Journal of Hydrology 679 (2026) 136201

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

A dependency-guided symbolic regression framework for overcoming error compensation in hybrid AI–physical models

Qingliang Li [a,*], Mujie Wu [a], Cheng Zhang [b], Zhongwang Wei [c], Wei Shangguan [c], Jinlong Zhu [b], Jing Wang [b], Xiaoning Li [a], Yuguang Yan [a], Zijian Zhang [b], Wenzong Dong [c], Xiaoming Shi [d], Yunfeng Lv [e], Yongjiu Dai [c]

[a] School of Artificial Intelligence, Changchun Normal University, Changchun 130032, China
[b] College of Computer Science and Technology, Jilin University, Changchun 130032, China
[c] School of Atmospheric Sciences, Sun Yat-Sen University, Zhuhai, Guangdong 519082, China
[d] Division of Environment and Sustainability, The Hong Kong University of Science and Technology, Hong Kong 999077, China
[e] College of Geographical Sciences, Changchun Normal University, Changchun 130032, China

ARTICLE INFO

This manuscript was handled by Dan Lu, Editor-in-Chief, with the assistance of Peishi Jiang, Associate Editor

Keywords:
Symbolic regression
Hybrid AI-physical model
Hydrological modeling
Error compensation
Evapotranspiration
Streamflow simulation

ABSTRACT

Integrating AI-discovered formulas into process-based Earth system models is difficult. Earth functions as a connected system with directional dependencies from upstream to downstream. When components are trained in isolation and later embedded, error compensation can occur: downstream modules absorb upstream errors, local metrics appear strong, yet system performance and physical consistency decline. We use hydrologic modeling as a tractable testbed. Deep symbolic regression (DSR) provides transparent evapotranspiration (ET) formulas. A knowledge guided variant (KG-DSR) builds runoff relations that preserve links among surface runoff, interflow, and baseflow. These expressions are first trained and evaluated with observed meteorology, then inserted into the HBV model under several integration strategies, including a sequential scheme that stabilizes ET before calibrating runoff. ET formulas from DSR outperform the original HBV scheme in offline tests. More than 70% of basins show higher Kling–Gupta efficiency (KGE) and lower RMSE, with better representation of vegetation controls on ET. In coupled runs, adding only the ET module yields limited benefit. The sequential strategy improves system behavior: relative to embedding runoff alone, ET KGE rises by 16.1% and streamflow KGE by 4.9%; relative to the HBV baseline, gains reach 11.6% and 10.1%. Performance for extreme flows and baseflow indicators is also highest under the sequential strategy. Stabilizing upstream ET reduces error compensation and strengthens cross-process consistency in coupled hydrologic models. A stepwise inclusion of explicit, interpretable parameters offers a general path to transparent, AI-guided modeling of complex Earth system processes, with implications for drought forecasting, flood risk assessment, and low flow management.

1. Introduction

A central challenge in modern Earth system modeling is to integrate AI-discovered formulas into process-based models without degrading performance or physical consistency. Earth functions as a tightly coupled system in which energy, water, and carbon exchanges move from upstream to downstream. Land-surface parameterizations supply boundary fluxes to the atmosphere, and photosynthesis regulates ET through stomatal conductance (Sellers et al., 1997). When components are tuned in isolation and later inserted into the full model, error compensation often occurs: downstream modules absorb upstream errors, local metrics can look strong, and the overall system loses identifiability and cross-process coherence. Without structural priors and dependency-aware optimization, such compensation is hard to diagnose or control.

We use hydrologic modeling as a tractable and decision-relevant testbed for strategies that reduce error compensation and improve cross-process consistency, with the broader goal of extending these ideas to more complex coupled systems. Hydrologic models support flood forecasting (Rogelis and Werner, 2018; Teja and Manikanta, 2023; Xu et al., 2024), drought monitoring (Han et al., 2023; Mei et al., 2023; Vieira and Stadnyk, 2023), and reservoir management (Abeshu et al.,

* Corresponding author.
E-mail address: liqingliang@ccsfu.edu.cn (Q. Li).

https://doi.org/10.1016/j.jhydrol.2026.136201
Received 26 February 2026; Received in revised form 29 June 2026; Accepted 6 August 2026
Available online 11 August 2026

2023; Beça et al., 2023; Vora et al., 2024), all of which require reliable predictions and interpretable internal states (Clark et al., 2017; Coron et al., 2014; Menard et al., 2021). Traditional frameworks simulate storage components such as snowpack, soil moisture, groundwater, and surface water, as well as fluxes such as ET, surface runoff, subsurface flow, and baseflow (Pomeroy et al., 2022). Many machine learning approaches, in contrast, target streamflow or other outcomes directly and provide limited process insight (Li et al., 2025; Li et al., 2022a; Li et al., 2022b). They can perform well under historical conditions, yet their value declines under change when decision makers need to understand how individual processes respond (Shen et al., 2023).

Even in process-based models, predictive skill is often limited by parameter and structural uncertainty. Key parameters such as hydraulic conductivity and storage capacities are rarely observed and must be inferred, which introduces ambiguity and equifinality in heterogeneous catchments (Beven, 2001). Calibration can mask internal errors through compensation (Seibert and Vis, 2012). Empirical assumptions and simplified representations add structural bias; because structure and parameters are interdependent, joint optimization is difficult (Clark et al., 2017; Draper, 1995; Loucks and van Beek, 2017). Calibrating only to streamflow can produce high outlet skill while misrepresenting ET, infiltration, and runoff generation. For example, in Iran's Karkheh Basin the Hydrologiska Byråns Vattenbalansavdelning (HBV) model achieved a high Nash–Sutcliffe efficiency for total runoff (NSE = 0.86) yet performed poorly for ET (NSE = 0.48), with seasonal biases between wet and dry periods. Such discrepancies argue for calibration and optimization that treat the hydrologic system as a whole rather than focusing only on outlet performance (Rientjes et al., 2013; Tsai et al., 2021).

Recent deep learning results are strong for soil moisture(Li et al., 2022a,b, 2024), runoff (Cai et al., 2024; Frame et al., 2022a; Kratzert et al., 2024), and water temperature (Rahmani et al., 2021a; Rahmani et al., 2021b). Hybrid designs embed physical constraints in learning architectures and improve both skill and physical consistency (Feng et al., 2023b; Read et al., 2019; Xie et al., 2021). Differentiable frameworks allow direct implementation of simplified hydrologic processes and joint optimization of physical parameters with neural components (Feng et al., 2022a; Kochkov et al., 2024; Rahmani et al., 2023). These ideas reduce structural uncertainty and exploit shared patterns across basins, yet many learned modules remain black boxes. In Earth system science, accurate predictions are not sufficient; scientific progress requires clear links between model behavior and process mechanisms (Häfnera et al., 2023).

Symbolic regression (SR) offers a path to interpretability by discovering analytical expressions that relate inputs to outputs (Bartlett et al., 2023; Lemos et al., 2022; Udrescu and Tegmark, 2020). Integrations with modern learning include problem simplification (Keren et al., 2023), end-to-end methodologies (Alnuqaydan et al., 2023; Kamienny et al., 2022), and neural architectures with embedded symbolic structure (Kim et al., 2020; Zhang et al., 2022). Deep symbolic regression (DSR) uses reinforcement learning to recover concise formulas even under noise and holds promise for geoscience applications (Matsubara et al., 2022; Petersen et al., 2021). This innovative approach provides significant insights for applications in Earth science(Tenachi et al., 2023; DiPietro and Zhu, 2022; Usama, 2022). For example, Häfnera et al. (2023) distilled a neural wave model into a closed-form expression that preserved skill and aligned with wave theory. Our prior work (Li et al., 2024b) introduced knowledge-guided DSR for Penman–Monteith, improving surface resistance estimates while honoring key constraints.

Most SR efforts in hydrology and Earth system science, however, target isolated datasets or single components (Häfnera et al., 2023; Li et al., 2024b). The unresolved problem is functional integration: can AI-derived expressions be inserted into coupled models in ways that improve structure, predictive skill, and process consistency rather than amplifying error compensation? When SR modules sit upstream, their errors can propagate and be absorbed downstream, erasing gains seen in offline tests.

HBV provides a suitable platform to address this gap. The model has a clear structure with modules for snow accumulation and melt, soil moisture, runoff generation, and baseflow. It captures essential interactions among hydrologic processes while remaining computationally efficient. This makes it a practical setting to examine how AI-derived formulas can be embedded and how dependency-aware optimization can reduce error compensation

Our study is organized around three questions.

Q1. Can DSR, guided by the HBV process structure, improve ET and runoff parameterizations within a hybrid LSTM-HBV model relative to the original schemes?

Q2. Do the symbolic regression derived formulas improve performance after they are integrated into the original HBV model?

Q3. If direct insertion is not sufficient, which integration and training strategies can make DSR effective at the system level, and what do these results suggest for more complex coupled models?

To address Q1, we simulate ET and total runoff using observed meteorological forcings and catchment attributes within a hybrid framework that couples a LSTM with HBV. This design builds on evidence that domain-informed DSR can outperform traditional hydrologic formulations. We use standard DSR to derive a new ET function and a knowledge-guided variant to derive a total runoff function that preserves the physically coherent relationships among surface runoff, interflow, and baseflow (Li et al., 2024b). The resulting symbolic expressions are evaluated against observed discharge and the MOD16A2-based ET reference.

For Q2, we embed the discovered ET and runoff formulas into the original HBV model one at a time. We then simulate ET and streamflow under the same observed forcings and compare results with those from the unmodified HBV and the LSTM-driven HBV baseline.

For Q3, we test a sequential, dependency-aware training strategy that stabilizes the ET component before calibrating runoff parameters. We evaluate each configuration using overall predictive skill and its representation of key hydrologic behaviors, with emphasis on surface runoff and baseflow. We then identify the configuration that produces an improved HBV model with SR-derived process representations while preserving process dependencies from upstream to downstream. This approach illustrates a path toward transparent and physically grounded AI components for coupled hydrologic and Earth system models.

2. Methodology

2.1. Overall framework

Earth system models (ESMs) represent energy, water, and carbon cycles through interconnected process modules. Vegetation interception and throughfall partition rainfall, which controls infiltration and runoff generation. Incoming shortwave and longwave radiation set net radiation and determine how energy is divided into sensible, latent, and soil heat fluxes. These coupled processes govern both ET and streamflow: ET is co-limited by available energy and canopy water, and streamflow arises from infiltration, infiltration- and saturation-excess runoff, and routing. In this study, error compensation refers to the situation in which biases in ET, water storage states, and runoff generation offset one another, allowing a model to reproduce outlet streamflow without necessarily improving the realism of its internal hydrologic processes. This problem is closely related to the concept of equifinality (Beven and Freer, 2001), where different model structures or parameter sets can yield similar observable responses in complex environmental systems. As a result, good streamflow performance alone does not ensure that the simulated internal processes are unique, physically consistent, or more realistic. When streamflow is used as the primary calibration target, errors in ET or storage dynamics may be hidden by compensating errors in runoff generation, leading to apparently satisfactory discharge simulations but potentially misleading process representations. Motivated by this, we concentrate on the hydrologic components and embed a data-driven approach within the HBV rainfall–runoff model.

To maintain physical realism, we first evaluated baseline HBV state variables against observations and remote-sensing products. Based on this assessment, we identified ET and streamflow as the primary targets for optimization. We then decomposed HBV into two coupled submodules: an ET module driven by atmospheric forcings, and a runoff module driven by surface soil moisture after ET. All other HBV processes, including snow dynamics, soil storage, and groundwater, remained in their original form. Only the ET and runoff parameterizations were replaced with data-driven formulations. In summary, our framework integrates DSR into a differentiable HBV model (Aghakouchak and Habib, 2010; Beck et al., 2020; Seibert and Vis, 2012) to learn new symbolic equations for ET and runoff while preserving the underlying process structure.

2.2. Dsr-driven HBV model

We implemented the HBV model and the DSR modules in a differentiable computing framework (PyTorch) to support end-to-end optimization. The integrated system uses deep learning to propose explicit symbolic expressions for selected processes while retaining the modular physical structure of HBV. In this setup, the DSR network acts as a submodel that generates analytic formulas (expression trees) for ET and runoff. A recurrent neural network (RNN) decoder produces sequences of tokens in prefix notation (operators, variables, and constants) to construct each candidate equation. The token set includes input forcings such as precipitation and temperature, HBV state variables such as soil moisture, and model parameters, combined through standard mathematical operators (for example, addition, multiplication, and logarithms).

We steer the symbolic search with knowledge-based priors to maintain physical plausibility and limit complexity. Expression depth and allowable operators are constrained, and known parameter ranges are imposed. This knowledge-guided DSR (KG-DSR) balances predictive accuracy and interpretability by penalizing overly complex formulas. Network weights and the constants inside the generated expressions are optimized jointly to minimize prediction error on the training data. In this way, the DSR-driven HBV framework learns explicit equations for ET and runoff that enhance model performance without discarding established process relationships. Fig. 1 shows the overall model structure, with default HBV equations in gray and the DSR-derived ET and runoff components highlighted. Table 1 summarizes the core model parameters and their roles.

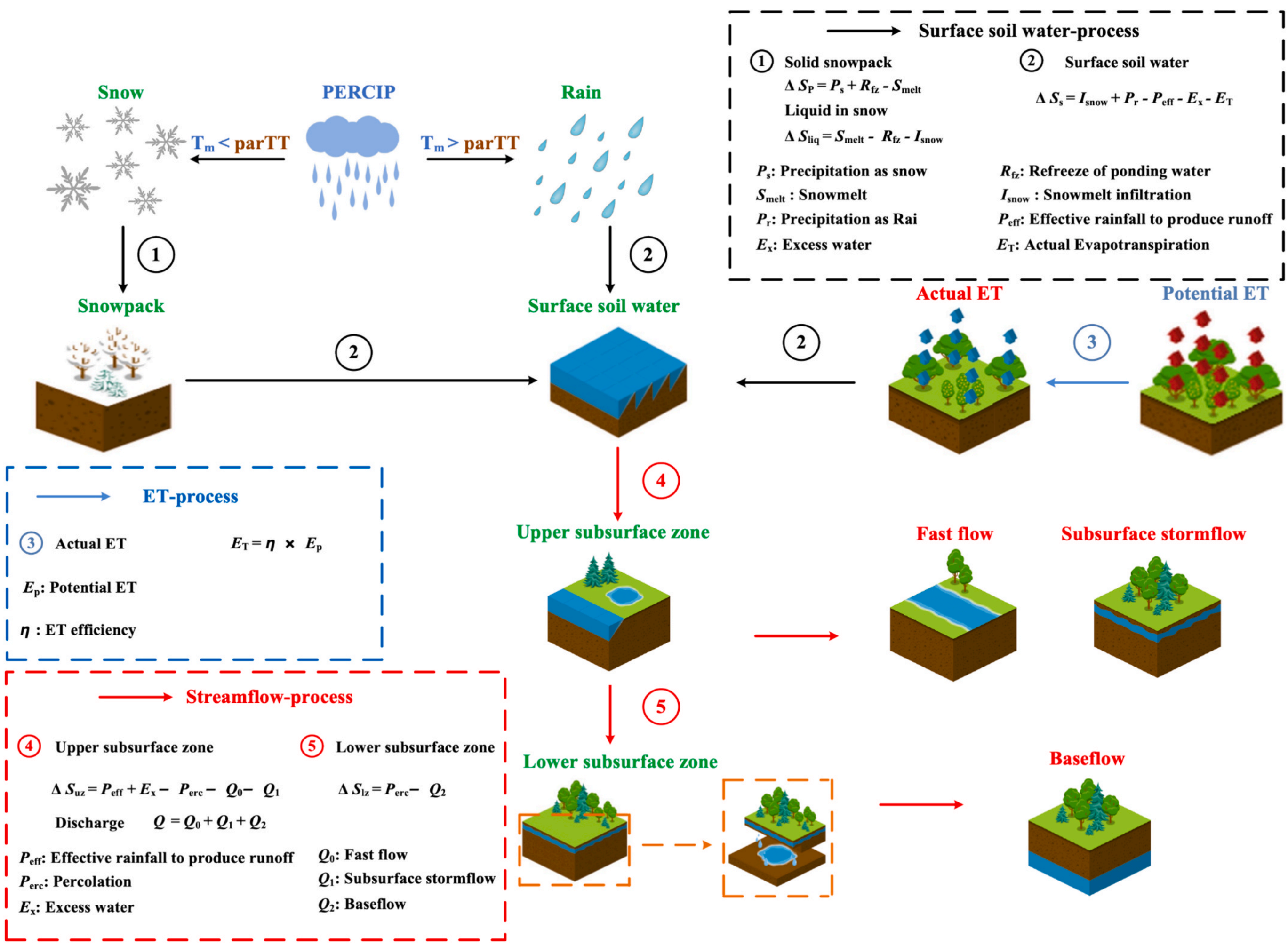

Fig. 1. Overview of the HBV Hydrologic Model Process. Blue represents input variables for the HBV model, brown denotes model parameters, green indicates intermediate variables simulated by the model, and red identifies variables validated in this study. Arrows illustrate the progression of variable transformations, with each pathway representing a specific physical process in the HBV model. These processes are described within color-coded dashed rectangles. The numbered circles along the arrows correspond to the expressions for each physical process. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Key Equations for the Original and DSR-optimized Time-Discrete Models. In the HBV model, θ denotes physical parameters. Ä indicates changes in a daily time step.

Structural priors	Default HBV equations	DSR-optimized schemes
Snow	Solid snowpack $\Delta \boldsymbol{S_p}$ $R_{fz} = (\theta_{TT} - T)\theta_{DD}\theta_{rfz}S_{smelt} = (T - \theta_{TT})\theta_{DD}$Liquid in snow $\Delta \boldsymbol{S_{lip}} I_{snow} = S_{lip} - \theta_{CWH} \times S_p$	Unchanged
Evapotranspiration (ET)	Actual ET $\boldsymbol{E_T}\eta = \min(S_s/(\theta_{FC}\theta_{LP}), 1)$	$\boldsymbol{E_T} = \boldsymbol{f_{DSR}}(\boldsymbol{x_{forcing}}, \boldsymbol{x_{Attributes}})$
Surface soil water	Surface soil water $\Delta \boldsymbol{S_s} P_{eff} = W(P_r + I_{snow})W = \min\left((S_s/\theta_{FC})^{\beta}, 1\right)E_x = (S_s - \theta_{FC})$	Unchanged
Streamflow	Upper subsurface zone $\Delta \boldsymbol{S_{uz}}$ $P_{erc} = \min(\theta_{perc}, S_{uz})Q_0 = \theta_{K0}(S_{uz} - \theta_{uzl})Q_1 = \theta_{K1}S_{uz}$Lower subsurface zone $\Delta \boldsymbol{S_{lz}} Q_2 = \theta_{K1}S_{lz}\boldsymbol{Q} = Q_0 + Q_1 + Q_2$	$\boldsymbol{Q} = \boldsymbol{f_{KG-DSR}}(\boldsymbol{x_{forcing}}, \boldsymbol{x_{Attributes}})$

2.2.1. HBV hydrological model

The HBV model is a lumped conceptual rainfall–runoff model that simulates the daily water balance at the basin scale (Fig. 1). It includes storage components for snowpack, soil moisture, and groundwater, along with modules for precipitation partitioning, evaporation, and runoff generation. At each time step, precipitation and temperature update the storage terms: snow accumulates or melts depending on temperature, soil moisture is recharged by infiltration, and groundwater receives recharge from excess soil water. Potential ET is computed with a standard formulation (for example, Thornthwaite or Penman) and is limited by soil moisture availability. Water that is not stored or evaporated generates runoff, with surface (quick) runoff arising from excess soil moisture and baseflow produced by linear groundwater reservoirs. Total streamflow is the sum of surface runoff and baseflow.

In this study, we retain the original HBV formulations for snow, soil moisture, and groundwater. We modify only the ET and runoff modules. Each watershed is represented with a single-basin (lumped) HBV model. This choice simplifies channel routing but emphasizes interactions among internal processes and clarifies how structural priors and parameter changes influence system behavior. The routing from soil moisture to runoff remains unchanged; only the mathematical forms of the ET flux and the runoff generation functions are replaced by DSR-derived expressions.

2.2.2. Deep symbolic regression (DSR) model

We established a framework to systematically evaluate parameterizations of ET and runoff. The approach began with offline testing of two parameterizations: DSR-ET for ET and KG-DSR for runoff. Next, these parameterizations were integrated into the HBV model, both individually and in combination, to assess their influence on overall model performance. Evaluation was based on four standard metrics—R, RMSE, bias, and KGE. To examine extremes and process representation, we further assessed model performance using high- and low-flow quantiles ($Q5$ and $Q95$, corresponding to 5% and 95% exceedance probabilities) and the Baseflow Index (BFI) to evaluate the characterization of slow-response components. Finally, we analyzed cross-process interactions: changes in runoff performance after incorporating DSR-ET, and shifts in ET behavior when the runoff parameterization was replaced. All experiments were conducted with consistent datasets and identical training–validation splits to ensure that results are both robust and reproducible.

2.3. DSR based-driven HBV hydrological model

The DSR based-driven HBV hydrological model enhances differentiable HBV models by integrating the DSR method. This integration utilizes DL techniques to optimize process-based models, similar to how LSTM models enhance HBV parameter optimization. By combining DL with process-based modeling on the PyTorch platform, this framework allows for seamless integration and training as a unified model. A notable feature of the DSR approach is its ability to drive explicit physical equations through influencing factors while simultaneously optimizing both the structural priors and parameters of the process model. This strategy ensures that all components of the resulting model exhibit improved process transparency. Fig. 1 and Table 1 denotes the core assumptions of the DSR optimization framework, which we will assess to understand the impact of these optimization. Fig. 1 illustrates the overall HBV hydrological model process and highlights the default HBV equations aligned with structural priors. Table 1 provides supplementary details on parameters.

2.3.1. Deep symbolic regression model

The DSR model conceptualizes symbolic expressions as binary trees whose nodes represent variables, constants, or operators (Petersen et al., 2021). In a hydrologic setting, this structure provides a natural way to represent process equations, since it combines flexibility with interpretability while capturing water movement and storage dynamics. The variable set includes meteorological forcings and catchment attributes, and the constants include both standard HBV parameters and additional free parameters. Operators cover basic arithmetic and common nonlinear functions such as exponentials, logarithms, squares, and divisions, which together describe interactions among hydrologic components.

To generate process equations, DSR uses an RNN that sequentially emits tokens and updates its hidden state based on observations and previously generated tokens (Fig. 2). Tokens correspond to the building blocks of an expression: variables, operators, and constants. The RNN produces these tokens in prefix notation until a complete, syntactically valid formula is formed. At each step, a categorical distribution over the token set is updated, and domain-informed priors reshape this distribution to enforce constraints and to reduce the size of the search space. These priors control maximum tree depth, restrict nested operations, and discourage unnecessary complexity, which helps balance robustness and generalizability. In combination, sequence generation, RNN dynamics, and prior-adjusted categorical sampling provide a coherent mechanism for constructing compact, data-driven analytical expressions for hydrologic processes.

Because the symbolic expressions and their associated loss are not differentiable with respect to the discrete tokens, direct optimization with standard automatic differentiation is not feasible. Instead, DSR uses a neural policy that defines a categorical distribution over tokens. The policy is updated so that token sequences producing high quality expressions become more likely over time. In our case, the model uses a risk seeking policy gradient with entropy regularization, which encourages exploration while concentrating probability mass on well performing formulas. Evaluation of each candidate expression requires fitting its free constants, interpreted as hydrologic parameters. DSR employs an L-BFGS routine (Zhu et al., 1997) for this inner optimization. For complex hydrologic parameter spaces, this approach can be restrictive, so in our later HBV experiments we introduce an LSTM-based optimizer to adjust parameters within the DSR-driven HBV model (Section 2.2). As a result, DSR in this study serves primarily to reveal functional relationships among variables and to provide candidate process equations.

Each candidate expression f receives a scalar reward that reflects its fit to the data. The reward is defined as:

$$R = 1/(1 + NRMSE)$$

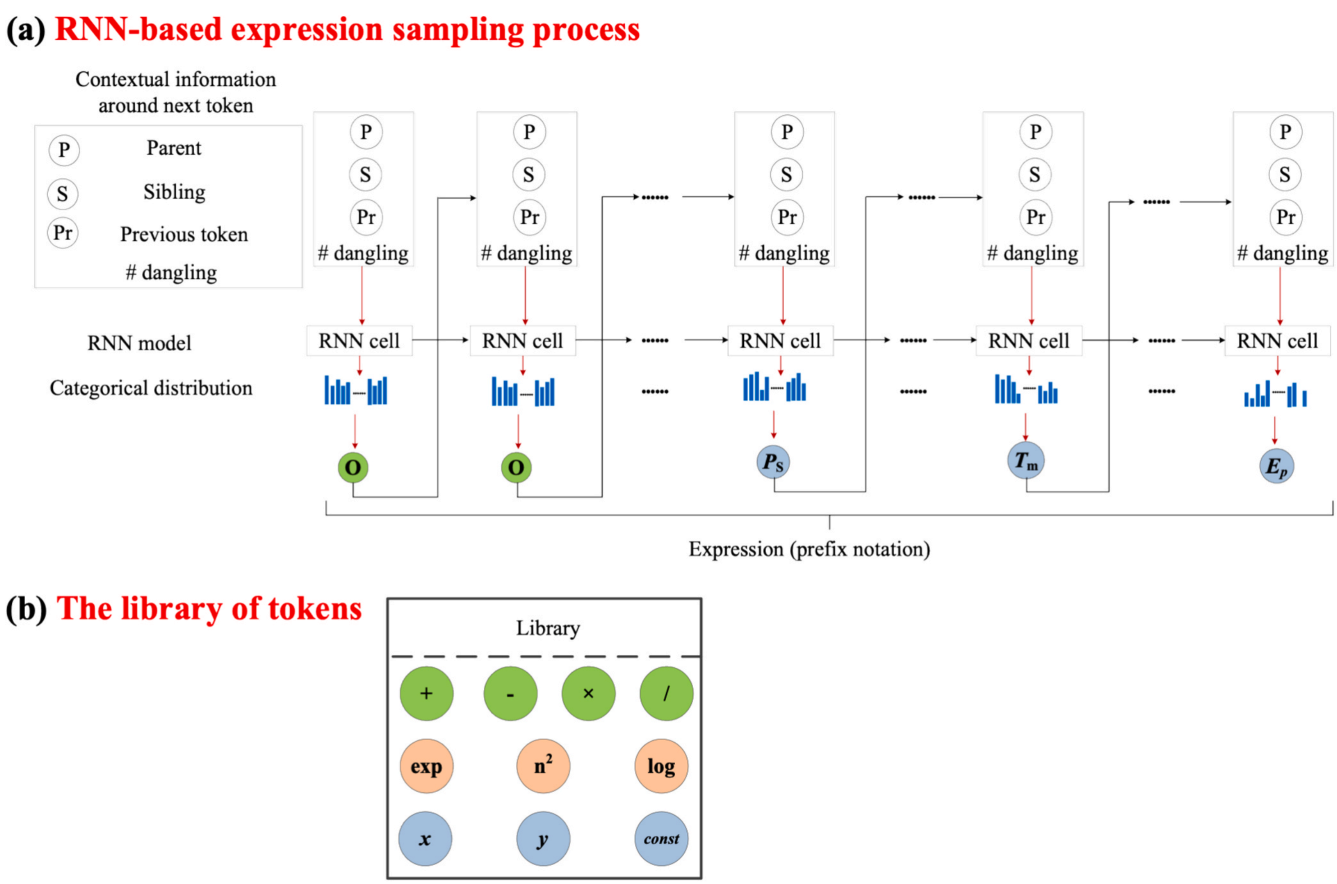

Fig. 2. (a). Illustration of the RNN-based expression sampling process. For each token, the RNN generates a categorical distribution, from which a token is sampled. The parent and sibling of the next token serve as inputs to the RNN. Tokens continue to be sampled autoregressively until all tree branches reach terminal nodes, completing the tree structure. This sequence of tokens represents a pre-order traversal, enabling reconstruction of the tree and instantiation of the corresponding expression. (b). The library of tokens.

where *NRMSE* is the normalized root mean square error between simulated ET and the MOD16A2 ET reference, or between simulated streamflow and observed discharge. The policy is trained with Adam using a risk seeking objective:

$$J_{risk}(\theta;\varepsilon) \doteq \mathbb{E}_{\tau\ p(\tau|\theta)}[R(\tau)\,|R(\tau) \geq R_{\varepsilon}(\theta)\,] \tag{1}$$

where θ denotes the policy parameters, τ a sampled program, and $R_{\varepsilon}(\theta)$ the reward threshold at the $(1-\varepsilon)$-quantile. The corresponding gradient is

$$\nabla_{\theta}J_{risk}(\theta;\varepsilon) = \mathbb{E}_{\tau\ p(\tau|\theta)}[(R(\tau) - R_{\varepsilon}(\theta)\,) \bullet \nabla_{\theta} logp(\tau|\theta)\,|R(\tau) \geq R_{\varepsilon}(\theta)\,] \tag{2}$$

In practice, this expectation is approximated with a Monte Carlo estimate over a batch of N samples:

$$\nabla_{\theta}J_{risk}(\theta;\varepsilon) \approx \frac{1}{\varepsilon N}\sum_{i=1}^{N}\left[\left(R\left(\tau^{(i)}\right) - \widetilde{R}_{\varepsilon}(\theta)\right) \bullet 1_{R\left(\tau^{(i)}\right) \geq \widetilde{R}_{\varepsilon}(\theta)} \nabla_{\theta} logp\left(\tau^{(i)}|\theta\right)\right] \tag{3}$$

where $\widetilde{R}_{\varepsilon}(\theta)$ is the empirical $(1-\varepsilon)$-quantile of the batch rewards and $1_{\{\bullet\}}$ is the indicator function.

2.3.2. Knowledge-guided deep symbolic regression model

A central challenge in streamflow modeling is that only total discharge is observed, while its components—surface runoff, subsurface stormflow, and baseflow—cannot be directly measured. A standard DSR model trained on total streamflow can represent only the aggregate response and does not recover the internal structure of these components. To address this, we use a knowledge-guided DSR (KG-DSR) model (Li et al., 2024b) that incorporates physical constraints into the symbolic search. These constraints organize the streamflow equation into three interacting branches that correspond to fast flow, subsurface stormflow, and baseflow, and they ensure that the sum of the branches remains consistent with observed total streamflow. The overall procedure is summarized in Fig. 3.

An RNN-based sampler generates prefix expressions one token at a time, and at each step it outputs a probability distribution over candidate tokens. This distribution is shaped by two types of priors. Hard priors enforce syntactic validity and dimensional consistency. Soft hierarchical priors encourage a three-branch structure aligned with the decomposition of streamflow into fast, intermediate, and slow components. Process structure is encoded through token sublibraries. For each subprocess i, we define a sublibrary L_i that contains binary operators $\{+, -, \times, Ã\cdot, \surd\}$, variables relevant to that subprocess, and a shared parameter library (Fig. 3a). Expression generation always begins with binary operators; during the first two steps, variables and unary operators are not allowed, and the root node is fixed as an additive operator with probability one. This design yields a three-branch additive root that mirrors the HBV partitioning (Fig. 4).

Activation of the sublibraries is controlled by tree depth. Let $Occ(d)$ denote the number of nodes already generated at depth d. When $Occ(1)\ = 1$, only the fast-flow library and the shared parameter library are active. After both children of the root are present $(Occ(1) = 1, cc(2) = 2\,)$, the subsurface stormflow library is activated. When the first two branches are complete $(Occ(1) = 2, cc(2) = 2\,)$, only binary operators are permitted so that the third branch connects to the existing structure. Once $Occ(1) = 2$ and $Occ(2)\ = 3$, the baseflow library and the shared parameter library become available. These rules guar-

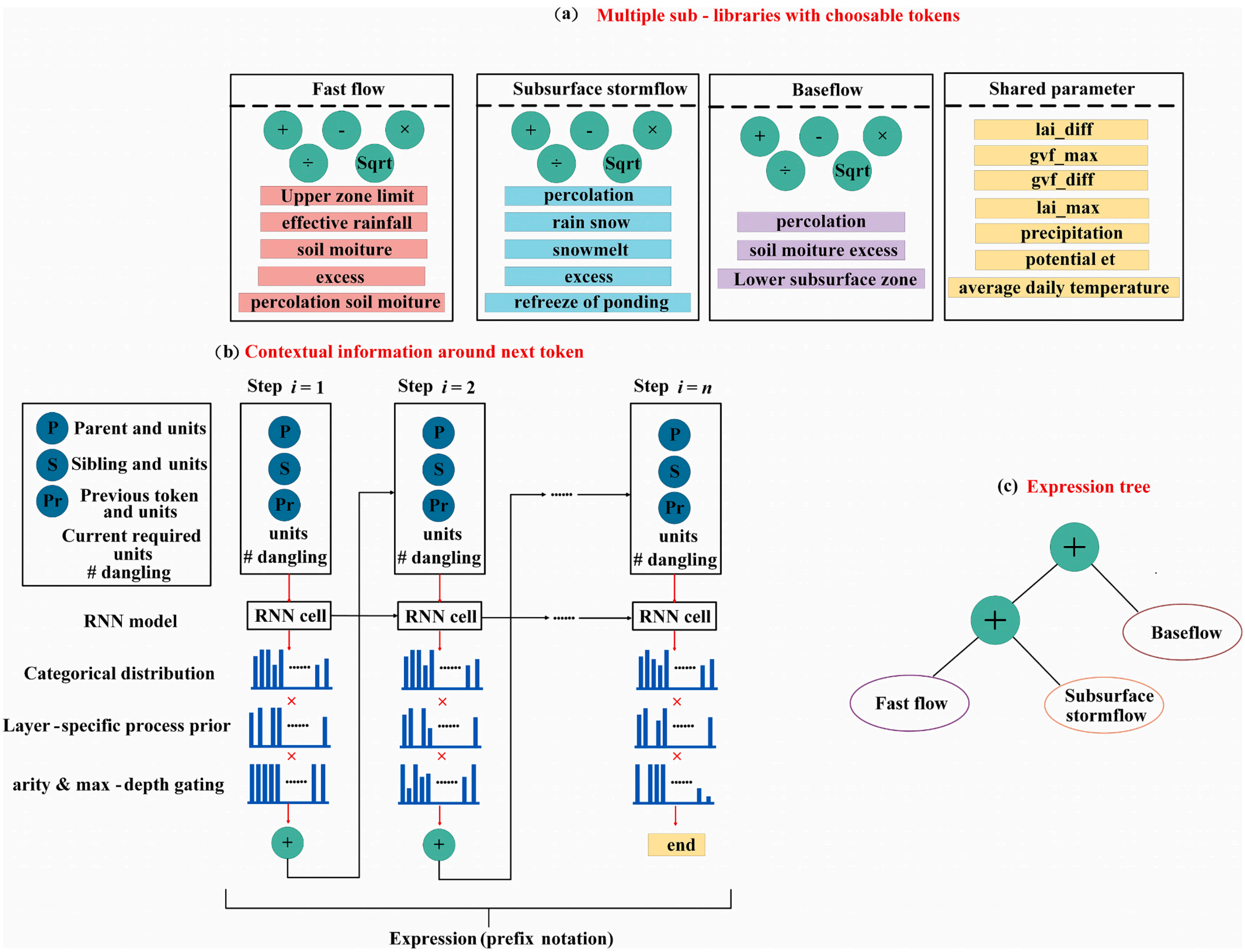

Fig. 3. (a). Includes three process-specific sub-libraries—Fast flow, Subsurface stormflow, and Baseflow—as well as a Shared-parameter sub-library. (b). Illustrates the expression sampling pipeline from the RNN: starting from the contextual inputs and the categorical distribution, then applying the arity & max-depth gating and the layer-specific process prior, until the prefix-form expression is produced. (c) The final expression is organized in a tree structure, with the root node as an additive aggregator and three subtrees corresponding to fast flow, subsurface storm runoff, and base flow, respectively, to achieve a structured and explicit decomposition of the total runoff.

antee that the sampled expression maintains a three-branch additive skeleton consistent with the hydrologic process ordering.

To balance structural growth and termination, token probabilities are further modulated with Gaussian weights. For a given sublibrary L_i, the weight at sampling step i is

$$Gaussian(i) = exp\left(\frac{(i - step_{loc})^2}{2 \times scale}\right)$$

where $step_{loc} = |L_i|$ and *scale* controls the spread. Early in generation ($i < step_{loc}$), this schedule reduces the probability of terminal symbols (variables and constants) and favors structural operators. Later ($i \geq step_{loc}$), operator probabilities are down-weighted, which encourages closure of expression branches with terminal symbols. A maximum tree depth prevents uncontrolled growth; once this depth is reached, all nonterminal symbols receive zero probability and only terminals can be chosen. Candidate expressions are trained with reinforcement learning. Each sampled program is evaluated against observed streamflow and the structural priors, and policy gradients update the RNN to maximize the combined reward.

For reproducibility, the KG-DSR procedure can be summarized as follows.

Define sub-libraries L_{fast}, $L_{sinsirface}$, L_{base}, and a shared parameter library.
Initialize an RNN policy that emits token logits at each step.
Constrain the first two steps to additive operators and disallowing variables/unary operators.
Use depth-based counts $Occ(d)$ to activate the appropriate sub-libraries and operator sets.
Apply Gaussian schedules within each sublibrary to balance operator and terminal selection.
Impose a maximum tree depth and restrict sampling to terminals beyond that limit.
Train the policy with reinforcement learning, combining accuracy on total streamflow with penalties and rewards that encode structural consistency.
Validate sampled expressions by checking dimensional consistency and verifying that the sum of the three components matches observed streamflow.

This framework produces compact, physically guided symbolic expressions for surface runoff, subsurface stormflow, and baseflow that remain consistent with observed discharge. Fig. 3 and Fig. 4 illustrate the sublibrary design and the depth-driven sampling procedure. By

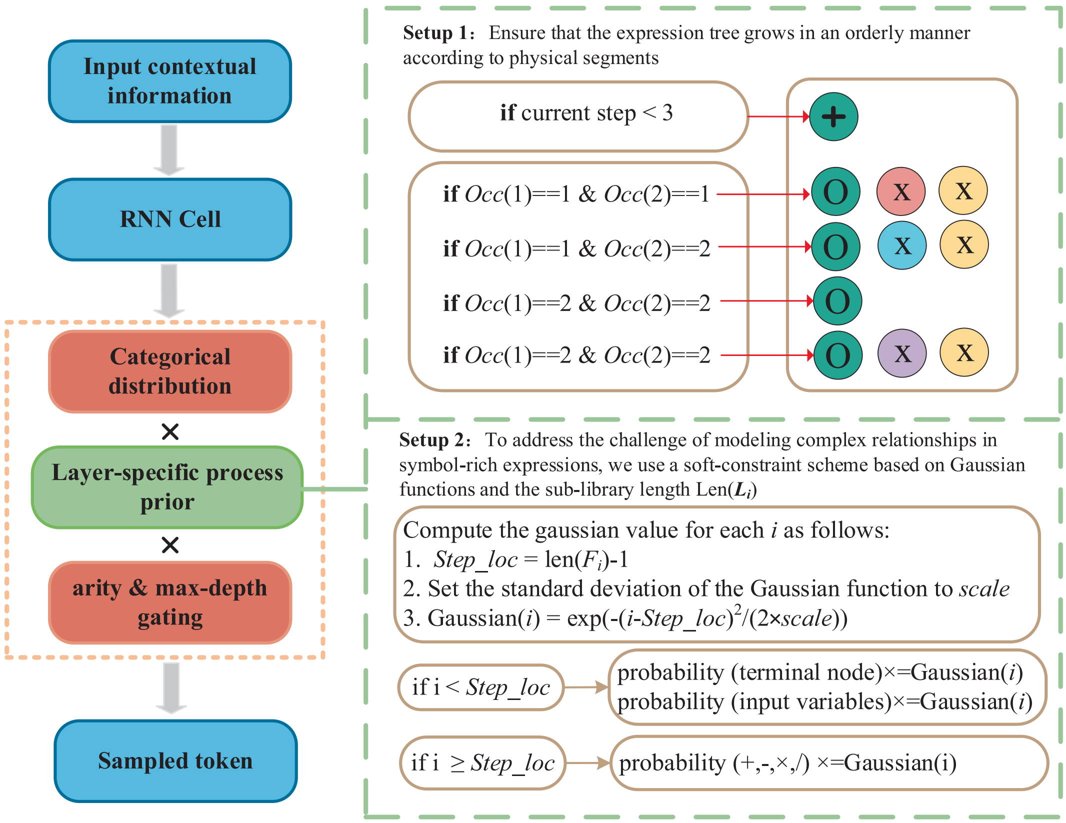

Fig. 4. Schematic diagram of the KG-DSR model and layer-specific process priors.

explicitly structuring the relationships among the three flow components, the KG-DSR model improves the physical consistency of streamflow simulations and yields a more explicit, structured representation of the underlying processes.

2.4. Experimental design

We designed three experiment sets to test whether DSR and KG-DSR can improve HBV parameterizations without changing the overall model structure, as illustrated in Fig. 5.

2.4.1. Experiment 1 (Q1): parameter derivation

We first used DSR and KG-DSR to derive new equations for ET and runoff from the training data. The DSR-ET module was trained to discover an explicit symbolic expression aligned with the MOD16A2-based ET reference. Since MOD16A2 is a model-based remote-sensing product, this step should not be interpreted as learning an unbiased or absolute representation of true ET. Rather, the goal is to derive a physically interpretable ET parameterization that links HBV state variables and meteorological forcings to the dominant ET variability captured by an independent remote-sensing reference. The learned ET expression is then embedded into HBV and evaluated as part of the full hydrological modeling chain. Therefore, its usefulness is not determined only by offline agreement with MOD16A2. More importantly, it is evaluated by whether it constrains the subsequent learning of runoff parameterization and improves the consistency of the overall hydrological simulation. Both models used the same meteorological forcings, basin attributes, and train–validation split as the baseline HBV. The resulting DSR-based ET and runoff formulations were then compared with the original HBV equations by evaluating simulated ET against MOD16A2 and simulated runoff against observed discharge.

2.4.2. Experiment 2 (Q2): ET embedding

Next, we selected the best performing DSR-derived ET formula from Experiment 1 and embedded it in HBV in place of the original ET module. All other model components, including runoff generation and routing, were kept identical to the baseline configuration. We retrained this hybrid model and evaluated whether the revised ET representation improved both ET and streamflow relative to the original HBV.

2.4.3. Experiment 3 (Q3): runoff improvement strategies

Finally, we tested two strategies for improving runoff:

Runoff first. KG-DSR was used to derive a new runoff equation while retaining the original HBV ET scheme.
ET first (staged). We first embedded and trained the DSR-based ET module from Experiment 2 to convergence. Using these updated ET outputs, we then applied KG-DSR to derive a new runoff formula. This staged setup tests whether correcting ET upstream enhances the subsequent runoff representation.

All experiments used the same datasets and identical training–validation splits to allow fair comparison. Model performance was assessed with root mean square error (RMSE), Pearson correlation (R), bias, and Kling–Gupta efficiency (KGE). We also evaluated extreme flow behavior by comparing the 5th and 95th percentile flow quantiles (Q5 and Q95) and the baseflow index (BFI) across models. Performance was examined for basins spanning a range of climates and hydrographic types to test generality. This experimental design quantifies the benefits and limitations of DSR and KG-DSR relative to the original HBV, identifies effective integration strategies, and reveals any compensatory fluxes or distortions introduced by the new parameterizations.

Detailed settings for Q1–Q3, including the DSR and KG-DSR configurations, ET embedding, and runoff optimization strategies, are described in Supplementary Texts S1–S3. Supplementary Tables S4–S7

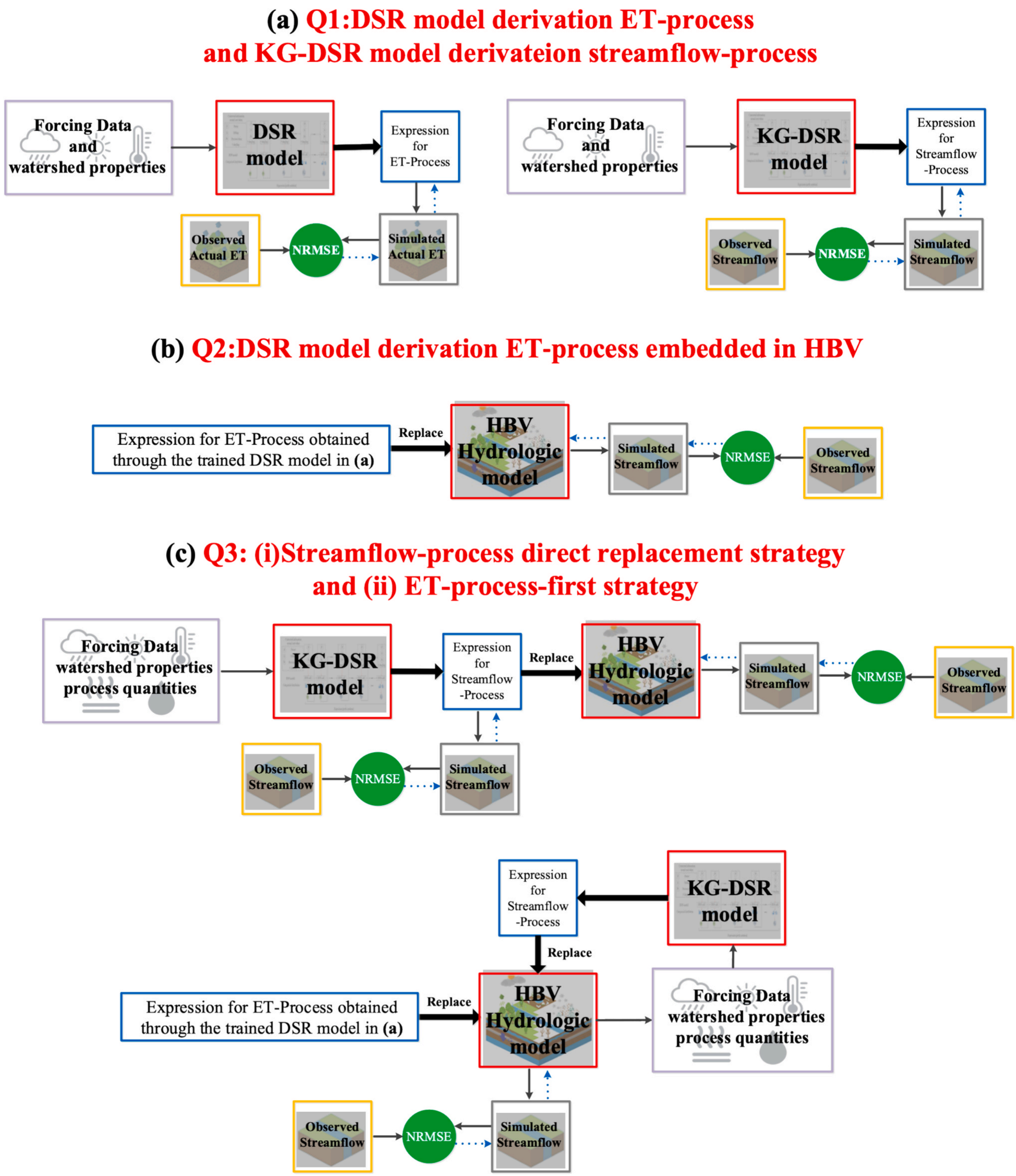

Fig. 5. Comparison of DSR/KG-DSR-based hydrologic modeling workflows and integration strategies (Q1–Q3). (a). The figure illustrates the overall workflow for optimizing the ET process with a DSR model and the streamflow process with a KG-DSR model. During training, the normalized root mean square error (NRMSE) between simulations and observations is used as the loss to update model parameters. Solid black arrows denote the forward pass toward the simulation targets, while blue dashed arrows indicate the backpropagation path used to adjust model weights. (b). This framework shows how the DSR-optimized ET process from panel (a) is embedded into the HBV model. (c). The figure illustrates two strategy workflows: (i) directly optimize the streamflow process using KG-DSR and embed it into the HBV model; (ii) first embed the DSR-optimized ET process (panel a) into HBV, use the updated intermediate outputs as inputs to KG-DSR, then optimize the streamflow process and embed it into HBV. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

provide additional details on the intermediate variables, formula structures, configuration settings, and coefficient combinations.

2.5. Reproducibility and symbolic-expression selection

Because both DSR and KG-DSR use sampling methods based on reinforcement learning, the symbolic search process can be sensitive to stochastic variation. To improve reproducibility, we fixed the random seeds in both methods and performed multiple independent runs for each symbolic regression task. This design reduced the influence of any single stochastic search path. In each run, the symbolic search continued until the reward showed no sustained improvement. After convergence, we retained the ten expressions with the highest rewards and evaluated them under the same training and validation setting. The final expression was selected based on validation performance, physical interpretability, and formula parsimony. This procedure ensured that the reported formulas were chosen from a set of consistently strong candidates rather than from a single randomly sampled expression.

2.6. Data sets

This study targets two key hydrologic variables: ET and streamflow. For streamflow, we used the CAMELS dataset (Addor et al., 2017; Newman et al., 2014), which provides meteorological forcings, daily discharge, and catchment attributes for 671 basins across the conterminous United States (CONUS). For ET, we used MOD16A2 (Running et al., 2017), which estimates ET over eight-day intervals using a modified Penman–Monteith formulation (Mu et al., 2011). MOD16A2 supplies spatially consistent ET estimates but should be viewed as a model product rather than a strict ground truth. MOD16A2 was used as the ET reference for discovering the DSR-based ET parameterization. We acknowledge that MOD16A2 is a model-based remote-sensing product and should not be interpreted as unbiased ground-truth ET. Therefore, the DSR-derived ET equation is not claimed to recover the true ET process in an absolute sense. Instead, MOD16A2 serves as a spatially continuous, temporally consistent, and discharge-independent reference for constraining the ET component of HBV. This choice was primarily motivated by scale compatibility and data availability. Eddy-covariance measurements provide valuable flux observations, but their footprints are local, dynamic, and sensitive to wind direction, atmospheric stability, measurement height, surface heterogeneity, and terrain. These observations are therefore difficult to compare directly with basin-averaged ET from lumped or semi-distributed hydrological models. This issue is particularly important for CAMELS catchments, which often contain substantial heterogeneity in land cover, soils, topography, and vegetation. A tower located within or near a catchment may still fail to represent basin-scale mean ET. MOD16A2 offers spatially continuous ET estimates at 500 m resolution and 8-day intervals, allowing aggregation to the basin scale. Although it includes algorithmic assumptions and product-specific uncertainties, it is built on the Penman–Monteith framework and integrates MODIS vegetation, land-cover, albedo, and meteorological information. Thus, MOD16A2 provides an ET-related process reference with hydrothermal and vegetation constraints, rather than a purely empirical label. Here, it is used as a scale-compatible external constraint that is independent of outlet-discharge calibration, not as a perfect representation of true ET (Chu et al., 2021; Schmid, 2002). After reconciling temporal coverage and basin boundaries between CAMELS and MOD16A2, we obtained 660 catchments with overlapping records for joint analysis. Eleven catchments were excluded because they lacked complete MOD16A2 ET records across the full training and testing periods. The excluded catchments are listed in Table S8.

To ensure consistency with previous large-sample hydrology work, we followed the established CAMELS training–testing basin partition (Frame et al., 2022b; Kratzert et al., 2018), but adopted a different time window. Data from 1 January 2007 to 31 December 2013 were used for training, covering all experiments (Q1 DSR and KG-DSR, Q2 HBV with embedded DSR, Q3 Strategies 1 and 2, and the baseline configurations). The period from 1 January 2014 to 31 December 2014 served as the common test interval.

Input sets were tailored to each model. The traditional HBV configuration used three core meteorological drivers: precipitation, temperature, and potential ET. The LSTM–HBV variant followed earlier studies (Feng et al., 2023a; Feng et al., 2022b), and incorporated an extended input suite that included 35 static catchment attributes (Table 2). The DSR model combined the three core meteorological inputs with a targeted subset of static attributes (Table S2): lai_max, lai_diff, gvf_max, and gvf_diff, which describe peak canopy density and seasonal greenness, both known to influence ET dynamics (Yang et al., 2023). For KG-DSR, Q1 used the three meteorological drivers plus a subset of static attributes (Table S3). In Q3, the same meteorological and static inputs were augmented with HBV process variables (Table S4) that summarize dynamic states of the watershed water cycle. Many previous studies have shown strong correlations between these process variables and runoff (Cadol et al., 2012; Dunne et al., 1991; Schaefli et al., 2013; Zhang et al., 2019). When supplied to KG-DSR, they provided physically consistent, information-rich inputs that narrowed the symbolic search space, reinforced process constraints, and enabled derivation of a new runoff formulation within the HBV two-layer groundwater framework.

Table 2
Description of the 35 static attributes.

Variables	Descriptions	Units
p_mean	Mean daily precipitation	mm/day
pet_mean	Mean daily PET	mm/day
p_seasonality	Seasonality and timing of precipitation	–
frac_snow	Fraction of precipitation falling as snow	–
Aridity	PET/P	–
high_prec_freq	Frequency of high precipitation days	days/year
high_prec_dur	Average duration of high precipitation events	days
low_prec_freq	Frequency of dry days	days/year
low_prec_dur	Average duration of dry periods	days
elev_mean	Catchment mean elevation	m
slope_mean	Catchment mean slope	m/km
area_gages2	Catchment area (GAGESII estimate)	km 2
frac_forest	Forest fraction	–
lai_max	Maximum monthly mean of the leaf area index	–
lai_diff	Difference between the maximum and minimum monthly mean of the leaf area index	–
gvf_max	Maximum monthly mean of the green vegetation	–
gvf_diff	Difference between the maximum and minimum monthly mean of the green vegetation fraction	–
dom_land_cover_frac	Fraction of the catchment area associated with the dominant land cover	–
dom_land_cover	Dominant land cover type	–
root_depth_50	Root depth at the 50th percentiles	m
soil_depth_pelletier	Depth to bedrock	m
soil_depth_statgso	Soil depth	m
soil_porosity	Volumetric soil porosity	–
soil_conductivity	Saturated hydraulic conductivity	cm/hr
max_water_content	Maximum water content	m
sand_frac	Sand fraction	%
silt_frac	Silt fraction	%
clay_frac	Clay fraction	%
geol_class_1st	Most common geologic class in the catchment	–
geol_class_1st_frac	Fraction of the catchment area associated with its most common geologic class	–
geol_class_2nd	Second most common geologic class in the catchment	–
geol_class_2nd_frac	Fraction of the catchment area associated with its 2nd most common geologic class	–
carbonate_rocks_frac	Fraction of the catchment area as carbonate sedimentary rocks	–
geol_porosity	Subsurface porosity	–
geol_permeability	Subsurface permeability	m 2

2.7. Outputs

In this study, DSR supplied structural priors to HBV in two forms. First, a DSR-derived ET equation identified through symbolic search in Q1 was embedded as the ET module in HBV (Q2). Second, within the constraints of the HBV two-layer groundwater structure, the knowledge-guided KG-DSR learned a new runoff formulation in Q3, which replaced the native HBV runoff module.

The resulting DSR/KG-DSR–HBV systems ran end-to-end at a daily time step to simulate both runoff and actual ET while tracking key hydrologic states and fluxes, including snow storage, soil moisture, infiltration, recharge, and upper and lower groundwater storage. This configuration enabled process-level diagnosis of model behavior and provided a clear basis for assessing how data-driven structural priors modify hydrological dynamics.

2.8. Experimental metrics

To evaluate the DSR- and KG-DSR-driven HBV models, we used an LSTM-optimized HBV configuration as the reference. This baseline represents a state-of-the-art approach that couples deep learning with a process-based hydrologic model, with parameter settings following Feng et al. (2022a) for consistency with prior work. All experiments used the same temporal split, with 2007–2013 for training and 2014 for validation, which allows a fair, direct comparison across models.

Model performance was assessed using four metrics: *RMSE*, *Bias*, *KGE*, and correlation coefficient *R*. These indicators jointly summarize error magnitude, systematic deviation, overall agreement, and temporal correlation. *RMSE* and *KGE* are used as the primary metrics. Let y_i be the observed value on day i, $\hat{y}_i$ the corresponding simulation, $\overline{y}$ and $\overline{\hat{y}}$ the mean observed and simulated values, and N the number of test days in a basin. Denote the standard deviations of observations and simulations by y_{std} and $\hat{y}_{std}$. The metrics are defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}} \tag{4}$$

$$KGE = 1 - \sqrt{(R-1)^2 + \left(\frac{\hat{y}_{std}}{y_{std}} - 1\right)^2 + \left(\frac{\overline{\hat{y}}}{\overline{y}} - 1\right)^2} \tag{5}$$

$$R = \frac{\sum_{i=1}^{N} (y_i - \overline{\hat{y}})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N} (y_i - \overline{\hat{y}})^2}\sqrt{\sum_{i=1}^{N} (y_i - \overline{y})^2}} \tag{6}$$

$$Bias = \frac{\sum (\overline{\hat{y}} - \overline{y})}{N} \tag{7}$$

3. Results

We reformulate the study using a dependency-guided symbolic regression framework that explicitly targets error compensation in coupled models. The framework orders model modules from upstream to downstream, learns explicit formulas for selected modules, and combines them with a training strategy that first stabilizes upstream components and then calibrates downstream ones. It then evaluates system behavior using diagnostics for extremes, baseflow, and residual patterns. Fig. 6 shows a schematic that is agnostic to any specific model structure. The same logic extends to other coupled systems; in this work, hydrology serves as the test case.

Calibrating only to outlet flow can mask substantial internal errors. In HBV, for instance, high streamflow skill can occur alongside strongly biased ET. A runoff expression learned solely from basic meteorological

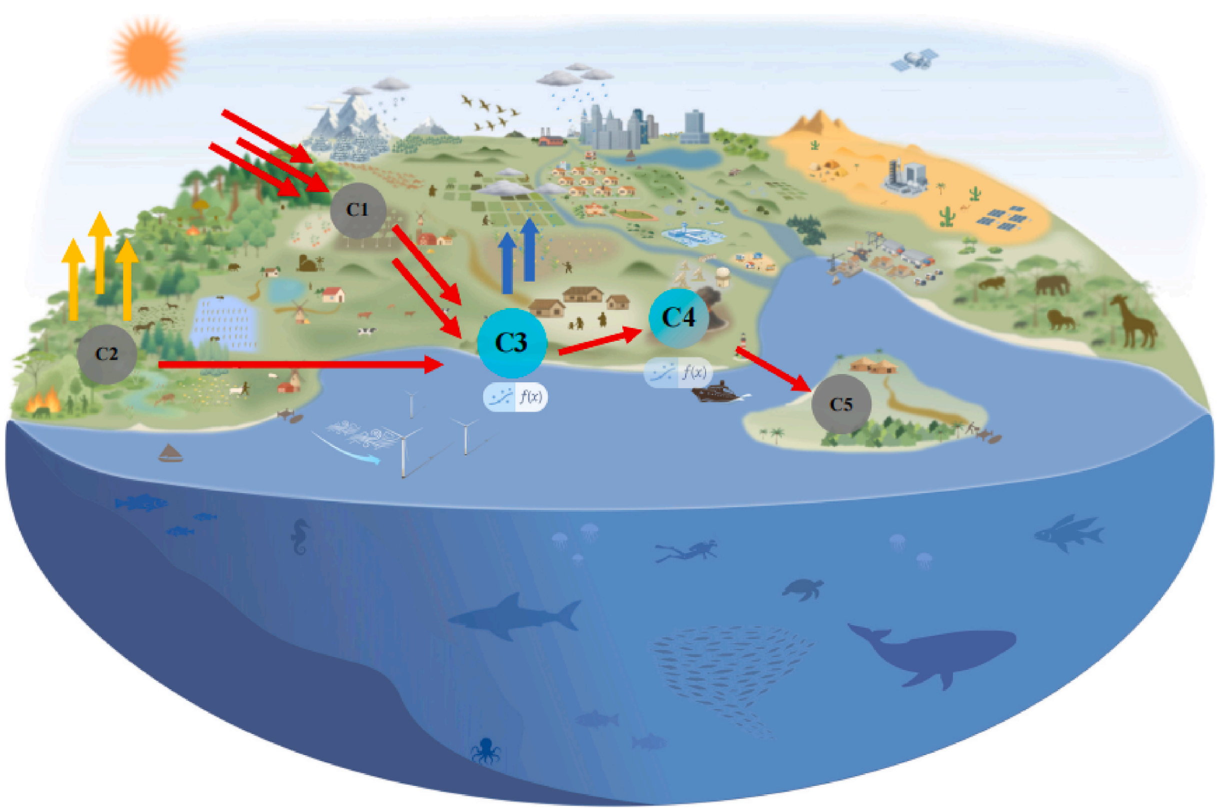

Fig. 6. Dependency-guided symbolic regression framework for a generic coupled system. The abstract system consists of five components ($C1$ through $C5$) arranged from upstream to downstream. Red arrows show the presence and direction of dependencies, and the $f(x)$ icons indicate components whose internal relationships are learned using symbolic regression. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

inputs also performs poorly, because it lacks key hydrologic state information. Together, these behaviors point to error compensation and highlight the limits of global fits that ignore dependencies among model modules.

3.1. Instantiation in hydrological modeling

We first asked whether AI-discovered formulas can identify physically interpretable relationships that are consistent with the selected hydrologic reference data when trained in isolation, before being coupled into the full model. Using observed meteorological forcings (precipitation, temperature, potential ET) and catchment attributes, we trained two symbolic models. The first used standard DSR to predict ET, targeting known vegetation and climate controls on ET. The second used a KG-DSR formulation to predict total runoff while explicitly linking its components (surface runoff, interflow, baseflow) in a way that preserves directional dependencies and mass balance. The resulting expressions for ET and runoff are listed in Table 3, with all variables defined in Supplementary Table S1. When evaluated offline against their corresponding reference data, the formulas showed high skill: ET relative to MOD16A2 and runoff relative to observed discharge: Fig. 7 demonstrates strong KGE and low bias across basins, consistent with Supplementary Fig. S1. The DSR-derived ET formulation improved the agreement between simulated ET and the MOD16A2-based reference. This result should be interpreted carefully. Since MOD16A2 is a model-based remote-sensing product, agreement with MOD16A2 does not prove that the learned equation represents unbiased real-world ET. It shows that the expression captures the main ET variability represented by an independent remote-sensing reference. This offline stage isolates the intrinsic performance of each AI-derived process representation without any influence from compensating errors in the coupled system.

Replacing only the ET component with the DSR-derived formulation already improved ET performance in more than 70% of the basins. Gains were concentrated in the central and western United States, whereas a small number of basins, mainly along portions of the western and southwestern coasts, showed modest declines (Fig. 7). These spatial patterns indicate that the AI-derived ET expression substantially reduces process-level errors in most regions and provides a more stable upstream component for subsequent runoff simulations.

To identify where the ET replacement had the strongest impact, we compared annual ET totals from the DSR-based formulation with those from the original HBV scheme across basins (Fig. 8). The largest improvements occurred in basins with dense vegetation, reflected by high NDVI, while sparsely vegetated catchments showed little change. In basins where overall ET performance did not improve, the model continued to overestimate ET, suggesting that the original scheme had been overly tuned to vegetation signals in settings where other controls, such as soil moisture (Berg and Sheffield, 2018) or atmospheric demand (Castaneda-Gonzalez et al., 2023; Jiang et al., 2021; Li et al., 2022c), dominate. This result indicates that the new ET formula adds substantial skill in many environments but that additional process factors may be needed to correct persistent biases in very arid or low-vegetation regions.

3.2. Mechanistic attribution and insights

We examined the learned runoff formula to verify that it captured realistic process controls. Using the final integrated model, we built a random forest surrogate and computed SHAP (SHapley Additive exPlanations) values to identify which inputs most strongly influence daily runoff (Fig. 9). In this analysis, positive SHAP values indicate that a variable increases the predicted runoff relative to the baseline prediction, whereas negative SHAP values indicate that it decreases the predicted runoff. The sign of a SHAP value therefore shows the direction of the contribution, and its magnitude indicates the strength of that contribution. The leading predictors were all hydrologic state variables rather than raw meteorological forcings. In particular, upper subsurface storage, percolation rate, upper zone storage capacity, and the soil wetness factor emerged as the most influential inputs, while precipitation and temperature played secondary roles. These dominant predictors encode catchment memory and nonlinear responses, indicating that the AI-derived runoff formulation is drawing on storage and threshold dynamics in a physically meaningful way.

The SHAP dependence plots clarify how each variable affects streamflow. When upper subsurface storage is low, SHAP values cluster near zero or slightly negative, indicating little contribution to runoff. At high storage levels, SHAP values become strongly positive, reflecting a marked increase in streamflow once the subsurface is sufficiently filled. Percolation shows a similar pattern: higher percolation rates correspond to larger positive SHAP values, consistent with enhanced runoff when more water moves through the soil column. Precipitation behaves as expected as well. Small events often yield negative SHAP contributions, for example when light rain is largely absorbed by dry soil, whereas large storms produce strongly positive values that sometimes exceed those of any other predictor. For the upper zone storage capacity, SHAP values increase gradually with the magnitude of this threshold, implying that catchments with a deeper upper soil layer tend to generate more runoff, all else being equal.

The soil wetness factor exerts a distinctly nonlinear influence. At low to moderate soil moisture, SHAP values are negative and act to dampen runoff, but as the soil approaches saturation they turn positive, indicating much more efficient runoff generation under wet conditions. Overall, these patterns align with hydrologic understanding and confirm that the learned runoff module is driven primarily by internal state information, including storages, thresholds, and wetness, rather than by spurious correlations with forcings. Asking a single global model to infer all of these behaviors directly from meteorological inputs would greatly increase the search space and likely reduce robustness. By contrast, the modular design focuses symbolic regression on state informed relationships and successfully captures the relevant process dependencies.

3.3. End-to-end evaluation and robustness

We then embedded the learned expressions in the process-based HBV model to examine how different integration strategies affect full-system performance and error compensation. Four configurations were considered. The Baseline configuration used the original HBV model calibrated to streamflow with an LSTM-aided procedure. The ET-only DSR-HBV configuration replaced only the ET module with the DSR-derived ET formula, thereby testing the effect of improving the

Table 3
Explicit ET parameterizations obtained from the DSR model. Coefficients are $a1 = 0.0849$, $a2 = 0.0034$, $a3 = 194.0153$, $a4 = 1$,$a5 = 0.1243$.

Models	Parameterization solution
DSR (for ET)	$E_p \times \left(a_1 - a_2\left(E_p - gvf\_max(1 + gvf\_max \times a_3)\right)\right)^{0.5} \times a_4$
KG-DSR (for streamflow)	$a_5 \times \left(soil\_depth\_statsgo + p \times \left((p\_mean - 1) \times (low\_prec\_dur \times (p\_mean - soil\_depth\_statsgo))\right)^{0.5}\right)$

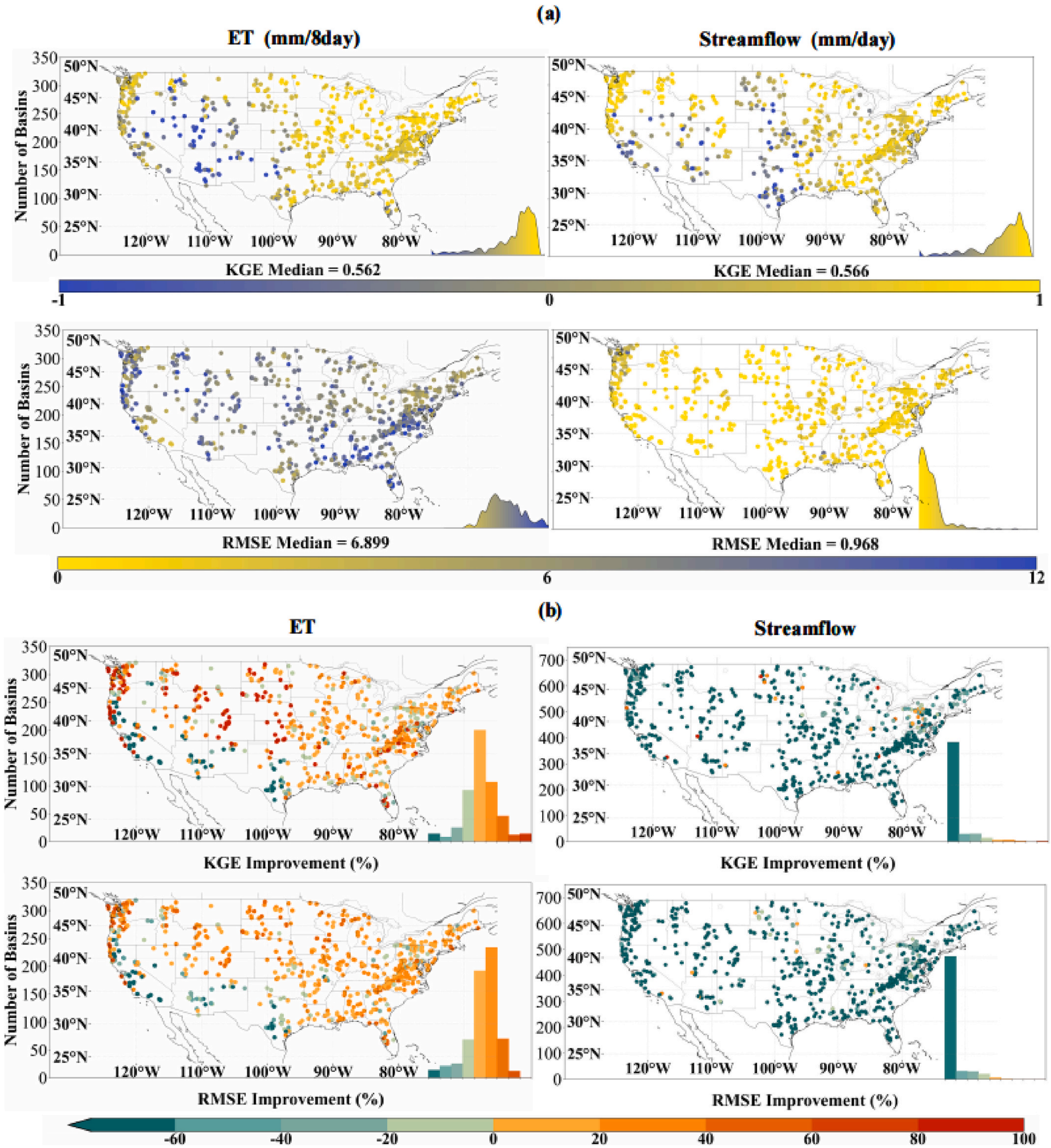

Fig. 7. Evaluation of DSR-based parameterizations. (a) Spatial maps and basin-scale histograms of KGE and RMSE for ET and streamflow, where DSR-based ET parameterizations are evaluated against MOD16A2, while runoff parameterizations are evaluated against observed discharge. (b) Spatial maps and histograms of changes in KGE and RMSE between the DSR-based and original HBV parameterizations for ET and streamflow, highlighting basins with improved or degraded performance.

upstream process alone. Strategy 1 retained the original ET scheme and replaced only the runoff module with the KG-DSR runoff expression, providing a downstream-only improvement that does not first correct ET. Strategy 2 applied the proposed stepwise integration: ET was first replaced by the DSR formulation, followed by replacement of the runoff module with the KG-DSR expression.

Replacing only the ET component improved process-level behavior but led to relatively modest gains at the full-system scale. In the configuration where only the DSR-based ET module was inserted (the ET-only DSR–HBV case), about 48% of basins showed reduced streamflow RMSE and 51% showed increased KGE relative to the baseline (Fig. 10a). Complementary performance metrics, including bias and R, are provided (Fig. S2). Large-scale baseflow patterns remained similar to those of the original model, and the distribution of residual errors

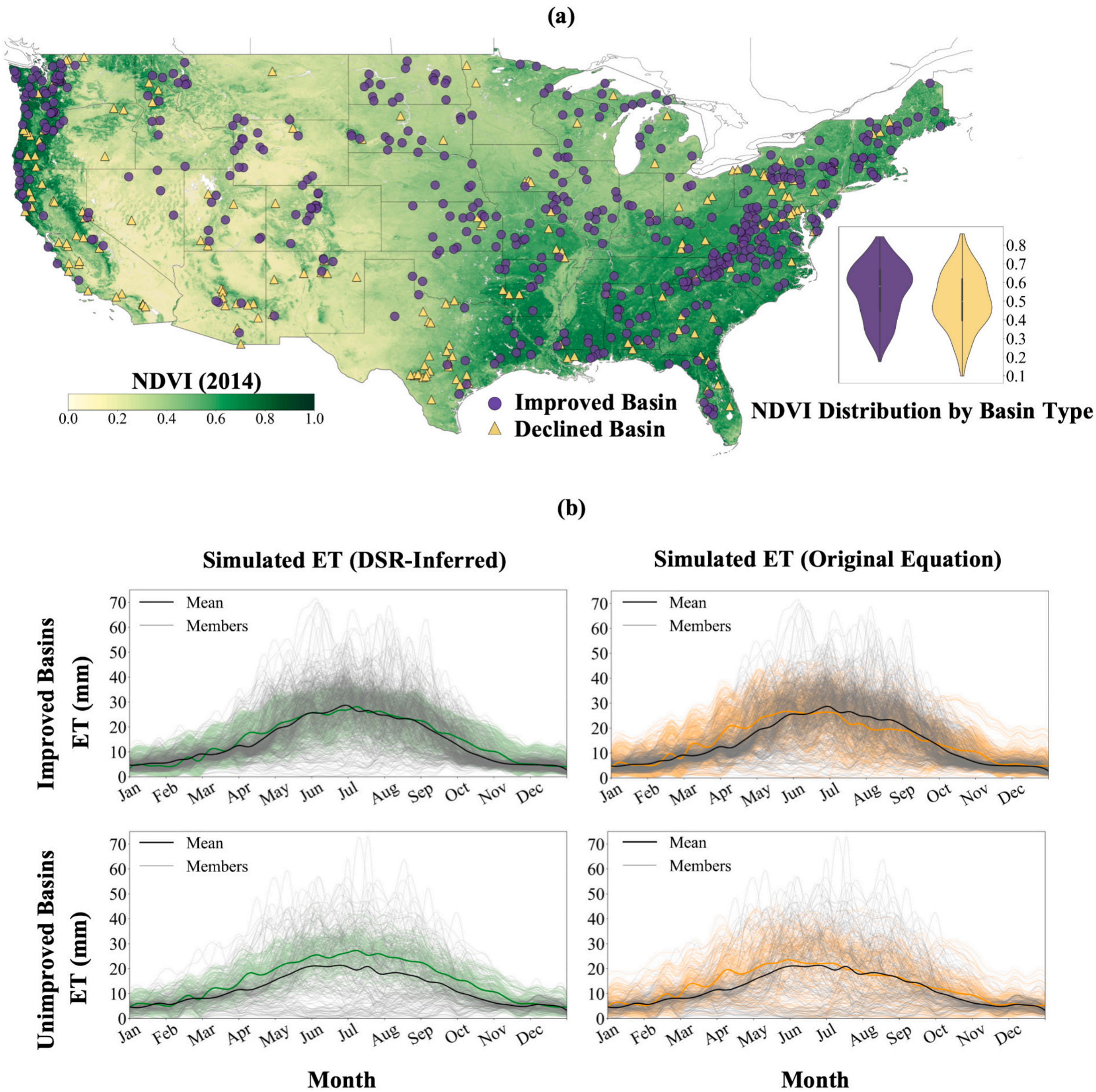

Fig. 8. (a) Spatial pattern of NDVI across the conterminous United States in 2014, used as a proxy for vegetation cover. (b) Annual ET simulated by the DSR-derived and original HBV formulations in basins with and without performance improvement. Thin colored lines show individual basin simulations, and thick colored lines show the multi-basin mean.

changed very little (Fig. 10b). Simply upgrading the ET scheme therefore did not substantially improve overall performance, which indicates that downstream components, particularly runoff generation, continued to compensate for upstream improvements. This outcome motivated a fully dependency-aware integration of both learned components.

Applying the dependency-guided integration confirmed its benefits. Both Strategy 1 and Strategy 2 outperformed the baseline in simulating ET and streamflow, and Strategy 2 delivered the largest overall gains (Fig. 11a). For ET, Strategy 2 achieved a median KGE that was 16.1% higher than Strategy 1 and about 11.6% higher than the baseline, along with higher correlations and lower RMSE. For streamflow, KGE under Strategy 2 exceeded Strategy 1 by about 4.9% and the baseline by about 10.1%. These improvements translated into many more basins reaching high-performance thresholds. With a streamflow criterion of KGE > 0.7, Strategy 2 achieved high skill in 236 basins, compared with 177 for Strategy 1 and 168 for the baseline (Fig. 11c), which is roughly a 40% increase over the original model. The number of basins with very low ET error (daily ET RMSE < 4) also more than tripled, from 25 under the baseline to 77 under Strategy 2. Across the dataset, Strategy 2 consistently produced higher ET agreement and smaller errors than either the baseline or Strategy 1, indicating a more robust overall system response. Complementary performance metrics, including bias and R, are provided (Fig. S3).

These improvements should be interpreted carefully because MOD16A2 is a model-based remote-sensing ET product. Agreement with MOD16A2 therefore does not prove that the learned equation represents unbiased real-world ET. Rather, it indicates that the DSR-derived ET expression captures the dominant ET variability represented by an independent remote-sensing reference. More importantly, the usefulness of the learned ET structure was evaluated within the full HBV modeling chain, not only through offline agreement with MOD16A2. An equation that merely reproduced MOD16A2-specific assumptions or biases could fit the ET reference offline without improving the coupled representation of soil–water storage, ET, and runoff generation. The fact that

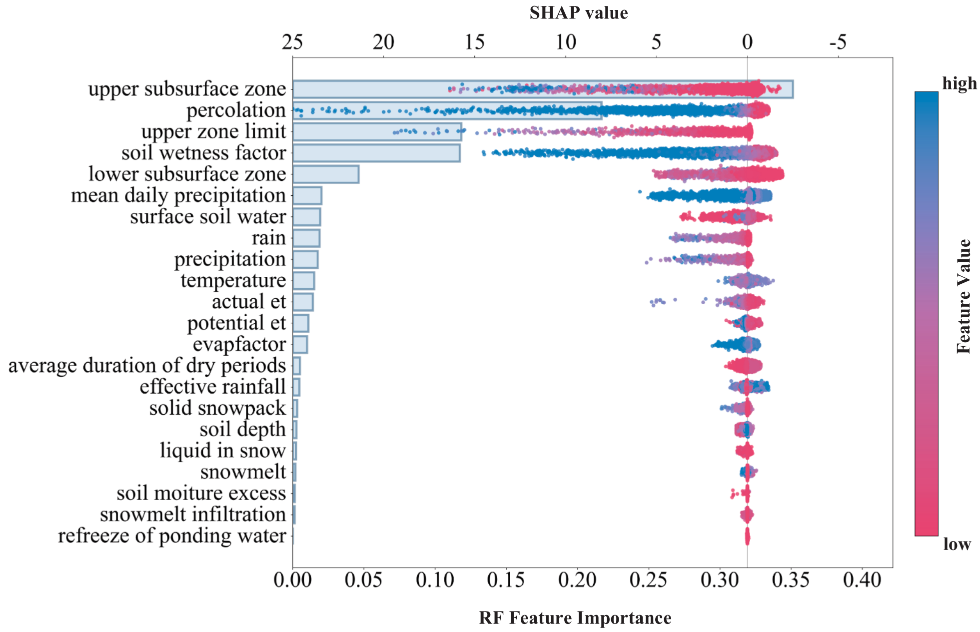

Fig. 9. Feature importance attribution obtained from the KG-DSR formula.

Strategy 2 improved both ET agreement and streamflow simulation suggests that the MOD16A2-constrained ET structure provided useful process constraints for subsequent runoff-generation parameterization learning, rather than functioning only as an offline approximation of MOD16A2.

We further conducted a regime-conditioned diagnostic analysis for extremely wet and extremely dry basins during the independent test period. The purpose was to evaluate whether Strategy 2 behaves consistently under contrasting water-availability conditions. Hydrological model evaluation should move beyond aggregate metrics and examine whether simulated behavior agrees with process understanding (Gupta et al., 2008; Yilmaz et al., 2008). Because basins differ in how they partition precipitation into ET, storage, and runoff generation, stratifying the test basins by wet and dry conditions provides a focused assessment of the regime-dependent physical consistency of the learned ET and runoff formulations (Euser et al., 2013; Sawicz et al., 2011).

For the 129 extremely wet basins, Strategy 2 improved streamflow performance in most cases(Fig.S5). Streamflow KGE increased in 100 basins, or 77.5%, and streamflow RMSE decreased in 102 basins, or 79.1%. Among the improved basins, streamflow KGE increased by 0.264 on average, with a median increase of 0.182, while streamflow RMSE decreased by 32.3% on average, with a median reduction of 27.6%. ET simulations showed similar improvements. ET KGE increased in 86 basins, or 66.7%, and ET RMSE decreased in 102 basins, or 79.1%. Among the improved basins, ET KGE increased by 0.276 on average, with a median increase of 0.218, while ET RMSE decreased by 20.7% on average, with a median reduction of 18.8%.

These results indicate that, under humid conditions where water supply substantially exceeds atmospheric evaporative demand, Strategy 2 improves both streamflow and ET simulations in most basins. The simultaneous improvement in both variables suggests that the performance gain is not simply due to runoff compensating for ET errors, but instead reflects improved consistency in the coupled ET and runoff processes. In the 5 extremely dry basins, Strategy 2 also improved streamflow performance, with both streamflow KGE and RMSE improving in 4 basins. ET improvements, however, were less stable. ET KGE increased in 2 basins, and ET RMSE decreased in only 1 basin. This result suggests that Strategy 2 can improve runoff simulation under severe water limitation, whereas its ET benefits are more uncertain in extremely dry regimes, partly because of the limited number of dry basins available for evaluation

Our regime-conditioned analysis revealed a clear contrast between extremely wet and extremely dry basins. In extremely wet basins, Strategy 2 improved both streamflow and ET in most cases. This joint improvement indicates that the learned formulations did not enhance streamflow simulation by degrading ET representation. Instead, it suggests a more coherent adjustment of the coupled ET and runoff processes. When water supply is sufficient, a better representation of ET may help regulate how precipitation is partitioned into atmospheric water loss, soil storage, and runoff generation.

This result is consistent with established understanding in catchment hydrology and ecohydrology. In humid basins, vegetation activity is generally stronger, and ET is more closely associated with canopy structure, available energy, and transpiration. Within the Budyko framework, vegetation cover, leaf area, and root water uptake capacity can influence the partitioning of precipitation between ET and runoff (Donohue et al., 2007; Li et al., 2013; Zhang et al., 2001). Global evidence also shows that transpiration represents a major component of terrestrial ET, indicating that vegetation-mediated water loss plays an important role in humid regions with active vegetation growth (Jasechko et al., 2013). Therefore, the simultaneous improvement in ET and streamflow under extremely wet conditions is physically plausible and supports the view that Strategy 2 improved water-partitioning consistency rather than relying on compensating runoff errors.

Extremely dry basins showed a different response. Strategy 2 still improved streamflow in most of these basins, but ET improvements were less stable. This contrast suggests that, under severe water limitation, improved runoff simulation does not necessarily require a comparable improvement in annual ET metrics. In these basins, ET is likely controlled less by sustained vegetation transpiration and more by intermittent water availability, short-term soil moisture pulses, root-zone moisture constraints, and atmospheric evaporative demand.

This interpretation is supported by previous studies showing that soil moisture directly constrains ET under water-limited conditions and can shift ET from energy limitation to water limitation as soils dry down (Seneviratne et al., 2010; Teuling et al., 2006). Arid and semi-arid ecosystems also often exhibit precipitation-pulse behavior, in which rainfall briefly replenishes soil moisture and triggers short-lived soil evaporation, plant water uptake, and limited transpiration before these responses decline rapidly as water is depleted (Kurc and Small, 2004;

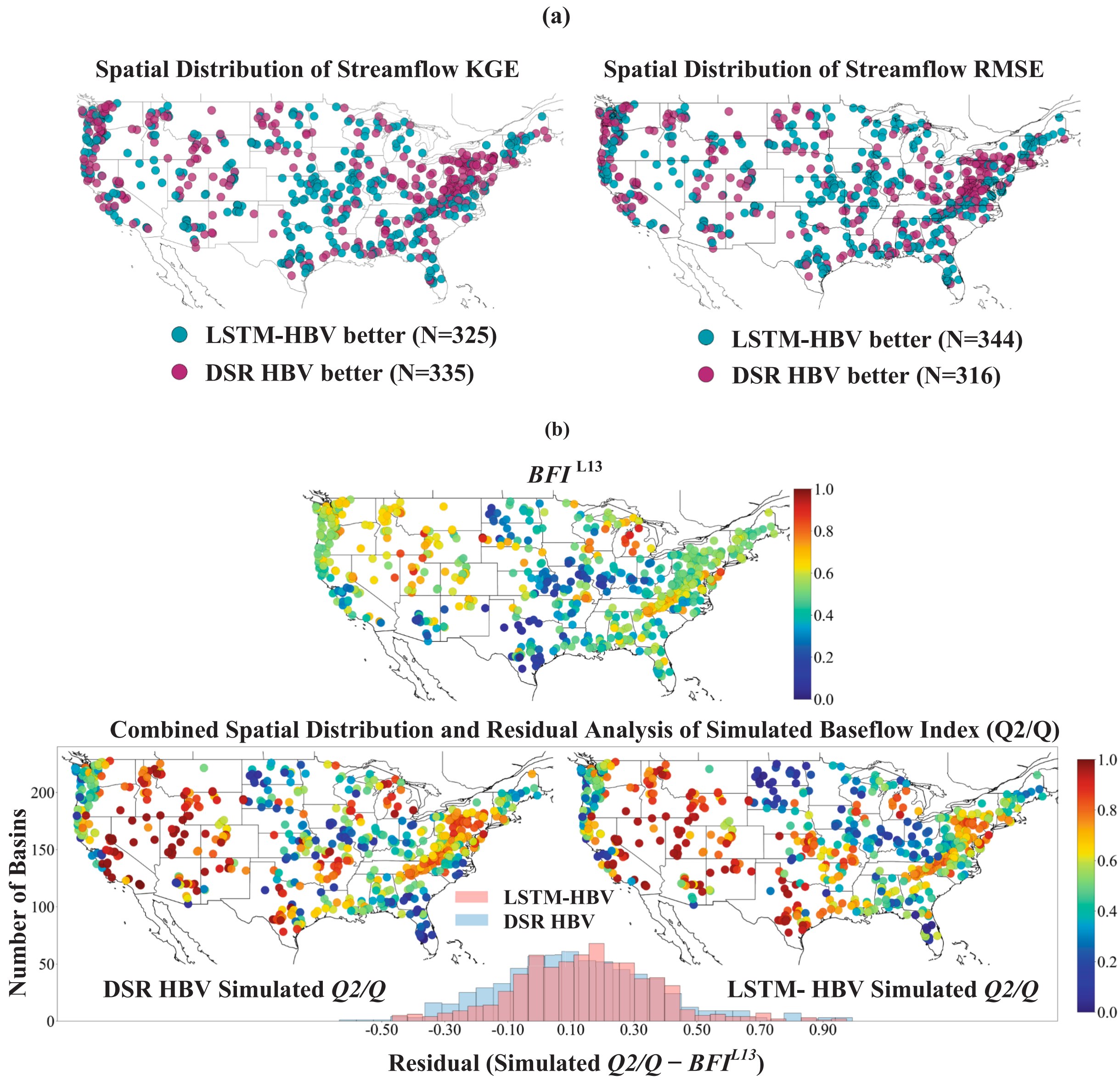

Fig. 10. Streamflow simulation performance of the original HBV and DSR-HBV models across 660 basins in the contiguous United States. (a) Spatial differences in KGE and RMSE, where teal dots mark basins in which the original HBV performs better and magenta dots mark basins in which DSR-HBV performs better. (b) Top panels show the spatial distribution of the baseflow index (BFI) from USGS baseflow separation. The bottom panel shows simulated Q_2/Q ratios from the DSR-HBV and original HBV models, together with a histogram of residuals between simulated Q_2/Q and BFIL13 for each model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Nagler et al., 2007; Reynolds et al., 2004). Thus, the weaker ET improvement in extremely dry basins does not necessarily imply that the learned equations are physically implausible. Rather, it reflects the fact that ET in these basins is an intermittent, water-limited process that may not be fully captured by annual-scale ET diagnostics. This explanation is also consistent with our NDVI-based analysis (Fig. 8). DSR-based ET improvements were concentrated mainly in high-NDVI, densely vegetated basins, whereas responses were weaker in low-NDVI, sparsely vegetated basins.

Together, these findings suggest that Strategy 2 is most effective in improving ET representation where vegetation-mediated transpiration is an important and persistent component of the water balance. In extremely dry basins, its main benefit appears to lie in improving the representation of storage states, wetness thresholds, and runoff-generation behavior, which explains why streamflow improvements were more stable than ET improvements.

KG-DSR is not intended to arbitrarily decompose total streamflow into three mathematical fitting terms. Instead, based on existing hydrological knowledge, it introduces structural constraints corresponding to fast, intermediate, and slow runoff responses, and searches for optimal explicit parameterization expressions within this physically informed framework. Therefore, the three branches should be interpreted as runoff response branches with hydrological functional meaning, rather than as independently observable physical fluxes that can be directly validated. Although the original HBV runoff branches are not regarded as observational truth for real runoff components, they provide a structural reference for examining whether KG-DSR preserves the basic dynamic distinctions among fast, intermediate, and slow runoff responses. The comparative results show that the three branches learned by KG-DSR do not degenerate into arbitrary additive terms. Instead, they maintain clear hydrological response differences(Fig.S4). The fast branch mainly exhibits event-driven peak responses and contributes primarily during high-flow events; the intermediate branch is smoother and has lower peaks than the fast branch, indicating a certain degree of

(a)

Median values of ET evaluation metrics for different models.

Model name	KGE	Bias(mm/8day)	R	RMSE(mm/8day)
S1	0.539	-0.615	0.809	6.545
S2	0.626	-0.545	0.851	5.686
DSR-HBV	0.595	0.257	0.833	6.385
LSTM-HBV	0.561	1.136	0.802	6.899

(b)

Median values of streamflow evaluation metrics for different models.

Model name	KGE	Bias(mm/day)	R	RMSE(mm/day)
S1	0.594	0.013	0.806	1.020
S2	0.623	0.003	0.833	0.910
DSR-HBV	0.550	0.063	0.812	0.980
LSTM-HBV	0.566	-0.006	0.803	0.967

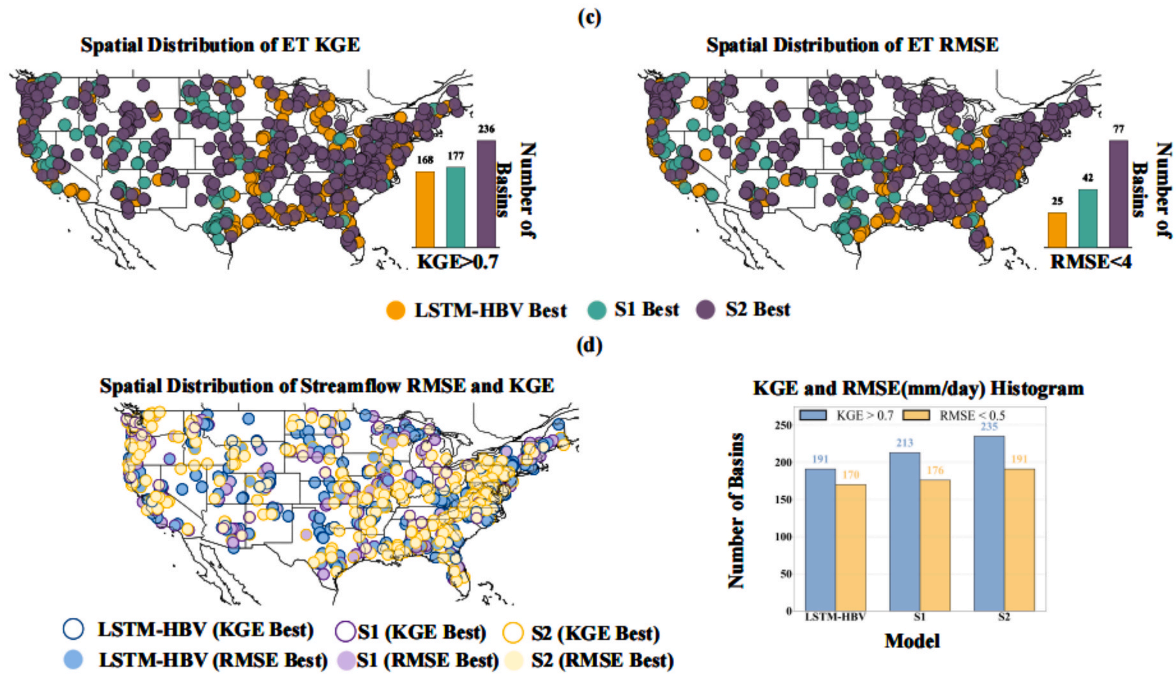

Fig. 11. Performance comparison of LSTM-HBV, S1, S2, and DSR-HBV for ET and streamflow across 660 watersheds in the conterminous United States. (a, b) Median KGE, bias, R, and RMSE for ET (a) and streamflow (b), with bias and RMSE reported in mm per 8 days. (c) Basin-wise best ET model, with LSTM-HBV in gold, S1 in green, and S2 in purple gray. The bar chart shows the number of basins with KGE greater than 0.7 and RMSE below 4 mm per day. (d) Basin-wise best streamflow model based on KGE in the outer ring and RMSE in the inner circle. The histogram shows the number of basins with KGE greater than 0.7 in blue and RMSE below 0.5 mm per day in yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

lag; and the slow branch shows stronger persistence and low-frequency variability, maintaining contributions during non-event periods and recession periods. These results are consistent with the typical functional understanding of surface runoff, interflow, and baseflow in hydrology.

Diagnostics for extreme flows and baseflow supported these conclusions (Fig. 12). For low flows ($Q5$), Strategy 2 markedly improved performance: the KGE increased from 0.61 in the baseline to 0.74, and the $Q5$ estimates clustered tightly around the 1:1 line, indicating minimal bias and reduced spread. Strategy 1 also improved low flows (KGE around 0.68) but did not match the performance of Strategy 2. For high flows ($Q95$), the baseline already performed very well (KGE around 0.87), leaving limited room for further gains. Strategy 1 reached a KGE of 0.96 and Strategy 2 reached 0.94. Although Strategy 1 was slightly better at the very highest peaks, Strategy 2 still captured flood magnitudes with high accuracy while providing a more balanced overall simulation. Spatial error maps show that, in most basins, the best configuration yields near-zero bias for both $Q5$ and $Q95$, with only isolated regions of underestimation or overestimation.

Although the 2014 test year provides a standard forward-in-time evaluation, Q5 and Q95 diagnostics may be sensitive to the hydrological conditions of that particular year. Low-flow results depend on whether dry-period behavior is sufficiently represented, while high-flow results depend on the occurrence and representativeness of high-flow events. We therefore added an independent five-year retrospective validation period from 2002 to 2006 to evaluate the temporal robustness of the Q5 and Q95 results. This period was excluded from model training and was used only as an additional multi-year evaluation window.

The five-year retrospective validation confirmed the flow-diagnostic patterns observed in the 2014 test period (Fig. 12 and Fig. S6). Strategy 2 outperformed LSTM-HBV for both low-flow and high-flow characteristics. For Q5, KGE increased from 0.37 to 0.56, indicating improved low-flow representation during 2002–2006. For Q95, KGE increased from 0.92 to 0.95, with predictions generally aligned with the 1:1 line, indicating strong performance for high-flow magnitudes. Spatial error maps showed that the best configuration produced near-zero Q5 and Q95 errors across many basins. Together, these results suggest that the gains achieved by Strategy 2 are not confined to the 2014 test year, but remain evident across a longer independent validation period.

Baseflow behavior improved in a similarly consistent way. Both integration strategies moved simulated baseflow closer to the observed baseflow index benchmark BFI^{L13}, and Strategy 2 provided the closest match while better reproducing the spatial variability of baseflow contributions. At the same time, the residual error distribution, which had been strongly right-skewed under the ET-only configuration, became much more symmetric under Strategy 2. This shift indicates that Strategy 2 removed a large portion of the systematic bias and produced a more reliable and largely unbiased simulation of the coupled hydrologic system.

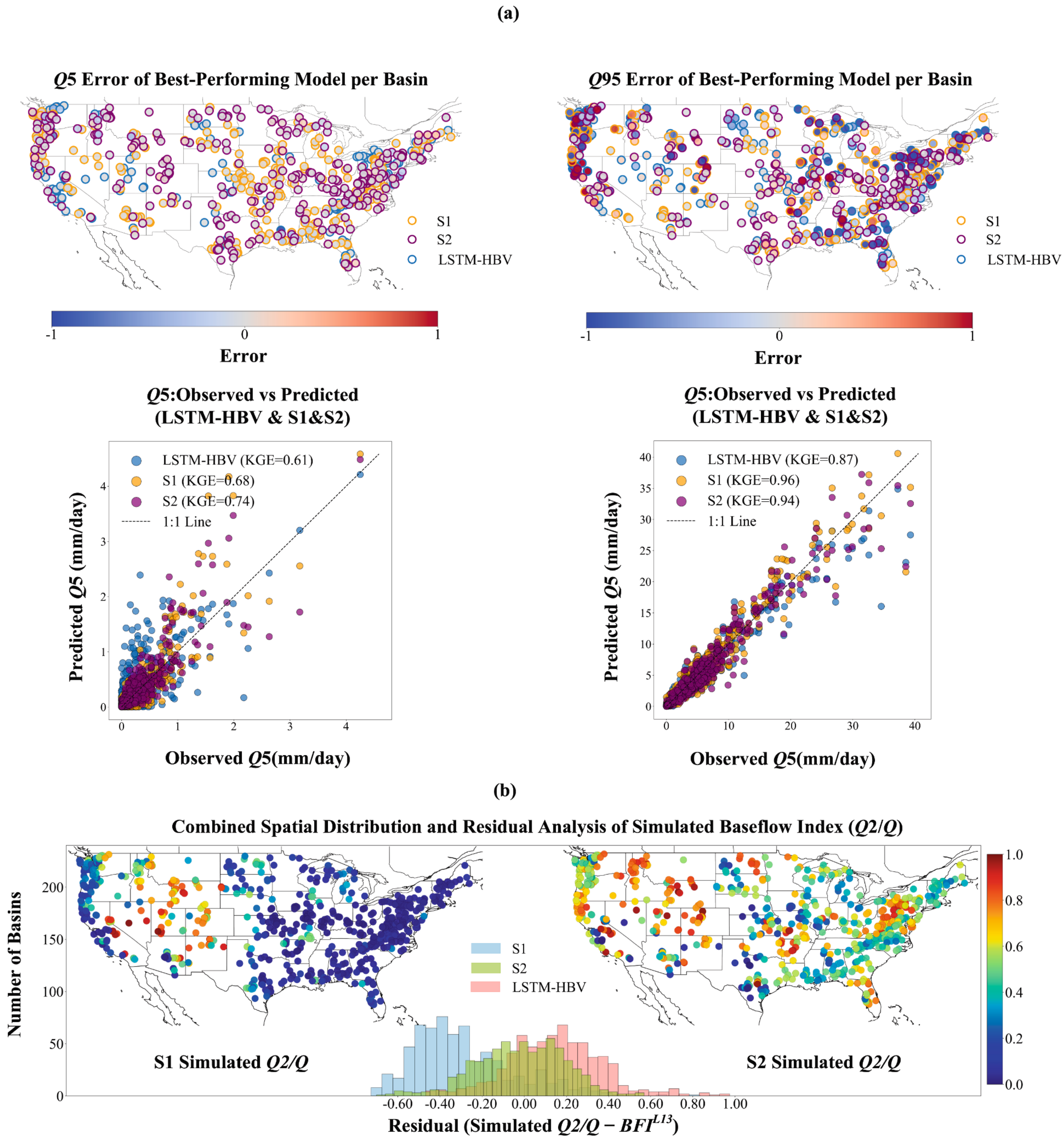

Fig. 12. Spatial and statistical assessment of $Q5$ and $Q95$ streamflow performance across 660 basins in the conterminous United States for three models (LSTM-HBV, S1, and S2). (a) Upper panels show the best-performing model in each basin for $Q5$ (left) and $Q95$ (right). Circle outlines identify the model (LSTM-HBV in blue, S1 in orange, S2 in purple), and fill color shows the simulation error (simulated minus observed). Lower panels show scatterplots of simulated versus observed $Q5$ and $Q95$ for all basins. Each point represents one basin, and the 1:1 line marks perfect agreement. KGE values for each model are given in the legend. (b) Upper panels show the spatial distribution of the Q_2/Q ratio, which is an indicator of baseflow contribution, simulated by S2 (left) and S1 (right) and compared with the BFIL13 reference. Lower panels show histograms of residuals between simulated Q_2/Q and BFIL13 for each model, highlighting differences in baseflow accuracy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

Most AI applications in hydrology focus on fitting a single process variable, such as ET, soil moisture, or runoff, to observations or reanalysis products. These standalone models often perform very well when evaluated in isolation. However, when they are embedded in a full process-based model, compensation effects and internal inconsistencies frequently emerge. Improving one component can cause other components to drift, and the overall system balance can deteriorate.Our experiments with a DSR ET model illustrate this behavior. The DSR ET formula performs strongly in offline tests, yet simply replacing the native ET module in the HBV model with this formula degrades streamflow accuracy. Strong skill in one process therefore does not guarantee system-level improvement.

The use of MOD16A2 should be interpreted with caution. MOD16A2 is a model-based ET product and may include algorithmic assumptions, retrieval uncertainties, and product-specific biases. Thus, the DSR-derived ET equation may partially inherit uncertainties from MOD16A2. We therefore do not claim that the learned equation represents unbiased true ET. At the same time, acknowledging this uncertainty does not mean that the learned equation is only a hydrologically meaningless replica of MOD16A2. The issue involves two distinct questions. First, the DSR-ET equation may inherit assumptions or biases from MOD16A2, which we explicitly acknowledge. Second, even with this uncertainty, the equation may still provide useful process constraints within HBV. The second question is the focus of this study. For this reason, we evaluated the DSR-ET equation within the full modeling chain, rather than only through offline agreement with MOD16A2. The learned ET expression was embedded into HBV, followed by runoff-generation parameterization learning under the ET-constrained structure. The resulting model was then evaluated using streamflow performance, runoff behavior, and consistency between ET representation and runoff-generation processes. If the equation only reproduced MOD16A2-specific assumptions or biases, it might fit the ET reference offline but would not necessarily improve runoff learning or streamflow simulation within HBV. The observed improvements suggest that the learned ET structure provides process-relevant information for HBV, even though it should not be viewed as an unbiased representation of true ET.

To address this challenge, we developed a dependency-aware symbolic integration framework that optimizes components in sequence while respecting the hydrological dependency structure. The framework is implemented as a two-stage optimization on an HBV backbone. In the first stage, we focus on ET, which depends directly on meteorological forcings and has limited feedback from upstream states. Calibrating this upstream process first reduces structural uncertainty early in the model chain. In the second stage, we optimize more tightly coupled components, specifically runoff and streamflow, which depend on both forcings and upstream states such as soil moisture and ET. Fig. 10 shows that this sequential strategy yields clear system-level gains. Configurations that combine well-structured formulations for both ET and streamflow consistently outperform any configuration that retains a poorly parameterized module. The improvement from Strategy 2 was not limited to streamflow simulation. Under the same number of runoff-generation-related optimizable parameters, Strategy 2 improved both discharge and ET. If the runoff improvement mainly resulted from compensation for ET-related errors, improved discharge would likely come at the cost of degraded or unstable ET performance. The concurrent improvement in both variables therefore suggests that Strategy 2 enhances ET and runoff consistency, rather than simply shifting ET biases into the runoff component. These results highlight the need to maintain structural consistency across submodules in order to improve the coupled system.

The main contribution of this work is a unified training framework that moves beyond searching for a single best parameter set or an isolated model for one process. Instead, it uses AI-generated, explicit process expressions to update multiple interdependent parameterizations in a coordinated manner. Embedding symbolic regression–based formulations within a dependency-aware training strategy provides a stable foundation for improving HBV and other complex hydrological models. Although we demonstrate the approach on the relatively simple HBV structure, the same ideas extend naturally to more advanced land surface schemes. For example, the Community Land Model (Dai et al., 2004) is a natural candidate, as are emerging symbolic regression architectures. Methods such as the Symmetric Q-network (Tian et al., 2025), which uses reinforcement learning and human-in-the-loop design to refine equation trees, universal neural symbolic regressions (Hu et al., 2025), which couples deep learning with pre-trained equation inference, and LASR (Grayeli et al., 2024), which employs large language model–guided concept discovery to accelerate evolutionary symbolic regression, all provide mechanisms for discovering governing equations. Integrating these tools into our dependency-aware framework would offer a practical pathway toward transparent, physically grounded AI in Earth system modeling.

For more complex models like CoLM, our results suggest following the internal process chain and optimizing from upstream to downstream. In CoLM, radiation and turbulence driven by meteorological forcings determine net radiation, aerodynamic resistance, and surface conductance. These energy and turbulence fields then control canopy interception, soil evaporation, and transpiration, which update snow and soil moisture storage. The updated water and energy states feed infiltration, percolation, and runoff schemes that partition water into fast and slow flow paths, ultimately shaping baseflow and streamflow. Errors or compensating biases in any upstream component can propagate through this chain and degrade both runoff predictions and land–atmosphere exchange estimates.

Guided by our HBV experiments, we suggest a dependency-guided, upstream-first strategy for optimizing CoLM in practice: Firstly, optimize net radiation closure and stabilize aerodynamic resistance and surface conductance under observed meteorological conditions. This step tightens the energy budget at the top of the process chain; Secondly, refine canopy interception and surface evaporation by tuning parameters such as canopy storage, soil evaporation rates, and transpiration conductance, subject to explicit surface water balance constraint; Finally, Adjust subsurface processes—including infiltration, percolation, interflow, and baseflow—under soil water conservation requirements, using recession characteristics and storage–discharge relationships to constrain both fast and slow flow components.

At each stage, AI-derived symbolic formulas can be introduced one module at a time, with explicit checks that any new expression respects energy and water budgets and remains consistent with upstream states. This dependency-guided optimization preserves the physical structure of the model throughout training. By following the internal causal chain of processes, the strategy demonstrated with HBV can be carried over to CoLM and related models. In doing so, our framework provides both a conceptual template and a practical route for integrating machine learning process formulations without violating mass and energy balance, thereby improving overall model performance in a transparent, physically grounded manne.

5. Conclusions

This study shows that strong standalone accuracy of an AI-derived process module does not guarantee better end-to-end performance after it is embedded in a coupled hydrologic model. When modules interact, error compensation and internal inconsistencies can mask weaknesses during offline evaluation and can even degrade system behavior once a single component is replaced.

To address this, we introduced a dependency-aware symbolic integration framework that improves modules sequentially along the hydrologic dependency chain. Using HBV as a testbed, we found that training from upstream to downstream produces reliable system-level gains. In practice, stabilizing ET first and then optimizing runoff and streamflow improved overall consistency and avoided the performance drops that occur when swapping in one module on its own. The framework also offers a transparent pathway for integrating AI-generated, explicit process equations into process-based models through staged, module-by-module deployment and conservation-based consistency checks.

This study has two main limitations. First, MOD16A2 is a model-based remote-sensing ET product, so the DSR-derived ET equation may partially inherit its assumptions, uncertainties, or biases. Therefore, the learned symbolic equation should not be interpreted as an unbiased representation of true ET. It should instead be viewed as an explicit ET parameterization constrained by a scale-compatible and discharge-independent remote-sensing reference. Second, the DSR-ET equation was not directly trained or validated using site-level flux-tower observations. Although flux towers provide valuable information on land–atmosphere water vapor exchange, their footprints are local, dynamic, and often inconsistent with basin-averaged ET simulations in CAMELS catchments. Future work should test the learned ET parameterization against multiple complementary references, including scale-matched FLUXNET observations, water-balance-based basin ET estimates, alternative satellite ET products, reanalysis datasets, and land data assimilation products.

Because many land surface and Earth system models are organized around directional process dependencies, the dependency guided strategy offers a general pathway for embedding equations discovered by AI into more complex modeling frameworks such as CoLM. Compared with HBV, CoLM represents a wider range of coupled land processes, including surface energy balance, vegetation physiology, stomatal conductance, soil heat and water transport, snow dynamics, and land atmosphere feedbacks. Extending the proposed framework to CoLM would therefore require an implementation tailored to the model structure, with careful treatment of high dimensional state variables, stronger parameter interactions, and feedbacks among water, energy, vegetation, and soil processes.

For such applications, the principle of staged integration remains essential. Equations discovered by AI should be introduced along physically meaningful dependency pathways and evaluated not only by their accuracy for individual processes, but also by their consistency with coupled water and energy budgets. In CoLM, for example, the framework could first target relatively localized processes such as stomatal conductance or canopy transpiration, and then assess how these changes propagate to latent heat, sensible heat, soil moisture, and runoff. This staged strategy would allow AI discovered equations to be incorporated into complex Earth system models in a transparent and physically constrained manner, while preserving consistency among interacting land surface processes.

CRediT authorship contribution statement

Qingliang Li: Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Mujie Wu:** Software, Project administration, Data curation. **Cheng Zhang:** Formal analysis. **Zhongwang Wei:** Methodology. **Wei Shangguan:** Resources. **Jinlong Zhu:** Investigation. **Jing Wang:** Validation. **Xiaoning Li:** Investigation. **Yuguang Yan:** Investigation. **Zijian Zhang:** Validation. **Wenzong Dong:** Visualization. **Xiaoming Shi:** Writing – review & editing. **Yunfeng Lv:** Writing – review & editing. **Yongjiu Dai:** Writing – review & editing.

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.

Acknowledgments

The study was partially supported by Guangdong Major Project of Basic and Applied Basic Research under grant 2021B0301030007, the National Natural Science Foundation of China, grant number 42575159, 42275155, 42375144, 62206028, by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China, grant number JYB2025XDXM902.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2026.136201.

Data availability

Data and code availability are described in the manuscript (Data and Code Availability section).

References

Abeshu, G.W., Tian, F., Wild, T., Zhao, M., Turner, S., Chowdhury, A.F.M.K., Vernon, C. R., Hu, H., Zhuang, Y., Hejazi, M., Li, H.-Y., 2023. Enhancing the representation of water management in global hydrological models. Geosci. Model Dev. 16 (18), 5449–5472. https://doi.org/10.5194/gmd-16-5449-2023.

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. Catchment attributes for large-sample studies. CO: UCAR/NCAR, Boulder. https://doi.org/10.5065/D6G73C3Q.

Aghakouchak, A., Habib, E., 2010. Application of a conceptual hydrologic model in teaching hydrologic processes. Int. J. Eng. Educ. 26 (4), 963–973.

Alnuqaydan, A., Gleyzer, S., Prosper, H., 2023. SYMBA: symbolic computation of squared amplitudes in high energy physics with machine learning. Mach. Learn.: Sci. Technol. 4 (1), 015007. https://doi.org/10.1088/2632-2153/acb2b2.

Bartlett, D.J., Desmond, H., Ferreira, P.G., 2023. Exhaustive symbolic regression. IEEE Trans. Evol. Comput. Early Access. https://doi.org/10.1109/TEVC.2023.3280250.

Beça, P., Rodrigues, A.C., Nunes, J.P., Diogo, P., Mujtaba, B., 2023. Optimizing reservoir water management in a changing climate. Water Resour. Manag. 37, 3423–3437. https://doi.org/10.1007/s11269-023-03508-x.

Beck, H.E., Pan, M., Lin, P., Seibert, J., Dijk, A.I.J.M.v., Wood, E.F., 2020. Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchment. J. Geophys. Res.: Atmos. 125(17):e2019JD031485 Doi: 10.1029/2019JD031485.

Berg, A., Sheffield, J., 2018. Soil moisture–evapotranspiration coupling in CMIP5 models: relationship with simulated climate and projections. J. Clim. 31 (12), 4865–4878. https://doi.org/10.1175/JCLI-D-17-0757.1.

Beven, K., 2001. How far can we go in distributed hydrological modelling? Hydrol. Earth Syst. Sci. 5 (1), 1–12. https://doi.org/10.5194/hess-5-1-2001.

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249 (1–4), 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8.

Cadol, D., Kampf, S., Wohl, E., 2012. Effects of evapotranspiration on baseflow in a tropical headwater catchment. J. Hydrol. 462, 4–14. https://doi.org/10.1016/j.jhydrol.2012.04.060.

Cai, K., He, J., Li, Q., Shangguan, W., Li, L., Hu, H., 2024. Meta-LSTM in hydrology: advancing runoff predictions through model-agnostic meta-learning. J. Hydrol. 639, 131521. https://doi.org/10.1016/j.jhydrol.2024.131521.

Castaneda-Gonzalez, M., Poulin, A., R.R.L., Turcotte, R., 2023. Weighting climate models for hydrological projections: effects on contrasting hydroclimatic regions. J. Hydrol. 625, 130098. https://doi.org/10.1016/j.jhydrol.2023.130098.

Chu, H., Luo, X., Ouyang, Z., Chan, W.S., Dengel, S., Biraud, S.C., Torn, M.S., Metzger, S., Kumar, J., Arain, M.A., 2021. Representativeness of eddy-covariance flux footprints for areas surrounding AmeriFlux sites. Agric. For. Meteorol. 301, 108350. https://doi.org/10.1016/j.agrformet.2021.108350.

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 (7), 3427–3440. https://doi.org/10.5194/hess-21-3427-2017.

Coron, L., Andréassian, V., Perrin, C., Bourqu, M., Hendrickx, F., 2014. On the lack of robustness of hydrologic models regarding water balance simulation: a diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments. Hydrol. Earth Syst. Sci. 18 (2), 727–746. https://doi.org/10.5194/hess-18-727-2014.

Dai, Y., Dickinson, R.E., Wang, Y.-P., 2004. A two-big-leaf model for canopy temperature, photosynthesis, and stomatal conductance. J. Clim. 17 (12), 2281–2299. https://doi.org/10.1175/1520-0442(2004)017<2281:ATMFCT>2.0.CO;2.

Donohue, R., Roderick, M., McVicar, T.R., 2007. On the importance of including vegetation dynamics in Budyko's hydrological model. Hydrol. Earth Syst. Sci. 11 (2), 983–995. https://doi.org/10.5194/hess-11-983-2007.

Draper, D., 1995. Assessment and propagation of model uncertainty. J. R. Stat. Soc. Ser. B 57 (1), 45–70. https://doi.org/10.1111/j.2517-6161.1995.tb02015.x.

Dunne, T., Zhang, W., Aubry, B.F., 1991. Effects of rainfall, vegetation, and microtopography on infiltration and runoff. Water Resour. Res. 27 (9), 2271–2285. https://doi.org/10.1029/91WR01585.

Euser, T., Winsemius, H.C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., Savenije, H., 2013. A framework to assess the realism of model structures using hydrological signatures. Hydrol. Earth Syst. Sci. 17 (5), 1893–1912. https://doi.org/10.5194/hess-17-1893-2013.

Feng, D., Beck, H., Lawson, K., Shen, C., 2023a. The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment. Hydrol. Earth Syst. Sci. 27 (12), 2357–2373. https://doi.org/10.5194/hess-27-2357-2023.

DiPietro, D.M., Zhu, B., 2022. Symplectically Integrated Symbolic Regression of Hamiltonian Dynamical Systems. arXiv:2209.01521.

Feng, D., Liu, J., Lawson, K., Shen, C., 2022a. Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. Water Resour. Res., 58:e2022WR032404 Doi: 10.1029/2022WR032404.

Feng, D., Liu, J., Lawson, K., Shen, C., 2022b. Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. Water Resour. Res., 58(10):e2022WR032404 doi: 10.1029/2022WR032404.

Feng, D., Tan, Z., He, Q., 2023b. Physics-informed neural networks of the Saint-Venant equations for downscaling a large-scale river model. Water Resour. Res. 59 (2), e2022WR033168. https://doi.org/10.1029/2022WR033168.
Frame, J.M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L.M., Gupta, H.V., Nearing, G.S., 2022a. Deep learning rainfall–runoff predictions of extreme events. Hydrol. Earth Syst. Sci. 26, 3377–3392. https://doi.org/10.5194/hess-26-3377-2022.
Frame, J.M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L.M., Gupta, H.V., Nearing, G.S., 2022b. Deep learning rainfall–runoff predictions of extreme events. Hydrol. Earth Syst. Sci. 26 (13), 3377–3392. https://doi.org/10.5194/hess-26-3377-2022.
Grayeli, A., Sehgal, A., Costilla Reyes, O., Cranmer, M., Chaudhuri, S., 2024. Symbolic regression with a learned concept library. Adv. Neural Inf. Proces. Syst. 37, 44678–44709. https://doi.org/10.52202/079017-1419.
Gupta, H.V., Wagener, T., Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrol. Process.: Int. J. 22 (18), 3802–3813. https://doi.org/10.1002/hyp.6989.
Häfnera, D., Gemmrichc, J., Jochum, M., 2023. Machine-guided discovery of a real-world rogue wave model. PNAS 120 (48), e2306275120. https://doi.org/10.1073/pnas.2306275120.
Han, Z., Huang, S., Peng, J., Li, J., Leng, G., Huang, Q., Zhao, J., Yang, F., He, P., Meng, X., Li, Z., 2023. GRACE-based dynamic assessment of hydrological drought trigger thresholds induced by meteorological drought and possible driving mechanisms. Remote Sens. Environ. 298, 113831. https://doi.org/10.1016/j.rse.2023.113831.
Hu, J., Cui, J., Yang, B., 2025. Learning interpretable network dynamics via universal neural symbolic regression. Nat. Commun. 16 (1), 6226. https://doi.org/10.1038/s41467-025-61575-7.
Jasechko, S., Sharp, Z.D., Gibson, J.J., Birks, S.J., Yi, Y., Fawcett, P.J., 2013. Terrestrial water fluxes dominated by transpiration. Nature 496 (7445), 347–350. https://doi.org/10.1038/nature11983.
Jiang, F., Xie, X., Liang, S., Wang, Y., Zhu, B., Zhang, X., Chen, Y., 2021. Loess Plateau evapotranspiration intensified by land surface radiative forcing associated with ecological restoration. Agric. For. Meteorol. 311, 108669. https://doi.org/10.1016/j.agrformet.2021.108669.
Kamienny, P.-A., d'Ascoli, S., Lample, G., Charton, F., 2022. End-to-end symbolic regression with transformers. Adv. Neural Inf. Proces. Syst. 35, 10269–10281. https://doi.org/10.48550/arXiv.2205.15764.
Keren, L.S., Liberzon, A., Lazebnik, T., 2023. A computational framework for physics-informed symbolic regression with straightforward integration of domain knowledge. Sci. Rep. 13, 1249. https://doi.org/10.1038/s41598-023-28328-2.
Kim, S., Lu, P., Mukherjee, S., Gilbert, M., Jing, L., Ceperi, V., M.S.Ć, 2020. Integration of neural network-based symbolic regression in deep learning for scientific discovery. IEEE Trans. Neural Networks Learn. Syst. 32 (9), 4166–4177. https://doi.org/10.1109/tnnls.2020.3017010.
Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Matthew Willson, M.P., Brenner, Hoyer, S., 2024. Neural general circulation models for weather and climate. Nature 632, 1060–1066. https://doi.org/10.1038/s41586-024-07744-y.
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 (17), 4187–4201. https://doi.org/10.5194/hess-28-4187-2024.
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 (11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018.
Kurc, S.A., Small, E.E., 2004. Dynamics of evapotranspiration in semiarid grassland and shrubland ecosystems during the summer monsoon season, central New Mexico. Water Resour. Res. 40 (9). https://doi.org/10.1029/2004WR003068.
Lemos, P., Jeffrey, N., Cranmer, M., Ho, S., Battaglia, P., 2022. Rediscovering orbital mechanics with machine learning, vol 4. Mach. Learn.: Sci. Technol.
Li, D., Pan, M., Cong, Z., Zhang, L., Wood, E., 2013. Vegetation control on water and energy balance within the Budyko framework. Water Resour. Res. 49 (2), 969–976. https://doi.org/10.1002/wrcr.20107.
Li, Q., Jin, X., Zhang, C., Shangguan, W., Wei, Z., Li, L., Liu, P., Dai, Y., 2025. Improving global soil moisture prediction based on meta-learning model leveraging Koppen-Geiger climate classification. Catena 250, 108743. https://doi.org/10.1016/j.catena.2025.108743.
Li, Q., Li, Z., Shangguan, W., Wang, X., Li, L., Yu, F., 2022a. Improving soil moisture prediction using a novel encoder-decoder model with residual learning. Comput. Electron. Agric. 195, 106816. https://doi.org/10.1016/j.compag.2022.106816.
Li, Q., Zhang, C., Wei, Z., Jin, X., Shangguan, W., Yuan, H., Zhu, J., Li, L., Liu, P., Chen, X., Yan, Y., Dai, Y., 2024b. Advancing symbolic regression for earth science with a focus on evapotranspiration modeling. npj Clim. Atmos. Sci., 7(321) Doi:10.1038/s41612-024-00861-5.
Li, Q., Xiao, Q., Zhang, C., Zhu, J., Chen, X., Yan, Y., Liu, P., Shangguan, W., Wei, Z., Li, L., Dong, W., Dai, Y., 2024. Improving global soil moisture prediction through cluster-averaged sampling strategy. Geoderma 449, 116999. https://doi.org/10.1016/j.geoderma.2024.116999.
Li, Q., Zhu, Y., Shangguan, W., Wang, X., Li, L., Yu, F., 2022b. An attention-aware LSTM model for soil moisture and soil temperature prediction. Geoderma 409, 115651. https://doi.org/10.1016/j.geoderma.2021.115651.
Li, S., Wang, G., Zhu, C., Lu, J., Ullah, W., Hagan, D.F.T., Kattel, G., Peng, J., 2022c. Attribution of global evapotranspiration trends based on the Budyko framework. Hydrol. Earth Syst. Sci. 26 (13), 3691–3707. https://doi.org/10.5194/hess-26-3691-2022.
Loucks, D.P., van Beek, E., 2017. Water Resource Systems Planning and Management: An Introduction to Methods, Models, and Applications. Springer, 624. doi:10.1007/978-3-319-44234-1.
Matsubara, Y., Chiba, N., Igarashi, R., Ushiku, Y., 2022. Rethinking symbolic regression datasets and benchmarks for scientific discovery. Neurips 2022 AI for Science Workshop.
Mei, Y., Mai, J., Do, H.X., Gronewold, A., Reeves, H., Eberts, S., Niswonger, R., Regan, R. S., Hunt, R.J., 2023. Can hydrological models benefit from using global soil moisture, evapotranspiration, and runoff products as calibration targets? Water Resour. Res. 59(2):e2022WR032064 Doi:10.1029/2022WR032064.
Menard, C.B., Essery, R., Krinner, G., Arduini, G., Bartlett, P., Boone, A., Brutel-Vuilmet, C., Burke, E., Cuntz, M., Dai, Y., Decharme, B., Dutra, E., Fang, X., Fierz, C., Gusev, Y., Hagemann, S., Haverd, V., Kim, H., Lafaysse, M., Marke, T., Nasonova, O., Nitta, T., Niwano, M., Pomeroy, J., Schädler, G., Semenov, V.A., Smirnova, T., Strasser, U., Swenson, S., Turkov, D., Wever, N., Yuan, H., 2021. Scientific and human errors in a snow model intercomparison. Bull. Am. Meteorol. Soc. 102 (1), E61–E79. https://doi.org/10.1175/BAMS-D-19-0329.1.
Mu, Q., Zhao, M., Running, S.W., 2011. Improvements to a MODIS global terrestrial evapotranspiration algorithm. Remote Sens. Environ. 115 (8), 1781–1800. https://doi.org/10.1016/j.rse.2011.02.019.
Nagler, P., Glenn, E., Kim, H., Emmerich, W., Scott, R., Huxman, T., Huete, A., 2007. Relationship between evapotranspiration and precipitation pulses in a semiarid rangeland estimated by moisture flux towers and MODIS vegetation indices. J. Arid Environ. 70 (3), 443–462. https://doi.org/10.1016/j.jaridenv.2006.12.026.
Newman, A.J., Sampson, K., Clark, M.P., Bock, A., Viger, R.J., Blodgett, D., 2014. A large-sample watershed-scale hydrometeorological dataset for the contiguous USA. In: UCAR/NCAR (ed).
Petersen, B.K., Larma, M.L., Mundhenk, T.N., Santiago, C.P., Kim, S.K., Kim, J.T., 2021. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In: The International Conference on Learning Representations, Vienna, Austria, 4, May 2021.
Pomeroy, J.W., Browna, T., Fang, X., Shook, K.R., Pradhananga, D., Armstrong, R., Harder, P., Marsh, C., Costa, D., Krogh, S.A., Aubry-Wake, C., Annand, H., Lawford, P., Hea, M.K.Z., Moreno, J.I.L., 2022. The cold regions hydrological modelling platform for hydrological diagnosis and prediction based on process understanding. J. Hydrol. 615, 128711. https://doi.org/10.1016/j.jhydrol.2022.128711.
Rahmani, F., Appling, A., Feng, D., Lawson, K., Shen, C., 2023. Identifying structural priors in a hybrid differentiable model for stream water temperature modeling. Water Resour. Res. 59, e2023WR034420. https://doi.org/10.1029/2023WR034420.
Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., Shen, C., 2021a. Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins. Hydrol. Process. 35 (11), e14400. https://doi.org/10.1002/hyp.14400.
Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., Shen, C., 2021b. Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. Environ. Res. Lett. 16 (2), 024025. https://doi.org/10.1088/1748-9326/abd501.
Read, J.S., Oliver, S.K., Watkins, W., Jia, X., Willard, J., Appling, A.P., Karpatne, A., Hansen, G.J.A., Steinbach, M., Kumar, V., 2019. Process-guided deep learning predictions of lake water temperature. Water Resour. Res. 55, 9173–9190. https://doi.org/10.1029/2019WR024922.
Reynolds, J.F., Kemp, P.R., Ogle, K., Fernández, R.J., 2004. Modifying the 'pulse–reserve'paradigm for deserts of North America: precipitation pulses, soil water, and plant responses. Oecologia 141 (2), 194–210. https://doi.org/10.1007/s00442-004-1524-4.
Rientjes, T.H.M., Muthuwatta, L.P., Bos, M.G., Booij, M.J., Bhatt, H.A., 2013. Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration. J. Hydrol. 505, 276–290. https://doi.org/10.1016/j.jhydrol.2013.10.006.
Rogelis, M.C., Werner, M., 2018. Streamflow forecasts from WRF precipitation for flood early warning in mountain tropical areas. Hydrol. Earth Syst. Sci. 22 (1), 853–870. https://doi.org/10.5194/hess-22-853-2018.
Running, S., Mu, Q., Zhao, M., 2017. MOD16A2 MODIS/terra net evapotranspiration 8-day L4 global 500m SIN grid V006. In: Center, N. E. L. P. D. A. A. (ed).
Sawicz, K., Wagener, T., Sivapalan, M., Troch, P.A., Carrillo, G., 2011. Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. Hydrol. Earth Syst. Sci. 15 (9), 2895–2911. https://doi.org/10.5194/hess-15-2895-2011.
Schaefli, B., Rinaldo, A., Botter, G., 2013. Analytic probability distributions for snow-dominated streamflow. Water Resour. Res. 49 (5), 2701–2713. https://doi.org/10.1002/wrcr.20234.
Schmid, H.P., 2002. Footprint modeling for vegetation atmosphere exchange studies: a review and perspective. Agric. For. Meteorol. 113 (1–4), 159–183. https://doi.org/10.1016/S0168-1923(02)00107-7.
Seibert, J., Vis, M.J.P., 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. Hydrol. Earth Syst. Sci. 16 (9), 3315–3325. https://doi.org/10.5194/hess-16-3315-2012.
Sellers, P.J., Dickinson, R., Randall, D., Betts, A.K., Hall, F.G., Berry, J.A., Collatz, G., Denning, A., Mooney, H.A., Nobre, C.A., 1997. Modeling the exchanges of energy, water, and carbon between continents and the atmosphere. Science 275 (5299), 502–509. https://doi.org/10.1126/science.275.5299.502.
Seneviratne, S.I., Corti, T., Davin, E.L., Hirschi, M., Jaeger, E.B., Lehner, I., Orlowsky, B., Teuling, A.J., 2010. Investigating soil moisture–climate interactions in a changing climate: a review. Earth Sci. Rev. 99 (3–4), 125–161. https://doi.org/10.1016/j.earscirev.2010.02.004.

Shen, C., Appling, A.P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, 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. Nature Reviews Earth & Environment 4 (8), 552–567. https://doi.org/10.1038/s43017-023-00450-9.

Teja, K.N., Manikanta, V., Das, J., Umamahesh, N.V., 2023. Enhancing the predictability of flood forecasts by combining numerical weather prediction ensembles with multiple hydrological models. J. Hydrol., 625, Part B:130176 Doi:10.1016/j.jhydrol.2023.130176.

Tenachi, W., Ibata, R., Diakogiannis, F.I., 2023. Deep symbolic regression for physics guided by units constraints: toward the automated discovery of physical laws. Astrophys J 959 (2), 99. https://doi.org/10.3847/1538-4357/ad014c.

Teuling, A.J., Seneviratne, S.I., Williams, C., Troch, P.A., 2006. Observed timescales of evapotranspiration response to soil moisture. Geophys. Res. Lett. 33 (23). https://doi.org/10.1029/2006GL028178.

Tian, Y., Zhou, W., Viscione, M., Dong, H., Kammer, D.S., Fink, O., 2025. Interactive symbolic regression with co-design mechanism through offline reinforcement learning. Nat. Commun. 16 (1), 3930. https://doi.org/10.1038/s41467-025-59288-y.

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., Shen, C., 2021. From calibration to parameter learning: harnessing the scaling effects of big data in geoscientific modeling. Nat. Commun. 12 (1), 5988. https://doi.org/10.1038/s41467-021-26107-z.

Udrescu, S.-M., Tegmark, M., 2020. AI Feynman: a physics-inspired method for symbolic regression. Sci. Adv. 6, eaay2631. https://doi.org/10.1126/sciadv.aay2631.

Usama, M., Lee, I.-Y., 2022. Data-driven non-linear current controller based on deep symbolic regression for SPMSM. sensors 22(21):8240 Doi:10.3390/s22218240.

Vieira, M.J.F., Stadnyk, T.A., 2023. Leveraging global climate models to assess multi-year hydrologic drought. npj Clim. Atmos. Sci. 6 (179). https://doi.org/10.1038/s41612-023-00496-y.

Vora, A., Cai, X., Chen, Y., Li, D., 2024. Coupling reservoir operation and rainfall-runoff processes for streamflow simulation in watersheds. Water Resour. Res., 60(6): e2023WR035703 Doi:10.1029/2023WR035703.

Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., Shen, C., 2021. Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. J. Hydrol. 603, 127043. https://doi.org/10.1016/j.jhydrol.2021.127043.

Xu, W., J. Chen, G. Corzo, C.-Y. Xu, X. J. Zhang, L. Xiong, D. Liu & J. Xia, 2024. Coupling deep learning and physically based hydrological models for monthly streamflow predictions. Water Resour. Res., 60(2):e2023WR035618 Doi:10.1029/2023WR035618.

Yang, Y., Roderick, M.L., Guo, H., Miralles, D.G., Zhang, L., Fatichi, S., Luo, X., Zhang, Y., McVicar, T.R., Tu, Z., 2023. Evapotranspiration on a greening Earth. Nature Reviews Earth & Environment 4 (9), 626–641. https://doi.org/10.1038/s43017-023-00464-3.

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 (9). https://doi.org/10.1029/2007WR006716.

Zhang, J., Zhou, L., Ma, R., Jia, Y., Yang, F., Zhou, H., Cao, X., 2019. Influence of soil moisture content and soil and water conservation measures on time to runoff initiation under different rainfall intensities. Catena 182, 104172, 10.1016j.catena.2019.104172.

Zhang, L., Dawes, W., Walker, G., 2001. Response of mean annual evapotranspiration to vegetation changes at catchment scale. Water Resour. Res. 37 (3), 701–708. https://doi.org/10.1029/2000WR900325.

Zhang, M., Kim, S., Lu, P.Y., M.S.ć, 2022. Deep learning and symbolic regression for discovering parametric equations. IEEE Trans. Neural Networks Learn. Syst. 99, 1–13. https://doi.org/10.1109/TNNLS.2023.3297978.

Zhu, C., Byrd, R.H., Lu, P., Nocedal, J., 1997. L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Trans. Math. Softw. 23 (4), 550–560. https://doi.org/10.1145/279232.279236.