



1 Reliability diagnostics for learned input-to-parameter relationships in differentiable 2 hybrid models: RELIP v1.0 with a large-sample hydrologic evaluation

3 Xin Jing¹, Na Wei¹, Xue Yang¹, and Jungang Luo¹

4 ¹State Key Laboratory of Water Engineering Ecology and Environment in Arid Area, Xi'an University of Technology, Xi'an,
5 Shaanxi 710048, China

6 Correspondence to: Na Wei (weina@xaut.edu.cn) and Jungang Luo (jgluo@xaut.edu.cn)

7 **Abstract.** Predictive performance remains the usual basis for assessing geoscientific models, but neural-parameterized
8 mechanistic models expose another object that may be interpreted: the learned mapping from inputs to parameters. RELIP
9 v1.0 evaluates whether that mapping is reproducible. It proceeds through four linked levels: predictive adequacy, stability
10 across random seeds and loss functions, consistency of dominant controls across parameter-learning formulations, and
11 compactness of the full relationship matrix. We apply RELIP to differentiable HBV parameter learning for 531 CAMELS-US
12 catchments, comparing deterministic, Monte Carlo dropout, and distributional formulations under three losses and five seeds.
13 The three formulations produced closely comparable streamflow skill, so predictive metrics alone offered little basis for
14 choosing among them. RELIP distinguished them at the level of relationship structure. The distributional formulation had
15 lower cross-seed variability in Spearman correlations, more reproducible dominant controls, and the most compact relationship
16 matrix. Seven of fourteen mechanistic parameters formed a shared dominant-control core. Parameter spread also contained
17 structured information, although several gradients were coupled to parameter means or boundary proximity. These results
18 identify learned parameterization structure as a separate target of model assessment and provide a practical diagnostic for
19 formulations that appear similar in prediction.

20 **Keywords.** Model assessment; Differentiable hybrid models; Learned parameterization; Reproducibility diagnostics;
21 Relationship reliability; CAMELS-US

22 1 Introduction

23 Hydrologic model assessment usually begins with observed and simulated discharge. Predictive scores and
24 regionalization tests evaluate rainfall-runoff models and parameter transfer across basins (Hrachowitz et al., 2013; Sivapalan
25 et al., 2003). Attribute-based regionalization asks whether climate, terrain, vegetation, soil, and geology contain transferable
26 information about model parameters (Parajka et al., 2005; Samaniego et al., 2010). Large-sample data sets such as CAMELS-
27 US make these tests possible across diverse catchments (Addor et al., 2017; Newman et al., 2015). These evaluations are
28 essential, but they do not test every model quantity that is later interpreted.

29 Neural parameterization changes what must be assessed. Purely data-driven hydrologic models can predict streamflow
30 accurately, yet their internal representations do not map cleanly to hydrologic states, fluxes, or parameters (Kratzert et al.,
31 2018, 2019; Shen, 2018). Differentiable hydrologic models retain a mechanistic simulator in the computational graph and use
32 neural networks for parameterization functions or selected components. This design preserves process-related outputs while
33 supporting competitive prediction (Feng et al., 2022, 2023; Reichstein et al., 2019). Differentiable parameter learning extends
34 the same idea to spatially varying geoscientific parameters (Tsai et al., 2021). Once those learned quantities are interpreted,
35 predictive skill is no longer a complete evaluation.

36 Most evaluations of differentiable parameter-learning models focus on predictive skill, spatial generalization, untrained
37 physical variables, or the coherence of inferred parameter and flux fields (Feng et al., 2022, 2023; Tsai et al., 2021). They
38 rarely ask whether the same input-to-parameter structure is recovered when training is perturbed. Two parameterizations can
39 fit discharge equally well and still encode different descriptor relationships. We examine three established choices: point



estimates, dropout-based stochastic approximation (Gal & Ghahramani, 2016; Kendall & Gal, 2017), and a variational distributional representation (Hüllermeier & Waegeman, 2021; Kingma & Welling, 2022). Their spread measures are treated as properties of the learned mapping, not as complete Bayesian posteriors (Abdar et al., 2021; Hüllermeier & Waegeman, 2021). The question is whether these formulations learn relationships that persist under the same controlled perturbations.

HBV provides a useful benchmark for this question. Its conceptual parameters reproduce discharge while summarizing storage, runoff generation, snow, and recession behavior (Lindström et al., 1997; Seibert & Bergström, 2022). Those parameters are already subject to equifinality, process compensation, bounds, data limitations, and objective-function choice (Beven & Freer, 2001). A learned descriptor-to-parameter mapping adds another source of variation. It may support accurate simulation while changing with initialization, loss function, or parameter representation. Controlled perturbations are therefore needed to test whether the recovered relationships persist (Clark et al., 2011; Clark & Kavetski, 2010; Fenicia et al., 2011). Related work found that sequence encoders with similar predictive accuracy can yield time-varying parameter trajectories with different temporal consistency (Jing, Luo, et al., 2026). Here the target is different: static descriptor-to-parameter relationships and their reproducibility across seeds, losses, and formulations.

RELIP v1.0 addresses this assessment gap. It is a model-agnostic diagnostic method for learned input-to-parameter relationships in differentiable hybrid models. Reliability is the overall objective; reproducibility is the operational test. Prediction first has to be adequate. RELIP then examines stability across seeds and losses, agreement in dominant controls across formulations, and compactness of the complete relationship matrix. The ordering reflects a basic model-evaluation principle: interpretation should rest on behavior that survives plausible modeling choices (Clark et al., 2011; Refsgaard et al., 2007).

We apply RELIP to differentiable HBV parameter learning for the standard 531-basin CAMELS-US benchmark. HBV provides a compact mechanistic structure for snow, soil moisture, evapotranspiration limitation, percolation, groundwater recession, and routing, while CAMELS-US spans broad hydroclimatic and physiographic variation (Addor et al., 2017; Seibert & Bergström, 2022). The model setting follows recent work in process-guided neural hydrology (Feng et al., 2022, 2023; Tsai et al., 2021). The assessment target, however, is the mapping from static descriptors to mechanistic parameters rather than discharge skill alone.

The HBV backbone, basin sample, forcing data, and temporal split were held fixed, while the formulations differed in parameter representation and associated stochastic or regularization mechanisms. The deterministic formulation returned one parameter vector per basin, MC-dropout generated stochastic samples by retaining dropout during inference, and the distributional formulation learned a latent parameter distribution with KL regularization. This controlled design therefore compares the reproducibility of the resulting relationship structures under otherwise consistent experimental conditions.

RELIP contributes an ordered diagnostic protocol for learned mechanistic parameterizations, a controlled comparison of three common parameter representations, and an evidence hierarchy that limits interpretation when results depend on formulation, parameter bounds, or mean-spread coupling. The scope is deliberately modest. HBV parameters are behavioral summaries of a chosen model structure, not direct measurements of physical properties, and stable descriptor relationships may reflect broad environmental gradients rather than isolated causes. Within those limits, reproducibility provides defensible evidence about model behavior.

2 Methods and assessment framework

RELIP is the methodological focus of this study. The differentiable HBV experiment serves as a controlled benchmark: the model backbone, basin sample, forcing data, temporal split, and evaluation records are fixed, whereas parameter-learning formulation, training objective, and random initialization are varied. The quantity under assessment is the learned input-to-parameter relationship structure.



2.1 RELIP v1.0 assessment framework

RELIP asks whether a learned parameterization is reproducible enough to interpret. It requires prediction records, learned parameter fields, static descriptors, and run identifiers from repeated training. In this benchmark, the controlled perturbations are random initialization, hydrologic loss function, and parameter-learning formulation.

The assessment has four ordered levels (Table 1). Level 1 establishes that the coupled simulator predicts adequately. Level 2 measures relationship stability across seeds and losses. Level 3 checks whether formulations recover the same dominant descriptors and directions. Level 4 evaluates the full input-by-parameter matrix. Each level closes a different loophole: adequate prediction does not imply stable relationships, stable pairs can remain formulation-specific, and agreement on selected controls does not guarantee a compact overall structure.

The outputs are pair-level stability summaries, parameter-level dominant-control classes, matrix-level compactness measures, and evidence labels. Robust evidence is supported throughout the ordered assessment. Conditional evidence fails one or more cross-formulation or diagnostic safeguards. Diagnostic evidence records sensitivity, boundary effects, or mean-spread coupling and is interpreted as behavior of the learning system rather than as a stable environmental control.

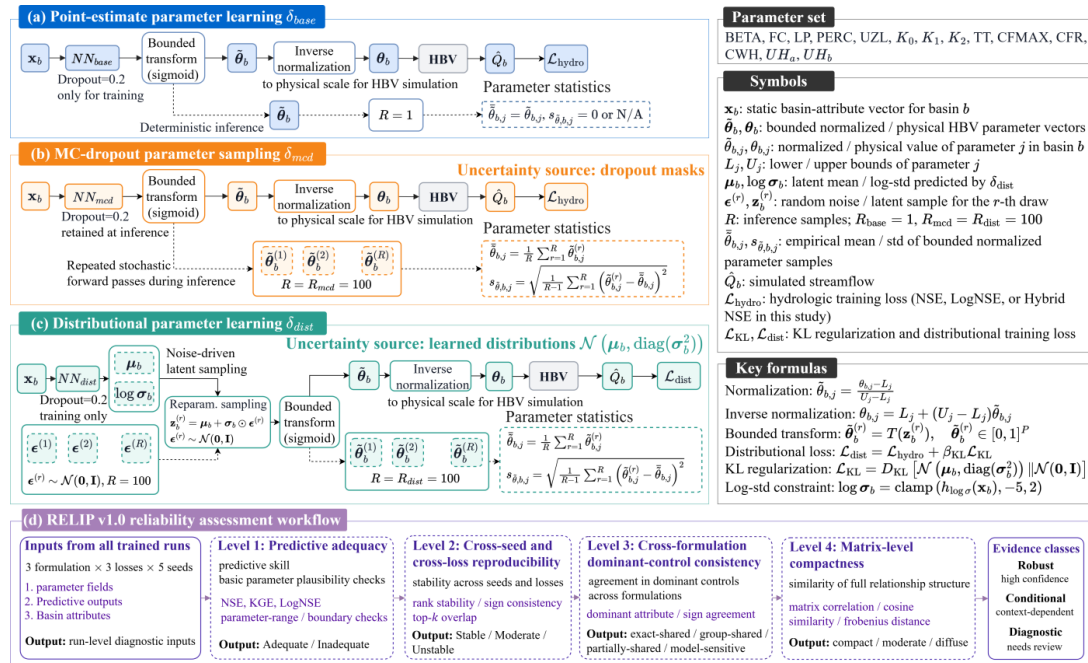


Figure 1. Schematic overview of the three parameter-learning formulations and the RELIP v1.0 assessment workflow. Panel (a) presents the deterministic point-estimate formulation, δ_{base} , in which basin descriptors are mapped to a single bounded, normalized parameter vector. Panel (b) presents the Monte Carlo dropout formulation, δ_{mcd} , which retains dropout during inference to generate stochastic parameter samples. Panel (c) presents the distributional formulation, δ_{dist} , in which the network predicts latent distribution parameters and bounded, normalized parameter samples are generated through reparameterization and a sigmoid transform. In all three formulations, normalized parameters are mapped to their physical HBV ranges before entering the differentiable simulator, while the parameter statistics used in the relationship and gradient analyses are calculated on the bounded normalized scale. Panel (d) summarizes the RELIP v1.0 workflow, linking the 45 controlled runs to four sequential assessment levels—predictive adequacy, cross-seed and cross-loss reproducibility, cross-formulation consistency of dominant controls, and matrix-level compactness—and to the resulting evidence classes.

Table 1. RELIP v1.0 four-level assessment workflow.

Level	Assessment target	Perturbation axis	Existing metrics	Failure mode addressed	Output
1	Predictive adequacy of the coupled simulator	Formulation and loss	NSE, KGE, absolute bias, absolute percent bias	Interpretation of an inadequately performing simulator	Adequacy status
2	Pair-level relationship reproducibility	Random seed and loss	Spearman correlation SD and range, sign consistency, top-k overlap	Initialization- or objective-specific relationships	Stable candidate relationships



3	Dominant-control consistency	Parameter-learning formulation	Modal dominant attribute and sign agreement	Formulation-specific interpretation	Exact-shared, group-shared, partially shared, or model-sensitive class
4	Full relationship structure	All complete runs	Matrix correlation, cosine similarity, Frobenius distance	Selection of isolated stable pairs from an unstable matrix	Within-formulation compactness and evidence constraint

2.2 Differentiable HBV benchmark and parameter-learning formulations

2.2.1 Differentiable parameter-learning setting

In differentiable parameter learning, a neural network maps static descriptors to parameters of a differentiable mechanistic simulator. The mapping is learned jointly across basins instead of calibrating an independent vector for each one. For basin b , let $\mathbf{x}_b$ denote the static attribute vector, and let m , l , and s index formulation, loss function, and random seed. This index order is used throughout. The learned mapping is

$$\tilde{\boldsymbol{\theta}}_b^{(m,l,s)} = f_m(\mathbf{x}_b; \boldsymbol{\omega}^{(m,l,s)}) \quad (1)$$

where $\tilde{\boldsymbol{\theta}}_b^{(m,l,s)}$ is the learned bounded normalized parameter vector and $\boldsymbol{\omega}^{(m,l,s)}$ denotes the trained neural-network weights. The normalized output is mapped to the physical HBV parameter scale before hydrologic simulation using the prescribed lower and upper bounds of each parameter. The hydrologic simulation driven by the corresponding physical parameter vector $\boldsymbol{\theta}_b^{(m,l,s)}$ remains inside the computational graph, so the neural mapping can be trained directly against streamflow observations through backpropagation (Feng et al., 2022; Tsai et al., 2021).

2.2.2 HBV mechanistic backbone and parameter scale

The mechanistic backbone is an HBV-type conceptual model implemented in a differentiable framework. It represents snow accumulation and melt, soil-water storage, evapotranspiration limitation, percolation, upper- and lower-zone groundwater storage, and runoff generation (Lindström et al., 1997; Seibert & Bergström, 2022). A gamma unit hydrograph provides routing. The 14 analyzed quantities comprise 12 HBV process parameters and two routing parameters: BETA, FC, LP, PERC, UZL, K_0 , K_1 , K_2 , TT, CFMAX, CFR, CWH, UH_a , and UH_b . Supplement Sect. S1.1 gives the equations, Table S1 lists definitions and bounds, and Table S2 summarizes the forcings, target, and basin attributes.

For analyses that compare parameters with different units and search ranges, parameter values are normalized as

$$\tilde{\theta}_{b,j} = \frac{\theta_{b,j} - L_j}{U_j - L_j} \quad (2)$$

where $\theta_{b,j}$ is the physical value of parameter j , and L_j and U_j are its lower and upper bounds. The learned parameters are interpreted as behavioral parameters of the adopted differentiable HBV model, not as direct observations of true physical properties.

2.2.3 Parameter-learning formulations

All three formulations use the same differentiable HBV backbone (Figure 1a-c). They share the same differentiable HBV backbone but differ in parameter representation, stochastic mechanism, and associated regularization. Unconstrained neural outputs or latent samples pass through a sigmoid transform to the normalized interval $[0, 1]$ and are then mapped to the physical HBV ranges.

The deterministic formulation provides a point-estimate baseline. A multilayer perceptron maps the static descriptor vector $\mathbf{x}_b$ to one bounded normalized parameter vector,

$$\tilde{\boldsymbol{\theta}}_b = f_{\text{det}}(\mathbf{x}_b; \boldsymbol{\omega}_{\text{det}}) \quad (3)$$



Each trained deterministic model returns one parameter vector per basin for a given loss and seed. No parameter spread is represented explicitly. Before HBV simulation, the normalized vector is mapped to the prescribed physical range,

$$\theta_{b,j} = L_j + (U_j - L_j)\tilde{\theta}_{b,j} \quad (4)$$

The Monte Carlo dropout (MC-dropout) formulation uses dropout during training and retains dropout during inference and relationship analysis (Gal & Ghahramani, 2016; Kendall & Gal, 2017). Repeated stochastic forward passes produce bounded normalized samples,

$$\tilde{\theta}_b^{(r)} = f_{\text{mcd}}(\mathbf{x}_b; \boldsymbol{\omega}_{\text{mcd}}, \mathbf{d}^{(r)}), \quad r = 1, \dots, R_{\text{mcd}} \quad (5)$$

Here $\mathbf{d}^{(r)}$ is the dropout mask for pass r . We used $R_{\text{mcd}} = 100$ stochastic passes per basin. Their empirical mean provides the basin-level parameter estimate, and their empirical standard deviation measures dropout-induced spread. This quantity reflects stochasticity in the neural parameterizer; it is not an explicitly learned parameter distribution.

The distributional formulation predicts a latent location and log-standard deviation for each basin and parameter (Kingma & Welling, 2022),

$$\boldsymbol{\mu}_b, \log \boldsymbol{\sigma}_b = g_{\text{dist}}(\mathbf{x}_b; \boldsymbol{\omega}_{\text{dist}}) \quad (6)$$

The constraint was implemented as a numerical clamp of the predicted log-standard deviation to the interval $[-5, 2]$ before sampling to avoid unrealistically small or large latent spreads. Latent samples are generated using the reparameterization form

$$\mathbf{z}_b^{(r)} = \boldsymbol{\mu}_b + \boldsymbol{\sigma}_b \odot \boldsymbol{\epsilon}^{(r)}, \quad \boldsymbol{\epsilon}^{(r)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (7)$$

and transformed to the bounded normalized parameter scale by $T(\cdot)$, i.e., $\tilde{\theta}_b^{(r)} = T(\mathbf{z}_b^{(r)})$, where $T(\cdot)$ denotes the sigmoid transform applied element-wise. During training, one latent parameter sample is drawn for each forward simulation. The distributional loss combines the hydrologic loss with KL regularization,

$$\mathcal{L}_{\text{dist}} = \mathcal{L}_{\text{hydro}} + \beta_{\text{KL}} D_{\text{KL}}[\mathcal{N}(\boldsymbol{\mu}_b, \text{diag}(\boldsymbol{\sigma}_b^2)) \| \mathcal{N}(\mathbf{0}, \mathbf{I})] \quad (8)$$

The KL coefficient was 0.001 and increased linearly during the first 10 epochs. A standard multivariate normal prior was used in latent space. At inference, 100 bounded parameter samples were drawn to estimate the empirical mean and standard deviation. The latter is reported as parameter-scale diagnostic spread. It is neither a complete Bayesian posterior nor direct evidence of physical identifiability (Abdar et al., 2021; Hüllermeier and Waegeman, 2021).

2.3 Data and controlled perturbation experiment

2.3.1 CAMELS-US benchmark sample and model inputs

CAMELS-US provides the forcing, daily streamflow, static descriptors, and basin locations needed for the benchmark (Addor et al., 2017; Newman et al., 2015). We used the standard 531-basin subset distributed with the dMG CAMELS configuration and widely adopted in large-sample hydrologic modeling (Kratzert et al., 2019). Basin identifiers came from the archived 531sub_id.txt file. No basin was added or removed on the basis of model performance.

Figure 5a shows the basin locations. The sample spans humid eastern catchments, arid and semi-arid western basins, snow-influenced mountain catchments, and lowlands with little snow influence. Elevation, slope, aridity, precipitation seasonality, snow fraction, vegetation, soil, and geology also vary widely. This cross-basin variation supports the descriptor-to-parameter analysis.

Daily precipitation and mean air temperature came from the Daymet v2 CAMELS extraction; potential evapotranspiration was calculated with the Hargreaves method. The dMG loader converted USGS discharge from cubic feet



per second to basin-area-normalized millimeters per day, the output unit of the HBV model. Static descriptors were supplied only to the neural parameterizer, so all descriptor-to-parameter relationships are cross-basin rather than temporal. Every formulation used the same archived normalization statistics and preprocessing configuration.

The adopted HBV and routing structure contains 14 learned parameters. Their number was fixed by the mechanistic model, not selected from the data. Across 531 basins, each complete run therefore contains 7,434 basin-parameter entries. Supplement Sect. S1.1 gives the equations, Table S1 lists parameter definitions and bounds, and Table S2 documents the forcing and descriptor schema.

2.3.2 Temporal split and warm-up

The model was trained from 1 January 1989 to 31 December 1998 and tested from 1 January 1999 to 31 December 2009. A 365-day warm-up preceded evaluation. Supplement Sect. S2 documents the data protocol and preprocessing.

2.3.3 Controlled perturbation design

The experiment crossed three formulations, three hydrologic losses, and five fixed seeds (111, 222, 333, 444, and 555), producing 45 complete runs. The NSE loss emphasizes overall fit and high flows, logNSE places more weight on low flows, and the hybrid loss balances the two. Losses and seeds were perturbations of the learned parameterization, not a pool from which to select a favorable run. Supplement Sect. S2 gives the formulas and run inventory.

2.3.4 Implementation and inference settings

All formulations used two hidden ReLU layers with 128 units each. Training used Adam, an initial learning rate of 0.01, a step scheduler (20-epoch step, 0.5 decay), batches of 100 basins, and 100 epochs. MC-dropout used a rate of 0.2 and 100 inference passes; the distributional model used 100 bounded samples. The HBV backbone, basin sample, temporal split, and preprocessing were identical across the 45 runs. Hardware, software versions, and computational costs are reported in Supplement Sect. S2.3.

2.4 Diagnostic metrics, evidence assignment, and supporting checks

2.4.1 Level 1: predictive adequacy and parameter-value safeguards

Predictive adequacy was summarized with Nash-Sutcliffe efficiency, Kling-Gupta efficiency, absolute bias, and absolute percent bias. Metrics were calculated for each basin over the test period after warm-up and then summarized across basins and runs (Gupta et al., 2009; Nash & Sutcliffe, 1970). They were not used to choose the formulation for interpretation.

Across seeds, parameter-value stability was summarized by the standard deviation and range of normalized values within each formulation, loss, basin, and parameter. These measures describe numerical stability only. Boundary saturation used a primary threshold of $\tilde{\theta} \leq 0.02$ or $\tilde{\theta} \geq 0.98$, with 0.05/0.95 and 0.10/0.90 used for sensitivity checks. Distance to the nearest boundary was defined as

$$d_{b,j} = \min(\tilde{\theta}_{b,j}, 1 - \tilde{\theta}_{b,j}) \quad (9)$$

For MC-dropout and distributional runs, the 90% interval width $q_{0.95} - q_{0.05}$ was calculated from 100 bounded samples and used as the main stochastic spread summary.

2.4.2 Level 2: reproducibility across seeds and loss functions

Level 2 focuses on individual descriptor-to-parameter relationships. Pair-level metrics measure sensitivity to seed and loss; ranked-overlap metrics test whether the same high-ranking descriptors recur. Supplement Sect. S3 documents the full workflow.



For model formulation m , loss function l , seed s , basin descriptor a , and HBV parameter j , the primary relationship statistic was the basin-level Spearman correlation

$$\rho_{a,j}^{(m,l,s)} = \text{corr}_s(x_{b,a}, \bar{\theta}_{b,j}^{(m,l,s)})_{b=1}^B, \quad B = 531 \quad (10)$$

Spearman correlation was chosen because the relationships may be monotonic without being linear (Spearman, 1904).

Cross-seed variability within each formulation and loss was calculated as

$$\text{SD}_{\text{seed}}(\rho_{a,j}^{(m,l)}) = \sqrt{\frac{1}{S-1} \sum_{s=1}^S (\rho_{a,j}^{(m,l,s)} - \bar{\rho}_{a,j}^{(m,l)})^2} \quad (11)$$

where $S = 5$. Cross-loss variability was computed after seed aggregation,

$$\text{SD}_{\text{loss}}(\rho_{a,j}^{(m)}) = \sqrt{\frac{1}{L-1} \sum_{l=1}^L (\bar{\rho}_{a,j}^{(m,l)} - \bar{\rho}_{a,j}^{(m)})^2} \quad (12)$$

where $L = 3$. Seed and loss ranges provide complementary sensitivity summaries.

Sign consistency is the fraction of nonzero correlations that share the majority sign. Candidate relationships were selected by absolute Spearman correlation after pooling all formulations, losses, and seeds, which avoids defining the comparison around a single formulation. Pooling can still favor pairs that are large in only one formulation, so the main ordering was checked across alternative top-k sets. Top 10 pairs per parameter were used for broad stability summaries and top 3 for dominant-control classification.

2.4.3 Level 3: cross-formulation consistency of dominant controls

Dominant controls were identified as the modal highest-magnitude relationship for each parameter after aggregation across losses and seeds. The descriptor groups, including climate/hydroclimate, soil, topography/scale, geology, and vegetation/land cover, were fixed before the comparison. Exact-shared relationships have the same dominant attribute and sign in all three formulations. Group-shared relationships retain a common sign, with dominant attributes belonging to the same predefined descriptor group across formulations. Partially shared relationships agree in dominant attribute and sign for two formulations but differ in the third. All remaining cases are classified as model-sensitive. Ties are resolved using a fixed ranking rule, without post-hoc hydrologic similarity judgements. These categories were prespecified. Group-shared was retained in the protocol, although no parameter met that criterion in this experiment.

2.4.4 Level 4: matrix-level compactness and evidence assignment

Each run is represented by its complete input-by-parameter Spearman matrix. Matrix-level reproducibility was assessed with Frobenius distance, cosine similarity, and matrix correlation. Frobenius distance is

$$D_F(\mathbf{R}_p, \mathbf{R}_q) = \|\mathbf{R}_p - \mathbf{R}_q\|_F \quad (13)$$

Cosine similarity and matrix correlation were calculated from the finite entries of the vectorized matrices. Within-formulation compactness is the mean pairwise matrix correlation among runs; larger values indicate greater internal consistency.

Evidence labels summarize the four levels. Robust evidence combines adequate prediction, low seed and loss sensitivity, consistent dominant control and sign, and a compact surrounding matrix. Conditional evidence is reproducible but fails at least one cross-formulation or diagnostic safeguard. Diagnostic evidence records inconsistent controls, boundary effects, or mean-



231 spread coupling. These labels limit the strength of interpretation within this experiment; they do not establish causality or
232 physical identifiability.

233 2.4.5 Supporting environmental and spread diagnostics

234 Environmental-gradient and parameter-spread analyses follow the four RELIP levels; they are not a fifth level. The
235 gradient analysis checks whether selected relationships have coherent expression across basins. The spread analysis screens
236 mean coupling and parameter-bound effects before any interpretation of parameter-scale variability.

237 Gradient analyses used Spearman correlations and rank-based tercile contrasts. For the headline parameter-mean
238 gradients, distributional parameter means under HybridNseBatchLoss were averaged across the five seeds within each basin.
239 Spearman correlations were then computed across all 531 basins, and basins were assigned to rank-based low, middle, and
240 high terciles for the contrast analysis. For each gradient–parameter pair, the high-minus-low tercile contrast was defined as

$$\Delta_{H-L}(a, j) = \text{median}(\tilde{\theta}_{b,j}; b \in H_a) - \text{median}(\tilde{\theta}_{b,j}; b \in L_a) \quad (14)$$

241 where H_a and L_a denote the high and low terciles of attribute a . Medians and interquartile ranges summarize each group.
242 High-minus-low tests are descriptive contrasts, not causal tests.

243 Parameter-scale diagnostic spread was analyzed on the normalized parameter scale. For basin and parameter, normalized
244 spread was defined as

$$u_{b,j} = \frac{s_{\theta,b,j}}{U_j - L_j} \quad (15)$$

245 where the stochastic standard deviation is expressed in physical parameter units. Mean-spread coupling is the Spearman
246 correlation between normalized parameter mean and standard deviation. Boundary sensitivity combines the correlation
247 between boundary distance and standard deviation with the fraction of near-boundary basins. A spread relationship is flagged
248 as mean-coupled when the first absolute correlation is at least 0.5, and as boundary-sensitive when the second is at least 0.4 or
249 the near-boundary share is at least 0.25. These predefined thresholds separate less-confounded gradients from cautionary
250 patterns. Supplement Sect. S6.1 gives the full definitions, and Figure S2 checks the boundary threshold. The resulting flags
251 restrict claims to diagnostic structure rather than physical identifiability (Abdar et al., 2021; Hüllermeier & Waegeman, 2021).

252 3 Results

253 3.1 Predictive adequacy and preliminary parameter diagnostics

254 All three formulations produced adequate streamflow simulations across the 531 basins (Figure 2; Table 2). Across
255 complete runs, median NSE was 0.611 for δ_{base} , 0.624 for δ_{mcd} , and 0.611 for δ_{dist} ; median KGE was 0.626, 0.639, and 0.625.
256 Under the hybrid loss, median NSE ranged from 0.632 to 0.636 and median KGE from 0.675 to 0.679. The performance
257 distributions overlapped closely. The fractions of basins above NSE = 0.5 were 0.74, 0.74, and 0.73 for δ_{base} , δ_{mcd} , and δ_{dist} ,
258 while the corresponding KGE fractions were 0.83, 0.81, and 0.83. Basin-level comparisons against δ_{base} also lay close to the
259 one-to-one line for δ_{mcd} and δ_{dist} . The relationship diagnostics therefore compare parameter fields learned under broadly
260 similar predictive conditions.

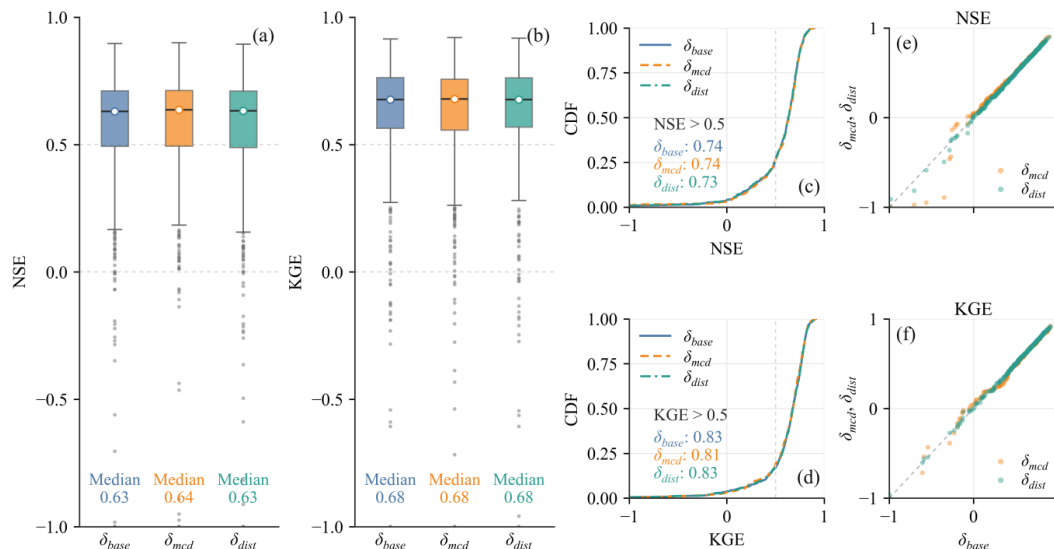


Figure 2. Predictive performance of the three parameterization formulations. Panels (a) and (b) compare basin-level NSE and KGE for δ_{base} , δ_{mcd} , and δ_{dist} under the reference hybrid loss, with medians printed inside each boxplot. Panels (c) and (d) show empirical CDFs and the fractions of basins exceeding NSE or KGE of 0.5. Panels (e) and (f) compare the probabilistic formulations against the deterministic formulation using basin-wise quantile-quantile summaries. The three formulations have similar performance distributions, so the metrics are used as adequacy checks for the 531-basin analysis rather than as the basis for selecting an interpretive formulation.

Loss-specific results followed the same pattern (Figure S1). Under the NSE loss, median NSE was 0.65, 0.66, and 0.65 for δ_{base} , δ_{mcd} , and δ_{dist} ; median KGE was also 0.65, 0.66, and 0.65. Under logNSE, median NSE fell to 0.56, 0.58, and 0.56 and median KGE to 0.53, 0.58, and 0.53. Hybrid-loss medians were 0.63, 0.64, and 0.63 for NSE and 0.68 for KGE in all three formulations. Loss-to-loss differences were as large as, or larger than, formulation differences. MC-dropout had slightly higher median NSE and KGE but also slightly greater absolute bias. Median absolute percent bias remained approximately 46–47 %, indicating that adequacy here is comparative and does not imply uniformly low water-balance error across basins. Prediction was therefore used as an adequacy screen, not as the criterion for choosing an interpretive formulation.

Table 2. Predictive performance of the three parameterization formulations across 531 CAMELS-US basins, summarized across complete runs and the reference hybrid loss.

Summary level	Metric	δ_{base}	δ_{mcd}	δ_{dist}
All complete runs	Median NSE	0.611	0.624	0.611
All complete runs	Median KGE	0.626	0.639	0.625
Hybrid loss	Median NSE	0.632	0.636	0.632
Hybrid loss	Median KGE	0.675	0.679	0.675
All complete runs	Fraction NSE (>0.5)	0.74	0.74	0.73
All complete runs	Fraction KGE (>0.5)	0.83	0.81	0.83
All complete runs	Median absolute bias($mm\ d^{-1}$)	0.736	0.790	0.732
All complete runs	Median absolute percent bias(%)	46.397	47.384	46.497

Raw parameter diagnostics gave a different warning (Figure 3). Cross-seed variability depended on parameter and loss, and several apparently stable parameters lay near their prescribed bounds. Boundary saturation was most evident for snow, recession, and routing parameters. Once boundary-sensitive parameters were separated, formulation differences in raw variability became smaller and no longer favored δ_{dist} consistently. Table S3 also identifies cases in which low parameter variability coexisted with unstable descriptor relationships, including deterministic CFR, CWH, and UH_a and dropout CFR, CWH, and K_2 . Several were strongly boundary-saturated. Numerical parameter stability and relationship reproducibility were therefore kept separate. The same interpretation held under alternative boundary thresholds (Figure S2).

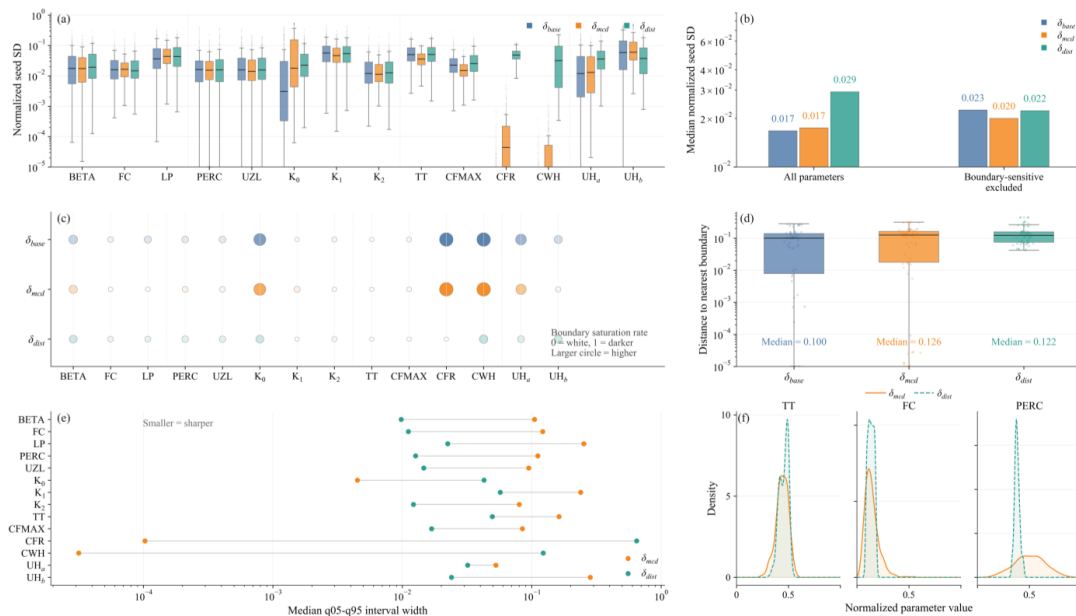


Figure 3. Cross-seed parameter stability, boundary saturation, and stochastic interval behavior. Panels (a) and (b) summarize normalized seed variability by parameter and formulation, including pooled summaries before and after excluding boundary-sensitive parameters. Panels (c) and (d) show boundary saturation and distance-to-boundary diagnostics, indicating where low apparent variability can partly reflect bounded transformations. Panels (e) and (f) compare 5th-95th percentile interval widths for the stochastic formulations and representative parameter-sample distributions. Raw parameter stability is therefore interpreted together with boundary behavior and interval width.

MC-dropout's median-NSE advantage was about 0.013. This difference was smaller than or comparable to the variation among losses and did not produce a distinct performance distribution. All three formulations passed the adequacy screen with closely overlapping results. At this point, conventional predictive metrics offer little basis for choosing among them. The remaining analyses ask a different question: do the learned relationship structures reproduce equally well?

3.2 Reproducibility of learned input-to-parameter relationships across seeds, losses, and hydroclimatic groups

Relationship diagnostics separated the formulations more clearly than prediction or raw parameter variability (Figure 4; Table 3). Candidate pairs were selected from pooled absolute Spearman correlations across all formulations, losses, and seeds, so the comparison was not centered on δ_{dist} . Within this common set, median cross-seed SD of Spearman ρ was 0.0241 for δ_{dist} , 0.0347 for δ_{mcd} , and 0.0461 for δ_{base} . Median seed ranges were 0.0630, 0.0891, and 0.1122 for δ_{dist} , δ_{mcd} , and δ_{base} .

Table 3. Learned-relationship reproducibility diagnostics across formulations for candidate high-magnitude input-to-parameter relationships.

Metric	δ_{base}	δ_{mcd}	δ_{dist}
Median cross-seed SD of Spearman ρ	0.0461	0.0347	0.0241
Median cross-seed range of Spearman ρ	0.1122	0.0891	0.0630
Mean cross-seed SD of Spearman ρ	0.0663	0.0546	0.0308
Median dominant-attribute consistency	0.800	0.800	1.000
Median top-5 overlap	0.629	0.679	0.700
Mean cross-loss SD of Spearman ρ	0.142	0.168	0.129
Median hydroclimatic-group seed SD of Spearman ρ	0.0586	0.0541	0.0491

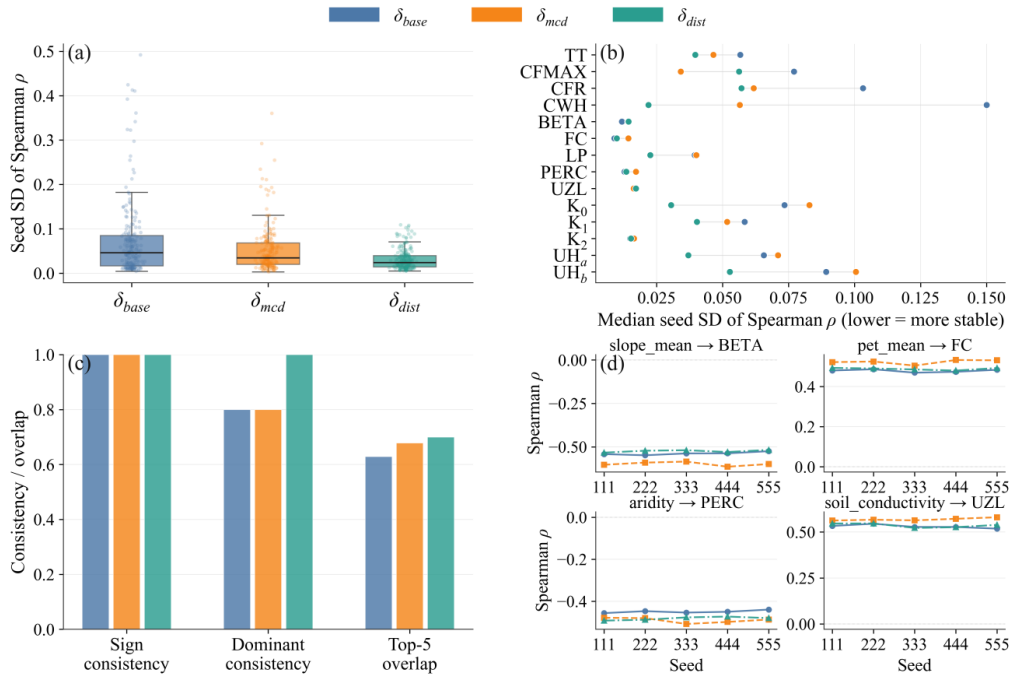


Figure 4. Cross-seed stability of learned input-to-parameter relationships. Candidate high-magnitude relationships are selected by mean absolute Spearman correlation across all formulations, losses, and seeds, avoiding selection based on δ_{dist} alone. Panels summarize pooled seed standard deviation, parameter-level seed variability, sign consistency, dominant-control consistency, top-5 overlap, and four representative robust pairs. Lower seed standard deviation and narrower seed ranges indicate more reproducible relationship recovery and are interpreted together with hydrologic and diagnostic context.

The full sensitivity summary retained this ordering. Mean cross-seed SD of Spearman ρ was 0.0308 for δ_{dist} , 0.0546 for δ_{mcd} , and 0.0663 for δ_{base} . Figure 4a shows a tighter distribution for δ_{dist} and longer upper tails for δ_{base} and δ_{mcd} . The difference under δ_{dist} was visible for several runoff-generation and storage parameters, including BETA, FC, PERC, and UZL, although snow, recession, and routing parameters remained more seed-sensitive. Reproducibility was formulation-dependent but not uniform across parameters.

Dominant-control summaries gave the same overall result. Median dominant-attribute consistency was 1.000 for δ_{dist} and 0.800 for both δ_{base} and δ_{mcd} . Median top-5 overlap was 0.700, 0.679, and 0.629 for δ_{dist} , δ_{mcd} , and δ_{base} , respectively (Figure 4c). The examples in Figure 4d—*Mean slope* – BETA, *Pet* – FC, *Aridity* – PERC, and *Soil cond.* – UZL—retain both sign and magnitude across the five seeds. They illustrate the candidate set rather than exhaust it.

Loss functions caused more variation than seeds, but the formulation ordering persisted. Mean cross-loss SD of Spearman ρ was 0.129 for δ_{dist} , 0.142 for δ_{base} , and 0.168 for δ_{mcd} (Table 3). The same qualitative ordering appeared for alternative top- k selections (Figure S3), indicating that it was not tied to one retained-set size.

The pattern also appeared within the seven hydroclimatic strata (Figure 5). Median group-wise seed SD of Spearman ρ was 0.0491 for δ_{dist} , 0.0541 for δ_{mcd} , and 0.0586 for δ_{base} . Top-5 overlap was comparable or higher for δ_{dist} in several groups, although the size of the difference varied. Tercile analyses based on aridity, snow fraction, and precipitation seasonality gave the same broad result (Figure S4).

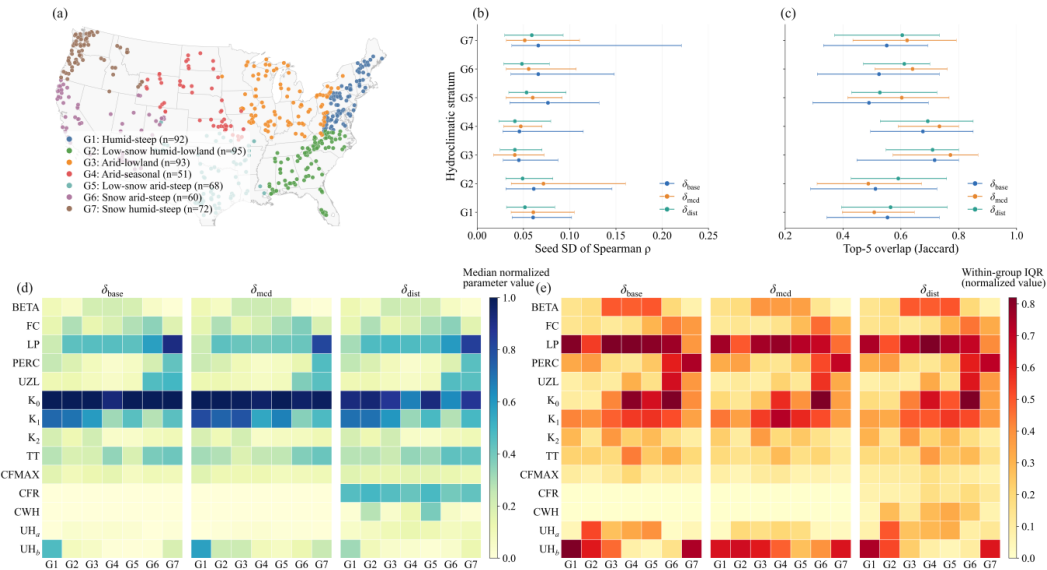


Figure 5. Basin-group reliability of learned input-to-parameter relationships. Panel (a) maps the 531 CAMELS-US basins in seven hydroclimatic groups used for robustness checks. Panels (b)-(e) compare group-wise pairwise variability, top-5 overlap, and normalized within-group parameter spread across formulations and hydroclimatic strata. The figure evaluates whether learned-relationship reliability persists across major basin environments; the groups are robustness strata rather than independent validation experiments.

The strata also differed in their learned parameter distributions (Figure 5d-e), especially for snow, storage, recession, and routing parameters. These summaries provide context but do not define the formulation comparison. Leave-one-group-out correlations remained within the operational flag $|\Delta\rho| \leq 0.15$ for all selected distributional relationships in the all-loss field, formed by averaging all three losses and five seeds within each basin (Figure S8; Table S6), so no single stratum controlled the full-sample result. Within-group estimates were less stable: FC-PET weakened and changed sign in one snow-arid group, while PERC-aridity changed sign in the low-snow arid steep group. Deduplication had negligible effect for δ_{dist} ; mean cross-seed SD changed from 0.0312 to 0.0313, and the main shared-control correlations shifted by less than 0.01.

3.3 Shared and formulation-sensitive input-to-parameter relationships

The formulations recovered related, but not identical, relationship matrices (Table 4 and Table 5). Seven of the 14 parameters had the same dominant descriptor and direction in all three formulations, six were partially shared, and one was model-sensitive. Figure 6 shows the distributional matrix and its descriptor-group summaries.

Table 4. Dominant basin controls, direction, and cross-formulation consistency for the 14 learned HBV and routing parameters, based on modal dominant attributes across losses and seeds.

Parameter	δ_{base}	δ_{mcd}	δ_{dist}	Direction	ρ_{dist}	Consistency class	Sign-stability flag
BETA	Mean slope	Mean slope	Mean slope	-	-0.583	Exact-shared	Consistent
FC	PET	PET	PET	+	0.509	Exact-shared	Consistent
LP	High-precip. dur.	LAI diff.	Precip. seasonality	-	-0.187	Model-sensitive	Consistent
PERC	Aridity	Aridity	Aridity	-	-0.594	Exact-shared	Consistent
UZZ	Soil cond.	Soil cond.	Soil cond.	+	0.570	Exact-shared	Consistent
K ₀	Snow fraction	Soil cond.	Snow fraction	-	-0.424	Partially shared	Consistent
K ₁	High-precip. dur.	GVP diff.	High-precip. dur.	-	-0.454	Partially shared	Consistent
K ₂	High-precip. dur.	High-precip. dur.	High-precip. dur.	-	-0.336	Exact-shared	Consistent
TT	Elevation	Elevation	Elevation	+	0.265	Exact-shared	Consistent
CFMAX	Clay	Snow fraction	Clay	-	-0.426	Partially shared	Consistent
CFR	Mean slope	Snow fraction	Snow fraction	-	-0.625	Partially shared	Consistent
CWH	Mean slope	Snow fraction	Snow fraction	-	-0.904	Partially shared	Consistent
UH _a	Mean slope	Mean slope	Mean slope	-	-0.426	Exact-shared	Consistent
UH _b	Aridity	Snow fraction	Aridity	-	-0.367	Partially shared	Inconsistent

Note: Direction and ρ refer to the dominant control under δ_{dist} . Exact-shared, group-shared, partially shared, and model-sensitive denote decreasing levels of cross-formulation agreement in dominant attribute and sign. Group-shared was predefined but not observed. For UH_b, sign inconsistency is an auxiliary formulation-loss flag, not a separate consistency class.



Four shared controls were prominent in the distributional matrix: Mean slope-BETA (-0.583), PET-FC (0.509), Aridity-
 PERC (-0.594), and Soil cond.-UZL (0.570) (Figure 6a; Table 4). Their Spearman ρ values span distinct descriptor groups,
 linking runoff-generation nonlinearity to terrain, storage capacity to evaporative demand, percolation to dryness, and upper-
 zone threshold behavior to soil hydraulic properties.

Figure 6b-c summarizes relationship strength by descriptor group. Snow parameters were associated mainly with climate
 and then topography (top-3 mean absolute correlations of 0.70 and 0.55). Soil parameters had similar contributions from
 climate, soil, and topography (0.47, 0.46, and 0.42). Production parameters drew more broadly on climate, soil, and vegetation
 (0.54, 0.51, and 0.49). Routing associations were moderate across these groups (0.40-0.46) and weaker for geology (0.28). At
 parameter level, CWH, PERC, and UZL had among the larger top- k mean relationship magnitudes, whereas LP, K_2 , and some
 routing entries were lower.

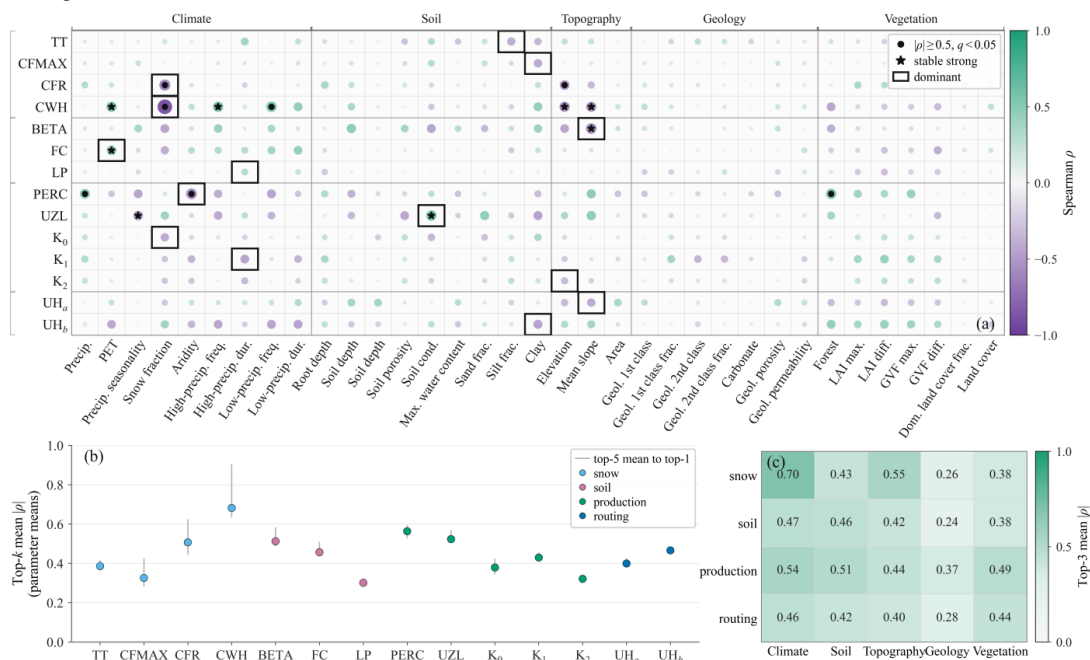


Figure 6. Distributional parameter-mean relationships with basin descriptors. Panel (a) is a 35-attribute by 14-parameter circle heatmap of Spearman correlations between δ_{dist} parameter means and basin descriptors, with attributes ordered consistently by climate/hydroclimate, soil, topography/scale, geology, and vegetation/land-cover groups. Color gives signed correlation and circle size gives absolute magnitude; black dots mark strong relationships, stars mark stable strong relationships, and boxes mark the dominant attribute for each parameter. Panels (b) and (c) summarize top-ranked relationship strength by parameter and by attribute group, linking the matrix view to hydrologic process categories.

The other parameters were less consistent across formulations. CFR, CWH, CFMAX, K_0 , K_1 , and UH_b were partially shared, and UH_b also changed sign in some formulation-loss summaries. LP was the only model-sensitive parameter. Figure S5 provides the deterministic and MC-dropout matrices; Figure S6 gives the complementary Integrated Gradients check.

Table 5. Matrix-level compactness and consistency of input-to-parameter relationship structures across formulations, using full input-by-parameter Spearman correlation matrices.

Metric	δ_{base}	δ_{mcd}	δ_{dist}	Interpretation
Within-formulation compactness	0.702	0.680	0.809	Larger values indicate more internally consistent relationship matrices
Mean within-formulation Frobenius distance	3.453	3.668	2.624	Smaller values indicate more similar relationship matrices
Mean within-formulation cosine similarity	0.701	0.679	0.808	Larger values indicate closer matrix orientation
Mean intermodel similarity with other formulations	0.682	0.677	0.679	Used to assess whether a formulation departs from the shared relationship structure

Matrix-level results agreed with the parameter-level classification (Table 5). Within-formulation compactness was 0.809 for δ_{dist} , 0.702 for δ_{base} , and 0.680 for δ_{mcd} . Mean Frobenius distance was also lowest for the distributional formulation



(2.624, compared with 3.453 for δ_{base} and 3.668 for δ_{mcd}). Intermodel similarity remained substantial, placing the distributional runs within the broad structure shared by all three formulations rather than in a separate regime. Figure 7 shows the spatial organization of the distributional parameter means; Figure S9 provides companion maps and spread summaries.

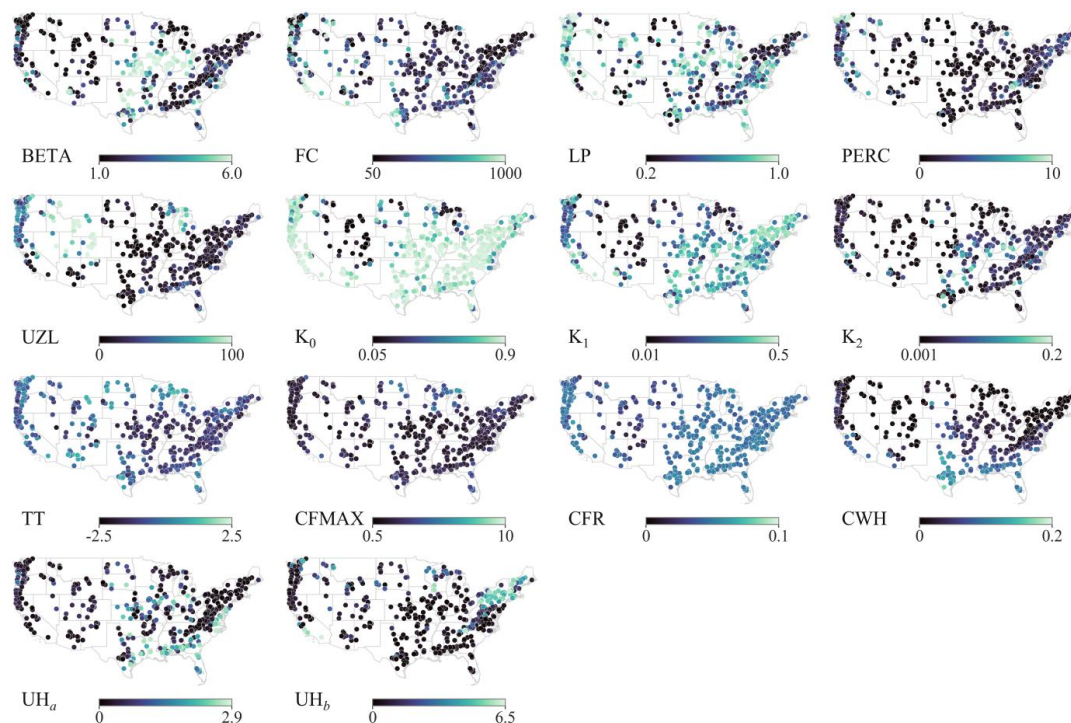


Figure 7. Spatial organization of learned HBV and routing parameter means. Maps show seed-averaged physical parameter means across the 531 CAMELS-US basins under the reference hybrid loss, with the distributional formulation shown here. Deterministic and MC-dropout companion maps are provided in Figure S9. Each panel uses the physical search range of its own parameter, so color scales should be interpreted within panels rather than compared across parameters. The maps provide a descriptive spatial check on the learned fields and support the subsequent attribute-gradient analyses; they are not a standalone test of spatial significance.

3.4 Environmental gradients and diagnostic parameter spread

Several dominant controls also appeared as coherent environmental gradients (Figure 8; Table S8). Under the reference distributional setting—HybridNseBatchLoss with parameter means averaged across five seeds at basin level—BETA decreased with mean slope, FC increased with PET, PERC decreased with aridity, and UZL increased with soil conductivity. Their Spearman correlations were -0.583, 0.509, -0.594, and 0.570, respectively, and their corresponding high-minus-low tercile contrasts were -0.403, 0.198, -0.229, and 0.260. These gradients provide a domain-coherence check on relationships already supported by the reproducibility diagnostics. PERC–aridity was the most loss-sensitive of these relationships: its Spearman correlation was -0.594 in the reference hybrid-loss field but -0.390 in the all-loss averaged field, indicating that its direction was more stable than its magnitude.

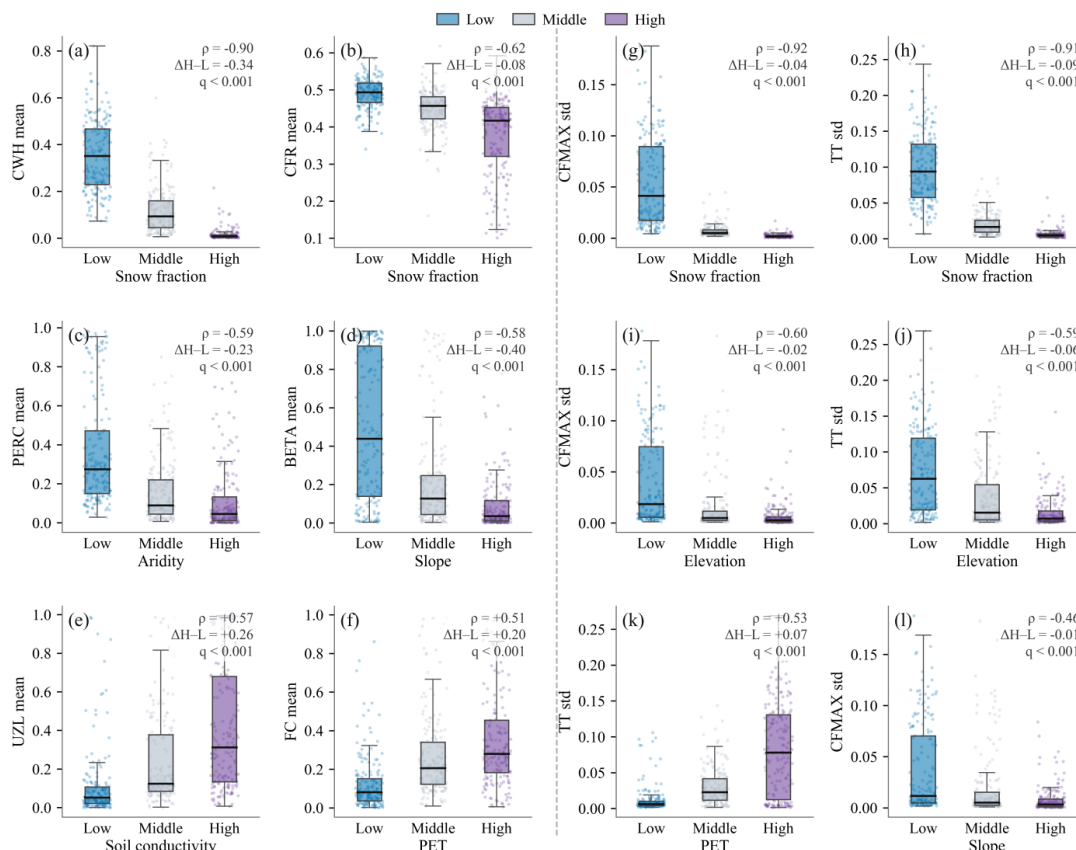


Figure 8. Selected parameter-mean and parameter-spread gradients. The four exact-shared controls are shown in panels (c)–(f): BETA decreased with mean slope (panel d), FC increased with PET (panel f), PERC decreased with aridity (panel c), and UZL increased with soil conductivity (panel e). Their high-minus-low median contrasts were -0.40 , $+0.20$, -0.23 , and $+0.26$, respectively. These four panels provide a domain-coherence check on relationships already supported by the reproducibility diagnostics. Panels (a)–(b) show partially shared snow-related mean gradients, while panels (g)–(l) summarize selected parameter-spread gradients. All parameter-mean gradients use the distributional formulation under HybridNseBatchLoss. Parameter means were averaged across the five seeds within each basin before Spearman correlations and rank-based tercile contrasts were computed across the 531 basins.

Snow-related means showed strong gradients as well. CWH and CFR both decreased with snow fraction, with Spearman $\rho = -0.90$ and $\rho = -0.62$ and high-minus-low differences of -0.34 and -0.08 , respectively (Figure 8). Their cross-formulation classification was only partially shared, so these gradients remain distributional evidence rather than part of the shared-control core.

Figure S11a shows that several mean relationships were monotonic without being linear. CWH declined sharply at low snow fraction and then levelled off; BETA changed most between low and moderate slopes; and UZL increased mainly over the lower and middle soil-conductivity range. These panels extend Figure 8 with secondary or more formulation-sensitive relationships and do not carry the same evidentiary weight as the shared controls.

Parameter-scale spread also varied with basin descriptors (Figure 8, Figure 9, and Figure S11b; Tables S7–S8). The clearest tercile patterns were snow-related: CFMAX and TT spread both decreased with snow fraction. Figure S11b gives the complete point-level and binned relationships, while Figures S12–S13 provide the MC-dropout comparisons.

The spread matrix was structured, but not every strong pattern was equally interpretable (Figure 9). CFMAX and TT were the least confounded selected cases because their snow-fraction gradients—CFMAX standard deviation and TT standard deviation—had weak mean coupling and little boundary sensitivity (Figure S14; Table S7). By contrast, selected gradients for



CWH, PERC, UZL, and UH_b were coupled to parameter means, boundary distance, or both. They are reported as properties of the learned mapping, not as independent uncertainty or identifiability signals.

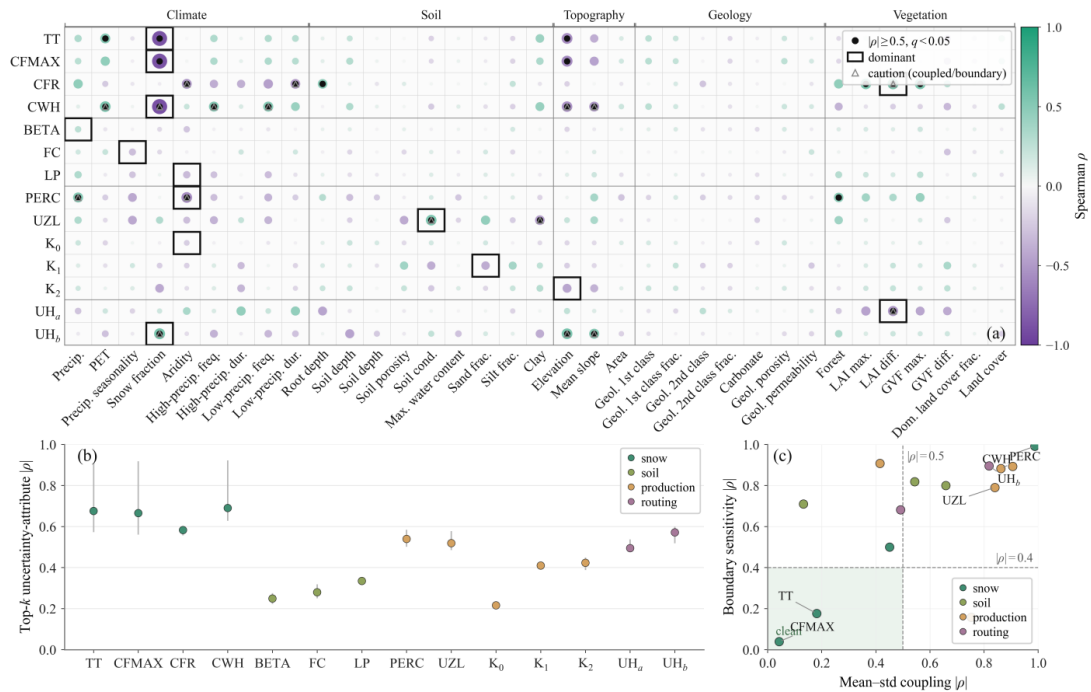


Figure 9. Distributional parameter-scale spread relationships and diagnostics. Panel (a) shows Spearman correlations between normalized δ_{dist} parameter standard deviations and basin descriptors, using the same shared attribute order and color-size encoding as the mean-relationship matrix. Panel (b) summarizes parameter-level spread-structure strength, and panel (c) compares mean-standard deviation coupling with boundary sensitivity. Grey caution markers identify strong spread relationships affected by these diagnostics, so the figure treats spread gradients as structured diagnostic signals rather than calibrated posterior uncertainty.

Figure 10 brings the large-sample gradients back to individual basins. The selected cases occupy different positions along snow fraction, aridity, slope, soil conductivity, PET, and area, and their parameter profiles differ across the full 14-parameter vector. The examples are illustrative only; they are not an additional source of inference.

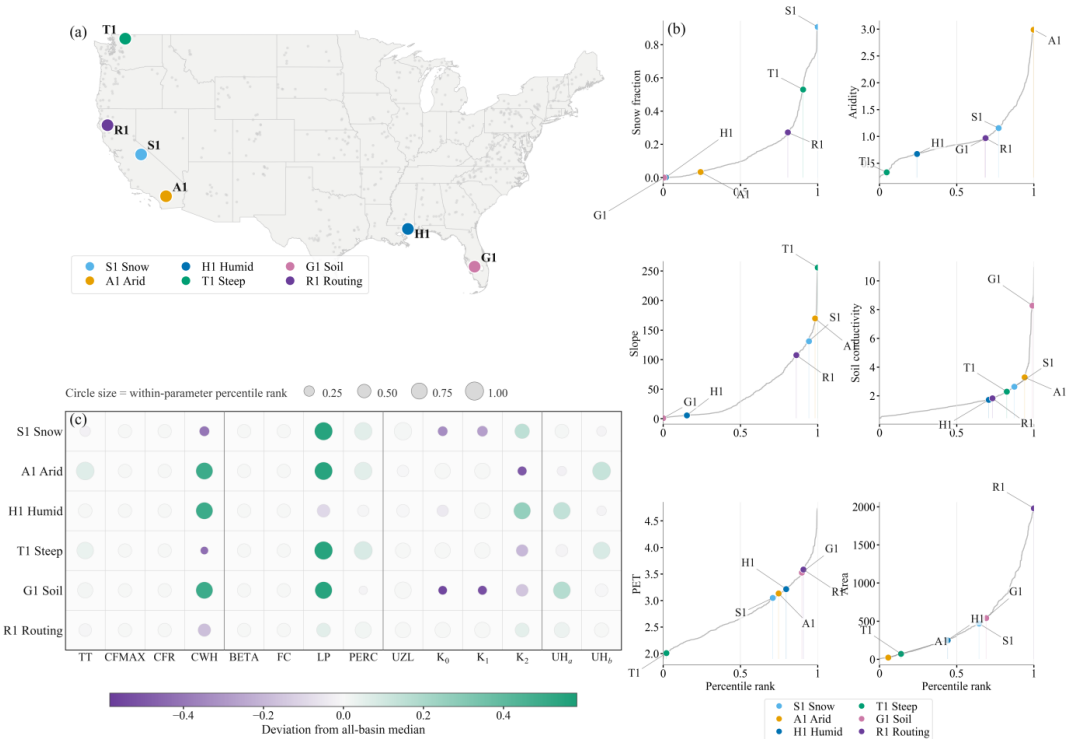


Figure 10. Representative basin cases and parameter regimes. Panel (a) maps the selected snow, arid, humid, steep, soil/storage, and routing-sensitive case basins. Panel (b) reports each case basin's percentile position along six environmental attribute distributions, using case IDs and arrows to separate nearby labels. Panel (c) compares parameter-mean deviations from the all-basin median, with circle size indicating within-parameter percentile rank. The cases show how the large-sample gradients appear in individual basins; they are not used as independent inferential evidence.

4 Discussion

4.1 Learned-relationship reliability as a diagnostic layer for differentiable hybrid models

The object of assessment is the learned descriptor-to-parameter structure, not discharge simulation alone. Traditional parameter transfer uses similarity rules, regressions, or transfer functions between basin descriptors and calibrated parameters (Beck et al., 2016; Parajka et al., 2005; Samaniego et al., 2010). Differentiable parameter learning replaces that post-calibration step with a jointly learned neural mapping (Feng et al., 2022; Tsai et al., 2021). Predictive skill, scalability, spatial generalization, and coherent parameter fields remain useful checks (Feng et al., 2022; Shen et al., 2023; Tsai et al., 2021), but none shows that a specific descriptor relationship can be recovered after training choices change. RELIP supplies that missing diagnostic layer.

Among the tested configurations, the distributional formulation showed the most reproducible relationship structure, not superior discharge skill. Deterministic and MC-dropout runs remain important because they show which controls persist across alternative parameter representations (Clark et al., 2011; Refsgaard et al., 2007). The comparison fixes the backbone, basins, inputs, periods, losses, seeds, and preprocessing, but the formulations are not capacity-matched: the distributional model learns latent location and scale with KL regularization, the deterministic model returns point estimates, and MC-dropout uses stochastic masks. Its greater compactness therefore cannot be attributed to the distributional representation alone; a matched KL ablation would be needed to isolate that effect.



The evidence classes define the strength of interpretation. Robust relationships pass the predictive, perturbation, cross-formulation, and matrix checks. Conditional relationships fail at least one safeguard, while diagnostic relationships are affected by instability, inconsistent controls, boundary effects, or mean-spread coupling. Because loss sensitivity generally exceeded seed sensitivity, relationships that persisted across objectives received greater weight.

The four RELIP levels address distinct failure modes: inadequate prediction, sensitivity to training choices, formulation-specific controls, and instability of the full relationship matrix. No single metric covers all four. RELIP does not establish causality, but it reduces the risk of treating recurrent training artifacts as model structure. A semi-synthetic benchmark with a known input-to-parameter map remains necessary to test recovery of the correct structure.

RELIP also distinguishes relationship reproducibility from parameter-value stability. Bounds, saturation, and parameter compensation may stabilize values while changing their environmental ordering, whereas moderate value variation can coexist with stable relationships. Hydrograph fit alone cannot resolve this distinction.

4.2 Interpretability boundaries and scale dependence of learned relationships

The shared controls provide a plausibility check on RELIP rather than a separate set of hydrologic discoveries. The reproducible patterns connect runoff-generation nonlinearity with slope, storage capacity with PET, percolation with aridity, and upper-zone threshold behavior with soil conductivity. These directions are consistent with the behavioral roles assigned to HBV parameters (Lindström et al., 1997; Seibert & Bergström, 2022). Their plausibility supports the view that RELIP retained structured model behavior instead of recurrent training noise, while the interpretation remains tied to the adopted HBV structure.

The descriptors themselves are correlated summaries of climate, terrain, vegetation, soil, and geology (Addor et al., 2017; Newman et al., 2015). Slope covaries with elevation and seasonality; aridity combines precipitation and evaporative demand; soil conductivity covaries with other soil properties. A dominant descriptor therefore represents a broader environmental gradient, not an isolated cause. Residualized checks retained the four headline directions but weakened some magnitudes (Table S5), which argues against both a single-descriptor explanation and a direct causal reading.

Scale further limits interpretation. A relationship that is coherent across 531 basins can weaken or reverse within a narrower hydroclimatic range. FC-PET, for example, represents a broad evaporative-demand gradient in the full sample but does not retain the same direction in every subgroup. Snow-related evidence differed between means and spreads. CWH and CFR means decreased with snow fraction but were only partially shared across formulations, whereas snow fraction mainly controlled CFMAX and TT spreads. These spread gradients were less confounded but do not indicate dominant controls of the corresponding means.

4.3 Diagnostic value and limits of distributional parameter spread

Distributional learning returns parameter-scale spread in addition to parameter means. We treat that spread as a diagnostic quantity, not as a calibrated posterior or a measure of physical identifiability (Abdar et al., 2021; Hüllermeier & Waegeman, 2021). Interpretation depends on whether a gradient can be separated from its mean and from boundary effects. The predefined screen labels a case less confounded when absolute mean-spread correlation is below 0.5, boundary-distance correlation below 0.4, and near-boundary share below 0.25. Other cases remain descriptive properties of the learned mapping.

The least confounded patterns were snow-related. CFMAX and TT spread decreased as snow fraction increased, which is consistent with stronger seasonal melt signals providing more information about melt-rate and threshold parameters. This remains a diagnostic interpretation, not proof of identifiability. Spread gradients involving aridity, soil conductivity, or vegetation were often coupled to the mean or to boundary distance. In those cases, the spread describes how the network uses its bounded parameter space more directly than it describes hydrologic uncertainty.



4.4 Transferability, implications, and limitations

The assessment problem extends beyond HBV. Land-surface models may learn spatially varying soil, vegetation, or exchange parameters; groundwater and water-quality models may learn hydraulic, storage, reaction, or transport coefficients across spatial units. In each case, good prediction does not show that the learned parameterization is reproducible.

The sequence of evidence transfers more readily than any particular statistic. A model first needs a domain-appropriate adequacy check. Learned relationships are then tested under controlled perturbations, compared across plausible formulations, and assessed as a complete structure. Spearman correlation can be replaced by monotonic association, rank agreement, attribution stability, sensitivity-pattern overlap, or another measure suited to the application. The invariant is the progression from adequacy to perturbation reproducibility, cross-formulation agreement, and structure-level compactness.

Transfer is meaningful only when learned quantities keep a stable semantic role, their admissible domains are defined, the model can be retrained under controlled perturbations, and the evaluation units span enough variation to estimate the relationships. Without these conditions, RELIP may still identify instability, but the evidence classes cannot support domain interpretation.

The present evaluation is not an ungauged-basin transfer test, and its relationships should not be extrapolated unchanged. It covers one HBV structure, one CAMELS-US sample, one descriptor set, and one temporal split. Alternative periods may change the available hydrologic signals and hence the relationship matrix. Pooled top-k selection reduces dependence on a single formulation but does not remove selection effects; the sensitivity analysis only shows that the main ordering is not tied to one value of k. Seeds and losses are perturbed, whereas architectures, forcing products, model structures, and input subsets are not.

Several extensions follow. Leave-region-out or ungauged-basin experiments would connect relationship reproducibility to transfer performance (Hrachowitz et al., 2013). Alternative neural architectures would test implementation sensitivity, especially for LP and routing parameters. A semi-synthetic benchmark with a known descriptor-to-parameter map would test whether RELIP recovers the correct structure, not only a reproducible one. These are separate questions from the present study. Here the evidence supports a narrower conclusion: predictive adequacy is necessary, but interpretation also requires stability across perturbations, agreement across formulations, compact matrix structure, coherent environmental expression, and checks for model-dependent artifacts.

5 Conclusions

RELIP v1.0 evaluates the reproducibility of learned input-to-parameter relationships in differentiable hybrid models. We demonstrated it with three formulations of differentiable HBV parameter learning across 531 CAMELS-US basins. Their streamflow skill was closely comparable, and median NSE or KGE did not identify a preferred formulation. The useful distinction emerged only at the level of learned relationship structure.

Across the tested seeds, losses, and formulations, the distributional model produced the most reproducible structure. It had lower relationship variability, more consistent dominant controls, and greater matrix compactness. This result concerns reproducibility under the stated perturbations. It does not establish that the formulation is more correct, more physically valid, or preferable in every application.

Seven of the fourteen HBV parameters formed a shared dominant-control core. BETA -slope, FC -potential evapotranspiration, PERC-aridity, and UZL-soil conductivity combined perturbation stability, cross-formulation agreement, and coherent environmental gradients. Their hydrologic plausibility supports the diagnostic method rather than constituting a separate geoscientific claim. Snow-related gradients remained partially shared and show why a plausible pattern should remain conditional when it does not pass the complete hierarchy.



Parameter spread added a narrower diagnostic result. CFMAX and TT showed structured snow-fraction gradients with relatively little mean or boundary coupling, whereas several other strong spread relationships were confounded. Spread should therefore be read as information about the learned mapping, not as calibrated posterior uncertainty or direct physical identifiability.

The main implication is methodological. Prediction is a prerequisite for interpretation, not its proof. RELIP adds tests of perturbation reproducibility, cross-formulation agreement, matrix compactness, domain coherence, and parameter-bound effects. Together, these tests provide a model-assessment layer for learned mechanistic parameterizations.

Code and data availability

The exact RELIP v1.0 release used to produce the results reported in this paper is permanently archived on Zenodo at <https://doi.org/10.5281/zenodo.21367213> (Jing et al., 2026). The archived release includes the RELIP source code and diagnostic implementation, configuration files, preprocessing, training, inference, and postprocessing scripts, trained checkpoints required by the released workflow, and machine-readable outputs. RELIP is released under the MIT License, a copy of which is included in the archived release. The current development version is available at <https://github.com/chooron/dmg-research/tree/master/project/parameterize>.

Hydrologic training and inference use dMG v1.4.1 and hydrod12 v1.3.4, which are distributed separately and remain subject to their respective upstream licence terms. The original CAMELS-US meteorological forcing, streamflow observations, and catchment attributes are available from the official NCAR repository at <https://ral.ucar.edu/dataset/camels-catchment-attributes-and-meteorology-large-sample-studies-dataset-downloads> (Addor et al., 2017; Newman et al., 2015). A preprocessed PKL-formatted copy compatible with the dMG workflow is available at <https://mhpi-spatial.s3.us-east-2.amazonaws.com/mhpi-release/data/1-camels.zip>. This copy is provided for convenient access; the preprocessing procedures and scripts required to generate the model-ready data from the original CAMELS-US data are included in the archived RELIP release (Jing et al., 2026).

Author contributions

Xin Jing: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. Na Wei: Data Curation, Writing - Review & Editing. Xue Yang: Supervision, Writing - Review & Editing. Jungang Luo: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing - Review & Editing.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

The authors gratefully acknowledge Dr. Chaopeng Shen and the MHPI development team for making the dMG and hydrod12 infrastructure and CAMELS processing resources available for research use.

Financial support

This work was supported by Science and Technology Program of Shaanxi Provincial Department of Water Resources (No. 2026SLKJ-4), and Doctoral Dissertation Innovation Fund of Xi'an University of Technology (No. BC202610).



555 **References**

- 556 Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A.,
557 Acharya, U. R., Makarekovic, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning:
558 Techniques, applications and challenges. *Information Fusion*, 76, 243–297.
559 <https://doi.org/10.1016/j.inffus.2021.05.008>
- 560 Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and
561 meteorology for large-sample studies. *Hydrology and Earth System Sciences*, 21(10), 5293–5313.
562 <https://doi.org/10.5194/hess-21-5293-2017>
- 563 Beck, H. E., Van Dijk, A. I. J. M., De Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., & Bruijnzeel, L. A. (2016).
564 Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, 52(5), 3599–3622.
565 <https://doi.org/10.1002/2015WR018247>
- 566 Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex
567 environmental systems using the GLUE methodology. *Journal of Hydrology*, 249(1–4), 11–29.
- 568 Clark, M. P., & Kavetski, D. (2010). Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and
569 efficiency of time stepping schemes. *Water Resources Research*, 46(10), 2009WR008894.
570 <https://doi.org/10.1029/2009WR008894>
- 571 Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological
572 modeling. *Water Resources Research*, 47(9). <https://doi.org/10.1029/2010WR009827>
- 573 Feng, D., Beck, H., Lawson, K., & Shen, C. (2023). The suitability of differentiable, physics-informed machine learning
574 hydrologic models for ungauged regions and climate change impact assessment. *Hydrology and Earth System
575 Sciences*, 27(12), 2357–2373. <https://doi.org/10.5194/hess-27-2357-2023>
- 576 Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, Learnable, Regionalized Process-Based Models With
577 Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy. *Water Resources Research*,
578 58(10), e2022WR032404. <https://doi.org/10.1029/2022WR032404>
- 579 Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling:
580 1. Motivation and theoretical development. *Water Resources Research*, 47(11), 2010WR010174.
581 <https://doi.org/10.1029/2010WR010174>
- 582 Gal, Y., & Ghahramani, Z. (2016). *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*
583 (arXiv:1506.02142). arXiv. <https://doi.org/10.48550/arXiv.1506.02142>
- 584 Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE
585 performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1), 80–91.
586 <https://doi.org/10.1016/j.jhydrol.2009.08.003>



- 587 Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T.,
588 Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A.,
589 Pande, S., Tetzlaff, D., ... Cudennec, C. (2013). A decade of Predictions in Ungauged Basins (PUB)—A review.
590 *Hydrological Sciences Journal*, 58(6), 1198–1255. <https://doi.org/10.1080/02626667.2013.803183>
- 591 Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to
592 concepts and methods. *Machine Learning*, 110(3), 457–506. <https://doi.org/10.1007/s10994-021-05946-3>
- 593 Jing, X., Luo, J., Yang, X., & Zuo, G. (2026). Benchmarking structured state space models for differentiable parameter learning:
594 A large-scale evaluation on accuracy and physical consistency. *Expert Systems with Applications*, 329, 133040.
595 <https://doi.org/10.1016/j.eswa.2026.133040>
- 596 Jing, X., Wei, N., Luo, J., & Yang, X. (2026). *RELIP v1.0: Reliability diagnostics for learned input-to-parameter relationships*
597 *in differentiable hybrid models*. <https://doi.org/10.5281/zenodo.21367213>
- 598 Kendall, A., & Gal, Y. (2017). *What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?*
599 (arXiv:1703.04977). arXiv. <https://doi.org/10.48550/arXiv.1703.04977>
- 600 Kingma, D. P., & Welling, M. (2022). *Auto-Encoding Variational Bayes* (arXiv:1312.6114). arXiv.
601 <https://doi.org/10.48550/arXiv.1312.6114>
- 602 Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term
603 Memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11), 6005–6022. [https://doi.org/10.5194/hess-](https://doi.org/10.5194/hess-22-6005-2018)
604 [22-6005-2018](https://doi.org/10.5194/hess-22-6005-2018)
- 605 Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional,
606 and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System*
607 *Sciences*, 23(12), 5089–5110. <https://doi.org/10.5194/hess-23-5089-2019>
- 608 Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed
609 HBV-96 hydrological model. *Journal of Hydrology*, 201(1–4), 272–288. [https://doi.org/10.1016/S0022-](https://doi.org/10.1016/S0022-1694(97)00041-3)
610 [1694\(97\)00041-3](https://doi.org/10.1016/S0022-1694(97)00041-3)
- 611 Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles.
612 *Journal of Hydrology*, 10(3), Article 3. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
- 613 Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J.
614 R., Hopson, T., & Duan, Q. (2015). Development of a large-sample watershed-scale hydrometeorological data set for
615 the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance.
616 *Hydrology and Earth System Sciences*, 19(1), 209–223. <https://doi.org/10.5194/hess-19-209-2015>
- 617 Parajka, J., Merz, R., & Blöschl, G. (2005). A comparison of regionalisation methods for catchment model parameters.
618 *Hydrology and Earth System Sciences*, 9(3), 157–171. <https://doi.org/10.5194/hess-9-157-2005>



- 619 Refsgaard, J. C., Van Der Sluijs, J. P., Højberg, A. L., & Vanrolleghem, P. A. (2007). Uncertainty in the environmental
620 modelling process – A framework and guidance. *Environmental Modelling & Software*, 22(11), 1543–1556.
621 <https://doi.org/10.1016/j.envsoft.2007.02.004>
- 622 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and
623 process understanding for data-driven Earth system science. *Nature*, 566(7743), 195–204.
624 <https://doi.org/10.1038/s41586-019-0912-1>
- 625 Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at
626 the mesoscale. *Water Resources Research*, 46(5). <https://doi.org/10.1029/2008WR007327>
- 627 Seibert, J., & Bergström, S. (2022). A retrospective on hydrological catchment modelling based on half a century with the
628 HBV model. *Hydrology and Earth System Sciences*, 26(5), 1371–1388. <https://doi.org/10.5194/hess-26-1371-2022>
- 629 Shen, C. (2018). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists.
630 *Water Resources Research*, 54(11), 8558–8593. <https://doi.org/10.1029/2018WR022643>
- 631 Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu,
632 X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., ... Lawson,
633 K. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews*
634 *Earth & Environment*, 4(8), 552–567. <https://doi.org/10.1038/s43017-023-00450-9>
- 635 Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDONNELL, J. J.,
636 Mendiondo, E. M., O'Connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., & Zehe, E. (2003).
637 IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological
638 sciences. *Hydrological Sciences Journal*, 48(6), 857–880. <https://doi.org/10.1623/hysj.48.6.857.51421>
- 639 Spearman, C. (1904). “General Intelligence,” objectively determined and measured. *The American Journal of Psychology*,
640 15(2), 201–292.
- 641 Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., & Shen, C. (2021). From calibration to parameter
642 learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, 12(1), 5988.
643 <https://doi.org/10.1038/s41467-021-26107-z>
- 644