

Supplementary Information

This Supplementary Information follows the sequence of the manuscript. Methods, audit tables, and supporting figures are grouped by topic so that each step of the RELIP workflow can be verified in context.

S1. Differentiable HBV formulation and parameter semantics

S1.1 Mathematical formulation of the differentiable HBV 1.1p model

The differentiable HBV model represents snow accumulation and melt, soil-moisture storage, evapotranspiration limitation, percolation, upper- and lower-zone groundwater storage, and routing. Daily inputs are precipitation P , mean air temperature T , and potential evapotranspiration E_p . The principal states are snowpack S_s , meltwater W_m , soil moisture S_m , upper-zone storage S_u , and lower-zone storage S_l ; simulated discharge is Q . The equations below define the benchmark implementation. Its parameters are behavioral quantities within the adopted HBV structure, not direct observations of isolated physical properties.

Precipitation is partitioned into liquid and solid components using the temperature threshold T_t :

$$P_l = P \mathbb{I}(T \geq T_t), \quad P_s = P \mathbb{I}(T < T_t) \quad (S1)$$

where $\mathbb{I}(\cdot)$ is the indicator function. Snowpack storage follows

$$\frac{dS_s}{dt} = P_s - M + RF \quad (S2)$$

with melt and refreezing represented as

$$\begin{aligned} M &= \min\{S_s, \max[0, \text{CFMAX}(T - T_t)]\}, \\ RF &= \min\{W_m, \max[0, \text{CFR CFMAX}(T_t - T)]\} \end{aligned} \quad (S3)$$

Meltwater storage evolves as

$$\frac{dW_m}{dt} = M - RF - I_s \quad (S4)$$

where liquid release from the snowpack and total liquid input to the soil are

$$I_s = \max(0, W_m - \text{CWH } S_s), \quad I = P_l + I_s \quad (S5)$$

Soil wetness and recharge are computed as

$$\omega = \min \left[1, \max \left(0, \left(\frac{S_m}{FC} \right)^{\text{BETA}} \right) \right], \quad R = I\omega \quad (\text{S6})$$

Soil-moisture excess, evapotranspiration limitation, and actual evapotranspiration are

$$EX = \max(0, S_m - FC), \quad f_e = \min \left(1, \frac{S_m}{LP \cdot FC} \right), \quad E_a = \min(S_m, E_p f_e) \quad (\text{S7})$$

The soil moisture balance is

$$\frac{dS_m}{dt} = I - R - EX - E_a + CAP \quad (\text{S8})$$

where capillary rise is represented diagnostically as

$$CAP = \min \left[S_l, C_{\text{cap}} S_l \left(1 - \min \left(1, \frac{S_m}{FC} \right) \right) \right] \quad (\text{S9})$$

The upper-zone storage balance is

$$\frac{dS_u}{dt} = R + EX - \text{PERC} - Q_0 - Q_1 \quad (\text{S10})$$

with percolation, quickflow, and interflow given by

$$Q_{\text{PERC}} = \min(S_u, \text{PERC}), \quad Q_0 = K_0 \max(0, S_u - \text{UZL}), \quad Q_1 = K_1 S_u \quad (\text{S11})$$

The lower-zone storage balance is

$$\frac{dS_l}{dt} = Q_{\text{PERC}} - Q_2 - CAP \quad (\text{S12})$$

with baseflow and total un-routed runoff defined as

$$Q_2 = K_2 S_l, \quad Q = Q_0 + Q_1 + Q_2 \quad (\text{S13})$$

Routing is performed using a gamma unit hydrograph:

$$Q_{\text{routed}} = \text{UH}(\text{UH}_a, \text{UH}_b) * Q \quad (\text{S14})$$

where * denotes convolution, Q_{routed} is routed discharge, and UH_a and UH_b are the unit-hydrograph shape and scale parameters.

S1.2 Parameter definitions, units, and bounds

Table S1 lists the parameter definitions and bounds used by the bounded transformation. Normalization permits comparison across quantities with different units and search ranges; it does not change their interpretation as behavioral HBV parameters.

Table S1. HBV and routing parameter definitions and bounds

Parameter	Process group	Hydrologic role	Unit	Lower bound	Upper bound	Interpretation note
BETA	Soil/runoff generation	Controls the nonlinearity of runoff generation from soil moisture storage	-	1.0	6.0	Behavioral runoff-generation parameter
FC	Soil moisture	Maximum soil water storage capacity	mm	50	1000	Behavioral storage-capacity parameter
LP	Evapotranspiration	Soil-moisture threshold controlling potential evapotranspiration limitation	-	0.2	1.0	Behavioral evapotranspiration-limitation parameter
PERC	Percolation	Maximum percolation from upper to lower groundwater storage	mm d ⁻¹	0	10	Behavioral percolation parameter
UZL	Upper groundwater	Upper-zone threshold for fast runoff response	mm	0	100	Behavioral upper-zone threshold parameter
K ₀	Recession	Fast runoff recession coefficient	d ⁻¹	0.05	0.9	Behavioral fast-recession parameter
K ₁	Recession	Intermediate runoff recession coefficient	d ⁻¹	0.01	0.5	Behavioral recession parameter
K ₂	Recession	Slow/baseflow recession coefficient	d ⁻¹	0.001	0.2	Behavioral slow-recession parameter
TT	Snow	Temperature threshold separating rainfall and snowfall/melt conditions	°C	-2.5	2.5	Behavioral snow-threshold parameter
CFMAX	Snow	Degree-day snowmelt factor	mm °C ⁻¹ d ⁻¹	0.5	10	Behavioral snowmelt parameter
CFR	Snow	Refreezing coefficient	-	0	0.1	Behavioral refreezing parameter
CWH	Snow	Liquid water holding capacity of snowpack	-	0	0.2	Behavioral snowpack-retention parameter
UH _a	Routing	Shape parameter of gamma unit hydrograph	-	0	2.9	Behavioral routing parameter
UH _b	Routing	Scale parameter of gamma unit hydrograph	d	0	6.5	Behavioral routing parameter

Note: Bounds are used for bounded parameter transformation and normalization. All parameters are interpreted as behavioral parameters of the adopted differentiable HBV model, not as direct observations of physical properties.

S2. Dataset, preprocessing, and experimental configuration

S2.1 CAMELS-US benchmark sample and preprocessing chain

The experiments used the established 531-basin CAMELS-US benchmark subset from earlier large-sample studies. The original benchmark excluded catchments with substantial drainage-area discrepancies between geospatial basin fabrics and catchments larger than 2000 km². The fixed gauge list is stored in 531sub_id.txt and was used unchanged for every formulation, loss, seed, diagnostic, and map.

Meteorological inputs came from the Daymet v2 CAMELS extraction distributed for dMG. Daily precipitation and mean air temperature were read from the archive, and potential evapotranspiration was calculated with the Hargreaves method. The source streamflow unit is ft³ s⁻¹; the loader converts it to basin-area-normalized mm d⁻¹, which is also the HBV output unit. No discharge denormalization is required before evaluation.

The dMG-compatible data layout uses camels_daymetv2 for the time series, gage_id.npy for gauge metadata, and 531sub_id.txt for the benchmark subset. Paths are supplied through configuration files rather than embedded as machine-specific locations. These data-loader conventions come from the public dMG/δHBV workflow; the mechanistic equations evaluated here are the HBV1.1p formulation in Sect. S1.

Static attributes were transformed with the archived training normalization statistics, using the same variables and order in all 45 runs. Missing target observations were removed by the validity mask used in the loss and metric calculations. No formulation-specific imputation or basin screening was applied. The public PKL-formatted CAMELS archive used by dMG is available at <https://mhpi-spatial.s3.us-east-2.amazonaws.com/mhpi-release/data/1-camels.zip>; the released configuration records the paths, gauge list, and normalization metadata used in this study.

Training covered 1 January 1989 to 31 December 1998, and testing covered 1 January 1999 to 31 December 2009. A 365-day warm-up preceded evaluation. These dates define the present experiment and differ from the temporal splits used in some earlier studies of the 531-basin benchmark.

S2.2 Dynamic inputs, static attributes, and target variables

Table S2 gives the released input schema. The 35 static descriptors were fixed before the reliability analysis and span climate/hydroclimate, soil, topography/scale, geology, and vegetation/land cover.

Table S2. CAMELS-US dynamic forcings, static catchment attributes, and target variables used in model training and evaluation

Data group	Abbrev. used	Variable/key	Description
Meteorological forcing	Daily T	tmean	2 m daily mean air temperature (°C)
Meteorological forcing	Daily P	prcp	Average daily precipitation (mm/day)
Meteorological forcing	Daily PET	pet	Potential evapotranspiration estimated using the Hargreaves equation
Target variable	Daily Q	streamflow	Observed daily streamflow at USGS gauges; converted to basin-area-normalized depth units for training and evaluation
Catchment attribute	Precip.	p_mean	Mean daily precipitation
Catchment attribute	PET	pet_mean	Mean daily potential evapotranspiration
Catchment attribute	Aridity	aridity	Ratio of mean PET to mean precipitation
Catchment attribute	Precip. seasonality	p_seasonality	Estimated by representing annual precipitation and temperature as sine waves. Positive values indicate summer precipitation peaks, negative values indicate winter precipitation peaks, and values near zero indicate nearly uniform annual precipitation
Catchment attribute	Snow fraction	frac_snow	Fraction of precipitation falling on days with temperature below 0 °C
Catchment attribute	High-precip. freq.	high_prec_freq	Frequency of days with precipitation at least 5 times mean daily precipitation
Catchment attribute	High-precip. dur.	high_prec_dur	Average duration of high-precipitation events, defined as consecutive days with precipitation at least 5 times mean daily precipitation
Catchment attribute	Low-precip. freq.	low_prec_freq	Frequency of dry days with precipitation below 1 mm/day
Catchment attribute	Low-precip. dur.	low_prec_dur	Average duration of dry periods, defined as consecutive days with precipitation below 1 mm/day
Catchment attribute	Elevation	elev_mean	Catchment mean elevation
Catchment attribute	Mean slope	slope_mean	Catchment mean slope
Catchment attribute	Area	area_gages2	Catchment area
Catchment attribute	Forest	frac_forest	Fraction of catchment covered by forest
Catchment attribute	LAI max.	lai_max	Maximum monthly mean leaf area index
Catchment attribute	LAI diff.	lai_diff	Difference between maximum and minimum monthly mean leaf area index
Catchment attribute	GVF max.	gvf_max	Maximum monthly mean green vegetation fraction
Catchment attribute	GVF diff.	gvf_diff	Difference between maximum and minimum monthly mean green vegetation fraction
Catchment attribute	Soil depth	soil_depth_pelletier	Depth to bedrock, capped at 50 m
Catchment attribute	Soil depth	soil_depth_statsgo	Soil depth, capped at 1.5 m
Catchment attribute	Soil porosity	soil_porosity	Volumetric porosity
Catchment attribute	Soil cond.	soil_conductivity	Saturated hydraulic conductivity
Catchment attribute	Max. water content	max_water_content	Maximum water content of the soil
Catchment attribute	Sand frac.	sand_frac	Fraction of sand in the soil
Catchment attribute	Silt frac.	silt_frac	Fraction of silt in the soil
Catchment attribute	Clay	clay_frac	Fraction of clay in the soil
Catchment attribute	Carbonate	carbonate_rocks_frac	Fraction of catchment area characterized as carbonate sedimentary rocks
Catchment attribute	Geol. permeability	geol_permeability	Surface permeability (log10)
Catchment attribute	Dominant land-cover frac.	dom_land_cover_frac	Fraction of the catchment occupied by the dominant land-cover class
Catchment attribute	Dominant land-cover class	dom_land_cover	Categorical identifier of the dominant land-cover class
Catchment attribute	Root depth	root_depth_50	Depth containing 50 % of the root biomass
Catchment attribute	Primary geology class	geol_1st_class	Most common geological class in the catchment
Catchment attribute	Primary geology frac.	glim_1st_class_frac	Fraction of catchment area occupied by the primary geological class
Catchment attribute	Secondary geology class	geol_2nd_class	Second most common geological class in the catchment
Catchment attribute	Secondary geology frac.	glim_2nd_class_frac	Fraction of catchment area occupied by the secondary geological class

Catchment attribute	Geol. porosity	geol_porosity	Catchment-scale geological porosity descriptor
---------------------	----------------	---------------	--

S2.3 Training objectives, implementation settings, and run inventory

The run inventory comprised 45 combinations of three formulations, three hydrologic losses, and five seeds. The formulations were deterministic δ_{base} , MC-dropout δ_{mcd} , and distributional δ_{dist} ; the seeds were 111, 222, 333, 444, and 555. Each processed run contained 14 HBV or routing parameters for all 531 basins.

The training objectives were NSE, logNSE, and an equally weighted hybrid of the two. NSE emphasizes overall hydrograph fit and high flows, whereas logNSE gives more weight to low flows. In RELIP, these objectives serve as structured perturbations of the information used for learning rather than as competing criteria for selecting a single best run.

The loss functions were implemented as batch-wise normalized squared-error criteria. Let Q and $\hat{Q}$ denote observed and simulated streamflow for time step t and basin b , and let Ω denote the set of non-missing basin-time pairs used in a batch. The basin-specific reference standard deviation is

$$\sigma_b = \text{sd}_t(Q_{t,b}) \quad (\text{S15})$$

The batch NSE loss is

$$\mathcal{L}_{\text{NSE}} = \frac{1}{|\Omega|} \sum_{(t,b) \in \Omega} \frac{(\hat{Q}_{t,b} - Q_{t,b})^2}{(\sigma_b + \varepsilon)^2} \quad (\text{S16})$$

The log-transformed NSE loss is

$$\mathcal{L}_{\log\text{NSE}} = \frac{1}{|\Omega|} \sum_{(t,b) \in \Omega} \frac{[\log\{\max(\hat{Q}_{t,b}, 0) + \varepsilon_{\log}\} - \log\{\max(Q_{t,b}, 0) + \varepsilon_{\log}\}]^2}{(\sigma_b^{\log} + \varepsilon)^2} \quad (\text{S17})$$

where the log-transformed basin-specific reference standard deviation is

$$\sigma_b^{\log} = \text{sd}_t[\log\{\max(Q_{t,b}, 0) + \varepsilon_{\log}\}] \quad (\text{S18})$$

The hybrid loss is a weighted sum of the batch NSE and log-transformed NSE losses:

$$\mathcal{L}_{\text{hybrid}} = w_{\text{NSE}} \mathcal{L}_{\text{NSE}} + w_{\log\text{NSE}} \mathcal{L}_{\log\text{NSE}} \quad (\text{S19})$$

where ε is the residual-scaling stabilizer, $\varepsilon_{\log}$ is the log-transform stabilizer, and the implementation used equal default weights $w_{\text{NSE}} = w_{\log\text{NSE}} = 0.5$ for the hybrid loss.

All formulations used two hidden ReLU layers with 128 units each. The deterministic model returned one parameter vector per basin. MC-dropout used a rate of 0.2 and 100 stochastic inference passes. The distributional model returned location and spread and used 100 bounded samples to estimate parameter means and standard deviations.

Training used Adam, an initial learning rate of 0.01, a step scheduler with a 20-epoch step and 0.5 decay, batches of 100 basins, and 100 epochs. The HBV backbone, basin sample, forcing, descriptors, temporal split, and evaluation metrics were fixed. Variation in the relationship diagnostics can therefore be attributed to formulation, objective, and initialization within this design.

Before analysis, parameter outputs were checked for completeness, duplicate logical keys, and consistent basin-parameter coverage. Exact duplicates were collapsed to a single record. This operation had negligible effect: mean cross-seed SD for the distributional relationships changed from 0.03120 to 0.03131, and the main shared-control correlations changed by less than 0.01.

All experiments were conducted under Linux through WSL2 on an NVIDIA GeForce RTX 3080 Ti Laptop GPU with 16 GB memory, a 12th Gen Intel Core i9-12900HX CPU, and 16 GB RAM allocated to WSL2. The implementation used Python 3.10.20, PyTorch 2.9.1, and CUDA 12.8. Across the 15 runs for each formulation, the median wall-clock time for one 100-epoch training run was approximately 5.0 h for δ_{base} , 4.3 h for δ_{mcd} , and 4.9 h for δ_{dist} . The complete 45-run experiment required approximately 212.8 GPU h. Checkpoint-based extraction of 100 parameter samples required an additional 5.6 GPU h, while the complete RELIP diagnostic workflow required approximately 0.1 CPU h. These estimates exclude data download and one-time preprocessing.

S2.4 Full predictive-performance results

Figure S1 shows the complete formulation-by-loss distributions underlying the main-text adequacy summary. The similar performance persists across all three objective functions.

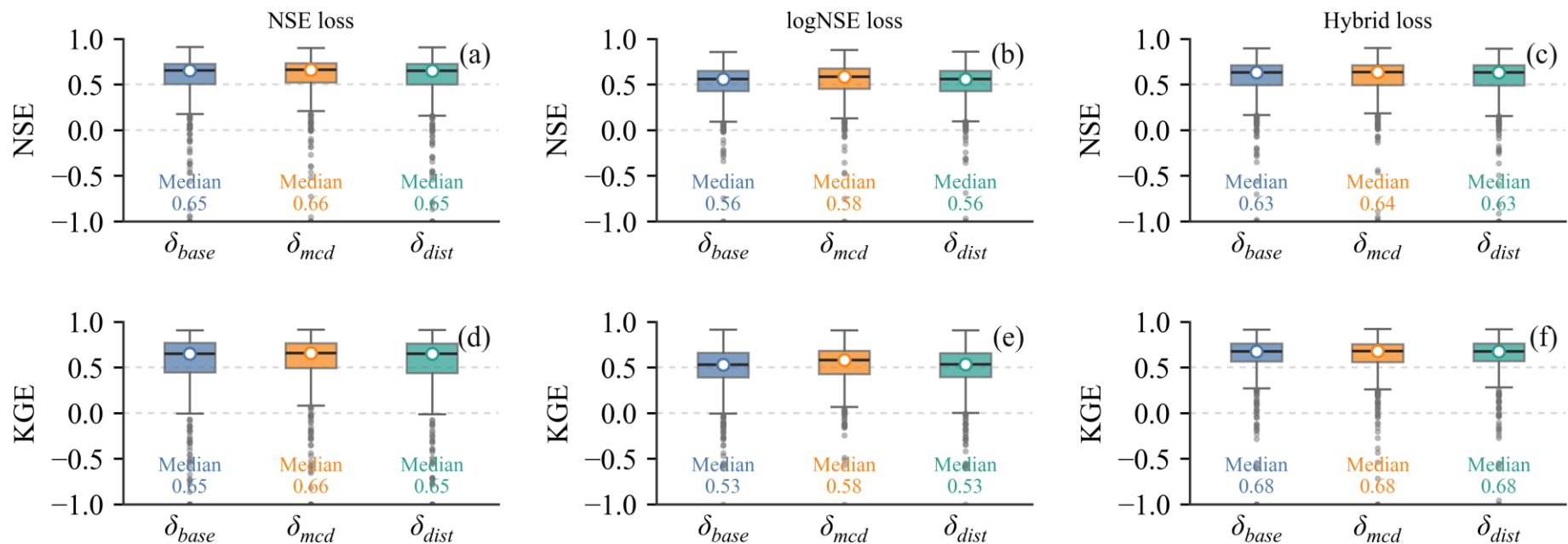


Figure S1. Full predictive-performance distributions across losses and parameter-learning formulations. Panels (a)-(c) show basin-level NSE for the NSE, logNSE, and hybrid losses, respectively; panels (d)-(f) show the corresponding KGE distributions for the same three losses. Within each panel, boxplots compare δ_{base} , δ_{mcd} , and δ_{dist} , with median values printed below the boxes and also marked by open points. The figure extends the main predictive-performance summary by showing the complete formulation-by-loss matrix, confirming that the relationship analyses are based on simulations with broadly comparable predictive adequacy rather than on a single favorable loss setting.

S3. Formal definition and sensitivity of RELIP v1.0

S3.1 Run-level data structure and aggregation order

RELIP operates on run-indexed predictions and parameter fields. The indices f, l, s, b, p, a, and t denote formulation, loss, seed, basin, mechanistic parameter, static attribute, and time step. Each run yields basin-level predictive metrics and a basin-by-parameter table. Attribute-parameter matrices are calculated for each formulation-loss-seed combination and then summarized across seeds, losses, and formulations.

Three aggregation estimands are used for different diagnostic purposes and are not interchangeable. Headline environmental gradients use a reference field: distributional parameter means under HybridNseBatchLoss are averaged across the five seeds within each basin, after which Spearman correlations and rank-based tercile contrasts are computed across 531 basins (Figure 8 and Table S8). Residualized-gradient sensitivity

is calculated separately for each of the 15 distributional loss-seed runs and reported as the arithmetic mean of the run-wise correlations (Table S5). Leave-one-group-out and within-group sensitivity use an all-loss field obtained by averaging all three losses and five seeds within each basin before calculating correlations (Figure S8 and Table S6). These estimands answer different questions and are not expected to be numerically identical.

The three perturbations test different sources of variation. Seeds probe initialization and optimization stochasticity; losses alter the parts of the hydrograph emphasized during training; formulations change the representation of the same input-conditioned parameter field. No axis is used to choose a favorable run before assessment.

S3.2 Four-level diagnostic workflow

RELIP applies four diagnostic levels in sequence. Predictive adequacy is assessed first, followed by pair-level stability across seeds and losses, cross-formulation agreement in dominant controls, and compactness of the complete relationship matrix. The final level prevents a few stable pairs from representing an otherwise unstable structure.

Relationship reproducibility is not the same as parameter-value stability. A bounded parameter may vary little while its descriptor ordering changes across runs; conversely, moderate value variation can coexist with a stable relationship. Parameter and boundary checks are safeguards, while relationship and matrix diagnostics carry the primary interpretive weight.

S3.3 Predefined dominant-control classification rules

Descriptor groups were fixed before the cross-formulation comparison. They are climate/hydroclimate (p_mean, pet_mean, p_seasonality, frac_snow, aridity, high_prec_freq, high_prec_dur, low_prec_freq, and low_prec_dur); soil (root_depth_50, soil_depth_pelletier, soil_depth_statsgo, soil_porosity, soil_conductivity, max_water_content, sand_frac, silt_frac, and clay_frac); topography/scale (elev_mean, slope_mean, and area_gages2); geology (geol_1st_class, glim_1st_class_frac, geol_2nd_class, glim_2nd_class_frac, carbonate_rocks_frac, geol_porosity, and geol_permeability); and vegetation/land cover (frac_forest, lai_max, lai_diff, gvf_max, gvf_diff, dom_land_cover_frac, and dom_land_cover).

Exact-shared controls have the same modal attribute and direction in all three formulations. Group-shared controls retain the same direction and fall within the same predefined descriptor group. Partially shared controls agree in two formulations but differ in the third. A control is model-sensitive when no two formulations agree in both attribute and direction. Ties are resolved by median absolute Spearman correlation across losses

and seeds; unresolved ties remain model-sensitive. The rules use stored summaries and no post-hoc hydrologic judgement. No parameter required the group-shared label, so the reported counts remain seven shared, six partially shared, and one model-sensitive.

S3.4 Parameter-value versus relationship-stability mismatch

Table S3 lists cases in which low parameter variability coexists with unstable descriptor relationships. They show why numerical stability is not used as a proxy for the RELIP target.

Table S3. Parameter-value stability versus relationship-stability mismatch diagnostics

Formulation	Parameter	Parameter group	Raw parameter cross-seed SD	Boundary saturation fraction	Relationship cross-seed SD of Spearman rho	Dominant-control consistency	Mismatch category	Diagnostic note
δ_{base}	CFR	Snow/canopy	0.0000	1.000	0.119	0.333	Stable value / less stable relationship	Low cross-seed parameter spread coincides with near-boundary estimates and weaker dominant-control consistency.
δ_{base}	CWH	Snow/canopy	0.0000	1.000	0.197	0.467	Stable value / less stable relationship	Low cross-seed parameter spread coincides with near-boundary estimates and weaker dominant-control consistency.
δ_{base}	K2	Recession	0.0144	0.277	0.018	0.600	Low variability / boundary saturation	Low cross-seed parameter spread coincides with elevated near-boundary estimates.
δ_{base}	UHa	Routing	0.0177	0.668	0.120	0.667	Stable value / less stable relationship	Low cross-seed parameter spread coincides with near-boundary estimates and weaker dominant-control consistency.
δ_{dist}	K2	Recession	0.0150	0.266	0.015	0.533	Low variability / boundary saturation	Low cross-seed parameter spread coincides with elevated near-boundary estimates.
δ_{mcd}	CFMAX	Snow/canopy	0.0148	0.096	0.049	0.333	Stable value / less stable relationship	Apparent value stability does not imply stable dominant controls.
δ_{mcd}	CFR	Snow/canopy	0.0000	1.000	0.062	0.600	Stable value / less stable relationship	Low cross-seed parameter spread coincides with near-boundary estimates and weaker dominant-control consistency.
δ_{mcd}	CWH	Snow/canopy	0.0000	1.000	0.057	0.467	Stable value / less stable relationship	Low cross-seed parameter spread coincides with near-boundary estimates and weaker dominant-control consistency.
δ_{mcd}	K2	Recession	0.0167	0.235	0.017	0.467	Stable value / less stable relationship	Apparent value stability does not imply stable dominant controls.

S3.5 Boundary-threshold sensitivity

The primary boundary threshold was evaluated at wider cutoffs to determine whether the interpretation depended on a single definition of boundary proximity.

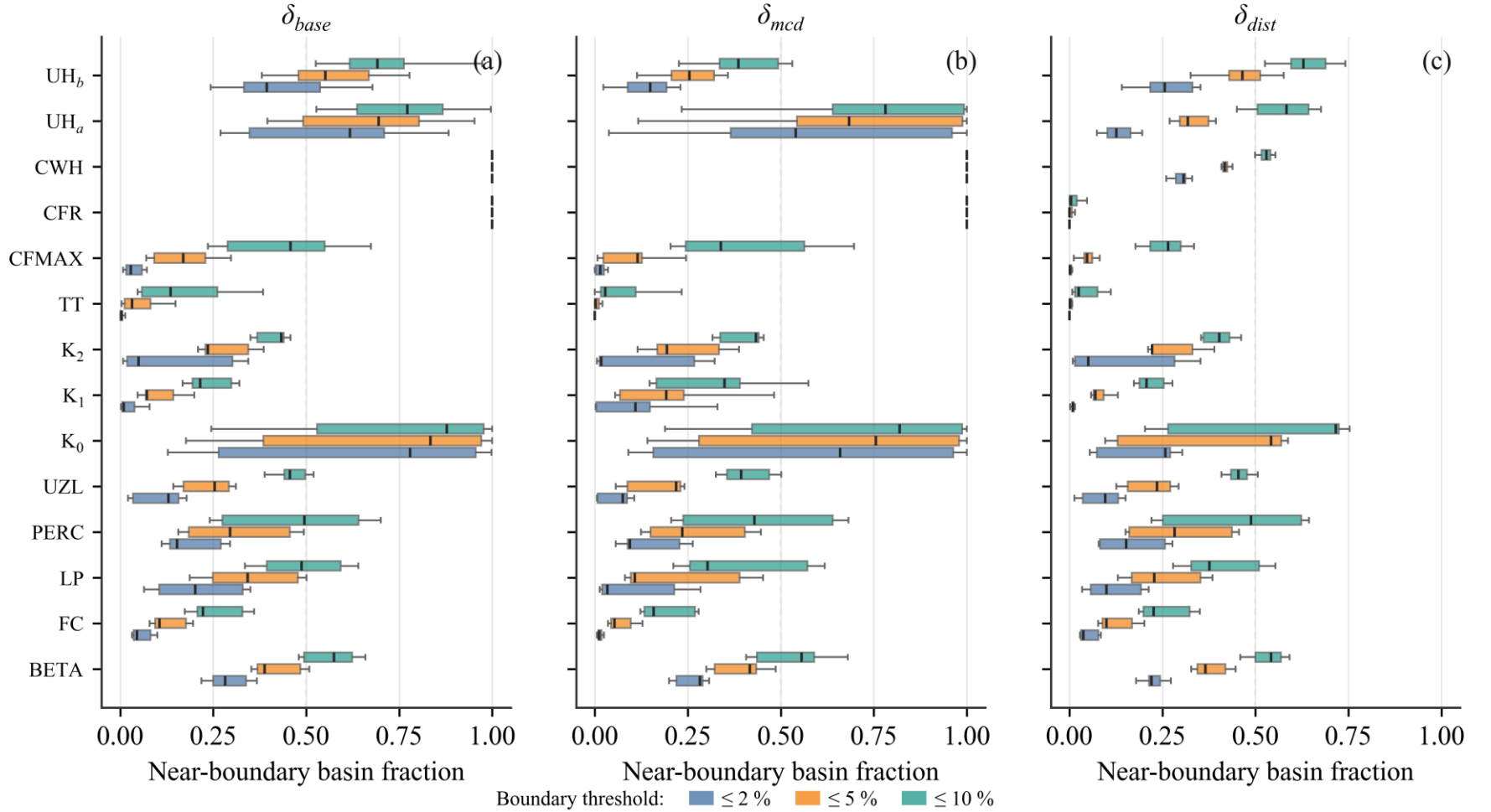


Figure S2. Sensitivity of near-boundary parameter behavior to the threshold used to define boundary proximity. Panel (a) summarizes the deterministic formulation, panel (b) summarizes MC-dropout, and panel (c) summarizes the distributional formulation. In each panel, grouped boxplots show the fraction of basins whose normalized parameter estimates fall within 2 %, 5 %, or 10 % of a parameter boundary, with parameters ordered consistently across panels. The comparison evaluates whether boundary-sensitive parameter exclusions and stability interpretations depend on a single near-boundary threshold.

S3.6 Top-k relationship-selection sensitivity

The number of retained candidate relationships was varied while the pooled selection rule was held fixed. The formulation ordering is considered stable only where it persists across these choices.

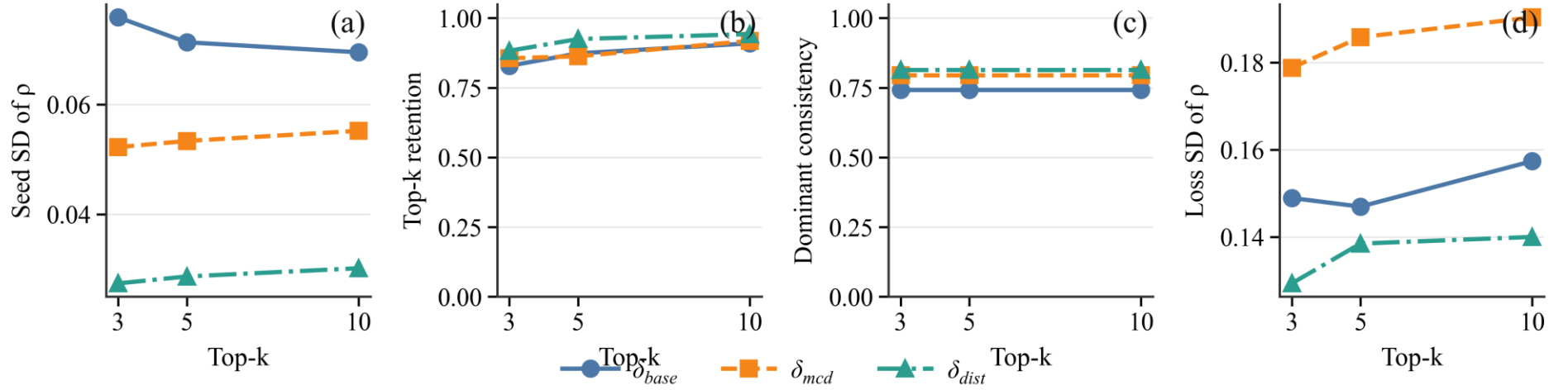


Figure S3. Sensitivity of relationship-stability summaries to the choice of top- k relationships. Panel (a) shows seed standard deviation of Spearman ρ , panel (b) shows top- k retention, panel (c) shows dominant-attribute consistency, and panel (d) shows loss standard deviation of Spearman ρ , each recomputed for $k = 3, 5$, and 10. Lines compare δ_{base} , δ_{mcd} , and δ_{dist} . Stable trends across k indicate that the manuscript conclusions about relationship reproducibility are not driven by a single top- k cutoff.

S3.7 Hydroclimatic-tercile sensitivity

Tercile stratification checks whether seed-level relationship variability is concentrated in a narrow part of the continental attribute space.

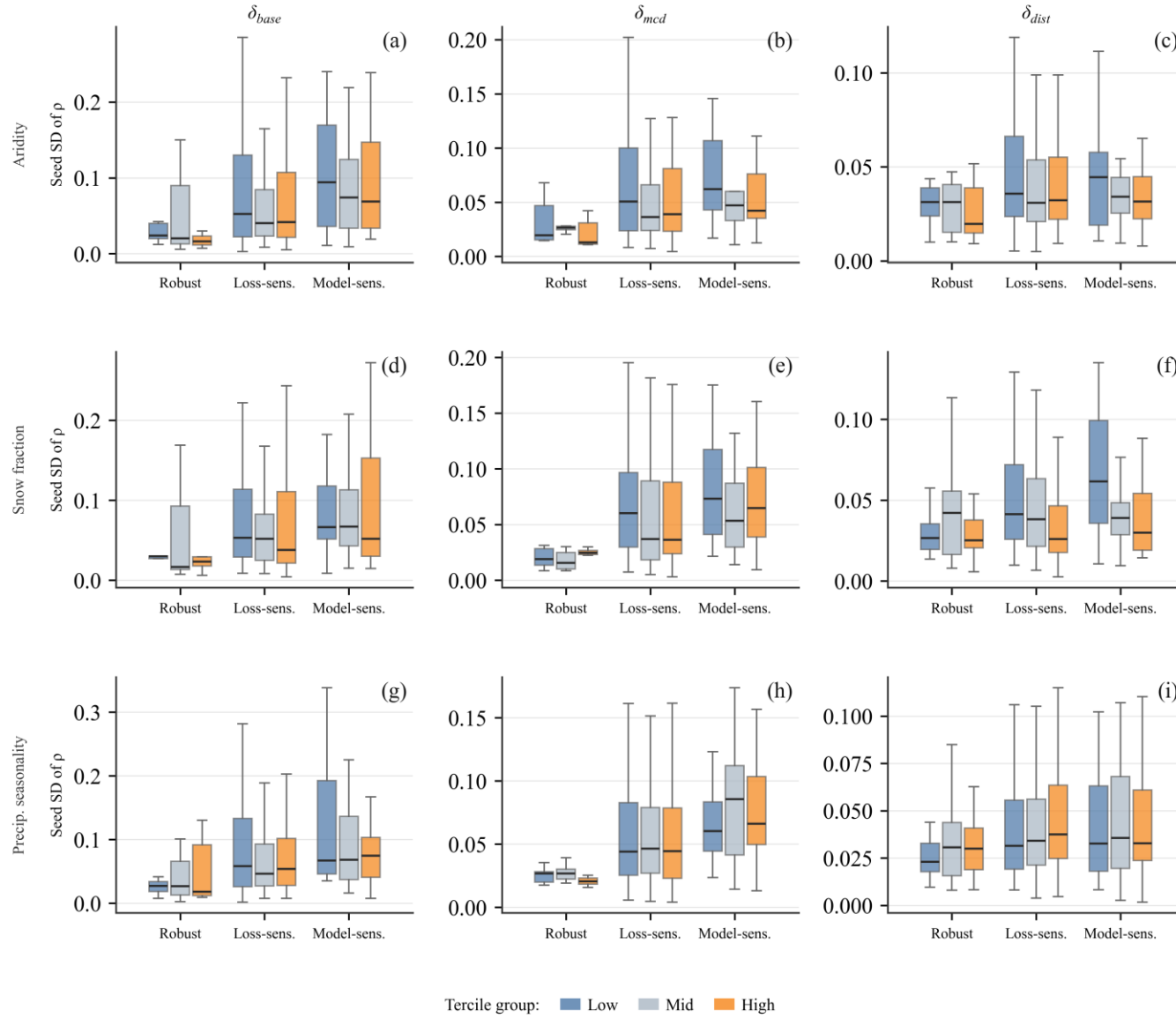


Figure S4. Regional stability of attribute-parameter relationships across hydroclimatic terciles. Panels (a)-(c) stratify by aridity for δ_{base} , δ_{mcd} , and δ_{dist} , respectively; panels (d)-(f) repeat the comparison for snow fraction; panels (g)-(i) repeat it for precipitation seasonality. Within each panel, relationship classes are grouped as robust, loss-sensitive, and model-sensitive, and colors distinguish low, middle, and high tercile groups. The figure tests whether seed-level relationship variability is concentrated in particular hydroclimatic regimes or remains broadly comparable across regional strata.

S4. Cross-formulation relationship structure and robustness

S4.1 Companion parameter-mean relationship matrices

Figure S5 supplies the deterministic and MC-dropout matrices that accompany the distributional matrix in the main text. All three use the same attribute order and visual definitions, so the cross-formulation labels can be checked directly.

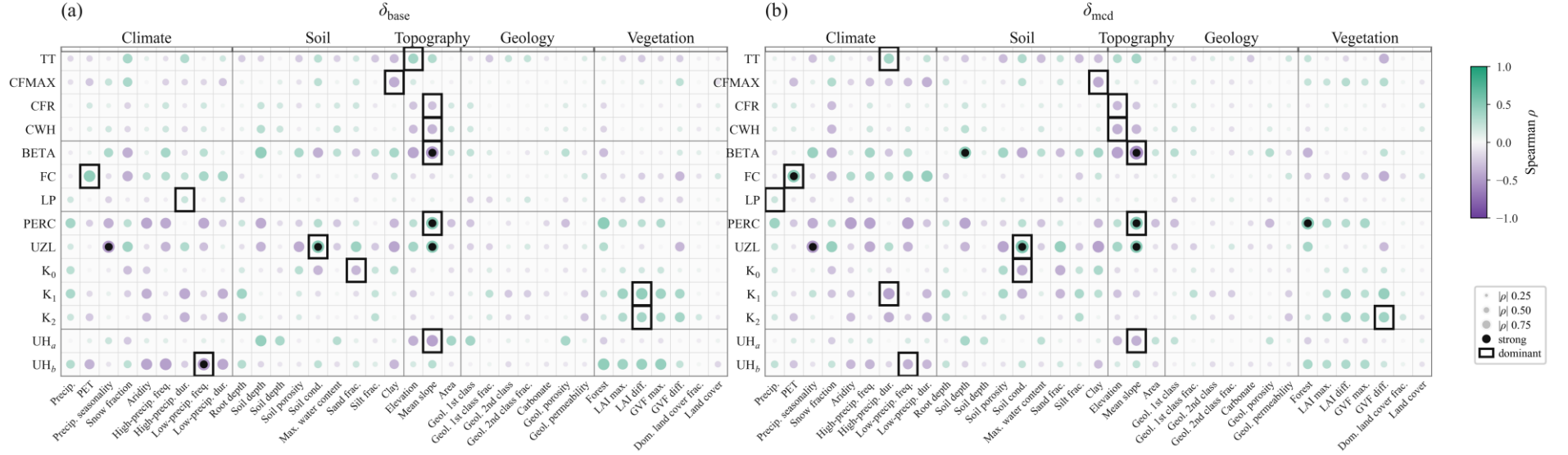


Figure S5. Baseline and MC-dropout parameter-mean relationships with basin attributes. Panel (a) shows the Spearman ρ matrix for δ_{base} , and panel (b) shows the corresponding matrix for δ_{mcd} . Rows are HBV and routing parameters, and columns are CAMELS basin attributes ordered by the same climate, soil, topography, geology, and vegetation groups used in the main distributional heatmap. Circle color gives signed Spearman ρ , circle size gives $|\rho|$, black dots mark strong relationships, and black boxes mark the dominant attribute for each parameter. These companion panels show which mean-parameter controls are shared with, or differ from, the δ_{dist} structure emphasized in the main text.

S4.2 Non-linear attribution check

Integrated Gradients provides a complementary nonlinear assessment of selected outputs. Agreement with the Spearman ranking provides convergent evidence, whereas disagreement identifies relationships requiring closer inspection. Neither method supports causal interpretation.

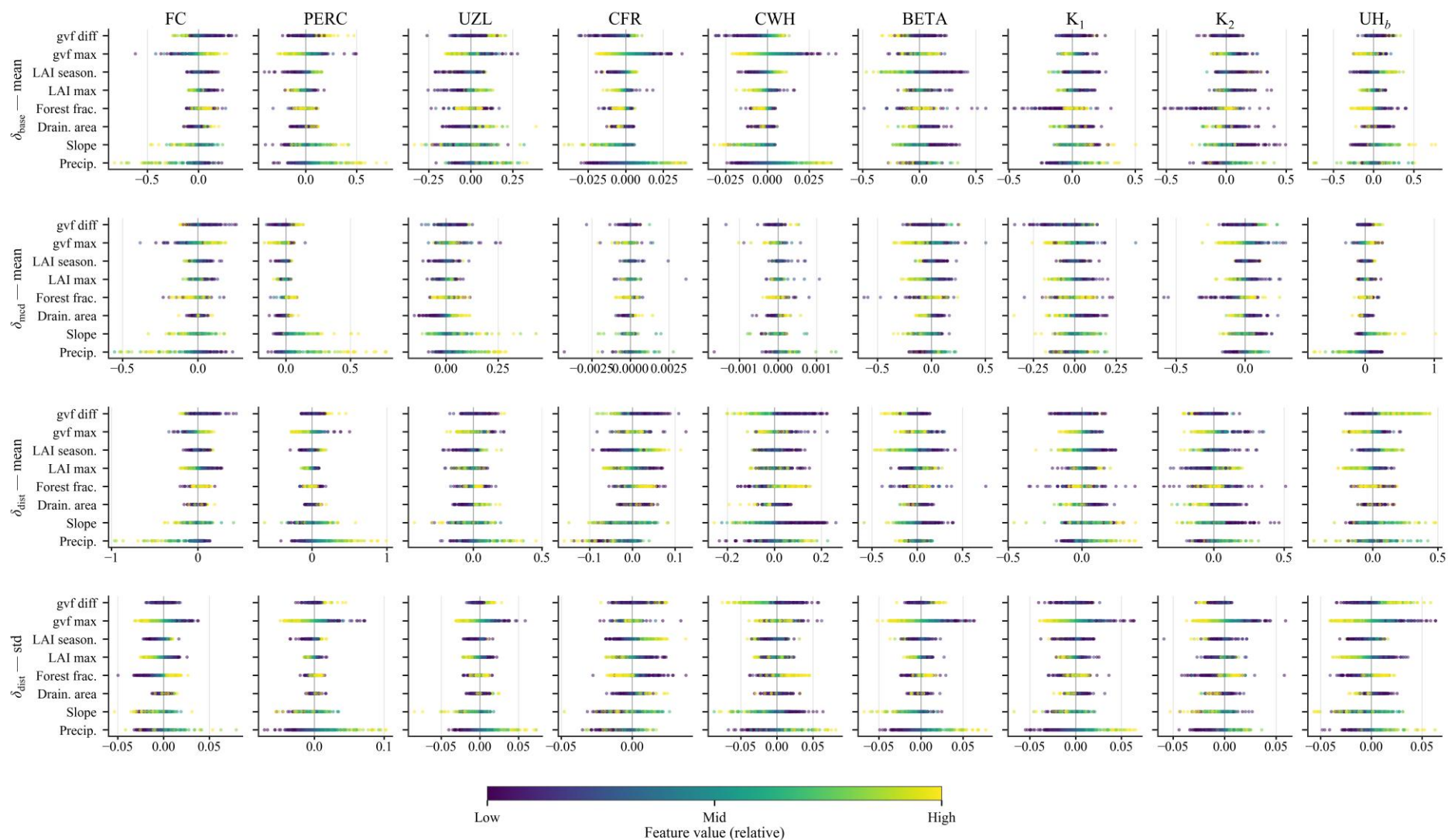
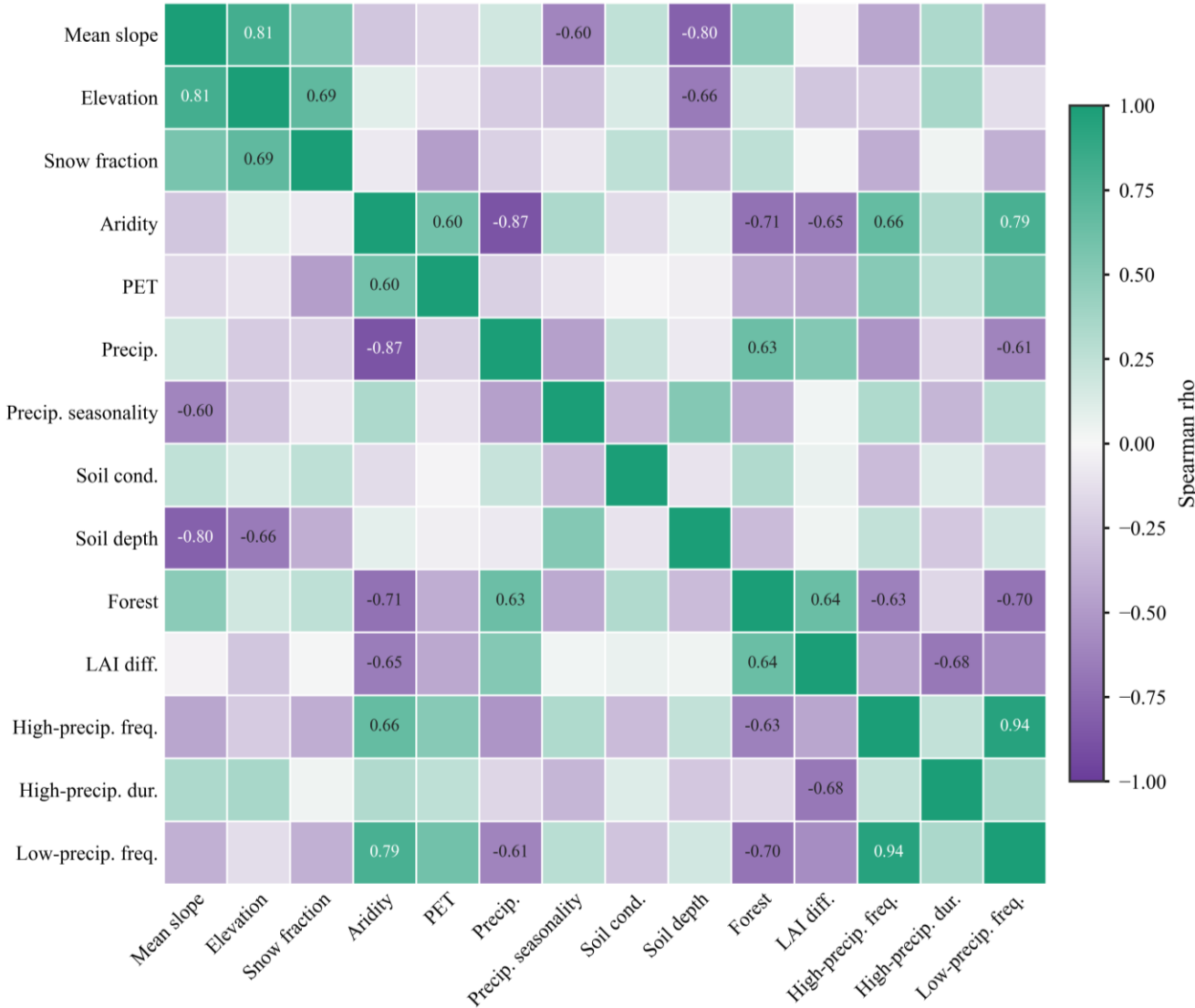


Figure S6. Supplementary Integrated Gradients explainability check. The figure is organized as four output rows by nine focal parameter columns. Row (a) summarizes δ_{base} mean outputs, row (b) summarizes δ_{mcd} mean outputs, row (c) summarizes δ_{dist} mean outputs, and row (d) summarizes δ_{dist} standard-deviation outputs. Columns correspond to FC, PERC, UZL, CFR, CWH, BETA, K_1 , K_2 , and UH_b . Each cell is a beeswarm of signed per-basin Integrated Gradients attributions for the top features of that parameter output, colored by relative feature value. Agreement and disagreement between attribution patterns and Spearman rankings are used as a non-linear diagnostic; neither Integrated Gradients nor Spearman correlation is interpreted as causal evidence.

S4.3 Attribute collinearity

Many CAMELS descriptors summarize overlapping environmental gradients. Figure S7 and Table S4 document this dependence so that a dominant descriptor is read as a representative of correlated conditions, not as an isolated cause.



Strong pairs ($|\rho| \geq 0.8$): High-precip. freq. vs Low-precip. freq.: $\rho = 0.94$; Aridity vs Precip.: $\rho = -0.87$; Mean slope vs Elevation: $\rho = 0.81$

Figure S7. Attribute collinearity among selected basin descriptors. The heatmap shows pairwise Spearman correlations among selected CAMELS attributes used in the parameter-gradient analyses, with a diverging color scale centered at zero. Numeric labels are printed for stronger correlations to highlight descriptor pairs with substantial collinearity, and the note below the heatmap lists pairs exceeding $|\rho| \geq 0.8$. This figure documents the attribute-dependence structure that must be considered when interpreting any single attribute-parameter gradient.

Table S4. Strong collinearity among selected CAMELS basin descriptors

Attribute 1	Attribute 2	Spearman rho	Absolute rho	Attribute group 1	Attribute group 2	Threshold class	Relevance to interpretation
High-precip. freq.	Low-precip. freq.	0.939	0.939	Climate	Climate	$ \rho \geq 0.8$	Relevant to precipitation-frequency gradient interpretation
Aridity	Precip.	-0.867	0.867	Climate	Climate	$ \rho \geq 0.8$	Relevant to PERC-Aridity and FC-PET interpretation
Mean slope	Elevation	0.807	0.807	Topography	Topography	$ \rho \geq 0.8$	Relevant to BETA-slope and snow-related gradient interpretation
Mean slope	Soil depth	-0.798	0.798	Topography	Soil	$0.6 \leq \rho < 0.8$	Relevant to UZL-soil and BETA-slope interpretation
Aridity	Low-precip. freq.	0.789	0.789	Climate	Climate	$0.6 \leq \rho < 0.8$	Relevant to PERC-Aridity interpretation
Aridity	Forest	-0.706	0.706	Climate	Vegetation	$0.6 \leq \rho < 0.8$	Relevant to climate-Vegetation type gradient interpretation
Forest	Low-precip. freq.	-0.701	0.701	Vegetation	Climate	$0.6 \leq \rho < 0.8$	Relevant to climate-Vegetation type gradient interpretation
Elevation	Snow fraction	0.687	0.687	Topography	Climate/snow	$0.6 \leq \rho < 0.8$	Relevant to snow-related gradients
LAI diff.	High-precip. dur.	-0.678	0.678	Vegetation	Climate	$0.6 \leq \rho < 0.8$	Relevant to climate-Vegetation type gradient interpretation
Elevation	Soil depth	-0.664	0.664	Topography	Soil	$0.6 \leq \rho < 0.8$	Relevant to topography-soil gradient interpretation
Aridity	High-precip. freq.	0.661	0.661	Climate	Climate	$0.6 \leq \rho < 0.8$	Relevant to PERC-Aridity interpretation
Aridity	LAI diff.	-0.651	0.651	Climate	Vegetation	$0.6 \leq \rho < 0.8$	Relevant to climate-Vegetation type gradient interpretation
Forest	LAI diff.	0.638	0.638	Vegetation	Vegetation	$0.6 \leq \rho < 0.8$	Relevant to Vegetation type-gradient interpretation
Forest	High-precip. freq.	-0.630	0.630	Vegetation	Climate	$0.6 \leq \rho < 0.8$	Relevant to climate-Vegetation type gradient interpretation
Precip.	Forest	0.629	0.629	Climate	Vegetation	$0.6 \leq \rho < 0.8$	Relevant to climate-Vegetation type gradient interpretation
Precip.	Low-precip. freq.	-0.613	0.613	Climate	Climate	$0.6 \leq \rho < 0.8$	Relevant to precipitation-frequency gradient interpretation
Mean slope	Precip. seasonality	-0.602	0.602	Topography	Climate	$0.6 \leq \rho < 0.8$	Relevant to slope-related interpretation
Aridity	PET	0.600	0.600	Climate	Climate	$0.6 \leq \rho < 0.8$	Relevant to FC-PET and PERC-Aridity interpretation

S4.4 Residualized-gradient sensitivity

Residualization tests whether the four headline directions persist after accounting for strongly correlated descriptors. This analysis constrains interpretation but does not establish causality.

Table S5. Residual sensitivity of the four main shared behavioral gradients. Original and residual Spearman correlations were calculated separately for each of the 15 distributional loss-seed runs; the reported values are arithmetic means of the run-wise correlations.

Parameter	Target attribute	Control attributes	Mean original run-wise Spearman ρ	Mean residual run-wise Spearman ρ	Direction retained?	Absolute change in ρ	Evidence interpretation
BETA	Mean slope	Elevation; Snow fraction	-0.523	-0.335	Yes	0.188	Direction retained after residualizing; magnitude weakened
FC	PET	Aridity; Precip.	0.487	0.281	Yes	0.206	Direction retained after residualizing; magnitude weakened
PERC	Aridity	PET; Precip.	-0.481	-0.221	Yes	0.260	Direction retained but more strongly affected by correlated climate attributes
UZL	Soil cond.	Soil depth; Mean slope; Forest	0.534	0.460	Yes	0.074	Direction and magnitude comparatively retained

S4.5 Leave-one-group-out and within-group sensitivity

Leave-one-group-out estimates test whether one hydroclimatic stratum drives a full-sample relationship. Within-group estimates show how the same continental gradient changes when the environmental range narrows. For this analysis, distributional parameter means were first averaged across all three losses and five seeds within each basin; full-sample, leave-one-group-out, and within-group correlations were then computed from the resulting 531-basin field.

(a)

Leave-one-group-out Spearman ρ

	All	G1	G2	G3	G4	G5	G6	G7
BETA-slope	-0.59	-0.62	-0.63	-0.61	-0.55	-0.56	-0.61	-0.54
FC-PET	0.54	0.48	0.52	0.49	0.56	0.54	0.58	0.55
PERC-aridity	-0.39	-0.29	-0.37	-0.44	-0.38	-0.42	-0.50	-0.30
UZL-soil cond.	0.61	0.65	0.63	0.52	0.56	0.64	0.59	0.64
CWH-snow frac.	-0.85	-0.85	-0.80	-0.85	-0.88	-0.81	-0.87	-0.85
CFR-snow frac.	-0.70	-0.73	-0.66	-0.72	-0.69	-0.69	-0.68	-0.67

(b)

Within-group Spearman ρ

	G1	G2	G3	G4	G5	G6	G7
BETA-slope	-0.41	-0.73	-0.51	-0.67	-0.52	-0.54	-0.57
FC-PET	0.61	0.10	0.42	0.07	0.38	-0.10	0.61
PERC-aridity	-0.51	-0.52	-0.42	-0.20	0.09	-0.39	-0.43
UZL-soil cond.	0.36	0.65	0.77	0.75	0.21	0.56	0.30
CWH-snow frac.	-0.92	-0.89	-0.82	-0.60	-0.75	-0.78	-0.77
CFR-snow frac.	-0.27	-0.49	-0.36	-0.65	-0.55	-0.69	-0.86

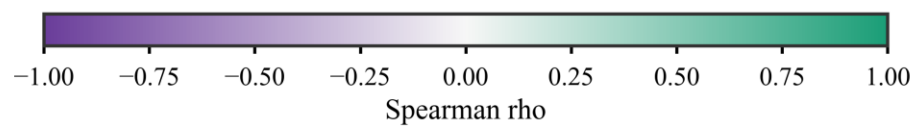
Groups follow Fig. 5 order (G1–G7); All = full-sample ρ .

Figure S8. Hydroclimatic-group sensitivity of selected distributional attribute-parameter relationships. Panel (a) compares the full-sample Spearman ρ with leave-one-group-out estimates for six selected relationships: BETA-slope, FC-PET, PERC-aridity, UZL-soil conductivity, CWH-snow fraction, and CFR-snow fraction. Panel (b) shows the same relationships estimated within each hydroclimatic group. Groups follow the Figure 5 ordering from G_1 to G_7 . The figure evaluates whether headline relationships are dominated by one group or whether their sign and magnitude are broadly retained across hydroclimatic subsets; it is a sensitivity diagnostic, not an ungauged-basin validation experiment. Distributional parameter means were first averaged across all three losses and five seeds within each basin; all displayed correlations were then computed from the resulting 531-basin field.

Table S6. Leave-one-group-out and within-group sensitivity for selected distributional relationships. Distributional parameter means were first averaged across all three losses and five seeds within each basin; all reported correlations were then computed from the resulting 531-basin field.

Relationship	Response type	All-basin Spearman rho	Leave-out group	Spearman rho after exclusion	Delta rho	Max absolute Delta rho across groups	Within-group rho range	Sign change within groups?	Diagnostic note
BETA-Mean slope	Mean	-0.587	G1	-0.615	-0.028	0.049	-0.731 to -0.409	No	Leave-out estimates are stable; within-group signs are retained.
BETA-Mean slope	Mean	-0.587	G2	-0.627	-0.040	0.049	-0.731 to -0.409	No	Leave-out estimates are stable; within-group signs are retained.
BETA-Mean slope	Mean	-0.587	G3	-0.606	-0.019	0.049	-0.731 to -0.409	No	Leave-out estimates are stable; within-group signs are retained.
BETA-Mean slope	Mean	-0.587	G4	-0.552	0.035	0.049	-0.731 to -0.409	No	Leave-out estimates are stable; within-group signs are retained.
BETA-Mean slope	Mean	-0.587	G5	-0.561	0.026	0.049	-0.731 to -0.409	No	Leave-out estimates are stable; within-group signs are retained.
BETA-Mean slope	Mean	-0.587	G6	-0.611	-0.024	0.049	-0.731 to -0.409	No	Leave-out estimates are stable; within-group signs are retained.
BETA-Mean slope	Mean	-0.587	G7	-0.538	0.049	0.049	-0.731 to -0.409	No	Leave-out estimates are stable; within-group signs are retained.
FC-PET	Mean	0.540	G1	0.482	-0.058	0.058	-0.096 to 0.615	Yes	Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
FC-PET	Mean	0.540	G2	0.521	-0.019	0.058	-0.096 to 0.615	Yes	Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
FC-PET	Mean	0.540	G3	0.492	-0.048	0.058	-0.096 to 0.615	Yes	Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
FC-PET	Mean	0.540	G4	0.561	0.021	0.058	-0.096 to 0.615	Yes	Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
FC-PET	Mean	0.540	G5	0.539	-0.001	0.058	-0.096 to 0.615	Yes	Leave-out estimates remain same-signed, but at least one within-group

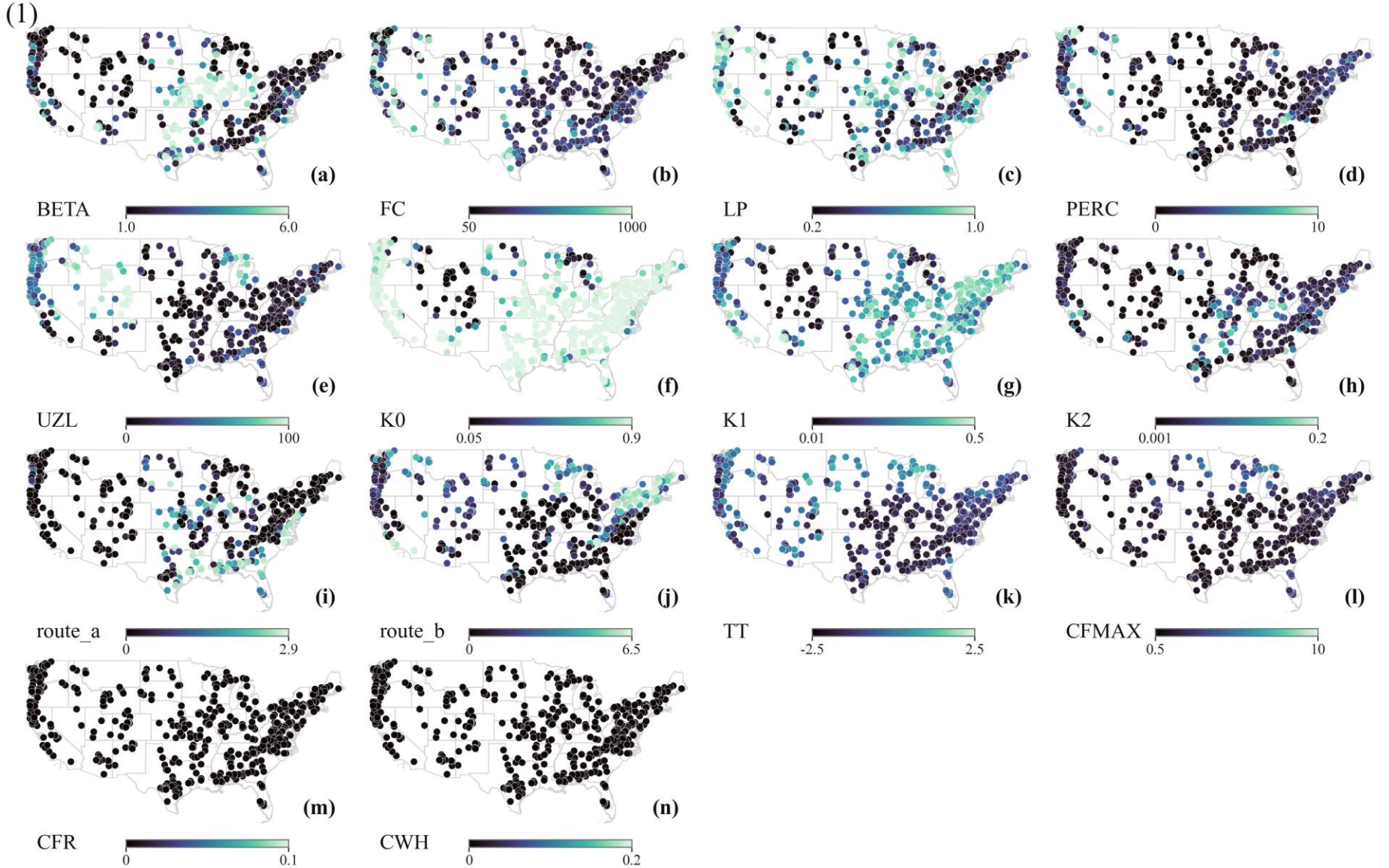
FC-PET	Mean	0.540	G6	0.584	0.044	0.058	-0.096 to 0.615	Yes	estimate changes sign or weakens strongly. Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
FC-PET	Mean	0.540	G7	0.546	0.006	0.058	-0.096 to 0.615	Yes	Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
PERC-Aridity	Mean	-0.390	G1	-0.291	0.099	0.114	-0.525 to 0.088	Yes	Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
PERC-Aridity	Mean	-0.390	G2	-0.369	0.021	0.114	-0.525 to 0.088	Yes	Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
PERC-Aridity	Mean	-0.390	G3	-0.437	-0.048	0.114	-0.525 to 0.088	Yes	Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
PERC-Aridity	Mean	-0.390	G4	-0.384	0.006	0.114	-0.525 to 0.088	Yes	Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
PERC-Aridity	Mean	-0.390	G5	-0.423	-0.034	0.114	-0.525 to 0.088	Yes	Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
PERC-Aridity	Mean	-0.390	G6	-0.503	-0.114	0.114	-0.525 to 0.088	Yes	Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
PERC-Aridity	Mean	-0.390	G7	-0.302	0.088	0.114	-0.525 to 0.088	Yes	Leave-out estimates remain same-signed, but at least one within-group estimate changes sign or weakens strongly.
UZL-Soil cond.	Mean	0.609	G1	0.652	0.043	0.086	0.213 to 0.768	No	Leave-out estimates are same-signed, with moderate within-group magnitude variation.
UZL-Soil cond.	Mean	0.609	G2	0.631	0.022	0.086	0.213 to 0.768	No	Leave-out estimates are same-signed, with moderate within-group magnitude variation.
UZL-Soil cond.	Mean	0.609	G3	0.522	-0.086	0.086	0.213 to 0.768	No	Leave-out estimates are same-signed, with moderate within-group magnitude variation.

UZL-Soil cond.	Mean	0.609	G4	0.561	-0.048	0.086	0.213 to 0.768	No	Leave-out estimates are same-signed, with moderate within-group magnitude variation.
UZL-Soil cond.	Mean	0.609	G5	0.639	0.030	0.086	0.213 to 0.768	No	Leave-out estimates are same-signed, with moderate within-group magnitude variation.
UZL-Soil cond.	Mean	0.609	G6	0.594	-0.014	0.086	0.213 to 0.768	No	Leave-out estimates are same-signed, with moderate within-group magnitude variation.
UZL-Soil cond.	Mean	0.609	G7	0.644	0.036	0.086	0.213 to 0.768	No	Leave-out estimates are same-signed, with moderate within-group magnitude variation.
CWH-Snow fraction	Mean	-0.847	G1	-0.851	-0.004	0.042	-0.915 to -0.604	No	Leave-out estimates are stable; within-group signs are retained.
CWH-Snow fraction	Mean	-0.847	G2	-0.805	0.042	0.042	-0.915 to -0.604	No	Leave-out estimates are stable; within-group signs are retained.
CWH-Snow fraction	Mean	-0.847	G3	-0.847	0.000	0.042	-0.915 to -0.604	No	Leave-out estimates are stable; within-group signs are retained.
CWH-Snow fraction	Mean	-0.847	G4	-0.882	-0.035	0.042	-0.915 to -0.604	No	Leave-out estimates are stable; within-group signs are retained.
CWH-Snow fraction	Mean	-0.847	G5	-0.814	0.034	0.042	-0.915 to -0.604	No	Leave-out estimates are stable; within-group signs are retained.
CWH-Snow fraction	Mean	-0.847	G6	-0.867	-0.020	0.042	-0.915 to -0.604	No	Leave-out estimates are stable; within-group signs are retained.
CWH-Snow fraction	Mean	-0.847	G7	-0.845	0.002	0.042	-0.915 to -0.604	No	Leave-out estimates are stable; within-group signs are retained.
CFR-Snow fraction	Mean	-0.697	G1	-0.728	-0.031	0.038	-0.860 to -0.274	No	Leave-out estimates are stable; within-group signs are retained.
CFR-Snow fraction	Mean	-0.697	G2	-0.659	0.038	0.038	-0.860 to -0.274	No	Leave-out estimates are stable; within-group signs are retained.
CFR-Snow fraction	Mean	-0.697	G3	-0.717	-0.020	0.038	-0.860 to -0.274	No	Leave-out estimates are stable; within-group signs are retained.
CFR-Snow fraction	Mean	-0.697	G4	-0.689	0.008	0.038	-0.860 to -0.274	No	Leave-out estimates are stable; within-group signs are retained.
CFR-Snow fraction	Mean	-0.697	G5	-0.694	0.003	0.038	-0.860 to -0.274	No	Leave-out estimates are stable; within-group signs are retained.
CFR-Snow fraction	Mean	-0.697	G6	-0.683	0.014	0.038	-0.860 to -0.274	No	Leave-out estimates are stable; within-group signs are retained.
CFR-Snow fraction	Mean	-0.697	G7	-0.675	0.022	0.038	-0.860 to -0.274	No	Leave-out estimates are stable; within-group signs are retained.

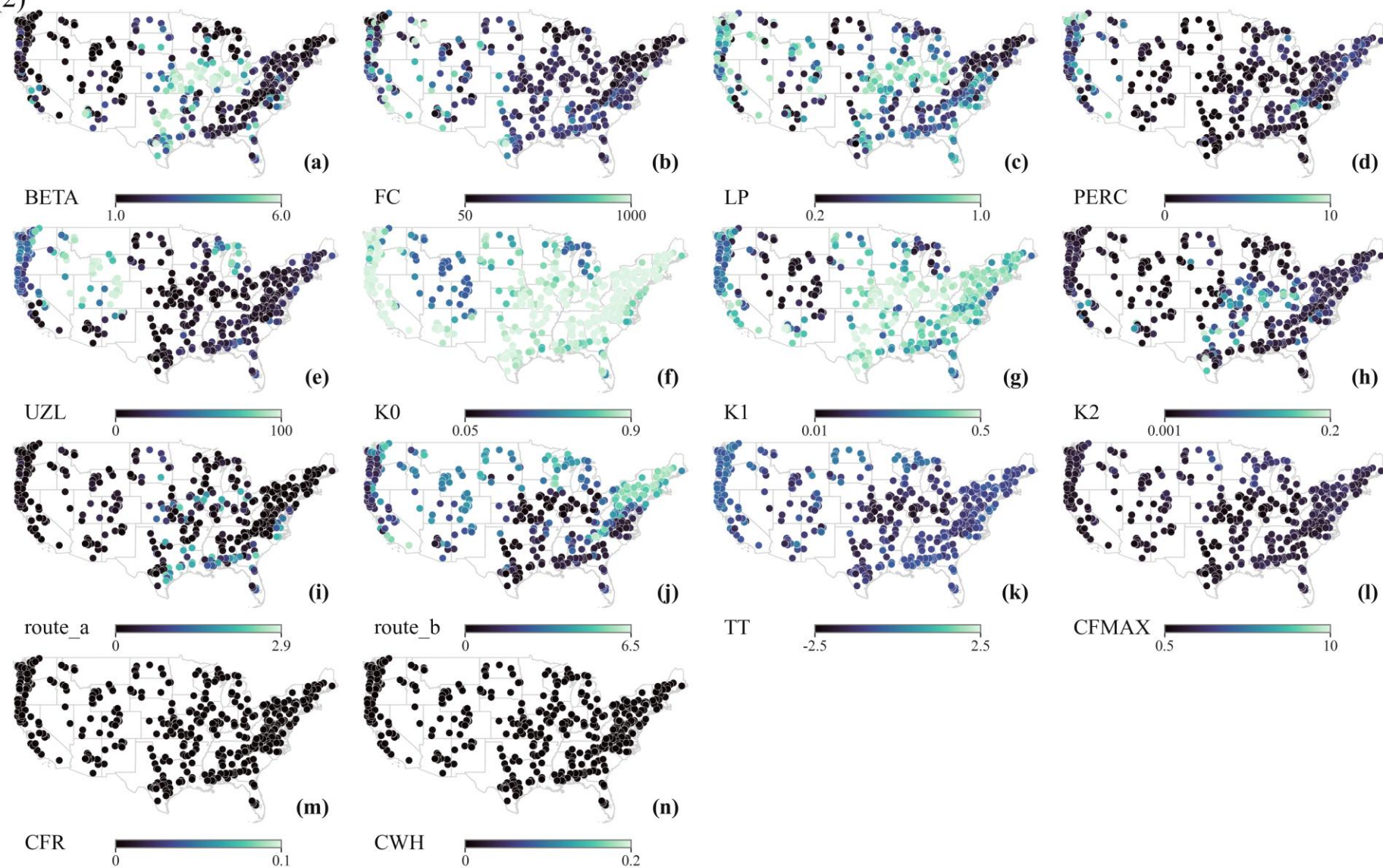
S5. Spatial organization and extended parameter-mean evidence

S5.1 Full spatial parameter summaries

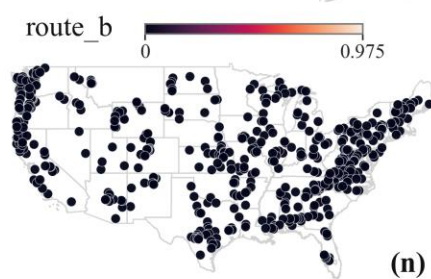
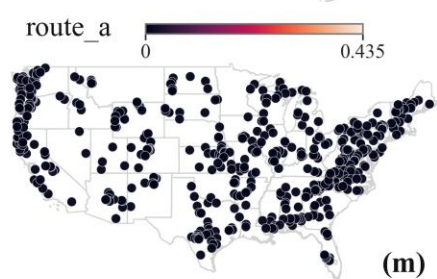
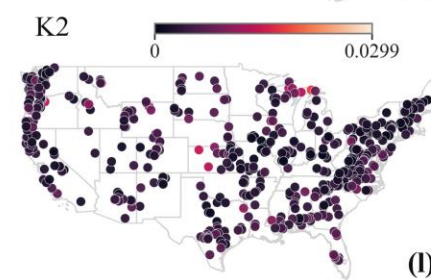
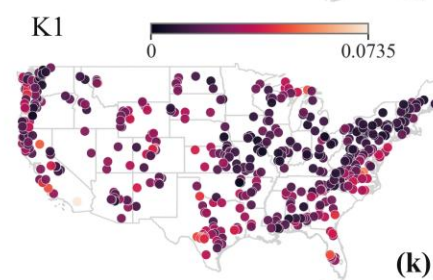
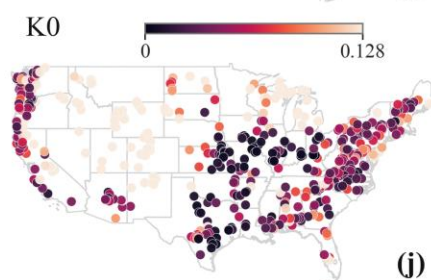
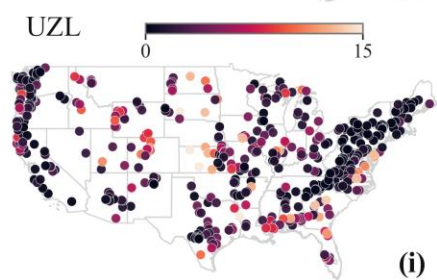
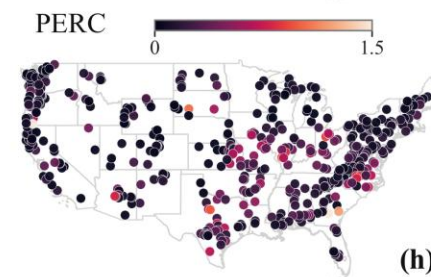
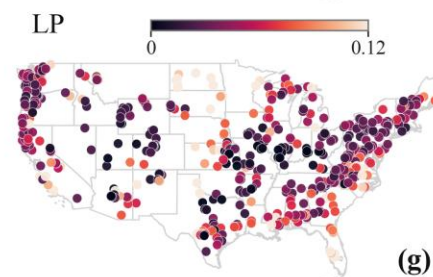
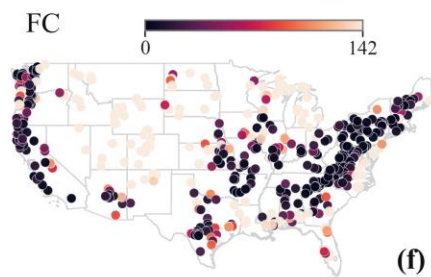
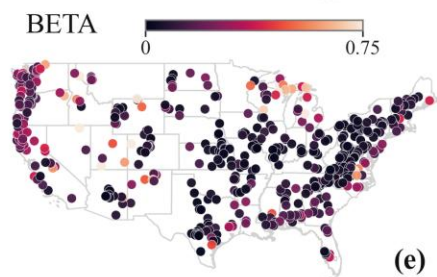
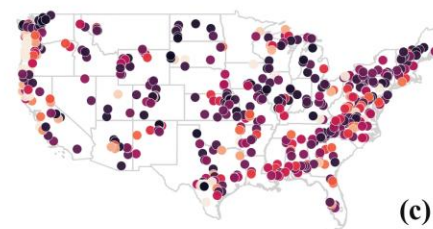
Figure S9 extends the spatial check to all 14 parameters, the companion formulations, and their spread summaries. Each parameter has its own scale, so the maps show within-parameter organization rather than differences in magnitude across parameters.



(2)



(3)



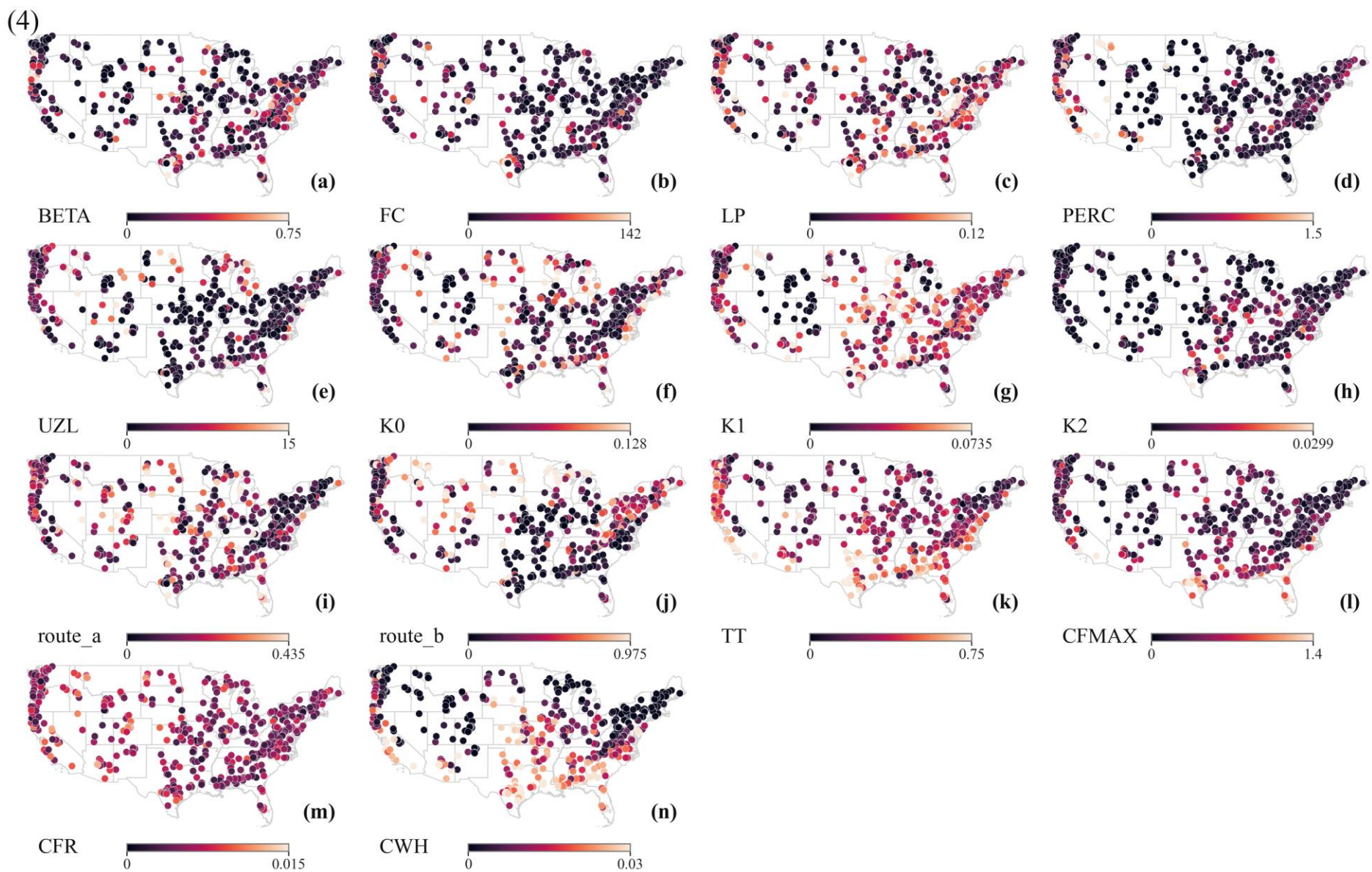


Figure S9. Spatial maps of companion learned HBV and routing parameter summaries for the deterministic, MC-dropout, and distributional formulations. Panel (1) maps δ_{base} seed-averaged parameter means, panel (2) maps δ_{mcd} seed-averaged parameter means, panel (3) maps δ_{mcd} parameter spread, and panel (4) maps δ_{dist} parameter-scale standard deviation. Within each multi-map panel, internal maps correspond to BETA, FC, LP, PERC, UZL, K_0 , K_1 , K_2 , TT, CFMAX, CFR, CWH, UH_a , and UH_b , with the final two grid cells left blank. Each parameter uses its own color scale, so colors should be read within panels rather than compared directly across different parameters.

S5.2 Correlated descriptor context

The pairwise panels situate the strongest parameter gradients within their correlated descriptor context and support the collinearity and residualization checks in Sect. S4.

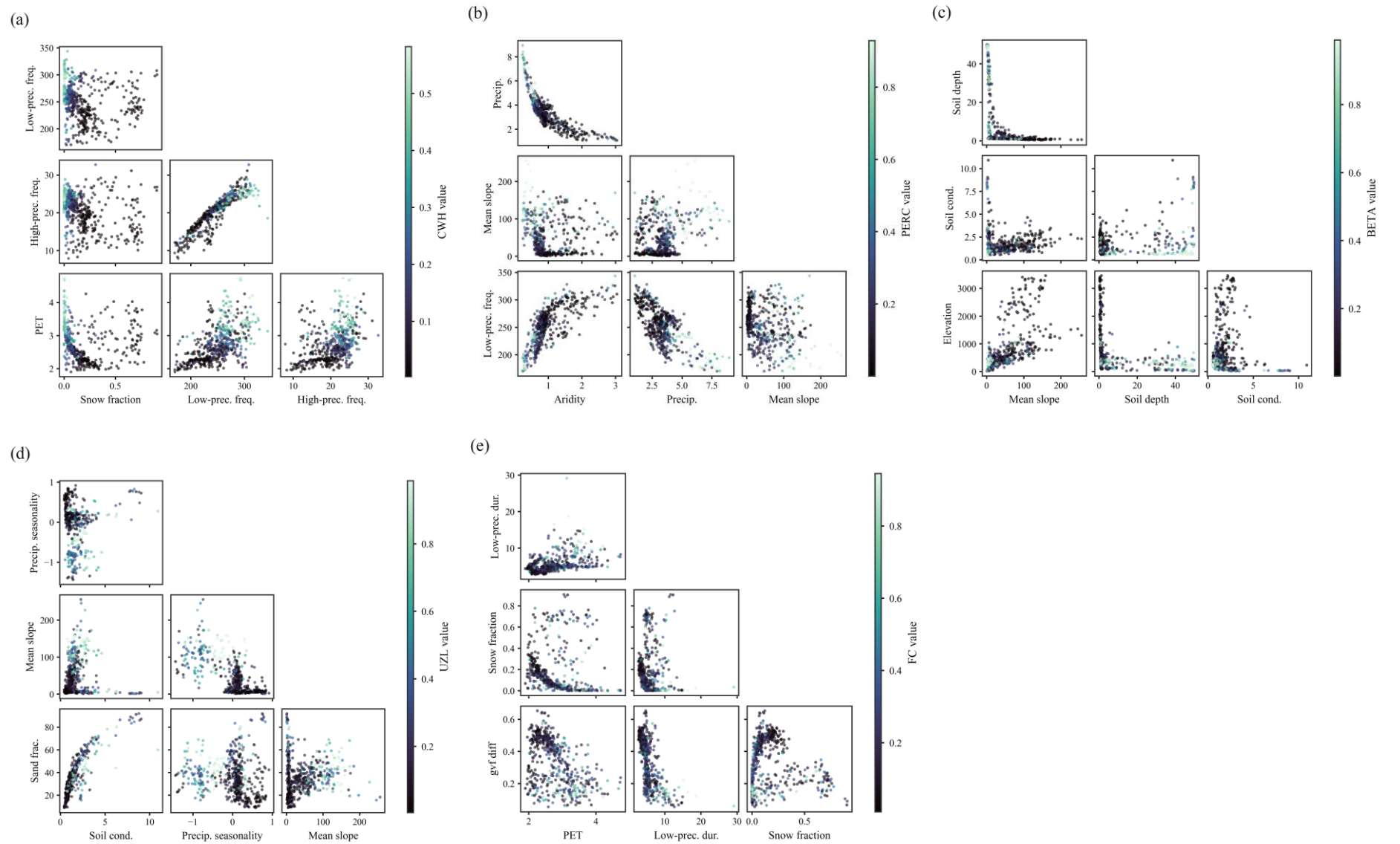
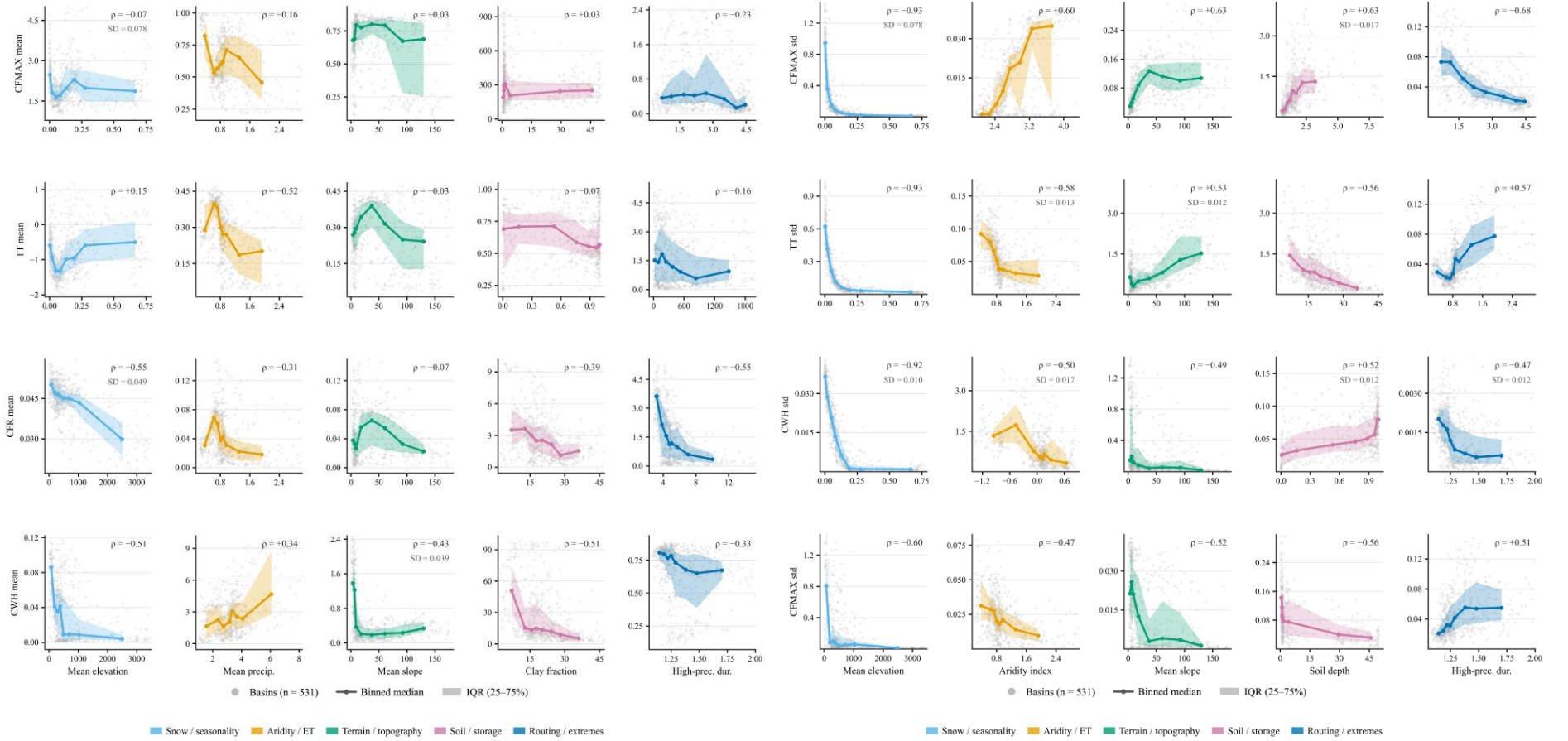


Figure S10. Pairwise attribute scatter matrices for the strongest distributional parameter-attribute relationships. Panels (a)-(e) show CWH with snow fraction, PERC with aridity, BETA with mean slope, UZL with soil conductivity, and FC with potential evapotranspiration, respectively. Each subfigure is a lower-triangle matrix of the four attributes most correlated with the focal parameter, with points colored by the normalized distributional parameter value. The panels assess whether each headline parameter gradient is embedded in a broader correlated attribute structure rather than being driven by one isolated descriptor.

S5.3 Extended environmental-gradient evidence

Figure S11 extends the selected main-text gradients. Panel (a) contains additional parameter-mean relationships; panel (b) contains the corresponding spread relationships discussed in Sect. S6. These panels broaden the descriptive evidence but do not define an additional evidence level.



(a)

(b)

Figure S11. Extended environmental-gradient evidence for distributional parameter means and standard deviations. Panel (a) shows 20 additional or alternative parameter-mean gradients, and panel (b) shows 20 corresponding parameter-standard-deviation gradients. Within each panel, columns group relationships by snow and seasonality, aridity and evapotranspiration, terrain and topography, soil and storage, and routing or extremes. Grey points are individual basins, colored curves are binned medians, shaded bands show interquartile ranges, and annotations report Spearman ρ and seed standard deviation where available. Together, the two panels provide the broader evidence base behind the selected main-text gradient examples.

S6. Parameter-spread diagnostics

S6.1 Boundary and diagnostic-spread definitions

Several HBV and routing parameters lie near their prescribed bounds, which can make a parameter appear stable simply because little feasible range remains. Boundary and spread checks separate this numerical effect from a reproducible environmental ordering.

Boundary saturation was evaluated using near-boundary thresholds on the normalized parameter scale. The primary threshold classified parameter values as near-boundary when the normalized value was no greater than 0.02 or no less than 0.98, with 0.05/0.95 and 0.10/0.90 thresholds used for sensitivity checks. Distance to the nearest boundary was defined as

$$d_{b,j} = \min(\tilde{\theta}_{b,j}, 1 - \tilde{\theta}_{b,j}) \quad (\text{S20})$$

Mean-standard deviation coupling and boundary-distance coupling were computed as

$$c_j^\mu = \text{corr}_S(\tilde{\theta}_{b,j}, s_{b,j}), \quad c_j^d = \text{corr}_S(d_{b,j}, s_{b,j}) \quad (\text{S21})$$

where $s_{b,j}$ is the normalized stochastic parameter standard deviation and corr_S denotes Spearman correlation across basins. Parameters were flagged as mean-coupled or boundary-sensitive using

$$|c_j^\mu| \geq 0.5, \quad |c_j^d| \geq 0.4 \text{ or } p_{\text{boundary},j} \geq 0.25 \quad (\text{S22})$$

where $p_{\text{boundary},j}$ is the near-boundary fraction for parameter j . Low variability is not treated as strong evidence when boundary saturation is high, because the apparent stability may reflect boundary locking.

Mean-spread coupling tests whether predicted spread follows the parameter mean. Boundary-distance coupling tests whether spread is governed by proximity to the bounds. Strong coupling does not invalidate a gradient, but it narrows the claim: the pattern describes the learned mapping rather than direct physical identifiability.

Distributional spread is not a calibrated Bayesian posterior. It depends on the neural parameterizer, the bounded transform, the hydrologic loss, the HBV structure, and the information in attributes and discharge. The screen identifies gradients that are less affected by mean coupling and boundary proximity. CFMAX and TT had the fewest diagnostic concerns; several other large gradients did not.

S6.2 MC-dropout spread structure

MC-dropout spread arises from stochastic masks. It is retained as a companion diagnostic and is not treated as equivalent to distributional parameter-scale spread.

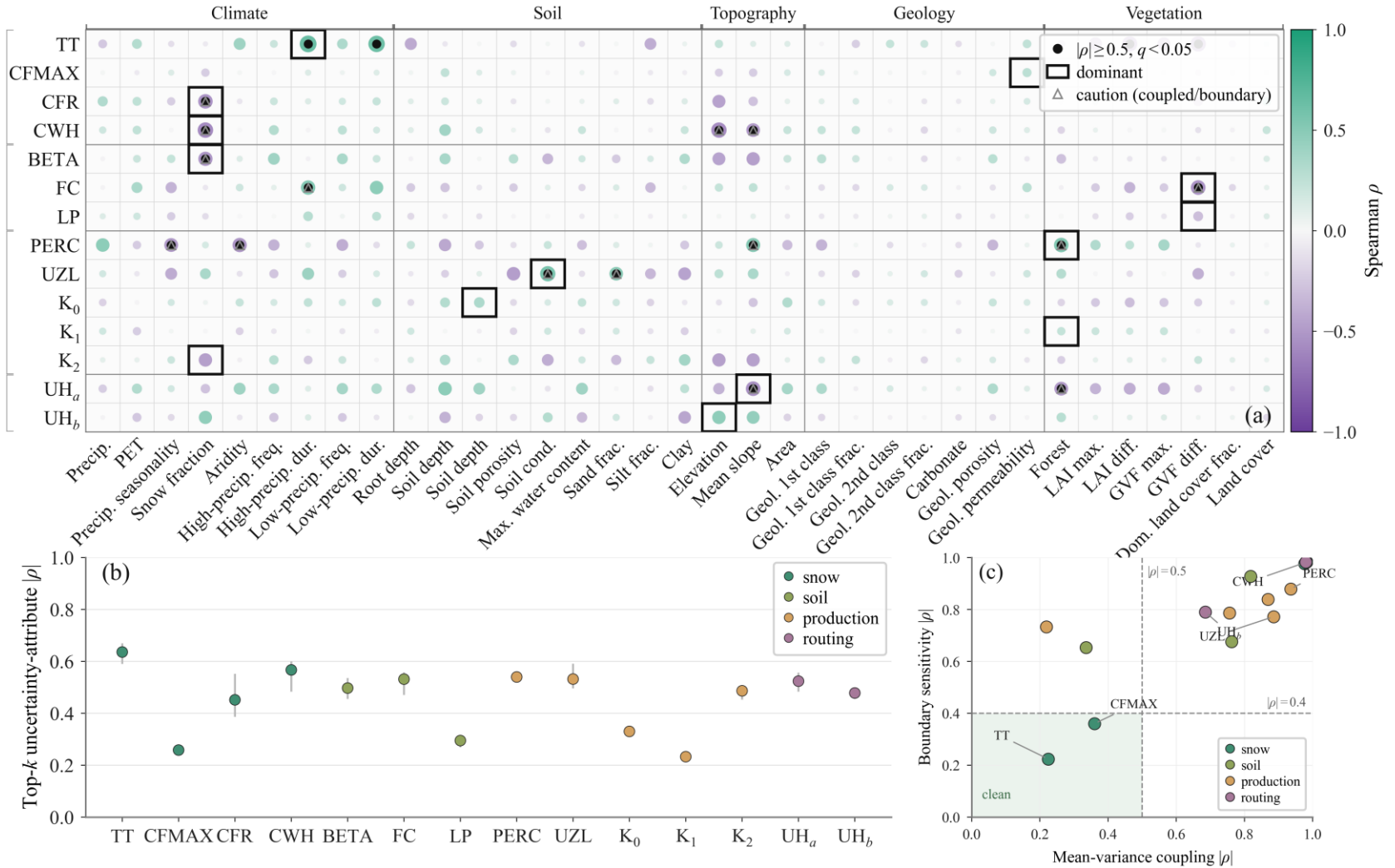


Figure S12. MC-dropout parameter-spread structure diagnostics. Panel (a) is a circle heatmap of Spearman correlations between MC-dropout parameter spread and basin attributes, using the same attribute ordering and visual encoding as the distributional parameter-spread figure. Panel (b) summarizes parameter-level spread-attribute structure by comparing top-ranked $|\rho|$ values across parameters. Panel (c) plots mean-spread coupling against boundary sensitivity for each parameter, with dashed threshold guides and process-group coloring. MC-dropout spread is computed from stochastic forward passes and averaged across seeds; it is therefore a dropout-induced spread diagnostic rather than the same quantity as the distributional formulation’s within-run parameter-scale spread.

S6.3 Direct spread-width comparison

The width comparison is descriptive because MC-dropout and distributional spread arise from different stochastic mechanisms. It does not constitute a posterior calibration test.

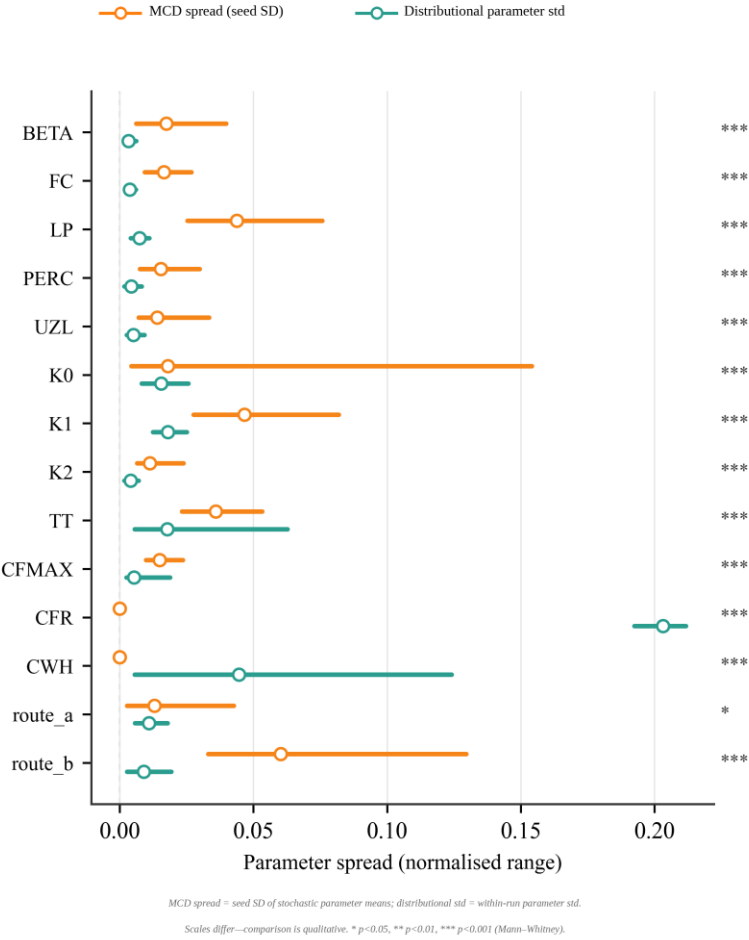


Figure S13. Direct comparison of parameter-spread width between δ_{mcd} and δ_{dist} . The horizontal point-range plot compares MC-dropout seed standard deviation of parameter means with distributional parameter-scale standard deviation for each parameter. Points show medians across basins, and horizontal ranges show interquartile ranges; significance markers report Mann-Whitney comparisons for each parameter. Because the two spread summaries are not identical statistical quantities, the figure is interpreted as a qualitative comparison of spread magnitude and sharpness rather than as a calibrated posterior comparison.

S6.4 Coupling and boundary classification

Figure S14 shows the predefined screening space, and Table S7 provides the values used to classify each parameter. The screen separates less-confounded spread from patterns dominated by mean coupling, boundary proximity, or both.

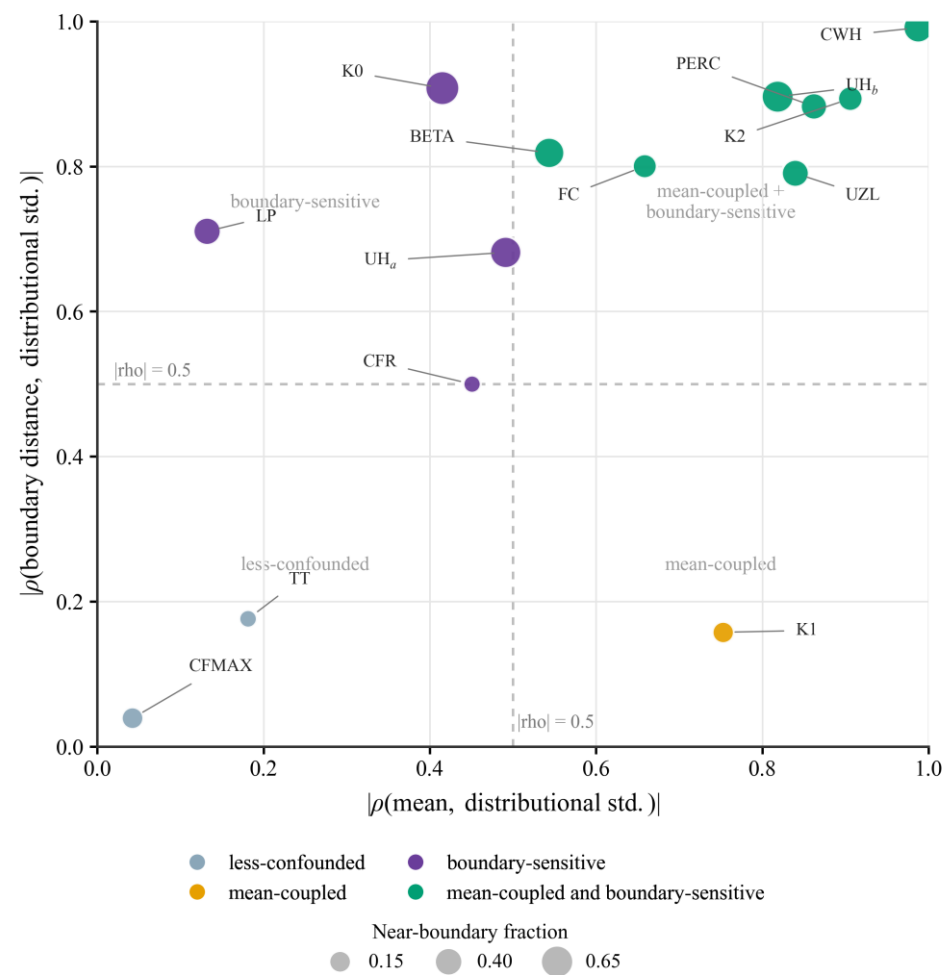


Figure S14. Diagnostic classification of distributional parameter-standard-deviation relationships. The scatterplot places each parameter by the absolute Spearman correlation between its mean and standard deviation on the x-axis and the absolute Spearman correlation between boundary distance and standard deviation on the y-axis. Dashed threshold lines divide less-confounded, mean-coupled, boundary-sensitive, and jointly coupled regions; point color indicates the diagnostic class, and point size indicates the near-boundary fraction. CFMAX and TT fall in the less-confounded region, whereas parameters such as CWH, PERC, UZL, and UH_b lie closer to the coupled or boundary-sensitive regions. The plotted diagnostics support the screening of spread gradients before hydrologic interpretation.

Table S7. Distributional parameter-spread diagnostic classification

Parameter std	Dominant attribute	Spearman rho	Mean-std coupling rho	Boundary-distance coupling rho	Near-boundary share	Diagnostic class	Evidence label	Diagnostic note
BETA std	Precip.	0.2750	0.544	0.818	0.539	Mean-coupled and boundary-sensitive	Cautionary spread gradient	Affected by mean-std coupling and boundary sensitivity
CFMAX std	Snow fraction	-0.9185	0.042	0.039	0.147	Less-confounded	Structured spread diagnostic	Low coupling and low boundary sensitivity
CFR std	LAI diff.	0.5891	0.451	0.500	0.000	Boundary-sensitive	Cautionary spread gradient	Affected by boundary sensitivity
CWH std	Snow fraction	-0.9221	0.988	0.991	0.503	Mean-coupled and boundary-sensitive	Cautionary spread gradient	Affected by mean-std coupling and boundary sensitivity
FC std	Precip. seasonality	-0.3182	0.659	0.800	0.241	Mean-coupled and boundary-sensitive	Cautionary spread gradient	Affected by mean-std coupling and boundary sensitivity
K0 std	Aridity	-0.2290	-0.415	0.908	0.738	Boundary-sensitive	Cautionary spread gradient	Affected by boundary sensitivity
K1 std	Sand frac.	-0.4172	0.753	0.157	0.137	Mean-coupled	Cautionary spread gradient	Affected by mean-std coupling
K2 std	Elevation	-0.4486	0.906	0.893	0.264	Mean-coupled and boundary-sensitive	Cautionary spread gradient	Affected by mean-std coupling and boundary sensitivity
LP std	Aridity	-0.3402	0.132	0.710	0.384	Boundary-sensitive	Cautionary spread gradient	Affected by boundary sensitivity
PERC std	Aridity	-0.5845	0.862	0.883	0.343	Mean-coupled and boundary-sensitive	Cautionary spread gradient	Affected by mean-std coupling and boundary sensitivity
TT std	Snow fraction	-0.9105	-0.181	-0.176	0.009	Less-confounded	Structured spread diagnostic	Low coupling and low boundary sensitivity
UH_a std	LAI diff.	-0.5369	0.491	0.681	0.589	Boundary-sensitive	Cautionary spread gradient	Affected by boundary sensitivity
UH_b std	Snow fraction	0.5987	0.818	0.896	0.629	Mean-coupled and boundary-sensitive	Cautionary spread gradient	Affected by mean-std coupling and boundary sensitivity

UZL std	Soil cond.	0.5779	0.840	0.790	0.373	Mean-coupled and boundary-sensitive	Cautionary spread gradient	Affected by mean-std coupling and boundary sensitivity
---------	------------	--------	-------	-------	-------	-------------------------------------	----------------------------	--

S6.5 Extended spread gradients

The full spread-descriptor panels appear in Figure **S11b**. Their interpretation follows the coupling and boundary screening in Sect. S6.4.

S6.6 Complete evidence hierarchy

Table S8 collects the existing seed, loss, cross-formulation, gradient, and spread diagnostics for selected relationships. It summarizes reported evidence and does not introduce a new pass/fail test.

Table S8. Evidence hierarchy for selected learned basin controls and parameter-spread gradients. For parameter-mean relationships, Spearman correlations and high-minus-low tercile contrasts were computed from distributional parameter means under HybridNseBatchLoss after averaging the five seeds within each basin. Seed SD and cross-loss SD are perturbation-sensitivity metrics calculated from run-specific correlations and therefore do not define the reference field. Standard-deviation relationships follow the distributional spread definitions in Sect. S6.

Relationship	Response type	Spearman rho	Seed SD	Cross-loss SD	High-minus-low difference	Cross-formulation class	Diagnostic flag	Evidence level
BETA-Mean slope	Mean	-0.583	0.017	0.049	-0.403	Shared	None / low	Reproducible behavioral gradient
FC-PET	Mean	0.509	0.012	0.108	0.198	Shared	None / low	Reproducible but sensitivity-affected behavioral gradient
PERC-Aridity	Mean	-0.594	0.015	0.260	-0.229	Shared	Collinearity sensitivity	Reproducible but sensitivity-affected behavioral gradient
UZL-Soil cond.	Mean	0.570	0.018	0.024	0.260	Shared	None / low	Reproducible behavioral gradient
CWH-Snow fraction	Mean	-0.904	0.012	0.138	-0.342	Partially shared	Formulation sensitivity	Distributional diagnostic gradient
CFR-Snow fraction	Mean	-0.625	0.063	0.052	-0.081	Partially shared	Formulation sensitivity	Distributional diagnostic gradient
CFMAX std-Snow fraction	Std	-0.919	0.011	0.016	-0.046	NA	Less-confounded	Structured spread diagnostic
TT std-Snow fraction	Std	-0.910	0.008	0.020	-0.083	NA	Less-confounded	Structured spread diagnostic
CWH std-Snow fraction	Std	-0.922	0.008	0.046	-0.149	NA	Mean-coupled / boundary-sensitive	Cautionary spread gradient
PERC std-Aridity	Std	-0.584	0.015	0.168	-0.005	NA	Mean-coupled / boundary-sensitive	Cautionary spread gradient
UZL std-Soil cond.	Std	0.578	0.033	0.019	0.009	NA	Mean-coupled / boundary-sensitive	Cautionary spread gradient
UH_b std-Snow fraction	Std	0.599	0.038	0.196	0.011	NA	Mean-coupled / boundary-sensitive	Cautionary spread gradient

Note: NA denotes non-applicable cross-formulation classes for standard-deviation relationships. Reference-field correlations and tercile contrasts match Figure 8; seed and cross-loss SDs quantify sensitivity. PERC–aridity should also be interpreted with the all-loss estimate in Table S6.