assembling the corresponding executable configuration and simulating it over time. Algorithm 1 summarizes this procedure. An individual's parameters and structural flags are first applied to the master configuration to yield an executable directed graph. Active flux identifiers are mapped to process equations using the resolver $\mathcal{E}$, and the model is advanced by time stepping while recording discharge, evapotranspiration, and storage trajectories (with optional internal flux histories). If unit-hydrograph routing is enabled, routing states are initialized from the selected routing type and delay, then updated at each time step using routing inflows from active fluxes.

Real storages are advanced using a fixed-step implicit Euler update, and dummy nodes are excluded from the state vector. Let $\mathbf{S}_{\text{real},t} \in \mathbb{R}^n$ denote the real storage vector at time t . Let $\mathbf{u}_t$ denote the meteorological forcing vector, and let θ denote the parameter set. For a given candidate configuration, flux equations are evaluated at the current estimate of the state vector $\mathbf{S}_{\text{new}}$, with dummy nodes held fixed. Because dummy nodes are algebraic junctions rather than storage states, they are excluded from the solved state vector $\mathbf{S}_{\text{real,new}}$, and appear only in the flux evaluation used to form $\Delta \mathbf{S}_{\text{real}}(\cdot)$. The resulting inflows and outflows are aggregated into the net storage-change operator $\Delta \mathbf{S}_{\text{real}}$. End-of-step storages are obtained by solving

$$\mathbf{r}(\mathbf{S}_{\text{real,new}}) = \frac{\mathbf{S}_{\text{real,new}} - \mathbf{S}_{\text{real,old}}}{\Delta t} - \Delta \mathbf{S}_{\text{real}}(\mathbf{S}_{\text{new}}; \mathbf{u}_t, \theta) = \mathbf{0}, \quad (1)$$

where $\mathbf{S}_{\text{real,old}}$ denotes the previous-step storage vector. The backward-Euler discretization is chosen here for its robust stability properties (although coarse Δt can introduce numerical diffusion/damping that smooths rapid transients and attenuates sharp discharge peaks). Fluxes evaluated at $\mathbf{S}_{\text{new}}$ update $\Delta \mathbf{S}_{\text{real}}$ and are also accumulated into external outputs. These include outlet discharge Q_t and actual evapotranspiration $E_{a,t}$, depending on the configured destination targets of each flux definition.

Eq. (1) is solved using an iterative nonlinear solver for $\mathbf{S}_{\text{real,new}}$. The Jacobian $\mathbf{J} = \partial \mathbf{r} / \partial \mathbf{S}_{\text{real}}$ is typically sparse. Each residual component depends only on storages that are connected through the active flux graph. Candidate structures change during the search, so the sparsity pattern is derived automatically for each candidate (Fig. A.1 shows an illustrative example). A binary dependency pattern $P_{ij} \in \{0, 1\}$ is constructed from the pruned and redirected configuration graph, and dummy nodes are collapsed so that dependencies are expressed over the real storage vector. This pattern is used to build a sparsity-aware finite-difference approximation of the Jacobian. Finite differences perturb only storage-residual pairs flagged as potentially dependent. Known-zero Jacobian

entries are not evaluated or stored. This reduces the number of residual evaluations and the memory footprint per solver iteration. As a result, candidate simulations run faster, which is useful when evaluating large populations over many generations. Because the pattern is computed automatically from the configuration graph, the implementation remains consistent across candidates as structures evolve. This improves scalability and leaves more of the computational budget for exploring structural and parameter diversity rather than for solver overhead.

3. Study design and experimental setup

To evaluate the framework, we conduct three experiments spanning from basic feasibility to broader applicability. The first is a proof-of-concept test in a single basin. The second examines whether a structure learned in that basin can be transferred to a hydrologically similar catchment. The third evaluates the framework across basins spanning a wide range of hydroclimatic conditions, with comparisons against established conceptual and deep learning benchmarks.

3.1. Dataset and benchmark baselines

The experiments in this study use the CAMELS-US dataset (Newman et al., 2015; Addor et al., 2017). The publicly available dataset, hosted by the US National Center for Atmospheric Research (NCAR), consists of daily hydrometeorological forcings and observed streamflow time series, along with comprehensive static catchment attributes for 671 minimally impacted basins across the contiguous United States (CONUS). It is one of the most widely used datasets for large-sample hydrological modelling and machine learning studies (Feng et al., 2023; Botterill and McMillan, 2023).

To compare our framework with established hydrological modelling approaches, we use the CAMELS benchmark simulations assembled by Kratzert (2019a). This collection contains model outputs for the CAMELS-US basins from several widely used hydrological models, including SAC-SMA coupled with the Snow-17 snow routine (Burnash et al., 1973; Anderson, 1973), VIC (Liang et al., 1994), several FUSE structures (Clark et al., 2008; Henn et al., 2015), HBV (Seibert and Vis, 2012), and mHM (Samaniego et al., 2010; Kumar et al., 2013). The simulations themselves were originally configured, calibrated, and run by independent research groups in previous large-sample studies over the CAMELS catchments (Newman et al., 2017; Mizukami et al., 2017, 2019; Seibert et al., 2018; Rakovec et al., 2019). Kratzert (2019a) collected these existing runs and converted them into a common per-basin format for benchmarking. Alongside the conceptual models, we also include the LSTM and entity-aware LSTM (EA-LSTM) results reported by Kratzert (2019b) as data-driven baselines.

In this benchmarking dataset, all conceptual and LSTM baseline models were calibrated or trained on the period 1 October 1999–30 September 2008 and evaluated on the period 1 October 1989–30 September 1999. We adopt the same train-test split to ensure a fair comparison with these baselines. In addition, we reserve the period 1 October 1980–30 September 1989 as a validation window for early stopping and model selection from the Pareto front. An identical validation period is used by Kratzert et al. (2024) to select the optimal epoch in the LSTM training phase. All benchmark models use the Maurer forcings (Wood et al., 2002) from CAMELS, and we adopt the same forcings in our experiments. Because our framework is built from MARRMoT components, it uses the same standard climate inputs: time series of precipitation P , potential evapotranspiration E_p , and air temperature T . Storages were initialized to zero, with a 365-day warm-up prior to evaluation.

3.2. Experiment 1: Single basin proof-of-concept test

The first experiment pilots HydroXGP by evaluating whether it can recover hydrologically consistent behaviour in a single catchment. This controlled setting allows us to examine the learned structure, parameter dynamics, and internal process representations. The selected catchment

Table 1

Comparison of key characteristics for the source basin and the hydrologically similar target basin used in Experiment 2.

Basin	Area (km ²)	Mean elev. (m)	Mean precip. (mm d ⁻¹)	Aridity (–)	Snow fraction (–)	Mean discharge (mm d ⁻¹)	FDC slope (–)
13011900	852	2683	2.38	1.14	0.70	1.47	0.80
13313000	562	2156	3.63	0.79	0.74	1.41	0.76

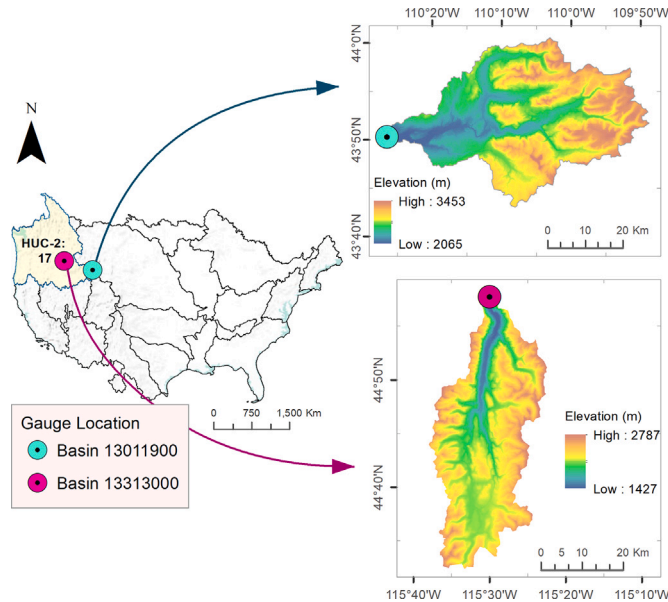


Fig. 4. Geographic setting of the two basins used in Experiment 2: Buffalo Fork above Lava Creek near Moran, Wyoming (13011900), and Johnson Creek at Yellow Pine, Idaho (13313000). Gauge locations are shown within CONUS and HUC2 region 17, and the corresponding basin DEMs are shown at right. DEMs are mosaicked from SRTM (NASA, 2013) 1 arc-second-resolution raster tiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

is Buffalo Fork above Lava Creek near Moran, Wyoming (Basin gauge ID 13011900), which has a drainage area of 851.8 km² and a well-defined seasonal flow regime driven by snowmelt. This basin was also used as an illustrative example in Kratzert et al. (2024) due to its strong forcing-response relationship, making it a useful reference for this study.

To account for stochasticity in the search process, we run five independent optimizations with different random seeds under identical settings and retain the best-performing realization for analysis, consistent with established hydrological calibration practice (Newman et al., 2015; Kratzert et al., 2018). We also examine parameter uncertainty around the best solution. First, we apply a Generalized Likelihood Uncertainty Estimation (GLUE) analysis (Beven and Binley, 1992) to assess parameter sensitivity and possible equifinality in this basin. In parallel, we run a small parameter fine-tuning step with the discovered structure fixed to assess whether additional calibration of the continuous parameters changes model behaviour in any meaningful way.

3.3. Experiment 2: Knowledge transfer to a hydrologically similar basin

The second experiment tests how well a structure learned for the Buffalo Fork basin (13011900) can be reused in a hydrologically similar basin. To select a suitable *target* catchment, we searched the CAMELS dataset for basins similar to Buffalo Fork in size, hydroclimatic regime, and physical characteristics. We first restricted candidates to the same broad hydrological region and to catchments of comparable scale. Within this screened set, we further required similar snow

influence and then computed a weighted Euclidean distance across a set of topographic, climatic, geological, soil, vegetation, and hydrological attributes (including elevation, slope, aridity, snow fraction, runoff ratio, and baseflow index). Among the closest basins identified through this procedure was Johnson Creek at Yellow Pine, Idaho (Basin gauge ID 13313000), which has a comparable drainage area (561.9 vs. 851.8 km²) and closely matches Buffalo Fork in two key hydrologic controls (Table 1): snow fraction (0.745 vs. 0.705) and mean streamflow (1.414 vs. 1.470 mm day⁻¹). At the same time, the basin is located several hundred kilometres away from Buffalo Fork (though within the same hydrologic region), providing sufficient geographic separation while still representing a closely related hydrologic setting (Fig. 4). We therefore used the Johnson Creek basin (13313000) as the transfer target.

Using the structure and parameters obtained in Experiment 1, we evaluated four transfer strategies with increasing adaptation: (i) direct transfer without modification, (ii) parameter-only recalibration with the structure fixed, (iii) a short *hot-start* GP search seeded with the source solution, and (iv) a full independently initialized GP search on the target basin as the strongest baseline. Together, these strategies test how transferable the discovered structure is, how much adaptation is required, and whether hot-starting offers advantages over learning from scratch.

3.4. Experiment 3: Framework application across diverse hydroclimatic regions

The third experiment evaluates framework performance across selected basins throughout CONUS. We use the USGS Hydrologic Unit Code level-2 (HUC2) classification, which divides the United States into 22 major water-resources regions, 18 of which lie within CONUS. These regions span diverse hydroclimatic settings, from humid low-elevation basins in the South Atlantic–Gulf to snow-dominated mountain catchments in the Pacific Northwest and the arid interior West. They are widely used in large-scale hydrologic studies because they group basins according to broad physiographic and climatic controls (Collins et al., 2022; Hughes et al., 2024).

To keep this experiment computationally manageable while still including every major region, we select one representative basin from each HUC2 region using a medoid-based approach. Here, a ‘medoid’ is defined as the basin whose attribute vector is, on average, closest to all other basins within the same region; that is, the basin that minimizes the total Euclidean distance to all others. Because HUC2 regions can contain substantial heterogeneity in topography, climate, soils, vegetation, and geology, the selected medoid should be viewed as a central point in the regional attribute space rather than a fully ‘typical’ basin. Each basin is described using 27 static attributes drawn from the CAMELS dataset, following the subset used in several large-sample hydrological modelling studies (Nai et al., 2024; Espinoza et al., 2025). These attributes span multiple environmental domains, including topography, vegetation, soils, geology, and climate. All attributes are standardized using a robust transformation (subtracting the median and dividing by the normalized median absolute deviation), and pairwise Euclidean distances (L2 norms) between all basins within each region are then computed in this scaled attribute space. The medoid basin for region r is defined as

$$\text{medoid}_r = \arg \min_{i \in r} \sum_{j \in r} \|z_i - z_j\|_2 = \arg \min_{i \in r} \sum_{j \in r} \sqrt{\sum_{k=1}^{27} (z_{ik} - z_{jk})^2}, \quad (2)$$

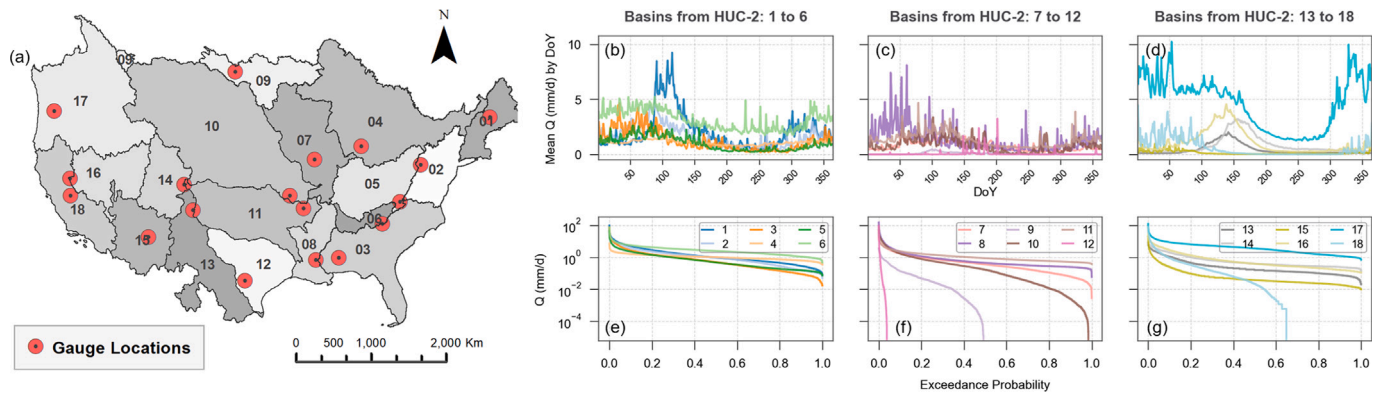


Fig. 5. Locations and discharge characteristics of the selected basins across the 18 CONUS HUC2 regions. (a) Spatial distribution of basin gauges. (b–d) Seasonality of mean daily discharge by day of year (DoY). (e–g) Flow-duration curves (FDCs) showing discharge exceedance probability. The climatological period is 1 October 1980–30 September 2008. Panels (b–g) are grouped by HUC2 indices (1–6, 7–12, and 13–18) for visualization only.

where z_i and z_j are the scaled attribute vectors for basins i and j , respectively, z_{ik} and z_{jk} denote their k th components, k indexes the 27 static attributes, and $i, j \in r$. This procedure provides a transparent and reproducible way to choose a single basin per region, while recognizing that a single basin will only approximate, and may not fully represent, the heterogeneity present within a HUC2 region.

The 18 selected basins span a wide range of hydroclimatic settings (Fig. 5, Supplementary Table S3), including low-lying and high-mountain catchments, humid and semi-arid climates, snowmelt-dominated and rainfall-dominated regimes, and strongly intermittent systems. The selected medoids therefore provide a compact but varied test bed for Experiment 3. For each medoid basin, we apply the same optimization protocol, search space, training period, population size, number of generations, and algorithm settings. We run five independent optimizations with different random seeds to account for stochasticity, as in Experiment 1. This uniform setup enables controlled comparison of framework performance across hydroclimatic regions.

4. Results and discussion

All results in this section (metrics and diagnostic plots) are computed on the held-out test period (1 Oct 1989–30 Sep 1999; out-of-sample). Benchmark comparisons are made over the same test window. Training/calibration-period results appear only where explicitly labelled.

4.1. Automated discovery of interpretable, hydrologically consistent structures

Experiment 1 tested whether HydroXGP could identify a suitable model structure for the Buffalo Fork catchment. Results suggest that it recovered a parsimonious structure with good predictive skill, process-based interpretability, and mass conservation.

4.1.1. Benchmarking predictive skill

The best-performing HydroXGP-derived structure achieved strong out-of-sample performance on the held-out test period (1989–1999), with $NSE = 0.849$, $KGE = 0.879$, and $\log NSE = 0.886$ (Fig. 6(a)). The simulated–observed scatter (Fig. 6(b)) shows a strong linear association ($R = 0.930$) and a regression line close to the 1:1 line, indicating that the model reproduces discharge magnitude with limited systematic bias over the test period. On their own, these scores would usually be considered high (Moriasi et al., 2007). However, in snow-dominated catchments, high efficiency values can sometimes be obtained by simple predictors because much of the discharge variability follows a repeatable seasonal cycle (Schaeffli and Gupta, 2007; Knoben et al., 2019b). For that

reason, Knoben (2024b) argues that model skill should be judged relative to simple benchmark models rather than interpreted in isolation. To test whether the HydroXGP-derived structure adds information beyond the catchment's climatology, we compare it against an ensemble of simple benchmark models generated with the HydroBM toolkit (Knoben, 2024a) for the same basin and period. This ensemble includes, among others, seasonality-only climatology baselines (e.g., daily mean flow) and precipitation-based baselines that use long-term rainfall–runoff ratio scaling.

Across the test period, the discovered HydroXGP model outperforms the seasonality-only daily climatology benchmark ($NSE = 0.58$) by $\Delta NSE \approx 0.27$ and exceeds all HydroBM benchmark models (NSE ranging from -0.36 to 0.73), indicating predictive skill beyond that explained by the seasonal cycle or other simple benchmark formulations. The high $\log NSE$ (0.886) further indicates that performance extends beyond high flows to recessions and low-flow dynamics, suggesting that the model captures subsurface-release behaviour during dry periods—a common challenge in conceptual modelling (Staudinger et al., 2011). Robustness of model performance across stochastic optimization is summarized in Supplementary Table S1; over five independent GP runs, test-period performance remains consistently high (NSE 0.722–0.849), suggesting that the search repeatedly finds high-performing regions of the structural space rather than relying on a single favourable initialization. In terms of solver behaviour, failures or non-convergence occurred predominantly in the initial random population ($\sim 25\%$) in generation 1, with all subsequent generations maintaining 500/500 valid individuals. Overall, infeasible evaluations contributed only ≈ 0.10 – 0.27% of total evaluations across runs.

4.1.2. Learned structure, process consistency, and uncertainty

A core motivation for process-based modelling, unlike many purely data-driven approaches, is that internal states can be inspected for hydrologic realism (Beven, 2020). The HydroXGP-discovered model (Fig. 8(a)) has interpretable internal states, and its simulated trajectories (Fig. 7) show plausible seasonal behaviour: snow accumulates through winter and depletes during spring melt, while soil moisture exhibits wetting and drying cycles broadly consistent with precipitation and seasonal evapotranspiration demand. These internal trajectories were not compared against independent observations of snow storage, soil moisture, or actual evapotranspiration, and are therefore interpreted as diagnostic checks on hydrological consistency rather than as evidence of process realism. In addition, the mass balance error remains negligible throughout the simulation (Fig. 7(e)), providing an internal consistency check.

Across the five independent GP runs, the selected models showed partial structural convergence rather than recovery of a unique model topology. Examination of the resulting structures showed a recurring

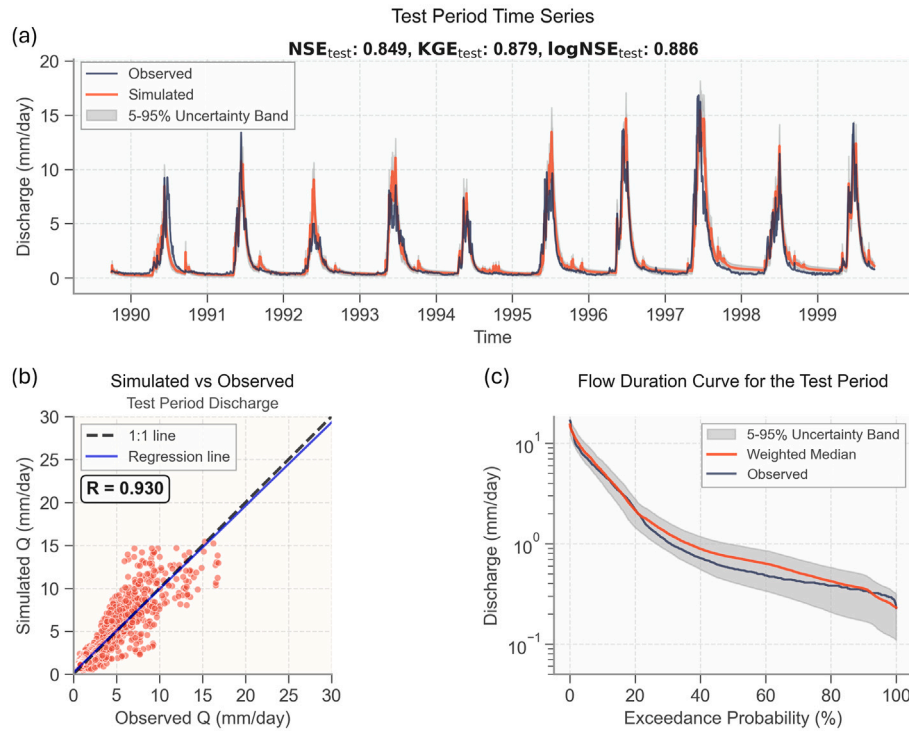


Fig. 6. Predictive performance over the held-out test period for Experiment 1 (Buffalo Fork; gauge 13011900). (a) Observed and simulated discharge, with NSE, KGE, and logNSE reported. (b) Simulated versus observed discharge, showing the 1:1 line and fitted regression line. (c) Flow-duration curve (FDC) with GLUE uncertainty. The grey band shows the 5–95% uncertainty interval on the test period from 5000 behavioural parameter sets accepted during calibration ($NSE_{train} \geq 0.8$).

snow–soil storage configuration across all five runs, consistent with the snowmelt-dominated hydrological regime of the basin (individual model structures shown in Supplementary Fig. S1). However, the high-performing structures differed in the activation of additional storages and in their routing and near-surface loss representations. Across pairwise comparisons among the five runs, the fraction of storage-activation decisions that agreed between two models averaged 0.74 (range 0.57–0.86), while similarity of the active process topology, quantified using the Jaccard index, averaged 0.46 (range 0.33–0.64). Parameter stability was also heterogeneous. Among parameters present in all five models, the degree-day factor and temperature threshold varied over only 6% and 14% of their respective allowed calibration ranges, whereas the primary soil-storage capacity spanned 90% (see Supplementary Fig. S2 for the full run-to-run stability analysis and metric definitions). This behaviour is consistent with the long-recognized problem of equifinality in conceptual hydrological modelling, whereby multiple structure–parameter combinations can provide similarly acceptable representations of the observed rainfall–runoff response (Clark et al., 2008; Khatami et al., 2019; Spieler et al., 2020). In the present case, the limitation appears to be that discharge alone is insufficient to uniquely identify all internal process representations, because outlet discharge provides only an integrated catchment response. In contrast to the more consistently recovered structural features, the remaining structural and process choices show greater variability and are only weakly identifiable under discharge-only evaluation. Importantly, all of these alternative configurations still achieved high predictive performance across the training and test periods ($NSE_{train} = 0.85$ – 0.89 ; $NSE_{test} = 0.72$ – 0.85 ; Supplementary Table S1), so the present results do not provide a basis for treating one routing or loss formulation as physically superior or as a more realistic representation of the catchment. Discriminating among them more rigorously would require additional constraints beyond outlet discharge, such as independent observations of internal states or fluxes. This limitation also points to an important

direction for future work, in which such observations could be incorporated as diagnostic constraints or additional calibration targets to assess whether discovered structures reproduce not only outlet discharge, but also process-relevant variables such as snow water equivalent, soil moisture, and actual evapotranspiration.

To complement efficiency scores, Table 2 summarizes a small set of hydrologically interpretable diagnostic signatures. The discovered model satisfies water balance to numerical precision: the maximum mass balance residual is 2.88×10^{-7} mm/d and the cumulative closure error is effectively zero, consistent with Fig. 7(e). Runoff partitioning between runoff generation and baseflow contribution is broadly consistent with observations, as reflected in similar runoff ratios and baseflow-index proxies.

Fig. 6(c) summarizes uncertainty in the test-period FDC using bands derived from behavioural parameter sets accepted during calibration ($NSE_{train} \geq 0.8$) and propagated to the held-out test period. These bands reflect parametric uncertainty across the 11 parameters of the discovered structure. The likelihood-weighted median closely follows the observed FDC across most exceedance probabilities, while the 5–95% band widens toward the low-flow tail, indicating greater uncertainty during sustained recessions and low-flow periods. The associated parameter–performance plots (Supplementary Fig. S3), constructed from the same behavioural set, show that snow and soil parameters concentrate in relatively narrow high-performance regions, whereas lag/routing parameters remain more diffuse, consistent with the wider spread in the low-flow tail. Further continuous-parameter fine-tuning of the best-discovered structure using Dynamically Dimensioned Search (DDS; Tolson and Shoemaker, 2007) yielded only marginal gains in training NSE (+0.0017) and slightly poorer test-period performance ($\Delta NSE = -0.0038$, $\Delta KGE = -0.0094$, $\Delta logNSE = -0.0078$), suggesting that the GP search had already identified a robust parameterization for the discovered topology and that further calibration mainly induced mild overfitting.

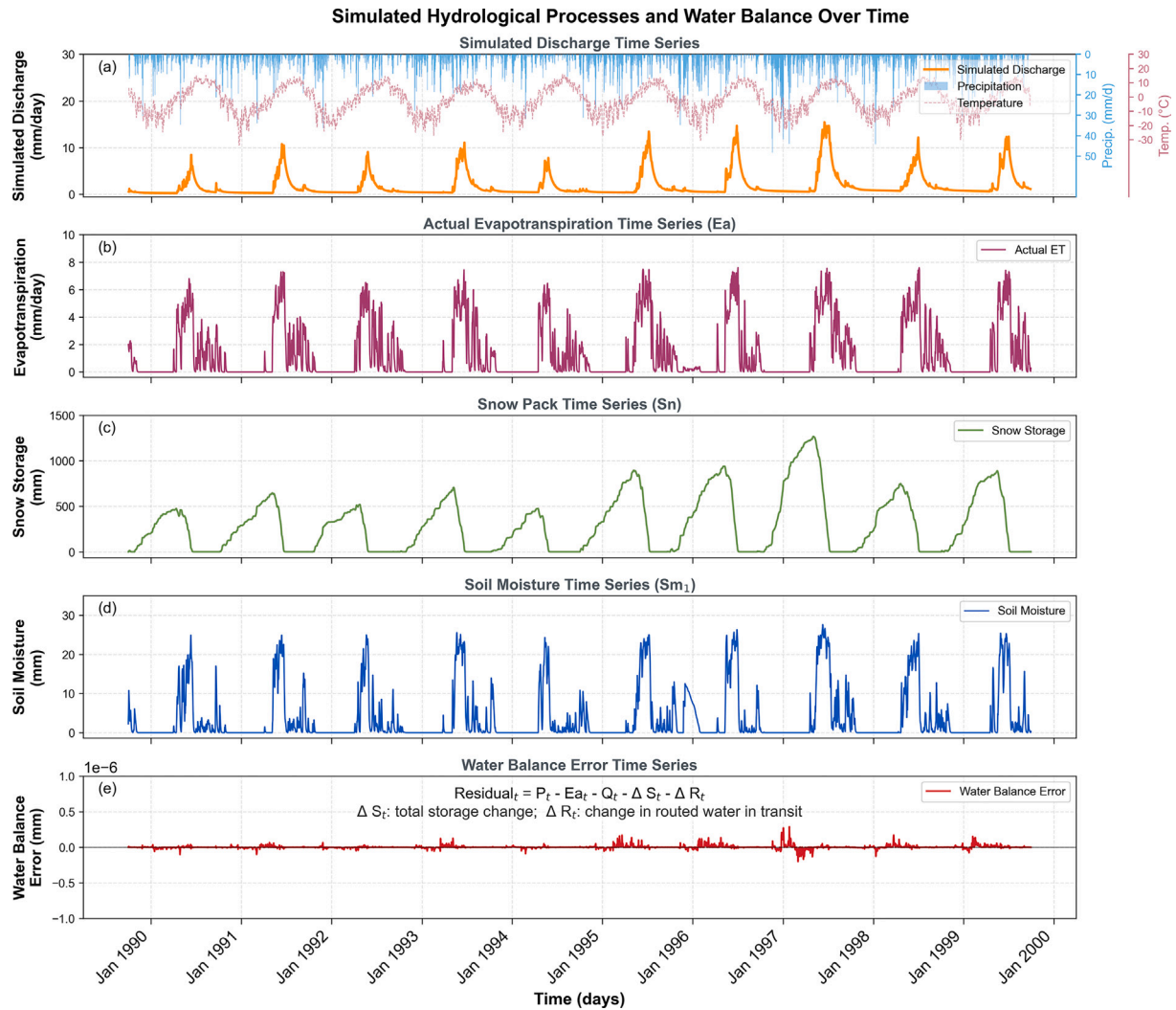


Fig. 7. Simulated internal states and mass balance over the out-of-sample test period for Experiment 1. Panels show (a) simulated discharge with meteorological forcing (precipitation and temperature), (b) simulated actual evapotranspiration, (c) simulated snow storage, (d) simulated soil moisture storage, and (e) water-balance error over time for the model shown in Fig. 8(a). The time series provide a consistency check on seasonal snow storage dynamics and numerical mass conservation.

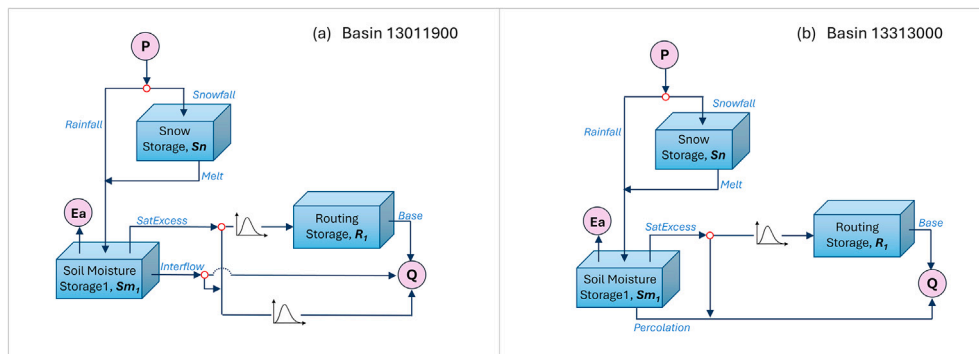


Fig. 8. Independently discovered structures for the two hydrologically similar basins used in Experiment 2: (a) Buffalo Fork (13011900) and (b) Johnson Creek (13313000). Structures are obtained from separate full GP optimizations on each basin and illustrate convergence toward a similar snow–soil–routing backbone.

4.2. Cross-basin transferability and optimization efficiency

Experiment 2 assesses cross-basin reuse by comparing direct transfer, parameter-only recalibration, and short-budget GP runs with and without hot-start initialization when moving from Buffalo Fork to Johnson

Creek—two basins with broadly comparable hydroclimatic attributes. It also tests whether transferred structural knowledge can accelerate evolutionary search in the target basin, analogous to transfer learning in deep learning (Ma et al., 2021; Yao et al., 2023). While cross-catchment reuse has been explored in other forms (Heřmanovský et al., 2017), we

Table 2

Hydrologic signature diagnostics for Experiment 1 over the out-of-sample test period; observed-discharge diagnostics are shown in parentheses where available.

Diagnostic	HydroXGP
Maximum mass balance residual, $ r_i $ (mm d ⁻¹)	2.88×10^{-7}
Mass balance closure, $\sum r_i / \sum P$ (%)	8.31×10^{-8}
Runoff volume bias, $100(\sum Q_{\text{sim}} - \sum Q_{\text{obs}}) / \sum Q_{\text{obs}}$ (%)	8.94
Runoff ratio, $\sum Q / \sum P$ (-)	0.566 (Obs: 0.519)
Baseflow index proxy, $\text{BFI} = \sum Q_b / \sum Q$ (-)	0.615 (Obs: 0.590)
Recession slope median, $\Delta \log Q$ (d ⁻¹)	-0.0133 (Obs: -0.0338)
Peak timing error (days; MAE)	1.80
Peak magnitude error (%; MAPE)	14.07
High-flow bias (FHV, %)	1.65
Low-flow bias (FLV, %)	-7.72
Mid-slope bias (FMS, %)	-12.39

Peaks: 10 largest observed peaks (matching window ± 3 d). BFI: Lyne and Hollick (1979) (three-pass, $\alpha = 0.925$; Ladson et al., 2013). Recession: median $\Delta \log Q$ on falling-limb days with $P \leq 0.01$ mm d⁻¹. FHV/FLV/FMS: FDC diagnostics (Yilmaz et al., 2008) (top 2%, bottom 30%, mid 20–70%).

are not aware of prior rainfall–runoff studies that hot-start GP-based structure discovery by seeding the target-basin search with donor-basin structures.

4.2.1. Baseline transfer and short-budget refinement

Direct transfer, in which the optimal structure and parameters from the source basin are applied to the target basin without recalibration, provides a ‘zero-shot’ baseline for structural adequacy. As detailed in Table 3, this uncalibrated transfer achieved a test-period NSE of 0.724, KGE of 0.677 and a logNSE of 0.715. This transfer is non-trivial because the daily precipitation series are only moderately correlated between basins (Pearson’s $r = 0.51$), meaning the forcings are distinct enough that the model is not simply memorizing an identical signal. This reduces the likelihood that performance is driven by near-identical inputs. The fact that the model retained substantial predictive skill without ever seeing the target data suggests that the GP optimization on the source basin successfully identified a reusable structural hypothesis that generalizes beyond the donor catchment, rather than overfitting to Buffalo Fork’s discharge series. Some degree of transferability is expected due to basin similarity, but donor-model skill also plays a role here (for context, transfer performance of the simple HydroBM benchmark models spans $\text{NSE} = -0.27$ to 0.61—below that achieved by the HydroXGP-derived model). This donor-to-target transfer is relevant to prediction in ungauged basins (PUB), where streamflow observations are often limited or absent and model structure must be informed by hydrologic understanding transferred from similar donor sites (Sivapalan, 2003; Hrachowitz et al., 2013). Retaining predictive skill during direct transfer to Johnson Creek suggests that the evolved model captures dominant, transferable hydrological behaviour. Although Johnson Creek is a validation case rather than a truly ungauged basin, the results suggest that structure-based transfer can provide a viable and interpretable starting point for ungauged modelling.

We subsequently tested whether this performance could be improved through strictly numerical calibration. Parametric fine-tuning was performed using DDS (with an evaluation budget of 5,500 iterations, i.e., 500 times the number of parameters being optimized). The transferred structure was held fixed while the continuous parameters were optimized using the target data. While the initial search/selection was multi-objective, the subsequent DDS-based parametric fine-tuning used a single calibration objective (NSE). The parameter refinement strategy yielded some improvement, but the gain was marginal. This suggests that the remaining error signal is structural in nature and cannot be resolved by parameter calibration alone.

However, the value of the transferred structure becomes more evident when it is used to seed the evolutionary search as a prior. When

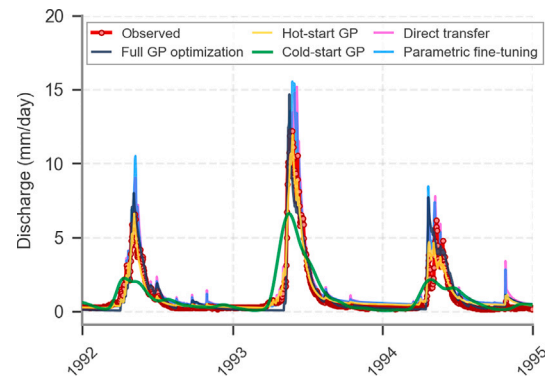


Fig. 9. Observed and simulated hydrograph for Experiment 2 showing a subset of the held-out 10-year test period. Direct transfer and short-budget hot-start refinement closely track the observed hydrograph and approach the performance of full GP optimization. In contrast, cold-start GP (random initialization), despite using the same short optimization budget, is visibly the poorest performer.

GP was initialized with a random population (cold-start) and restricted to a short budget of five generations, performance remained relatively low ($\text{NSE} = 0.581$), and it even underperformed the direct-transfer approach. Visual inspection of the hydrograph in Fig. 9 confirms that the cold-start model poorly captures peak timing and magnitude, consistent with a search that has not yet reached a high-performing region of the topological space. In contrast, the hot-start strategy, where the initial population was seeded with the structure transferred from the source basin, allowed the search to begin from an informed rather than a fully random candidate population and achieved a test NSE of 0.838 under the same five-generation budget. This represents a substantial performance gain ($\Delta \text{NSE} \approx +0.26$). By starting the search with a hydrologically plausible topology, the evolutionary algorithm can reduce expensive early exploration and focus more quickly on refining secondary processes to match local conditions. Fig. 9 also shows that hot-start closely approaches the full-optimization simulation, while direct transfer captures overall timing but deviates more in event magnitudes; parametric fine-tuning produces modest changes relative to direct transfer.

4.2.2. Convergent evolution and computational efficiency

When the framework was allowed to optimize fully on the target basin (Johnson Creek) without any prior knowledge of the source basin, it converged on a topology broadly consistent with that of the source basin (Fig. 8). Despite starting from random initialization, independent searches in both catchments identified a similar snow-soil-routing backbone. This cross-basin similarity suggests that some structural features are consistently favoured for representing the dominant runoff dynamics in these hydrologically similar catchments. Minor structural differences, such as the arrangement of the routing delay, may reflect basin-specific characteristics or equifinality in secondary processes, where the rainfall–runoff signal is insufficient to distinguish among subtle structural variants.

Table 3 further summarizes the skill–cost trade-off across all transfer and adaptation strategies. The efficiency of hot-start is evident when compared with the full GP optimization baseline, which was run for 200 generations. All experiments were executed under identical hardware configurations to make wall-clock runtime comparisons meaningful. Despite the large difference in search budget, hot-start achieved nearly 96% of the predictive skill of the full run ($\text{NSE} 0.838$ vs. 0.877) while requiring only about 3.7% of the computational time (412 s vs. 11,213 s). The efficiency gain from hot-starting is conceptually similar to pre-training: information learned in the donor basin is reused to initialize the target-basin search. Here, the transferred ‘knowledge’ is structural (topology and parameters) rather than neural network weights. In this setting, donor-basin solutions help constrain the hypothesis space

Table 3

Comparison of model skill and computational efficiency across alternative transfer/adaptation strategies for target basin (Johnson Creek; gauge 13313000) in Experiment 2. Results on the held-out test period are shown together with the corresponding optimization budget (generations or iterations) and wall-clock runtime.

Strategy	Budget	Run/optimization (s)	NSE _{test}	KGE _{test}	logNSE _{test}
Direct transfer (no optimization)	–	5.38	0.7235	0.6767	0.7150
Parametric fine-tuning (fixed structure)	5500 iterations	10,060.59	0.7502	0.6824	0.7131
Cold-start GP (random initialization)	5 generations	374.19	0.5813	0.6283	0.6408
Hot-start GP (seeded initialization)	5 generations	412.19	0.8376	0.8473	0.8028
Full GP optimization	200 generations	11,213.41	0.8771	0.8966	0.7578

Runtime environment: Intel Xeon 8452Y; 64 CPU cores used (no GPU); 8 GB RAM requested; peak memory usage <2 GB. Detailed computational resource-usage records are provided with the accompanying analysis data cited in the Data Availability section.

and accelerate model identification when full from-scratch discovery is computationally infeasible.

The runtimes in Table 3 highlight the practical constraint imposed by the computational cost of full GP optimization for structure discovery across large basin samples. However, the substantially lower runtime achieved by the hot-start strategy suggests a possible pathway towards larger-scale deployment: full structure discovery could be reserved for selected representative or data-rich basins, while transferred structures initialize shorter searches in hydrologically similar target basins. This could support computationally tractable scaling of HydroXGP to larger basin sets.

4.2.3. Further tests on transferability

To expand the transfer analysis beyond a single source–target pair, we performed two additional experiments, keeping the Buffalo Fork model as the donor and using geographically more distant basins from different HUC2 regions as targets. The targets include Taylor River, Colorado (09107000; HUC2–14), also a snow-influenced basin but with higher elevation, lower mean precipitation, and lower discharge than Buffalo Fork, and Satilpa Creek, Alabama (02469800; HUC2–03), a much lower-elevation, rain-dominated basin with no snow influence and substantially higher mean precipitation. The additional test results (Supplementary Table S2) show that direct transfer is strongly dependent on donor–target similarity. Direct transfer retained reasonable test performance for Taylor River (NSE = 0.66), but performed poorly for Satilpa Creek (NSE = 0.23), indicating that a structure calibrated for a snow-dominated watershed does not necessarily generalize directly to a climatically distinct catchment. Hot-start GP optimization improved both cases under the same short search budget of five generations, increasing NSE to 0.82 for Taylor River and 0.66 for Satilpa Creek. The larger improvement observed for Satilpa Creek suggests that, even when direct transfer performance is limited, the transferred structure can still provide a useful starting point for adaptation to local conditions. However, the stronger hot-start performance for Taylor River indicates that good donor selection remains important. These results support a limited but useful interpretation of HydroXGP’s transferability, which is that direct reuse is most effective between hydrologically similar catchments, while transferred structural knowledge through hot-start can substantially reduce the search effort required to identify good models even when the target basin is less similar.

4.3. Generalization across selected representative basins

Using the 18 HUC2 medoid basins as a representative set, we assess the framework’s performance consistency and generalization behaviour across the contrasting hydroclimatic regimes covered by these selected basins. In Experiment 3, we compare performance across basins, interpret basin-wise differences through the lens of regime characteristics, and identify where additional process representations may be needed.

4.3.1. Distribution of performance and robustness

Table 4 provides three complementary views of cross-basin performance: central tendency (median), average behaviour (mean),

and a robustness or shortfall indicator – the count of basins with metric values at or below 0.5. Although rigid evaluation thresholds have known limitations (Clark et al., 2021; Melsen et al., 2025), a score of 0.5 remains a widely recognized community benchmark for satisfactory model performance (Moriassi et al., 2007; Pumo and Noto, 2023), and is advocated as a stricter requirement than simply achieving non-negative skill (e.g., $NSE \geq 0$) (Coxon et al., 2019). Because the analysis spans only 18 basins, these summaries should be interpreted as indicative and diagnostic rather than statistically decisive. To quantify sampling uncertainty, we report nonparametric bootstrap 95% confidence intervals (CIs) for the medians (10,000 resamples of basins with replacement). HydroXGP attains a median NSE of 0.702 (95% CI: [0.651, 0.767]), a median KGE of 0.754 ([0.659, 0.810]), and a median logNSE of 0.701 ([0.670, 0.765]). Among the conceptual benchmarks, only basin-calibrated mHM reaches a comparable median NSE (0.705), while all other conceptual models fall below 0.67. The two static-input LSTM variants achieve higher median NSE (0.773 and 0.755), consistent with the benefits of regional training combined with static catchment attributes. This corresponds to a median NSE performance loss of 0.053–0.071 relative to the attribute-augmented LSTM variants. However, as a trade-off for the modest reduction in median accuracy, HydroXGP provides an explicit, interpretable structure. In contrast, the LSTM trained without static inputs attains a lower median NSE (0.685), suggesting that HydroXGP’s per-basin process structure can match or exceed the skill obtained from meteorological forcings alone. For a complementary pairwise perspective, we also report comparisons against the best representative baseline from each benchmark family, selected as the model with the highest median NSE across the 18 basins (mHM-basin for conceptual models and the static-input LSTM for deep learning; Supplementary Fig. S6).

A more informative comparison comes from the mean and the count of basins with metric values at or below 0.5, both of which are sensitive to tail-end breakdowns. Under NSE, HydroXGP has zero basins below 0.5, but this is also true for the two static-input LSTM variants (Table 4). Within the conceptual family, the best basin-calibrated models (HBV-upper and the FUSE variants) still exhibit two basins below the $NSE = 0.5$ threshold, while several conceptual baselines, particularly the CONUS-calibrated variants, show frequent breakdowns (e.g., VIC-CONUS has 16 basins with $NSE \leq 0.5$). Fig. 10 makes this regime sensitivity visually clear: HydroXGP sits at the upper end of the conceptual distributions with a relatively narrow spread, whereas multiple conceptual baselines (especially regionally parameterized ones) show both lower central tendency and substantially wider dispersion. Even basin-calibrated fixed-structure models retain non-trivial tail failures (e.g., VIC: 4 and SAC-SMA: 3 basins with $NSE \leq 0.5$), indicating that parameter calibration alone does not necessarily prevent regime-mismatch breakdowns when the structural hypothesis is fixed.

The spread statistics further emphasize this point. Using the basin-wise results, HydroXGP has the lowest cross-basin standard deviation for KGE (0.163; next-lowest SAC-SMA at 0.232) and logNSE (0.328; next-lowest FUSE-902 at 1.723, more than five times larger). For NSE, HydroXGP’s standard deviation (0.110) is comparable to the two best

Table 4

Performance comparison in Experiment 3 between HydroXGP and conceptual and LSTM benchmark models across the 18 selected HUC2 basins (one per region) over the held-out test period (1 Oct 1989–30 Sep 1999). For each efficiency metric, the median, mean, and number of basins with performance ≤ 0.5 are reported.

Model	NSE _{test}			KGE _{test}			logNSE _{test}		
	Median	Mean	$n(\leq 0.5)$	Median	Mean	$n(\leq 0.5)$	Median	Mean	$n(\leq 0.5)$
HydroXGP	0.702	0.706	0	0.754	0.717	2	0.701	0.617	4
HBV (lower)	0.469	0.367	11	0.392	0.092	13	0.550	−2.011	8
HBV (upper)	0.666	0.672	2	0.717	0.555	1	0.590	−1.469	7
SAC-SMA	0.615	0.589	3	0.653	0.585	5	0.585	−0.125	7
VIC (basin)	0.584	0.545	4	0.607	0.553	3	0.590	−1.037	7
VIC (CONUS)*	0.326	0.232	16	0.255	−0.109	15	0.333	−2.139	14
mHM (basin)**	0.705	0.670	2	0.717	0.509	1	0.721	−1.611	5
mHM (CONUS)**	0.521	0.403	6	0.534	0.194	8	0.425	−2.203	9
FUSE (900)	0.629	0.612	2	0.646	0.516	3	0.493	−1.360	9
FUSE (902)*	0.657	0.632	2	0.712	0.634	3	0.675	0.134	7
FUSE (904)	0.632	0.616	2	0.648	0.536	2	0.397	−0.878	12
LSTM (no static input)*	0.685	0.628	1	0.677	0.533	3	0.749	−1.079	6
EA-LSTM (static input)*	0.755	0.751	0	0.815	0.500	1	0.829	−1.600	3
LSTM (static input)*	0.773	0.765	0	0.800	0.563	1	0.838	−1.170	2

* One basin's result is missing from the dataset.

** Two basins' results are missing from the dataset.

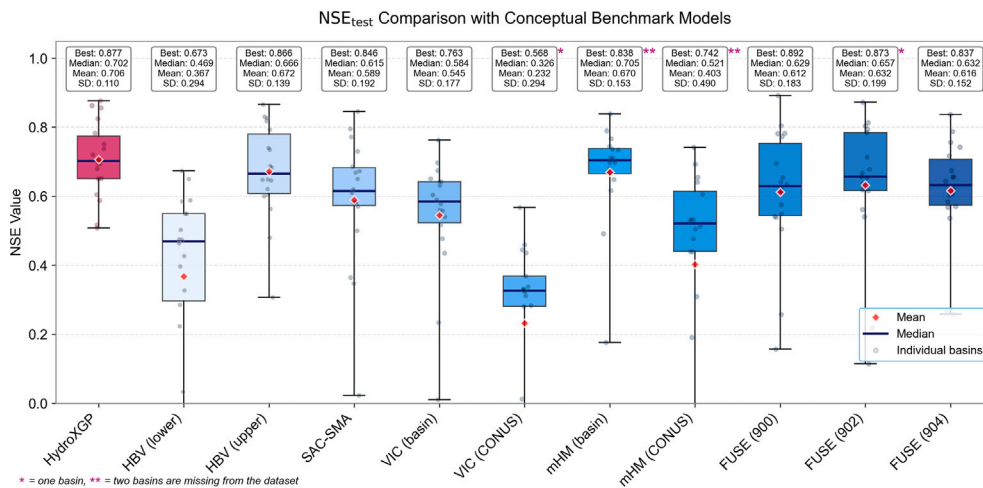


Fig. 10. NSE comparison between HydroXGP and conceptual benchmark models across the 18 representative basins. Boxplots summarize the across-basin distribution, highlighting HydroXGP's consistently high central tendency (mean/median) and low across-basin variability (SD) relative to the conceptual benchmarks. Asterisks indicate missing basin results due to unavailable benchmark-model outputs for those basins.

LSTM variants (0.100 and 0.103) and smaller than every conceptual benchmark. These patterns are visible in the boxplots of Figs. 10 and 11 and Supplementary Figs. S4–S5, where HydroXGP shows a narrow interquartile range and comparatively short lower whiskers, while several benchmark models show long lower tails. As a consequence, HydroXGP's mean closely tracks its median (mean–median gaps of +0.004 for NSE, −0.037 for KGE, and −0.084 for logNSE), whereas several benchmarks exhibit much larger gaps driven by a small number of extreme basins (e.g., EA-LSTM KGE: median 0.815 vs. mean 0.500; LSTM-static logNSE: median 0.838 vs. mean −1.170).

As described in Section 2.2, logNSE was omitted from the optimization objectives wherever frequent near-zero flows made the metric unsuitable for training. This affected four basins from HUC2–09, HUC2–10, HUC2–12, and HUC2–18; however, these basins were still included in the test-period evaluation results in Table 4. Three of the four had logNSE values below 0.5, indicating that these basins pulled down the aggregate HydroXGP logNSE summary. In practical terms, the static-input LSTMs achieve higher typical (median) performance, but low-flow-sensitive evaluation reveals heavier tails that strongly depress the mean. HydroXGP does not dominate the LSTMs on low-flow metrics

in typical basins, but it exhibits a different robustness profile, with more bounded tails under KGE/logNSE and the highest mean KGE (0.717) and mean logNSE (0.617) among the models in Table 4; only FUSE-902 also has a positive mean logNSE (0.134), while all other benchmarks are negative. Overall, the distributional results indicate that one distinguishing feature of HydroXGP in Experiment 3 is stability across basins: its mean performance remains close to its median because it avoids the extreme outliers that drive down mean KGE and logNSE in several benchmark families. This does not imply uniformly superior typical skill—static-input LSTMs achieve higher medians, but it does suggest that the HydroXGP framework's process library and search constraints provide a degree of robustness when moving across regimes. The basin-wise comparisons in Fig. 11 and the diagnostic cases discussed next clarify where this robustness is most beneficial, and where low-flow-sensitive behaviour remains the primary challenge.

4.3.2. Conceptual benchmarks: structural rigidity versus adaptive structures

The spatial bar charts in Fig. 12(a) and (b) show that HydroXGP exceeds the conceptual benchmark mean in all 18 basins for NSE and KGE. For logNSE (Fig. 12(c)), HydroXGP outperforms the mean conceptual

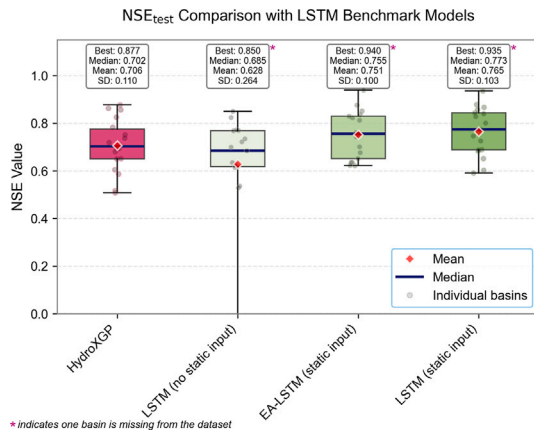


Fig. 11. NSE comparison between HydroXGP and LSTM benchmark variants across the 18 representative basins. LSTM with no static inputs underperforms HydroXGP, whereas the static-input variants achieve higher performance. HydroXGP exhibits a modest loss in skill relative to the best static-input LSTM benchmarks, while providing an interpretable, process-based structure. Asterisks indicate missing basin results due to unavailable benchmark-model outputs for those basins.

benchmark in 16 of 18 basins. The two exceptions are Rio Pueblo de Taos (HUC2-13) and Walker Creek (HUC2-05), where HydroXGP falls slightly below the conceptual mean. Because no single conceptual or LSTM benchmark is consistently best across basins and metrics (Supplementary Table S5), we use the benchmark-family mean as a stable basin-wise reference. Across basins, HydroXGP's improvements over the conceptual benchmark mean are not marginal. Averaged over the 17 basins with available comparisons (excluding HUC2-12 Seco Creek, where the conceptual benchmark logNSE is an extreme negative outlier), the mean gains are +0.17 in NSE, +0.18 in KGE and +0.31 in logNSE. However, the magnitude of these gains is not uniform and varies with hydroclimatic setting, as Fig. 12 suggests.

Two notable cases are Augusta Creek (HUC2-04) and Cherry Creek (HUC2-15), which occupy very different parts of the regime space. Augusta exhibits unusually persistent flow with a flat FDC and weak seasonality (Fig. 5), consistent with strong storage-release control. In this setting, insufficient groundwater memory is heavily penalized and matters more than fine seasonal timing. The mean conceptual benchmark NSE at Augusta is only 0.14, and several fixed-structure models produce negative NSE values (e.g., HBV-lower -0.46, mHM-CONUS -1.36), indicating a severe mismatch in persistence representation. HydroXGP achieves an NSE of 0.74, an improvement of +0.60 over the conceptual mean—the largest single-basin margin in the ensemble. This strong performance is consistent with the discovered model structure, which combines slow drainage pathways with delayed routing, supporting sustained flow during recessions. Cherry Creek, in contrast, is among the most water-limited basins (high aridity and low mean discharge; Supplementary Table S3). For this basin, flows are very low on most days (CAMELS-US daily Q , 1980–2008: median $Q \approx 0.036 \text{ mm d}^{-1}$) but not persistently zero (minimum $Q \approx 0.009 \text{ mm d}^{-1}$), consistent with a system that retains some baseflow yet exhibits pronounced event-driven variability. Here, the mean conceptual benchmark NSE is 0.45 while HydroXGP reaches 0.72 (gain $\approx +0.27$), and the conceptual-mean logNSE is negative (≈ -0.15) while HydroXGP remains positive (≈ 0.68), indicating improved low-flow sensitivity in a regime where fixed structures struggle to reconcile very low baseflows with episodic peaks (Fig. 12(a) and (c)).

Large relative gains also appear in highly intermittent systems. Seco Creek (HUC2-12) is a near-ephemeral flash-flood basin (median $Q = 0 \text{ mm d}^{-1}$; zero-flow days 96.22%). Under NSE, several conceptual models are effectively non-functional (e.g., SAC-SMA 0.02, VIC-basin

0.01), and the mean conceptual benchmark NSE is only 0.15. HydroXGP achieves an NSE of 0.51, a gain of +0.36 over the conceptual mean—still modest in absolute terms, and therefore better interpreted as stress-test performance rather than high predictive skill. The low-flow metric highlights why such basins dominate distributional behaviour: the mean conceptual benchmark logNSE at Seco is -28.25, while HydroXGP's logNSE is -0.50. HydroXGP is still poor in absolute low-flow fidelity here, but it remains bounded while the conceptual mean collapses under log-space evaluation. This single basin shows why conceptual-family mean logNSE values can become strongly negative even when medians remain moderate (Table 4; Supplementary Fig. S4). More broadly, this regime dependence is not unique to the benchmark set considered here: evaluations of national-scale distributed systems such as NOAA's National Water Model (NWM) also show weaker low-flow streamflow skill in parts of the arid western and southwestern U.S. (Hughes et al., 2024). A similar but less extreme pattern appears in Willow Creek (HUC2-09), another strongly intermittent basin (median $Q = 0 \text{ mm d}^{-1}$; zero-flow days 50.83%). HydroXGP improves substantially over the conceptual mean across all three metrics (e.g., NSE 0.52 vs. 0.29, KGE 0.40 vs. 0.18, logNSE 0.37 vs. -0.42), yet the absolute values remain only moderate, reflecting the inherent difficulty of modelling systems that are effectively dry for much of the year. In contrast, in 'well-behaved' humid rainfall-dominated basins the advantage is smaller but still consistent: across the rainfall-dominated set (HUC2-02, 03, 05, 07, 08, 10, 11), HydroXGP's NSE gain over the conceptual benchmark mean ranges from about +0.07 to +0.18 (Fig. 12(a)). These basin-wise patterns suggest that HydroXGP yields modest gains in typical humid basins, while the largest margins occur in regimes dominated by persistence/storage control or by strongly water-limited low-flow behaviour, consistent with the added flexibility of adapting structure to local hydrologic controls.

4.3.3. Deep learning baselines: accuracy-robustness trade-offs across regimes

The LSTM results of Kratzert et al. (2019), which we use as our deep learning baseline, were trained regionally on hundreds of CAMELS basins using the same Maurer forcings and the same train-test split as the conceptual benchmarks and HydroXGP. These configurations are implemented in the open-source NeuralHydrology library, which has become a common reference implementation for CAMELS-based deep learning studies (Lees et al., 2021; Clark and Jaffrés, 2025). Although this is a regional model and HydroXGP is trained per basin, it provides the most directly comparable single-forcing baseline under a consistent setup. This choice is conservative because regional training is not a disadvantage for LSTM rainfall-runoff models; rather, it is generally the preferred and stronger setup. Kratzert et al. (2024) demonstrated that, unlike traditional hydrological models, LSTMs typically achieve higher skill when trained across many basins than when trained separately on individual basins. The comparison is therefore made against a strong LSTM baseline. More recent studies (Kratzert et al., 2021) use similar architectures but combine multiple meteorological forcings (Daymet, Maurer, NLDAS), and are therefore not directly comparable to our single-forcing experiments; accordingly, we rely on Kratzert et al. (2019) for quantitative comparison.

Across the basins with available LSTM results (Willow Creek, HUC2-09, is not included in the Kratzert et al. (2019) set), Fig. 12 shows that the averaged LSTM benchmark (mean over the three variants) typically attains higher NSE than HydroXGP. This likely reflects the closer alignment of LSTM training with NSE, together with regional exposure to behaviours that are only sparsely represented in HydroXGP's per-basin training data and not yet well captured by its current process library, especially ephemeral, threshold-driven dynamics. The LSTM baselines exceed HydroXGP in certain humid eastern basins (Carrabasset River, HUC2-01; Walker Creek, HUC2-05), at Rio Pueblo de Taos (HUC2-13), and at Seco Creek (HUC2-12). At Seco Creek, the mean LSTM NSE (0.61) is higher than HydroXGP's (0.51), despite weak conceptual benchmark performance overall (mean NSE of

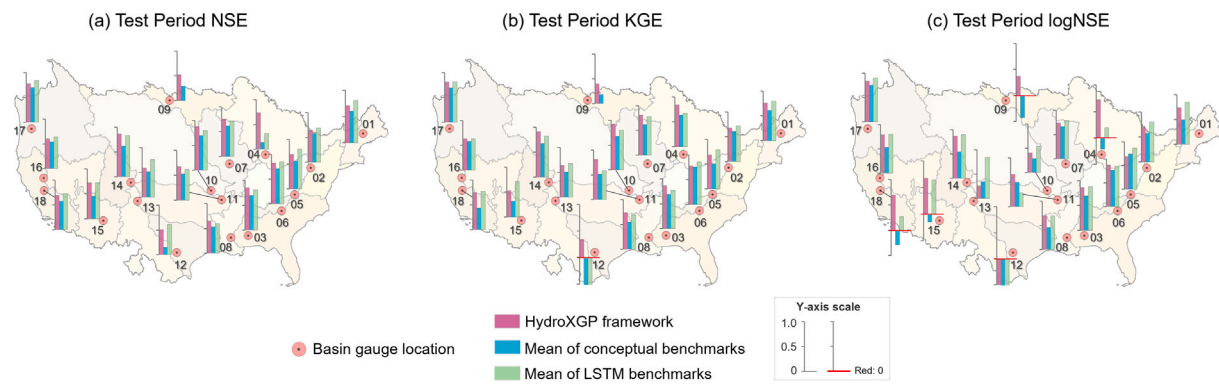


Fig. 12. Spatial comparison of basin-wise performance across the 18 representative HUC2 basins for (a) NSE, (b) KGE, and (c) logNSE, showing HydroXGP alongside the mean conceptual benchmark and mean LSTM benchmark performance at each basin. HydroXGP exceeds the mean conceptual benchmark for nearly all basins and metrics, with exceptions only for logNSE in HUC2-05 and HUC2-13, while performance relative to the mean LSTM benchmark is mixed across regions.

0.15). This difference could arise from the fact that the regional training across hundreds of basins provides sufficient exposure to strong intermittent and event-driven dynamics to learn generalizable threshold activation and event timing patterns that a continuous-storage process library may not yet express explicitly. The basins where HydroXGP exceeds the LSTM mean are therefore particularly informative, as they highlight conditions under which per-basin structural adaptation can compensate for the LSTM's data advantage. The largest HydroXGP gain in NSE occurs at Augusta Creek (HUC2-04; +0.42), the same persistence-dominated basin where conceptual models collapse; notably, the mean LSTM NSE there is only 0.32, with the no-static-input variant at -0.33 . HydroXGP also exceeds the mean LSTM at Cedar Creek (HUC2-10), North Fork River (HUC2-11), Homochitto River (HUC2-08), and Taylor River (HUC2-14). These basins span contrasting regimes—from humid rainfall-driven catchments with differing low-flow persistence (HUC2-08, HUC2-10, HUC2-11) to a cold, high-elevation basin with a dominant snowmelt season (HUC2-14).

For logNSE, the LSTM mean exceeds HydroXGP in most basins, and the differences can be large in specific cases, such as the Rio Pueblo de Taos basin within HUC2-13 (high-elevation, high aridity, strongly water-limited, with snowmelt-driven flow and summer monsoon spikes). In contrast, under KGE, HydroXGP exceeds the mean LSTM performance in a larger fraction of basins. The difference likely comes from objective alignment, since the LSTMs are trained with squared-error losses (MSE, or basin-averaged NSE) rather than KGE or low-flow objectives, whereas the HydroXGP models are selected through a multi-objective search that includes these criteria (although this alignment appears to benefit KGE more clearly than logNSE). As a result, LSTM training does not directly penalize systematic bias or variance mis-scaling even when timing and correlation are strong. By design, KGE penalizes these effects through its bias and variability-ratio terms, whereas NSE is dominated by correlation and high-flow timing. In a regional LSTM, good event timing can therefore coexist with basin-specific scaling or water-balance errors, while the per-basin GP search is additionally constrained by storage–flux relationships, which can stabilize bias and variability across regimes. This is consistent with HydroXGP's comparatively low cross-basin dispersion in KGE (SD 0.163), suggesting more stable bias/variability behaviour across basins even when NSE favours the LSTM.

4.3.4. Limitations and expanded robustness assessment

From a scalability perspective, the high computational demand poses a practical challenge for applying HydroXGP to large basin samples. Across the 18 medoid basins, the median wall-clock runtime per independent GP run was 13,752 s (3.82 h), limiting the feasibility of extending the 200-generation setup to hundreds or thousands of basins. As demonstrated in Section 4.2.2, transferring previously discovered

structures to hydrologically similar basins and using them to hot-start shorter GP searches provides one practical way to reduce this cost. The framework already supports parallel evaluation of individuals within each GP population and early stopping once the convergence criterion is reached; these capabilities can be leveraged to improve computational efficiency, particularly when larger populations and additional CPU resources are available. Further reductions in computational demand could be explored through surrogate-assisted optimization and progressive reduction of the search space in later generations, focusing the search on promising structural configurations. The scalability of such staged deployment and optimization strategies merits further investigation in future large-sample HydroXGP studies.

Model performance was notably weaker in some of the dry and strongly intermittent basins. In such settings, slowly responding storage components and groundwater processes can require longer periods to adjust, raising the question of whether the 365-day warm-up used in this study was sufficient. A one-year warm-up is commonly used in lumped conceptual rainfall–runoff modelling (De la Fuente et al., 2023), although previous studies have shown that the warm-up required for state convergence can vary across model structures, initial conditions, and catchments (Rahman and Lu, 2015; Kim et al., 2018), suggesting that no single warm-up duration is universally applicable. To assess sensitivity to warm-up length, we conducted a post hoc analysis in which the selected final models were re-simulated with longer warm-up periods. For each simulation, NSE was calculated over the same fixed test period used in the original evaluation. Extending the warm-up from 1 to 5 years had very little effect on test-period NSE, with changes no greater than 5.3×10^{-3} across all basins (Supplementary Table S4). We also assessed convergence across all conceptualized storage states. Some basins—including the strongly intermittent Seco Creek (HUC2-12) and Willow Creek (HUC2-09)—required more than 365 days for at least one storage state to reach convergence. However, the delayed convergence of these stores had little influence on test-period NSE, indicating that insufficient warm-up is unlikely to explain the weaker performance in these basins. The remaining limitations may instead reflect other factors, including structural constraints in the current model search space and the need to represent a broader range of hydrological processes. These findings therefore reinforce the motivation for enriching the HydroXGP library/search component set with additional threshold-based flux formulations and representations of processes such as channel transmission losses and groundwater exchange, which may be important for capturing intermittent and ephemeral flow dynamics (Shanfield and Cook, 2014; Gutiérrez-Jurado et al., 2019). Nevertheless, our sensitivity analysis was performed only on the final selected models and therefore did not directly assess whether using longer warm-up periods during training would affect the training-period objectives that drive model search and selection. The influence of warm-up length and alternative

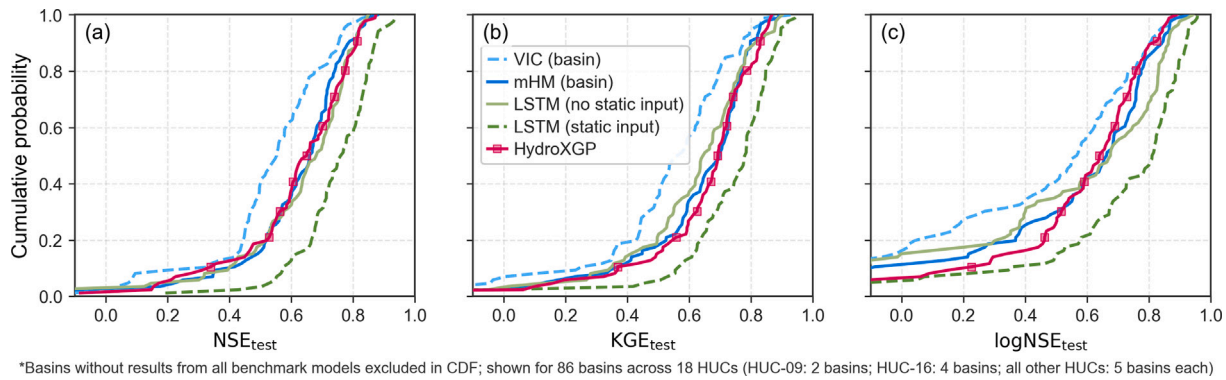


Fig. 13. Cumulative distribution functions (CDFs) of test-period (a) NSE, (b) KGE, and (c) logNSE for HydroXGP and selected benchmark models across the expanded robustness assessment.

state-initialization procedures on the optimization outcome could be examined more comprehensively in future work across basins and multiple independent runs.

Another limitation of Experiment 3 is that each HUC2 region was represented by a single medoid basin. This design was adopted to keep the experiment compact and computationally efficient while still covering all the major hydroclimatic regions. It also allowed basin-level interpretation of the results, including the identification of catchment characteristics in which HydroXGP performed relatively well or poorly, such as humid rainfall-dominated basins, snow-influenced catchments, arid systems, and highly intermittent low-flow regimes. However, as noted earlier, a single medoid basin can only approximate the regional attribute distribution and cannot fully represent the heterogeneity within a HUC2 region.

A related consideration concerns the sensitivity of the medoid-selection procedure to the choice of attribute scaling. Although the medoids were selected using a deterministic and reproducible method (Section 3.4), alternative standardization schemes can alter the geometry of the attribute space and therefore influence which basins are identified as the most representative. In the baseline analysis, attributes were transformed using a robust z-score based on median centering and median absolute deviation (MAD) scaling. This choice was intended to provide a stable representation of the attribute space, given the heterogeneous units, skewed distributions, and occasional extreme values present in CAMELS attributes. To assess whether the resulting medoids were unduly sensitive to this scaling choice, we repeated the complete HUC2 medoid analysis using classical z-score standardization and min-max normalization. The resulting medoid rankings were broadly consistent across methods. For 17 of the 18 HUC2 regions, the top-five candidate sets obtained from robust z-score scaling shared at least one basin with the corresponding top-five candidates identified using both alternative approaches, with a mean overlap of ≈ 3 basins per region. Furthermore, the robust-z-score rank-1 medoid typically remained among the highest-ranked candidates under the alternative scalings. The largest disagreement occurred in HUC2-10, where rank-1 medoids and top-five candidates diverged most across all three scalings, suggesting regional ambiguity rather than a single-method artefact. Overall, the sensitivity analysis indicates that medoid selection is not hypersensitive to the choice of scaling, although multiple attribute-similar candidate basins may exist within some regions.

To address these concerns, we performed an expanded robustness assessment using the five highest-ranked medoid candidates from each HUC2 region under the baseline robust z-score transformation. Candidate selection was carried out within the subset of 448 CAMELS basins for which complete benchmark results were available, giving a theoretical maximum of 90 basins across the 18 HUC2 regions. However, this subset contained only two basins from HUC2-09 and four basins

from HUC2-16; all other HUC2 regions contained at least five basins. The final CDF comparison therefore includes 86 basins. To keep the analysis computationally feasible, HydroXGP was run with a reduced optimization budget relative to the main Experiment 3 setup: 100 generations instead of 200, and three independent runs instead of five. The benchmark simulation results were kept unchanged from the original compiled dataset.

Fig. 13 presents cumulative distribution functions (CDFs) of NSE, KGE, and logNSE across the expanded basin set. To provide a more stringent comparison than benchmark-family means alone, the figure includes selected individual benchmarks: mHM (basin) and the LSTM with static inputs, which were the best-performing conceptual and LSTM benchmarks by median NSE in the original 18-basin analysis, respectively; and VIC (basin) and the LSTM without static inputs, which provide lower-performing references from each benchmark family. The expanded analysis broadly supports the conclusions from the original medoid-based experiment. HydroXGP outperforms VIC (basin) across the distribution and remains competitive with both mHM (basin) and the LSTM without static inputs. At the same time, the LSTM with static inputs remains the strongest overall performer. The predictive accuracy–interpretability trade-off between the LSTM with static inputs and HydroXGP was also evident in the earlier results. However, it is difficult to define a universal, quantitative accuracy threshold at which this trade-off becomes acceptable; rather, its acceptability depends primarily on the modelling objective and intended application of the model (Beven, 2020). For purely operational forecasting, where predictive accuracy is the primary objective, higher-performing LSTM models may be preferable, and sacrificing accuracy for interpretability may not be warranted (Nearing et al., 2021). On the other hand, for applications requiring transparent process representation, structural diagnosis, or the generation of testable hydrological hypotheses, a modest reduction in predictive skill may be justified by the additional interpretability afforded by an explicit process-based model structure.

The expanded CDF comparison indicates that the main performance patterns are stable when the evaluation is broadened beyond one basin per HUC2 region, strengthening the original medoid-based conclusions. Nevertheless, this analysis should be interpreted as a broader robustness check rather than as a comprehensive large-sample assessment across the full CAMELS basin set.

5. Conclusion

Hydrologically constrained genetic programming can serve as a process-informed machine learning approach, offering a practical middle ground between flexible conceptual modelling and modern data-driven methods. By embedding hydrologic building blocks and mass-conservation constraints within the model search space and the simulator, model discovery becomes neither purely manual hypothesis

selection nor unconstrained statistical fitting, but rather a structured exploration of plausible process configurations. In this sense, our proposed HydroXGP framework treats model structures themselves as learnable, inspectable hypotheses, rather than fixed templates or opaque parameterizations.

The experiments indicate that when the dominant runoff-generation mechanisms are represented within the component library, the evolutionary search tends to recover parsimonious backbones that are consistent with hydrological expectations. In snow-influenced settings, the recurrent emergence of a snow–soil–routing architecture without any direct constraint on seasonal state trajectories is consistent with the idea that process priors, coupled with stepwise simulation, may guide discovery toward plausible mechanisms. Beyond this shared backbone, alternative secondary formulations often remained competitive, with differences most apparent in routing and low-flow behaviour—areas where different formulations remained hard to distinguish because they produced similar outlet discharge hydrographs. The transfer results further support the interpretation that discovered structures capture more than basin-specific calibration artefacts: architectures evolved in one basin retained useful skill in hydrologically similar basins and, when used as priors, provided efficient starting points for subsequent adaptation. Reusing structural hypotheses offers a pragmatic way to scale discovery by starting from an informed candidate and refining it for the target basin, rather than starting from scratch, while still allowing basin-specific adjustment. This also suggests a potential pathway for data-scarce or ungauged applications, where transferred structures could be treated as prior hypotheses and updated as limited local information becomes available.

Across diverse hydroclimatic regimes, allowing structure to adapt within hydrologically constrained limits reduced the frequency of severe model breakdowns relative to fixed conceptual templates. Nevertheless, the framework remains bounded by the expressiveness of its component library. In highly intermittent or event-dominated systems, where threshold behaviour and transmission losses are critical, existing conceptual models already struggle. Expanding the library may improve representation of the dominant dynamics, but may not be sufficient on its own to fully resolve these limitations. Enriching the set of admissible flux formulations and activation mechanisms is a logical next step, but it raises the question of how to balance expressiveness against identifiability. A richer hypothesis space may capture more behaviours, yet it may also amplify equifinality unless additional observational constraints are incorporated.

This work points toward a modelling paradigm in which structural hypotheses are not static choices but evolving artefacts that can be revised as new data, diagnostics, or process understanding becomes available. Such evolution should not be interpreted as automated replacement of expert judgement, but as a systematic extension of it. By preserving interpretability and explicit structure definitions, HydroXGP enables transparent comparison, transfer, and refinement of competing explanations. In doing so, it contributes to a more iterative and evidence-driven exploration of model structure—one that retains the conceptual foundations of hydrology while embracing the adaptive capabilities of modern machine learning.

CRedit authorship contribution statement

Naila Matin: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis. **Herath Mudiyansele Viraj Vidura Herath:** Writing – review & editing, Visualization, Supervision, Methodology, Conceptualization. **Abhishek Saha:** Writing – review & editing, Visualization, Supervision, Methodology, Conceptualization. **Lucy Marshall:** Writing – review & editing,

Supervision, Conceptualization. **Vladan Babovic:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

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

Acknowledgements

The authors are grateful to the National Research Foundation of Singapore and PUB (Singapore's National Water Agency) for supporting this work done in the Coastal Protection and Flood Resilience Institute (CFI) Singapore under the Coastal Protection and Flood Management Research Programme (CFRP). The conclusions put forward in this publication reflect the views of the authors alone and not necessarily those of CFI Singapore or any of the supporting entities.

Computational resources for this study were partially supported by the [National University of Singapore](#) Information Technology (NUS IT) Research Computing group under grant number [NUSREC-HPC-00001](#).

Appendix A. Library of hydrologic building blocks and model assembly

Algorithm 1 Configuration-driven model assembly and stepwise simulation.

- 1: **Input:** master config C_0 , parameters/flags θ , forcings $\{P_t, E_{p,t}, T_t\}_{t=1}^T$
- 2: **Structures:** $\mathcal{L}$ (flux library), $\mathcal{E}$ (resolver), $\mathcal{A}$ (active flux definitions), $\mathcal{R}$ (routing/UH states), $\mathbf{P}$ (Jacobian sparsity pattern)
- Setup:*
- 3: $C \leftarrow \text{LOAD}(C_0)$
- 4: $C \leftarrow \text{MODIFYFROMFLAGS}(C, \theta)$
- 5: $C \leftarrow \text{PRUNEANDREDIRECT}(C)$
- 6: $\mathcal{A} \leftarrow \text{ACTIVEFLUXDEFS}(C)$ {store active flux configs; callables resolved via $\mathcal{E}$ }
- 7: $\mathbf{P} \leftarrow \text{DERIVESPARSITYPATTERN}(C)$ {binary dependencies $P_{ij} \in \{0, 1\}$ for $J = \partial \mathbf{r} / \partial \mathbf{S}_{real}$; dummy nodes collapsed}
- 8: initialize states $\mathbf{S}$ for active storages (define $\mathbf{S}_{real}$ excluding dummy nodes)
- 9: initialize routing states $\mathcal{R}$ for each routing flux (from routing-type, delay, Δt)
- 10: initialize nonlinear solver with pattern-aware finite-difference Jacobian (using $\mathbf{P}$); fallback solver if needed
- Time stepping:*
- 11: **for** $t = 1$ **to** T **do**
- 12: $\mathbf{S}_{real}^{old} \leftarrow \mathbf{S}_{real}$
- 13: define residual function $\mathbf{r}(\mathbf{S}_{real}^{new}) \in \mathbb{R}^n$:
- 14: reconstruct full $\mathbf{S}^{new}$ (dummy nodes fixed)
- 15: $(\Delta \mathbf{S}_{real}, Q_t, E_{a,t}, \text{inputs}^{route}, \text{hist}^{flux}) \leftarrow \text{EVALFLUXES}(C, \mathcal{A}, \mathcal{L}, \mathcal{E}, \mathbf{S}^{new}, P_t, E_{p,t}, T_t, \theta, \mathcal{R})$
- 16: for each active flux f : $g \leftarrow \mathcal{E}(\mathcal{L}, f.\text{function})$; $q \leftarrow g(\cdot)$ {ID $\rightarrow$ callable}
- 17: **return** $\mathbf{r}(\mathbf{S}_{real}^{new}) = (\mathbf{S}_{real}^{new} - \mathbf{S}_{real}^{old}) / \Delta t - \Delta \mathbf{S}_{real}$
- 18: $\mathbf{S}_{real}^{new} \leftarrow \text{SOLVE}(\mathbf{r} = 0)$ {nonlinear solver with $\mathbf{P}$; fallback if needed}
- 19: update $\mathbf{S}_{real}$; record $\mathbf{S}$, Q_t , $E_{a,t}$, and per-flux histories
- 20: $\mathcal{R} \leftarrow \text{UPDATEUH}(\text{inputs}^{route}, \mathcal{R})$
- 21: **end for**
- 22: **Output:** $Q_{1:T}$, $E_{a,1:T}$, state histories $\{\mathbf{S}_{1:T}^k\}$, and per-flux histories

Table A.1

Hydrological process equations and unit-hydrograph routing schemes used as building blocks in the HydroXGP framework, with their occurrences in MARRMoT model structures.

Process	Equation	MARRMoT model IDs
Baseflow	$flux_{out} = \theta_1 S$	2, 4, 6, 8, 9, 12, 13, 15, 16, 17, 18, 20, 21, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 40, 41, 43, 44, 45, 46
Baseflow	$flux_{out} = \min\left(\frac{S}{\Delta t}, \theta_1 \left(\frac{\max(0, S)}{S_{max}}\right)^{\theta_2}\right)$	22
Baseflow	$flux_{out} = \min\left(\frac{S}{\Delta t}, \theta_1 \max(0, S)^{\theta_2}\right)$	10, 19, 37, 39, 42, 43, 47
Capillary rise	$flux_{out} = \min\left(\theta_1 \left(1 - \frac{S_1}{S_{1,max}}\right), \frac{S_2}{\Delta t}\right)$	13, 15, 37
Evaporation	$E_a = \min\left(E_p, \frac{S}{\Delta t}\right)$	2, 6, 12, 13, 16, 17, 18, 23, 25, 26, 27, 33, 34, 36, 38, 39, 41, 42, 44, 45, 46, 47
Evaporation	$E_a = \min\left(E_p \frac{S}{\theta_1 S_{max}}, E_p, \frac{S}{\Delta t}\right)$	3, 11, 14, 21, 26, 34, 37, 42
Evaporation	$E_a = \min\left(E_p \cdot \max\left(0, \theta_1 \frac{S - \theta_2 S_{max}}{S_{max} - \theta_2 S_{max}}\right), \frac{S}{\Delta t}\right)$	15
Evaporation	$E_a = \min\left((1 - \theta_1) \frac{S}{S_{max}}, E_p, \frac{S}{\Delta t}\right)$	4, 8, 9, 16
Evaporation	$E_a = \min\left(\frac{S}{S_{max}}, E_p, \frac{S}{\Delta t}\right)$	1, 3, 8, 10, 11, 19, 22, 24, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 40, 41, 43, 44, 45
Evaporation	$E_a = \min\left(1, e^{2\left(1 - \frac{S}{S_1}\right)}\right) E_p$	5
Evaporation	$E_a = \min\left(\theta_1 E_p \cdot \text{logisticSmoother}_S(S_2, \theta_2), \frac{S_1}{\Delta t}\right)$	17, 25
Evaporation	$E_a = \min\left(\max\left(\theta_2, \min\left(\frac{S}{\theta_1}, 1\right)\right) E_p, \frac{S}{\Delta t}\right)$	28
Infiltration excess	$flux_{out} = flux_{in} - \min\left(\theta_1 \exp\left[-\theta_2 \frac{S}{S_{max}}\right], flux_{in}\right)$	18, 36, 44
Infiltration excess	$flux_{out} = flux_{in} - \min(flux_{in}, \theta_1)$	15, 23, 40, 44
Interception excess	$flux_{out} = flux_{in} [1 - \text{logisticSmoother}_S(S, S_{max})]$	16, 18, 22, 26, 34, 36, 39, 42, 44, 45, 46, 47
Interception excess	$flux_{out} = \max(flux_{in} - \theta_1, 0)$	2, 13, 15
Interception excess	$flux_{out} = \left(1 - \left(\min\left(1, \max\left(0, 1 - \frac{S}{S_{max}}\right)\right)\right)^{\theta_1}\right) flux_{in}$	2, 13, 22, 28, 29
Interflow	$flux_{out} = \theta_1 \frac{S}{S_{max}} flux_{in}$	18, 36
Interflow	$flux_{out} = \max(0, \theta_1 (S - \theta_2))$	3, 12, 20, 27, 38
Interflow	$flux_{out} = \min\left(\frac{S - \theta_2}{\Delta t}, (\theta_1 \max(0, S - \theta_2))^{\theta_3}\right)$	4, 9, 11, 16, 39, 47
Interflow	$flux_{out} = \theta_1 \frac{\max(0, S - \theta_2)}{\theta_3}$	14
Percolation	$flux_{out} = \min\left(\frac{S}{\Delta t}, \theta_1\right)$	37
Percolation	$flux_{out} = \min\left(\frac{S}{\Delta t}, \theta_1 \frac{S}{S_{max}}\right)$	21, 26, 34, 39
Recharge	$flux_{out} = \theta_1 S_1 \left[1 - \min\left(1, \frac{S_2}{\theta_2}\right)\right]$	20
Saturation excess	$flux_{out} = flux_{in} [1 - \text{logisticSmoother}_S(S, S_{max})]$	1, 3, 4, 6, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 20, 21, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 40, 41, 43, 44, 45, 46
Saturation excess	$flux_{out} = \left(1 - \left(\min\left(1, \max\left(0, 1 - \frac{S}{S_{max}}\right)\right)\right)^{\theta_1}\right) flux_{in}$	2, 13, 22, 28, 29
Saturation excess	$flux_{out} = \left(1 - \min\left(1, \left(\frac{\max(S, 0)}{\theta_1}\right)^{\theta_2}\right)\right) flux_{in}$	5
Saturation excess	$flux_{out} = flux_{in} \cdot \text{logisticSmoother}_S(S, 0)$	9
Saturation excess	$flux_{out} = \begin{cases} (0.5 - \theta_1)^{1-\theta_2} \max\left(0, \frac{S}{S_{max}}\right)^{\theta_3} \cdot \frac{S}{S_{max}} \leq 0.5 - \theta_1 + \\ 1 - (0.5 + \theta_1)^{1-\theta_2} \max\left(0, 1 - \frac{S}{S_{max}}\right)^{\theta_3} \cdot \frac{S}{S_{max}} > 0.5 - \theta_1, \end{cases} flux_{in}$	28
Snow partition	$flux_{out} = flux_{in} \cdot \text{logisticSmoother}_T(T, T_{threshold})$	6, 12, 30, 31, 32, 34, 35, 41, 43, 44, 45
Melt	$flux_{out} = \min\left(\frac{S}{\Delta t}, \max(0, \theta_1 (T - T_{threshold}))\right)$	6, 12, 30, 31, 32, 34, 35, 37, 43, 44, 45
Unit hydrograph	Triangular; linearly increasing	 13, 15, 21, 26, 34
Unit hydrograph	Triangular; increasing then decreasing	 16, 37, 47
Unit hydrograph	Exponentially decreasing	 5
Unit hydrograph	Pure time delay	 5

Notation follows MARRMoT toolbox. Symbol definitions, parameter ranges, and the exact implementations used in this study are provided in the accompanying codebase on HydroShare (Matin, 2026). For the original MARRMoT flux-equation and unit-hydrograph definitions, see the MARRMoT repository (github.com/wknoben/MARRMoT).

Table A.2

Continuous parameters of the HydroXGP process library: description, unit, and calibration bounds. The GP encodes each parameter as a normalized value $u_j \in [0, 1]$, mapped linearly onto these bounds when a model is instantiated (see Table A.3).

Parameter	Description	Unit	Bounds
<i>Global (shared across storages)</i>			
tt	Rain-snow threshold temperature	°C	[−3, 5]
ddf	Degree-day factor for snowmelt	mm °C ^{−1} d ^{−1}	[0, 20]
M	Forest-cover fraction	–	[0.05, 0.95]
p	Evaporation reduction coefficient	–	[0, 1]
lp	Wilting point as fraction of $S_{\max}$	–	[0.1, 0.8]
Fmax	Maximum infiltration rate	mm d ^{−1}	[0, 200]
Fscale	Infiltration-decay exponent	–	[0, 15]
<i>Interception store</i>			
inc_Dw	Interception capacity	mm	[0.01, 5]
inc_beta_w	Interception-excess non-linearity	–	[0, 6]
<i>Soil store 1</i>			
sm1_Smax	Maximum storage capacity	mm	[1, 2000]
sm1_beta_w	Saturation-excess non-linearity	–	[0, 10]
sm1_sat_inflect	Saturation-curve inflection offset	–	[−0.49, 0.49]
sm1_Kw	Baseflow runoff coefficient	d ^{−1}	[0, 1]
sm1_thresh	Threshold fraction of $S_{\max}$	–	[0.05, 0.95]
sm1_a	Interflow runoff coefficient	d ^{−1}	[0, 1]
sm1_b	Interflow/baseflow non-linearity	–	[1, 5]
sm1_perc_max	Maximum percolation rate	mm d ^{−1}	[0, 20]
sm1_SE_slow_frac	Slow-flow fraction of saturation excess	–	[0, 1]
sm1_Int_slow_frac	Slow-flow fraction of interflow	–	[0, 1]
<i>Soil store 2</i>			
sm2_Smax	Maximum storage capacity	mm	[1, 1000]
sm2_Kw	Baseflow runoff coefficient	d ^{−1}	[0, 1]
sm2_thresh	Threshold fraction of $S_{\max}$	–	[0.05, 0.95]
sm2_a	Interflow runoff coefficient	d ^{−1}	[0, 1]
sm2_b	Interflow/baseflow non-linearity	–	[1, 5]
sm2_C	Maximum capillary rise	mm d ^{−1}	[0, 4]
sm2_SE_slow_frac	Slow-flow fraction of saturation excess	–	[0, 1]
sm2_Base_slow_frac	Slow-flow fraction of baseflow	–	[0, 1]
<i>Soil store 3</i>			
sm3_Smax	Maximum storage capacity	mm	[1, 600]
sm3_Kw	Baseflow runoff coefficient	d ^{−1}	[0, 1]
sm3_thresh	Threshold fraction of $S_{\max}$	–	[0.05, 0.95]
sm3_a	Interflow runoff coefficient	d ^{−1}	[0, 1]
sm3_b	Interflow/baseflow non-linearity	–	[1, 5]
sm3_SE_slow_frac	Slow-flow fraction of saturation excess	–	[0, 1]
<i>Routing stores</i>			
r1_Kw	Baseflow runoff coefficient	d ^{−1}	[0, 1]
r2_Kw	Baseflow runoff coefficient	d ^{−1}	[0, 1]
r2_Rec	Recharge rate coefficient (to soil store 1)	d ^{−1}	[0, 1]
<i>Unit-hydrograph routing</i>			
fast_delay	Fast routing delay	d	[1, 120]
slow_delay	Slow routing delay	d	[1, 120]
Q_delay	Outlet routing delay	d	[1, 120]
<i>Numerical smoothing constants</i>			
rho_snow	Logistic smoothing scale (snow threshold)	–	[0.01, 0.02]
rho_soil	Logistic smoothing scale (soil thresholds)	–	[0.01, 0.05]

Parameter bounds adapted from the MARRMoT toolbox: github.com/wknoben/MARRMoT.

Table A.3

Example mapping from normalized GP entries to model configuration values. Binary entries are mapped using a threshold, categorical entries are mapped by dividing the interval [0, 1] into equal-width bins and selecting the corresponding admissible option from the ordered option list $\mathcal{O}_j$, and continuous entries are mapped linearly to their lower and upper parameter bounds (see Table A.2). The continuous-parameter bounds were adapted from the original MARRMoT formulations.

Entry type	Example entry	Mapping rule	Model value
Binary	sn_on	$b_j = \mathbb{I}(u_j > 0.5)$	active/inactive
Categorical	sm1_evap	$c_j = \mathcal{O}_j[q_j]; q_j = \min(\lfloor K_j u_j \rfloor, K_j - 1)$	selected process formulation; $c_j \in \mathcal{O}_j$
Continuous	sm1_Smax	$\theta_j = \theta_{j,\min} + u_j(\theta_{j,\max} - \theta_{j,\min})$	value between lower and upper bounds

$$\begin{aligned}
\mathbf{J}_{\text{dense}} &= \begin{bmatrix} J_{11} & J_{12} & J_{13} & J_{14} & J_{15} & J_{16} & J_{17} \\ J_{21} & J_{22} & J_{23} & J_{24} & J_{25} & J_{26} & J_{27} \\ J_{31} & J_{32} & J_{33} & J_{34} & J_{35} & J_{36} & J_{37} \\ J_{41} & J_{42} & J_{43} & J_{44} & J_{45} & J_{46} & J_{47} \\ J_{51} & J_{52} & J_{53} & J_{54} & J_{55} & J_{56} & J_{57} \\ J_{61} & J_{62} & J_{63} & J_{64} & J_{65} & J_{66} & J_{67} \\ J_{71} & J_{72} & J_{73} & J_{74} & J_{75} & J_{76} & J_{77} \end{bmatrix} \quad \text{Dense Jacobian} \\
&\quad \Downarrow \text{Masking operation} \\
&\quad \mathbf{J} * \mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} \quad \text{Jacobian sparsity pattern} \\
&\quad \begin{bmatrix} J_{11} & 0 & 0 & 0 & 0 & 0 & 0 \\ J_{21} & J_{22} & 0 & 0 & 0 & 0 & 0 \\ 0 & J_{32} & J_{33} & J_{34} & 0 & 0 & J_{37} \\ 0 & 0 & J_{43} & J_{44} & 0 & 0 & 0 \\ 0 & 0 & 0 & J_{54} & J_{55} & 0 & 0 \\ 0 & J_{62} & J_{63} & J_{64} & 0 & J_{66} & 0 \\ 0 & 0 & 0 & J_{74} & J_{75} & 0 & J_{77} \end{bmatrix} \begin{bmatrix} \delta S_1 \\ \delta S_2 \\ \delta S_3 \\ \delta S_4 \\ \delta S_5 \\ \delta S_6 \\ \delta S_7 \end{bmatrix} = \begin{bmatrix} -r_1 \\ -r_2 \\ -r_3 \\ -r_4 \\ -r_5 \\ -r_6 \\ -r_7 \end{bmatrix} \\
&\quad \text{Same Newton system, now with sparse } \mathbf{J}
\end{aligned}$$

State order
\$S_n\$
\$I_n\$
\$Sm_1\$
\$Sm_2\$
\$Sm_3\$
\$R_1\$
\$R_2\$

Jacobian / sparsity
\$\mathbf{J} = \partial \mathbf{r} / \partial \mathbf{S}_{\text{real}}\$: Jacobian
\$P_{ij} \in \{0, 1\}\$: binary sparsity mask
Newton solve
\$\mathbf{J}(\mathbf{S}_{\text{real},k}) \delta \mathbf{S}_{\text{real},k} = -\mathbf{r}(\mathbf{S}_{\text{real},k})\$
\$\mathbf{S}_{\text{real},k+1} = \mathbf{S}_{\text{real},k} + \delta \mathbf{S}_{\text{real},k}\$

Fig. A.1. Illustrative example of Jacobian sparsity construction for the master process space or model structure shown in Fig. 3. The sparsity pattern is model-specific: it varies across discovered structures and is generated automatically from the active dependency graph after removing inactive components and redirecting to bypass targets. Only real storages (dynamic states) are solved, whereas dummy split/routing junctions are excluded as unknowns and their connectivity is preserved through graph collapse/redirection; thus, if source $\rightarrow$ dummy and dummy $\rightarrow$ target, a direct source $\rightarrow$ target dependency is inserted.

Appendix B. Objective metrics used for model training and evaluation

Table B.1

Definitions, ranges, and desirable values of the performance metrics.

Metric	Definition	Range	Desirable value
NSE	$1 - \frac{\sum_i (Q_i - \hat{Q}_i)^2}{\sum_i (Q_i - \bar{Q})^2}$	$(-\infty, 1]$	Closer to 1
KGE	$1 - \sqrt{\frac{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}{3}}$	$(-\infty, 1]$	Closer to 1
logNSE	$1 - \frac{\sum_i (\log(Q_i + \delta) - \log(\hat{Q}_i + \delta))^2}{\sum_i (\log(Q_i + \delta) - \log(\bar{Q} + \delta))^2}$	$(-\infty, 1]$	Closer to 1

$r = \text{corr}(Q, \hat{Q})$; $\alpha = \sigma_{\hat{Q}}/\sigma_Q$; $\beta = \bar{\hat{Q}}/\bar{Q}$; $\delta = 10^{-6}$, a small constant introduced to ensure numerical stability of the logarithmic transform.

Appendix C. Supplementary data

Supplementary data to this article can be found online at doi: [10.1016/j.jhydrol.2026.136493](https://doi.org/10.1016/j.jhydrol.2026.136493).

Data availability

The HydroXGP source code and supporting analysis data and results are available on HydroShare (Matin, 2026): <http://www.hydroshare.org/resource/49b2089113724baf9faa5bb45246db2f>. Input and benchmark data are available from the sources cited in the main text.

References

- Addor, N., Melsen, L.A., 2019. Legacy, rather than adequacy, drives the selection of hydrological models. *Water Resour. Res.* 55, 378–390. <https://doi.org/10.1029/2018WR022958>
- Addor, N., Newman, A., Mizukami, N., Clark, M.P., 2017. Catchment attributes for large-sample studies. [Data set]. UCAR/NCAR, Boulder, CO, USA. <https://doi.org/10.5065/D6G73C3Q>
- Anderson, E.A., 1973. National Weather Service River Forecast System: Snow Accumulation and Ablation Model. Technical Report 17, U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration, National Weather Service, Office of Hydrology, Washington, D.C.
- Babovic, V., Keijzer, M., 2000. Genetic programming as a model induction engine. *J. Hydroinformatics* 2, 35–60. <https://doi.org/10.2166/hydro.2000.0004>

- Babovic, V., Keijzer, M., 2002. Rainfall runoff modelling based on genetic programming. *Hydrol. Res.* 33, 331–346. <https://doi.org/10.2166/nh.2002.0012>
- Babovic, V., Wu, Z., 1994. Calibrating hydrodynamic models by means of simulated evolution. In: Verwey, A., Minns, A.W., Babovic, V., Maksimovic, C. (Eds.), *Hydroinformatics '94: Proceedings of the First International Conference on Hydroinformatics*. A. A. Balkema, Rotterdam, Netherlands, pp. 193–200.
- Bayati, A., Ameli, A.A., Razavi, S., 2026. Evaluating the functional realism of deep learning rainfall-runoff models using catchment hydrology principles. *Water Resour. Res.* 62, e2025WR040076. <https://doi.org/10.1029/2025WR040076>
- Beven, K., 2020. Deep learning, hydrological processes and the uniqueness of place. *Hydrol. Process.* 34, 3608–3613. <https://doi.org/10.1002/hyp.13805>
- Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. *Hydrol. Process.* 6, 279–298. <https://doi.org/10.1002/hyp.3360060305>
- Botterill, T.E., McMillan, H.K., 2023. Using machine learning to identify hydrologic signatures with an encoder-decoder framework. *Water Resour. Res.* 59, e2022WR033091. <https://doi.org/10.1029/2022WR033091>
- Burnash, R.J.C., Ferral, R.L., McGuire, R.A., 1973. A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers. Technical Report, U.S. Department of Commerce, National Weather Service, and State of California, Department of Water Resources, Sacramento, California.
- Chadalawada, J., Babovic, V., 2019. Review and comparison of performance indices for automatic model induction. *J. Hydroinformatics* 21, 13–31. <https://doi.org/10.2166/hydro.2017.078>
- Chadalawada, J., Havlíček, V., Babovic, V., 2017. A genetic programming approach to system identification of rainfall-runoff models. *Water Resour. Manag.* 31, 3975–3992. <https://doi.org/10.1007/s11269-017-1719-1>
- Chadalawada, J., Herath, H.M.V.V., Babovic, V., 2020. Hydrologically informed machine learning for rainfall-runoff modeling: a genetic programming-based toolkit for automatic model induction. *Water Resour. Res.* 56, e2019WR026933. <https://doi.org/10.1029/2019WR026933>
- Chiew, F., McMahon, T., 1994. Application of the daily rainfall-runoff model MODHYDROLOG to 28 Australian catchments. *J. Hydrol.* 153, 383–416. [https://doi.org/10.1016/0022-1694\(94\)90200-3](https://doi.org/10.1016/0022-1694(94)90200-3)
- Clark, S.R., Jaffrés, J.B., 2025. Associations between deep learning runoff predictions and hydrogeological conditions in Australia. *J. Hydrol.* 651, 132569. <https://doi.org/10.1016/j.jhydrol.2024.132569>
- Clark, M.P., Slater, A.G., Rupp, D.E., Woods, R.A., Vrugt, J.A., Gupta, H.V., Wagener, T., Hay, L.E., 2008. Framework for understanding structural errors (FUSE): a modular framework to diagnose differences between hydrological models. *Water Resour. Res.* 44, W00B02. <https://doi.org/10.1029/2007WR006735>
- Clark, M.P., Kavetski, D., Fenicia, F., 2011. Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.* 47, W09301. <https://doi.org/10.1029/2010WR009827>
- Clark, M.P., Bierkens, M.F.P., Samaniego, L., Woods, R.A., Uijlenhoet, R., Bennett, K.E., Pauwels, V.R.N., Cai, X., Wood, A.W., Peters-Lidard, C.D., 2017. The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism. *Hydrol. Earth Syst. Sci.* 21, 3427–3440. <https://doi.org/10.5194/hess-21-3427-2017>
- Clark, M.P., Vogel, R.M., Lamontagne, J.R., Mizukami, N., Knoben, W.J.M., Tang, G., Gharari, S., Freer, J.E., Whitfield, P.H., Shook, K.R., Papalexiou, S.M., 2021. The

- abuse of popular performance metrics in hydrologic modeling. *Water Resour. Res.* 57, e2020WR029001. <https://doi.org/10.1029/2020WR029001>
- Clark, M.P., Thebault, C., Eythorsson, D., Vasquez, N., Knoben, W., Wood, A., 2025. What is the role of machine learning when we want to simulate hydrological processes? In: EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-12547. <https://doi.org/10.5194/egusphere-egu25-12547>
- Collins, E.L., Sanchez, G.M., Terando, A., Stillwell, C.C., Mitsova, H., Sebastian, A., Meentemeyer, R.K., 2022. Predicting flood damage probability across the conterminous United States. *Environ. Res. Lett.* 17, 034006. <https://doi.org/10.1088/1748-9326/ac40f0>
- Coxon, G., Freer, J., Lane, R., Dunne, T., Knoben, W.J.M., Howden, N.J.K., Quinn, N., Wagener, T., Woods, R., 2019. DECIPHER v1: dynamic fluxes and connectivity for predictions of hydrology. *Geosci. Model Dev.* 12, 2285–2306. <https://doi.org/10.5194/gmd-12-2285-2019>
- Craig, J.R., Brown, G., Chlumsky, R., Jenkinson, R.W., Jost, G., Lee, K., Mai, J., Serrero, M., Sgro, N., Shafii, M., Snowden, A.P., Tolson, B.A., 2020. Flexible watershed simulation with the Raven hydrological modelling framework. *Environ. Model. Softw.* 129, 104728. <https://doi.org/10.1016/j.envsoft.2020.104728>
- Dantzer, T.A., Kerkez, B., 2024. Generating interpretable rainfall-runoff models automatically from data. *Adv. Water Resour.* 193, 104796. <https://doi.org/10.1016/j.advwatres.2024.104796>
- De la Fuente, L.A., Gupta, H.V., Condon, L.E., 2023. Toward a multi-representational approach to prediction and understanding, in support of discovery in hydrology. *Water Resour. Res.* 59, e2021WR031548. <https://doi.org/10.1029/2021WR031548>
- De la Fuente, L.A., Ehsani, M.R., Gupta, H.V., Condon, L.E., 2024. Toward interpretable LSTM-based modeling of hydrological systems. *Hydrol. Earth Syst. Sci.* 28, 945–971. <https://doi.org/10.5194/hess-28-945-2024>
- Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* 6, 182–197. <https://doi.org/10.1109/4235.996017>
- Espinoza, E.A., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., Ehret, U., 2025. Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events. *Hydrol. Earth Syst. Sci.* 29, 1277–1294. <https://doi.org/10.5194/hess-29-1277-2025>
- Eythorsson, D., Clark, M., 2025. Toward automated scientific discovery in hydrology: the opportunities and dangers of AI augmented research frameworks. *Hydrol. Process.* 39, e70065. <https://doi.org/10.1002/hyp.70065>
- Feng, D., Liu, J., Lawson, K., Shen, C., 2022. Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resour. Res.* 58, e2022WR032404. <https://doi.org/10.1029/2022WR032404>
- Feng, D., Beck, H., Lawson, K., Shen, C., 2023. The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment. *Hydrol. Earth Syst. Sci.* 27, 2357–2373. <https://doi.org/10.5194/hess-27-2357-2023>
- Fenicia, F., Kavetski, D., Savenije, H.H.G., 2011. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resour. Res.* 47, W11510. <https://doi.org/10.1029/2010WR010174>
- Fenicia, F., Kavetski, D., Savenije, H.H.G., Clark, M.P., Schoups, G., Pfister, L., Freer, J., 2014. Catchment properties, function, and conceptual model representation: is there a correspondence? *Hydrol. Process.* 28, 2451–2467. <https://doi.org/10.1002/hyp.9726>
- Frame, J.M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L.M., Gupta, H.V., Nearing, G.S., 2022. Deep learning rainfall-runoff predictions of extreme events. *Hydrol. Earth Syst. Sci.* 26, 3377–3392. <https://doi.org/10.5194/hess-26-3377-2022>
- Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *J. Hydrol.* 377, 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>
- Gupta, H.V., Clark, M.P., Vrugt, J.A., Abramowitz, G., Ye, M., 2012. Towards a comprehensive assessment of model structural adequacy. *Water Resour. Res.* 48, W08301. <https://doi.org/10.1029/2011WR011044>
- Gupta, H.V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., Andréassian, V., 2014. Large-sample hydrology: a note to balance depth with breadth. *Hydrol. Earth Syst. Sci.* 18, 463–477. <https://doi.org/10.5194/hess-18-463-2014>
- Gutiérrez-Jurado, K.Y., Partington, D., Batelaan, O., Cook, P., Shanfield, M., 2019. What triggers streamflow for intermittent rivers and ephemeral streams in low-gradient catchments in Mediterranean climates. *Water Resour. Res.* 55, 9926–9946. <https://doi.org/10.1029/2019WR025041>
- Havlíček, V., Hanel, M., Máca, P., Kuraž, M., Pech, P., 2013. Incorporating basic hydrological concepts into genetic programming for rainfall-runoff forecasting. *Computing* 95, 363–380. <https://doi.org/10.1007/s00607-013-0298-0>
- He, L., Shi, L., Song, W., Shen, J., Wang, L., Hu, X., Zha, Y., 2024. Synergizing intuitive physics and big data in deep learning: can we obtain process insights while maintaining state-of-the-art hydrological prediction capability? *Water Resour. Res.* 60, e2024WR037582. <https://doi.org/10.1029/2024WR037582>
- Henn, B., Clark, M.P., Kavetski, D., Lundquist, J.D., 2015. Estimating mountain basin-mean precipitation from streamflow using Bayesian inference. *Water Resour. Res.* 51, 8012–8033. <https://doi.org/10.1002/2014WR016736>
- Herath, H.M.V.V., Chadalawada, J., Babovic, V., 2021a. Genetic programming for hydrological applications: to model or to forecast is the question. *J. Hydroinformatics* 23, 740–763. <https://doi.org/10.2166/hydro.2021.179>
- Herath, H.M.V.V., Chadalawada, J., Babovic, V., 2021b. Hydrologically informed machine learning for rainfall-runoff modelling: towards distributed modelling. *Hydrol. Earth Syst. Sci.* 25, 4373–4401. <https://doi.org/10.5194/hess-25-4373-2021>
- Heřmanovský, M., Havlíček, V., Hanel, M., Pech, P., 2017. Regionalization of runoff models derived by genetic programming. *J. Hydrol.* 547, 544–556. <https://doi.org/10.1016/j.jhydrol.2017.02.018>
- Horton, P., Schaeffli, B., Kaularic, M., 2022. Why do we have so many different hydrological models? A review based on the case of Switzerland. *WIREs Water* 9, e1574. <https://doi.org/10.1002/wat2.1574>
- Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., Fenicia, F., Freer, J., Gelfan, A., Gupta, H., Hughes, D., Hut, R., Montanari, A., Pande, S., Tetzlaff, D., Troch, P., Uhlenbrook, S., Wagener, T., Winsemius, H., Woods, R., Zehe, E., Cudennec, C., 2013. A decade of predictions in ungauged basins (PUB)—a review. *Hydrol. Sci. J.* 58, 1198–1255. <https://doi.org/10.1080/02626667.2013.803183>
- Hughes, M., Jackson, D.L., Unruh, D., Wang, H., Hobbins, M., Ogden, F.L., Cifelli, R., Cosgrove, B., DeWitt, D., Dugger, A., Ford, T.W., Fuchs, B., Glaudemans, M., Gochis, D., Quiring, S.M., Rafieei-Nasab, A., Webb, R.S., Xia, Y., Xu, L., 2024. Evaluation of retrospective National Water Model soil moisture and streamflow for drought-monitoring applications. *J. Geophys. Res. Atmos.* 129, e2023JD038522. <https://doi.org/10.1029/2023JD038522>
- Kavetski, D., Fenicia, F., 2011. Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. *Water Resour. Res.* 47, W11511. <https://doi.org/10.1029/2011WR010748>
- Khatami, S., Peel, M.C., Peterson, T.J., Western, A.W., 2019. Equifinality and flux mapping: a new approach to model evaluation and process representation under uncertainty. *Water Resour. Res.* 55, 8922–8941. <https://doi.org/10.1029/2018WR023750>
- Kim, K.B., Kwon, H.H., Han, D., 2018. Exploration of warm-up period in conceptual hydrological modelling. *J. Hydrol.* 556, 194–210. <https://doi.org/10.1016/j.jhydrol.2017.11.015>
- Kiraz, M., Coxon, G., Wagener, T., 2023. A priori selection of hydrological model structures in modular modelling frameworks: application to Great Britain. *Hydrol. Sci. J.* 68, 2042–2056. <https://doi.org/10.1080/02626667.2023.2251968>
- Kirchner, J.W., 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. *Water Resour. Res.* 42, W03504. <https://doi.org/10.1029/2005WR004362>
- Knoben, W.J.M., 2024a. HydroBM: simple benchmark models for hydrology. [Software]. Zenodo. <https://doi.org/10.5281/zenodo.13754737>
- Knoben, W.J.M., 2024b. Setting expectations for hydrologic model performance with an ensemble of simple benchmarks. *Hydrol. Process.* 38, e15288. <https://doi.org/10.1002/hyp.15288>
- Knoben, W.J.M., Trotter, L., 2022. MARRMoT: modular assessment of rainfall-runoff models toolbox. [Software]. Zenodo. <https://doi.org/10.5281/zenodo.6460624>
- Knoben, W.J.M., Freer, J.E., Fowler, K.J.A., Peel, M.C., Woods, R.A., 2019a. Modular assessment of rainfall-runoff models toolbox (MARRMoT) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. *Geosci. Model Dev.* 12, 2463–2480. <https://doi.org/10.5194/gmd-12-2463-2019>
- Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019b. Technical note: inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrol. Earth Syst. Sci.* 23, 4323–4331. <https://doi.org/10.5194/hess-23-4323-2019>
- Knoben, W.J.M., Freer, J.E., Peel, M.C., Fowler, K.J.A., Woods, R.A., 2020. A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. *Water Resour. Res.* 56, e2019WR025975. <https://doi.org/10.1029/2019WR025975>
- Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA.
- Kratzert, F., 2019a. CAMELS benchmark models. [Data set]. HydroShare. <https://doi.org/10.4211/hydroshare.474ec37e7db45baa425cbb4fc1b61e1>
- Kratzert, F., 2019b. Pretrained models + simulations for our HESSD submission “Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets”. [Data set]. HydroShare. <https://doi.org/10.4211/hydroshare.5312635e44dc824eeb99eda12f06>
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall-runoff modelling using long short-term memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* 22, 6005–6022. <https://doi.org/10.5194/hess-22-6005-2018>
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.* 23, 5089–5110. <https://doi.org/10.5194/hess-23-5089-2019>
- Kratzert, F., Klotz, D., Hochreiter, S., Nearing, G.S., 2021. A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall-runoff modeling. *Hydrol. Earth Syst. Sci.* 25, 2685–2703. <https://doi.org/10.5194/hess-25-2685-2021>
- Kratzert, F., Gauch, M., Klotz, D., Nearing, G., 2024. HESS opinions: never train a long short-term memory (LSTM) network on a single basin. *Hydrol. Earth Syst. Sci.* 28, 4187–4201. <https://doi.org/10.5194/hess-28-4187-2024>
- Kumar, R., Samaniego, L., Attinger, S., 2013. Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resour. Res.* 49, 360–379. <https://doi.org/10.1029/2012WR012195>
- Ladson, A.R., Brown, R., Neal, B., Nathan, R., 2013. A standard approach to baseflow separation using the Lyne and Hollick filter. *Australas. J. Water Resour.* 17, 25–34. <https://doi.org/10.7158/13241583.2013.11465417>
- Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., Dadson, S.J., 2021. Benchmarking data-driven rainfall-runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. *Hydrol. Earth Syst. Sci.* 25, 5517–5534. <https://doi.org/10.5194/hess-25-5517-2021>
- Liang, X., Lettenmaier, D.P., Wood, E.F., Burges, S.J., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res. Atmos.* 99, 14415–14428. <https://doi.org/10.1029/94JD00483>

- Lyne, V., Hollick, M., 1979. Stochastic time-variable rainfall-runoff modelling. In: *Hydrology and Water Resources Symposium*. Institution of Engineers Australia, Perth, Australia, pp. 89–92.
- Ma, K., Feng, D., Lawson, K., Tsai, W.P., Liang, C., Huang, X., Sharma, A., Shen, C., 2021. Transferring hydrologic data across continents – leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resour. Res.* 57, e2020WR028600. <https://doi.org/10.1029/2020WR028600>
- Matin, N., 2026. Repository of HydroXGP source code and analysis data supporting this study. [Software and data set]. HydroShare. <http://www.hydroshare.org/resource/49b2089113724baf9faa5bb45246db2f>
- Melsen, L.A., Puy, A., Torfs, P.J.J.F., Saltelli, A., 2025. The rise of the Nash-Sutcliffe efficiency in hydrology. *Hydrol. Sci. J.* 70, 1248–1259. <https://doi.org/10.1080/02626667.2025.2475105>
- Mendoza, P.A., Clark, M.P., Mizukami, N., Newman, A.J., Barlage, M., Gutmann, E.D., Rasmussen, R.M., Rajagopalan, B., Brekke, L.D., Arnold, J.R., 2015. Effects of hydrologic model choice and calibration on the portrayal of climate change impacts. *J. Hydrometeorol.* 16, 762–780. <https://doi.org/10.1175/JHM-D-14-0104.1>
- Mizukami, N., Clark, M.P., Newman, A.J., Wood, A.W., Gutmann, E.D., Nijssen, B., Rakovec, O., Samaniego, L., 2017. Towards seamless large-domain parameter estimation for hydrologic models. *Water Resour. Res.* 53, 8020–8040. <https://doi.org/10.1002/2017WR020401>
- Mizukami, N., Rakovec, O., Newman, A.J., Clark, M.P., Wood, A.W., Gupta, H.V., Kumar, R., 2019. On the choice of calibration metrics for “High-flow” estimation using hydrologic models. *Hydrol. Earth Syst. Sci.* 23, 2601–2614. <https://doi.org/10.5194/hess-23-2601-2019>
- Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* 50, 885–900. <https://doi.org/10.13031/2013.23153>
- Nai, C., Liu, X., Tang, Q., Liu, L., Sun, S., Gaffney, P.P.J., 2024. A novel strategy for automatic selection of cross-basin data to improve local machine learning-based runoff models. *Water Resour. Res.* 60, e2023WR035051. <https://doi.org/10.1029/2023WR035051>
- NASA, 2013. NASA shuttle radar topography mission global 1 arc second. [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), Sioux Falls, South Dakota, USA. <https://doi.org/10.5067/MEASURES/SRTM/SRTMGL1.003>
- Nash, J., Sutcliffe, J., 1970. River flow forecasting through conceptual models Part I — A discussion of principles. *J. Hydrol.* 10, 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
- Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Prieto, C., Gupta, H.V., 2021. What role does hydrological science play in the age of machine learning? *Water Resour. Res.* 57, e2020WR028091. <https://doi.org/10.1029/2020WR028091>
- Newman, A.J., Clark, M.P., Sampson, K., Wood, A., Hay, L.E., Bock, A., Viger, R.J., Blodgett, D., Brekke, L., Arnold, J.R., Hopson, T., Duan, Q., 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. *Hydrol. Earth Syst. Sci.* 19, 209–223. <https://doi.org/10.5194/hess-19-209-2015>
- Newman, A.J., Mizukami, N., Clark, M.P., Wood, A.W., Nijssen, B., Nearing, G., 2017. Benchmarking of a physically based hydrologic model. *J. Hydrometeorol.* 18, 2215–2225. <https://doi.org/10.1175/JHM-D-16-0284.1>
- Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.* 279, 275–289. [https://doi.org/10.1016/S0022-1694\(03\)00225-7](https://doi.org/10.1016/S0022-1694(03)00225-7)
- Pumo, D., Noto, L.V., 2023. Exploring the use of multi-gene genetic programming in regional models for the simulation of monthly river runoff series. *Stoch. Environ. Res. Risk Assess.* 37, 1917–1941. <https://doi.org/10.1007/s00477-022-02373-1>
- Pushpalatha, R., Perrin, C., Moine, N.L., Andréassian, V., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. *J. Hydrol.* 420–421, 171–182. <https://doi.org/10.1016/j.jhydrol.2011.11.055>
- Rahman, M.M., Lu, M., 2015. Model spin-up behavior for wet and dry basins: a case study using the Xinanjiang model. *Water* 7, 4256–4273. <https://doi.org/10.3390/w7084256>
- Rakovec, O., Mizukami, N., Kumar, R., Newman, A.J., Thober, S., Wood, A.W., Clark, M.P., Samaniego, L., 2019. Diagnostic evaluation of large-domain hydrologic models calibrated across the contiguous United States. *J. Geophys. Res. Atmos.* 124, 13991–14007. <https://doi.org/10.1029/2019JD030767>
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven earth system science. *Nature* 566, 195–204. <https://doi.org/10.1038/s41586-019-0912-1>
- Saavedra, D., Mendoza, P.A., Addor, N., Llaucá, H., Vargas, X., 2022. A multi-objective approach to select hydrological models and constrain structural uncertainties for climate impact assessments. *Hydrol. Process.* 36, e14446. <https://doi.org/10.1002/hyp.14446>
- Samaniego, L., Kumar, R., Attinger, S., 2010. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resour. Res.* 46, W05523. <https://doi.org/10.1029/2008WR007327>
- Schaeffli, B., Gupta, H.V., 2007. Do Nash values have value? *Hydrol. Process.* 21, 2075–2080. <https://doi.org/10.1002/hyp.6825>
- Schwemmler, R., Leistert, H., Steinbrich, A., Weiler, M., 2024. RoGeR v3.0.5 – a process-based hydrological toolbox model in Python. *Geosci. Model Dev.* 17, 5249–5262. <https://doi.org/10.5194/gmd-17-5249-2024>
- Seibert, J., Vis, M.J.P., 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrol. Earth Syst. Sci.* 16, 3315–3325. <https://doi.org/10.5194/hess-16-3315-2012>
- Seibert, J., Vis, M.J.P., Lewis, E., van Meerveld, H.J., 2018. Upper and lower benchmarks in hydrological modelling. *Hydrol. Process.* 32, 1120–1125. <https://doi.org/10.1002/hyp.11476>
- Selle, B., Muttil, N., 2011. Testing the structure of a hydrological model using genetic programming. *J. Hydrol.* 397, 1–9. <https://doi.org/10.1016/j.jhydrol.2010.11.009>
- Shanfield, M., Cook, P.G., 2014. Transmission losses, infiltration and groundwater recharge through ephemeral and intermittent streambeds: a review of applied methods. *J. Hydrol.* 511, 518–529. <https://doi.org/10.1016/j.jhydrol.2014.01.068>
- Shen, C., Appling, A.P., Gentile, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fénicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C.J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., Rahmani, F., Song, Y., Beck, H.E., Bindas, T., Dwivedi, D., Fang, K., Höge, M., Rackauckas, C., Mohanty, B., Roy, T., Xu, C., Lawson, K., 2023. Differentiable modelling to unify machine learning and physical models for geosciences. *Nat. Rev. Earth Environ.* 4, 552–567. <https://doi.org/10.1038/s43017-023-00450-9>
- Sivapalan, M., 2003. Prediction in ungauged basins: a grand challenge for theoretical hydrology. *Hydrol. Process.* 17, 3163–3170. <https://doi.org/10.1002/hyp.5155>
- Spieler, D., Schütze, N., 2024. Investigating the model hypothesis space: benchmarking automatic model structure identification with a large model ensemble. *Water Resour. Res.* 60, e2023WR036199. <https://doi.org/10.1029/2023WR036199>
- Spieler, D., Stadnyk, T., 2026. When calibration metrics choose your model: investigating model selection choices for different modelling purposes. EGU General Assembly 2026, Vienna, Austria, 3–8 May 2026, EGU26-15121. <https://doi.org/10.5194/egusphere-egu26-15121>
- Spieler, D., Mai, J., Craig, J.R., Tolson, B.A., Schütze, N., 2020. Automatic model structure identification for conceptual hydrologic models. *Water Resour. Res.* 56, e2019WR027009. <https://doi.org/10.1029/2019WR027009>
- Staudinger, M., Stahl, K., Seibert, J., Clark, M.P., Tallaksen, L.M., 2011. Comparison of hydrological model structures based on recession and low flow simulations. *Hydrol. Earth Syst. Sci.* 15, 3447–3459. <https://doi.org/10.5194/hess-15-3447-2011>
- Tolson, B.A., Shoemaker, C.A., 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resour. Res.* 43, W01413. <https://doi.org/10.1029/2005WR004723>
- Trotter, L., Knoben, W.J.M., Fowler, K.J.A., Saft, M., Peel, M.C., 2022. Modular assessment of rainfall–runoff models toolbox (MARRMoT) v2.1: an object-oriented implementation of 47 established hydrological models for improved speed and readability. *Geosci. Model Dev.* 15, 6359–6369. <https://doi.org/10.5194/gmd-15-6359-2022>
- Wagner, P., Knoben, W.J.M., Schütze, N., Spieler, D., 2026. Metrics that matter: objective functions and their impact on signature representation in conceptual hydrological models. *Hydrol. Earth Syst. Sci.* 30, 4867–4888. <https://doi.org/10.5194/hess-30-4867-2026>
- Wood, A.W., Maurer, E.P., Kumar, A., Lettenmaier, D.P., 2002. Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res. Atmos.* 107, ACL 6–1–ACL 6–15. <https://doi.org/10.1029/2001JD000659>
- Yao, Y., Zhao, Y., Li, X., Feng, D., Shen, C., Liu, C., Kuang, X., Zheng, C., 2023. Can transfer learning improve hydrological predictions in the alpine regions? *J. Hydrol.* 625, 130038. <https://doi.org/10.1016/j.jhydrol.2023.130038>
- Yilmaz, K.K., Gupta, H.V., Wagener, T., 2008. A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. *Water Resour. Res.* 44, W09417. <https://doi.org/10.1029/2007WR006716>