water

MDPI

*Article*

# A Differentiable Spatially Varying Ensemble Modelling Framework for Runoff Model Heterogeneity

**Yi Ding [1], Zhaohui He [1], Ningpeng Dong [2,*], Jintao Ding [1,*] and Ruxuan Ma [2]**

[1] Yalong River Hydropower Development Co., Ltd., Chengdu 610051, China
[2] State Key Laboratory of Water Cycle and Water Security, China Institute of Water Resources and Hydropower Research, Beijing 100038, China
* Correspondence: dongnp@iwhr.com (N.D.); dingjintao001@sdic.com.cn (J.D.)

**Abstract**

Conceptual rainfall–runoff models usually apply a single runoff generation structure to an entire catchment, although contrasts in soil, terrain, climate, and vegetation can produce different runoff responses within it. We developed a differentiable spatially varying ensemble (Spatial ENS) that combines GR4J, a vertically mixed runoff model (VMM), and HBV, and applied it to the Jiagenba catchment in the Yalong River Basin, southwestern China. A neural gating network maps gridded climate and land surface attributes to model weights for each cell, and the weighted runoff is routed through a shared Muskingum river network. Under leave-one-year-out cross-validation for 2017–2022, Spatial ENS achieved a flood season (June–October) Nash–Sutcliffe efficiency (NSE) of 0.78, compared with 0.77, 0.70, and 0.74 for GR4J, VMM, and HBV, respectively, and the highest Kling–Gupta efficiency (0.80). Its mean annual NSE gains were 0.02, 0.09, and 0.05. The learned weights varied systematically in space, and permutation diagnostics identified clay fraction, terrain slope, and potential evapotranspiration as the dominant inputs to the gating network. These results indicate that outlet discharge can constrain spatially structured model weights and that the resulting ensemble improves flood season simulation in this catchment.

**Keywords:** differentiable hydrological modeling; mixture of models; model structural uncertainty; spatial heterogeneity; GR4J; HBV; vertically mixed runoff model; leave-one-year-out validation

## 1. Introduction

Hydrological models are essential tools for understanding catchment water cycles, forecasting floods and droughts, and supporting water resources management. By representing how precipitation is partitioned, stored, evaporated, and routed through river networks, they convert limited observations into continuous information on catchment water fluxes and storage [1–3]. The credibility of such simulations depends not only on parameter values but also on whether the chosen process structure suits the climate, terrain, soil, and vegetation of the study area.

Many runoff generation models have been developed to describe different catchment responses. GR4J represents production and routing stores with a parsimonious formulation [4], HBV uses a soil moisture routine with upper- and lower-zone response reservoirs [5], and vertically mixed models combine infiltration-excess and saturation-excess runoff [6]. Because these models emphasize different storage, infiltration, and runoff release processes, their performance varies with hydroclimate and catchment properties,

Check for updates

Academic Editor: David Post

Received: 17 October 2026
Revised: 31 October 2026
Accepted: 6 November 2026
Published: 8 November 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

and the choice of model structure remains a major source of uncertainty in rainfall–runoff simulation. Equifinality further complicates this choice, because different structures and parameter sets can produce similar outlet hydrographs [7].

Combining several models is a common response to this structural uncertainty. Mixture-of-models approaches exploit complementary model responses to reduce dependence on any single structure, and some allow the model weights to vary with catchment state or other explanatory variables [8,9]. When model outputs are combined at the outlet, one set of weights represents the basin as a whole. Within a catchment, soils, terrain, vegetation, and storage conditions differ, and a structure that performs well in one part may be less suitable elsewhere. A catchment-scale ensemble can improve the outlet hydrograph while leaving this spatial mismatch unresolved, so structural errors at the grid scale may persist in the ensemble simulation and limit further gains in accuracy.

Parameter regionalization addresses spatial differences by relating model parameters to catchment attributes [10–16]. Differentiable hydrological modeling extends this idea: process equations are implemented as differentiable recurrent operators, and neural networks generate spatially varying parameters or other model quantities from attributes [17–19]. Streamflow errors can then be backpropagated through the complete runoff–routing system to update these mapping rules. Applications include parameter learning across large samples of basins [17,18], regionalization within distributed hydrological models [20], high-resolution national-scale modeling [21], and fully distributed rainfall–runoff simulation [22]. Li et al. [23] combined differentiable process-based and data-driven simulations from multiple meteorological forcing datasets. These studies show that differentiable methods can link sparse streamflow observations to spatially distributed model behavior while retaining an explicit process structure.

Spatially varying parameters describe differences in catchment properties within a prescribed runoff structure. Where runoff processes vary within a catchment, the relative suitability of different structures may also change. Relating model weights to local attributes extends regionalization to this structural dimension [8–11,17,18], allowing several runoff models to contribute in different proportions across the catchment. The central issue is whether outlet discharge can constrain these spatial weights well enough to improve streamflow prediction.

In this study, we combine GR4J, a vertically mixed runoff model (VMM), and HBV in a differentiable spatially varying ensemble (Spatial ENS) and apply it to the Jiagenba catchment in the Yalong River Basin, southwestern China. A gating network learns spatially varying model weights from gridded climate and land surface attributes, the three runoff outputs are combined in each runoff cell, and the combined runoff is routed through a shared Muskingum river network. As references, each model is also trained alone with its weight fixed to one. Using leave-one-year-out cross-validation for 2017–2022, we evaluate (1) the flood season discharge prediction of Spatial ENS relative to the individual models, (2) the stability of the learned spatial weight patterns across test years, and (3) the relationships between the model weights and environmental attributes.

## 2. Materials and Methods

### *2.1. Study AREA and Observational Data*

The Jiagenba catchment is located in the mountainous Yalong River Basin in southwestern China. Its gauging station lies at 101.551° E, 30.002° N (Figure 1a) and controls a drainage area of approximately 1654 km$^2$.

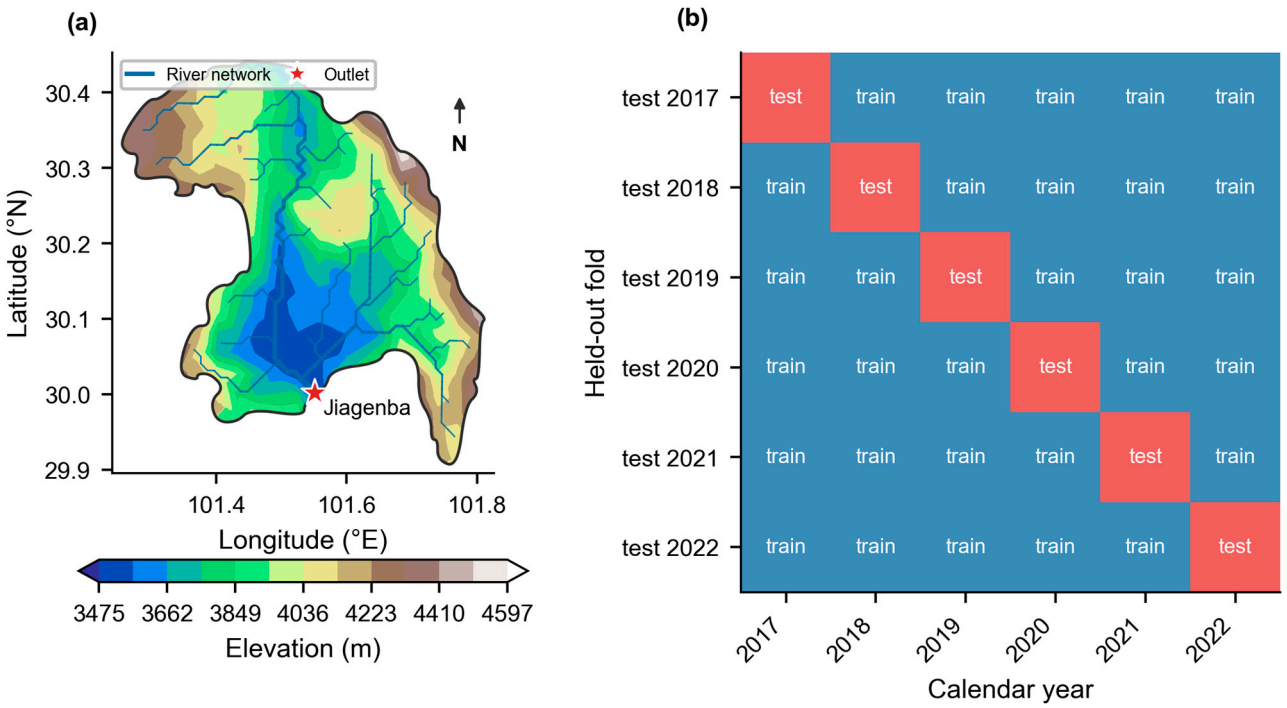



**Figure 1.** Study area and temporal design. (**a**) Catchment topography, the modeled river network, and the Jiagenba outlet; river line width is scaled by upstream area. (**b**) Leave-one-year-out design for the six test years. In each experiment, the displayed year is withheld for testing, the complete preceding calendar year is used for state warm-up, and June–October is used for the loss and the primary evaluation.

Hourly discharge observations for 2017–2022 were aggregated to daily mean flow after temporal interpolation of invalid hourly values. Twenty-five nonpositive records (0.05% of the hourly observations) were replaced; the longest consecutive gap was 13 h. The resulting daily series was used for model training and evaluation.

Daily precipitation for 2016–2022 was obtained from the gauge-based CHM_PRE V2 dataset at 0.1° resolution [24,25] and mapped to the active runoff cells. Air temperature was obtained from ERA5 [26], and daily potential evapotranspiration (PET) was calculated with the FAO-56 Penman–Monteith method [27].

### *2.2. Common Differentiable Modeling Framework*

The model uses a runoff grid of approximately 3 km and a 1 km river network. All four model configurations were built within the same differentiable framework (Figure 2). Ten standardized static attributes were supplied to the parameter networks of the runoff models and to the gating network: clay fraction, elevation, terrain slope, long-term mean precipitation, PET, and temperature, leaf area index (LAI), aridity, precipitation seasonality, and specific humidity. The climate attributes characterize water supply, evaporative demand, and seasonality; clay fraction, terrain, and LAI describe controls on infiltration, storage, and vegetation water use. Comparable attribute groups are widely used in large-sample hydrology [28,29]. Alternative attribute combinations were not systematically evaluated. The ten attributes were extracted from the national 1 km distributed-model database of Dong et al. [30], in which long-term precipitation is derived from CHM_PRE V2, temperature from ERA5-Land, and PET from GLEAM; elevation and slope from MERIT Hydro; clay fraction from HWSD v1.2; and LAI from the Global 1 km Land Surface Parameter dataset [31]. Specific humidity (kg kg$^{-1}$) describes atmospheric moisture conditions. The 1 km attribute layers were averaged over the contributing pixels of each runoff cell and

standardized using the national active-domain mean and standard deviation. Aridity was calculated as the ratio of long-term PET to precipitation, and precipitation seasonality as the coefficient of variation in the twelve long-term mean monthly precipitation totals. The climatic attributes represent averages over 1981–2010, and LAI over 2001–2005. Daily precipitation, PET, and temperature first passed through shared snow and evapotranspiration components. A trainable degree-day module, whose spatial parameters were generated from the same attributes, separated rainfall from snowfall and computed snowmelt. PET was adjusted by a dynamic multiplicative correction that depended on the attributes and the current meteorological conditions; a weak quadratic penalty (weight 0.001) discouraged unnecessarily large corrections.

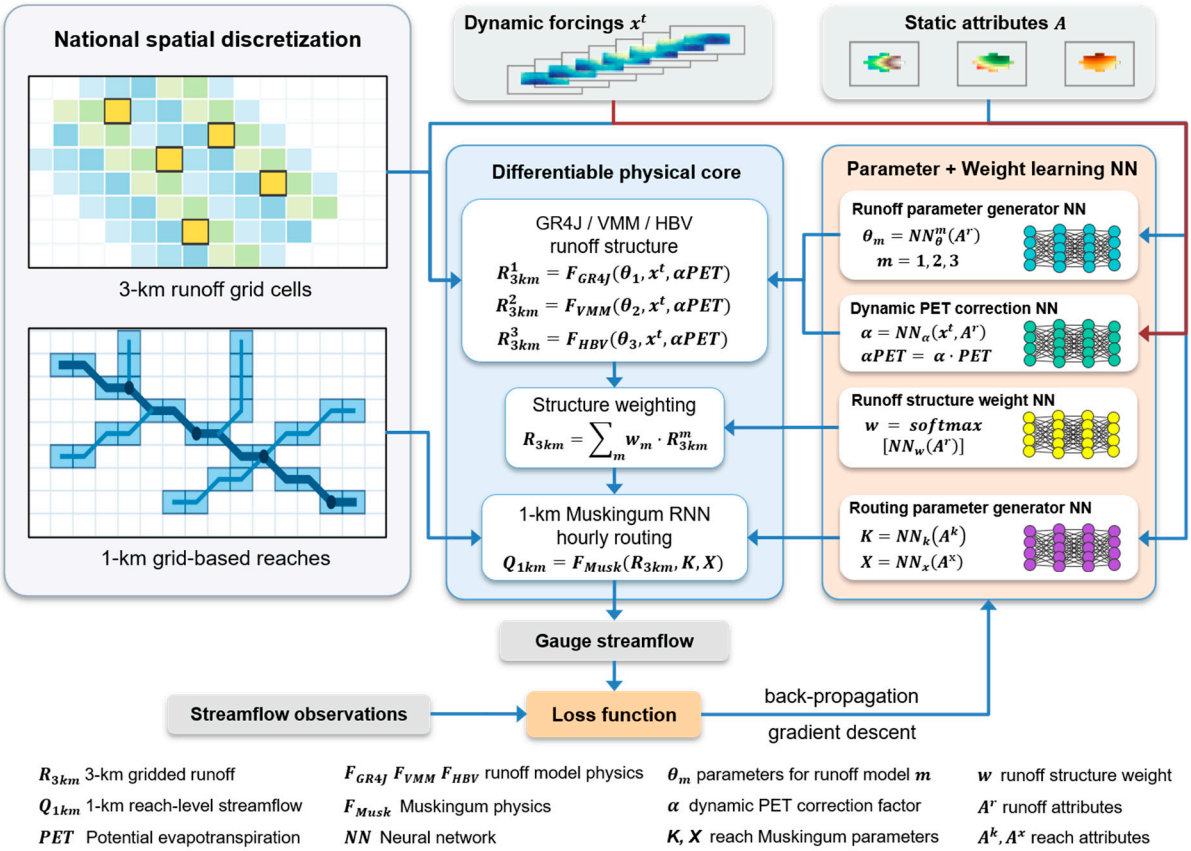



**Figure 2.** Differentiable spatial ensemble framework. The three runoff models run in parallel in each runoff cell. In Spatial ENS, the gating network maps ten static attributes to non-negative, time-invariant model weights; in the single-model experiments, the gate is replaced by a fixed one-hot vector. The mixed runoff is routed once through the shared Muskingum river network.

For model $m$ in runoff cell $i$ on day $t$, runoff was calculated as

$$q_{m,i,t} = H_m\left(P_{i,t}, E_{i,t}, T_{i,t}, \boldsymbol{\theta}_{m,i}, \mathbf{s}_{m,i,t-1}\right) \quad (1)$$

where $q_{m,i,t}$ is the cell runoff; $H_m$ is the differentiable recurrence of GR4J, VMM, or HBV; $P$, $E$, and $T$ are precipitation, PET, and temperature; $\theta_{m,i}$ is the parameter vector generated from the attributes; and $\mathbf{s}_{m,i,t-1}$ is the model state on the preceding day. Each model had its own multilayer perceptron (10–128–128) to generate spatial parameters, and bounded activation functions kept the parameters within physically admissible ranges.

The mixed cell runoff was mapped onto the shared river graph and routed once with an hourly Muskingum scheme [32]. Reach travel-time and weighting parameters were generated by a shared network from reach attributes and constrained to numerically admissible ranges. The internal unit hydrographs of the runoff models were disabled so

that routing was neither duplicated nor structurally different among models. The same river graph, outlet, snow module, evapotranspiration correction, and routing operator were used in every experiment.

### *2.3. Runoff Generation Models*

The three models represent distinct and interpretable hypotheses of runoff generation (Table 1).

**Table 1.** Runoff generation structures used in the controlled comparison.

| Method | Principal Stores and Fluxes | Spatially Generated Model Parameters | Ensemble Weight |
|---|---|---|---|
| GR4J | Production store, percolation, groundwater exchange, routing store | $X_1$ (production-store capacity), $X_2$ (exchange coefficient), $X_3$ (routing-store capacity), $C_{PET}$ | Fixed (1,0,0) in the individual model |
| VMM | Infiltration-excess split plus Xinanjiang-style saturation-excess and source separation | WUM/WLM/WDM, $K_{PET}$, CD, SM, KSS, KG, KKSS, KKGF; FC, KF, BF, D | Fixed (0,1,0) in the individual model |
| HBV | Soil bucket, beta recharge, upper and lower reservoirs | FC, LP, BETA, UZL, $K_0$, $K_1$, PERC, $K_2$, $C_{PET}$ | Fixed (0,0,1) in the individual model |
| Spatial ENS | All three models evaluated in parallel; cell runoff mixed before common routing | All model parameter fields plus a 10–32–32–3 gate | Learned softmax vector $\mathbf{w}_i$ |

The reduced GR4J model used a production store, nonlinear percolation, groundwater exchange, and a routing store [4]. Its spatial parameters were $X_1$ (production-store capacity), $X_2$ (groundwater exchange coefficient), $X_3$ (routing-store capacity), and $C_{PET}$ (static PET coefficient). The unit-hydrograph parameter $X_4$ was omitted because translation and attenuation were handled by the shared river network.

VMM adds an infiltration-excess component ahead of a Xinanjiang-type core that represents saturation-excess runoff, storage, and source separation [6,33]. The parameters of this core were WUM, WLM, and WDM (tension-water storage capacities of the upper, lower, and deep layers), $K_{PET}$ (static PET coefficient), CD (lower-layer evaporation partition coefficient), SM (free-water storage capacity), KSS and KG (interflow and groundwater recession coefficients), and KKSS and KKGF (recession-memory coefficients). The infiltration component added FC (stable infiltration capacity), KF (deficit amplification factor), BF (exponent of the infiltration-capacity distribution), and D (equivalent duration). Because the model runs at a daily time step, each positive daily net liquid input was treated as an equivalent rainfall event of duration D, so that D represents the sub-daily rainfall variability that daily totals cannot resolve.

The HBV model consisted of a soil moisture bucket, beta-function recharge, an upper response zone, percolation, and a lower response zone [5]. Its parameters were FC (soil-water storage capacity), LP (fraction of FC above which actual evapotranspiration reaches the potential rate), BETA (recharge exponent), UZL (upper-zone threshold), $K_0$ (near-surface response coefficient), $K_1$ (upper-zone response coefficient), PERC (percolation capacity), $K_2$ (lower-zone response coefficient), and $C_{PET}$ (static PET coefficient). The internal snow routine and MAXBAS transformation of HBV were disabled because snow and routing were handled by the shared framework.

### 2.4. Spatial Mixture of Models

The gating network was a multilayer perceptron with ten inputs, two hidden layers of 32 units, and three output logits. For the standardized attribute vector $\mathbf{a}_i$ of cell $i$, the weights were

$$w_{m,i} = \frac{\exp(g_m(\mathbf{a}_i))}{\sum_{k=1}^{3} \exp(g_k(\mathbf{a}_i))}, w_{m,i} \geq 0, \sum_{m=1}^{3} w_{m,i} = 1 \quad (2)$$

where $g_m(\mathbf{a}_i)$ is the $m$th output logit. The weights varied in space but were constant in time, because the gating network received no date, meteorological forcing, discharge, or model state. Cell runoff was the convex combination

$$q_{i,t}^{\mathrm{ENS}} = \sum_{m=1}^{3} w_{m,i} q_{m,i,t} \quad (3)$$

The gating network was trained jointly with the parameter networks of the three runoff models. In the experiments with individual models, the weights were fixed to the one-hot vector of the selected model, which was trained together with the shared snow, evapotranspiration, and routing modules. This design keeps the forcing, routing, and training procedure identical across experiments, so that the intended difference lies in the runoff structure.

### 2.5. Training and Leave-One-Year-Out Evaluation

Six experiments were conducted, each withholding one calendar year of 2017–2022 for testing and using the remaining five years for training (Figure 1b). Each annual simulation was preceded by a forcing-only warm-up over the complete previous calendar year. The loss and the primary evaluation metrics were both calculated for 1 June–31 October, the flood season of the catchment, and discharge in the held-out year was never used for optimization.

Training minimized the variance-normalized mean squared error, which equals one minus the Nash–Sutcliffe efficiency (NSE) for the sampled annual window:

$$L_Q = \frac{\frac{1}{n}\sum_t \left(Q_t - \hat{Q}_t\right)^2}{\frac{1}{n}\sum_t \left(Q_t - \overline{Q}\right)^2} \quad (4)$$

where $Q_t$ and $\hat{Q}_t$ are the observed and simulated outlet discharge on day $t$, $\overline{Q}$ is the mean observed discharge, and $n$ is the number of days in the window. At each update, one complete training year was sampled following a deterministic, balanced sequence. All models were trained with AdamW for 100 updates. Learning rates for the model and routing components were initially 0.001 and 0.0007 and were reduced to 0.0002 and 0.00015, respectively, after 70 updates. The final model was evaluated after the fixed training budget, without early stopping or selection using held-out discharge.

The primary experiments used a common random seed. Spatial ENS and GR4J were then trained from two additional independent initializations under the same sequence of training years and optimization settings.

### 2.6. Performance and Spatial Diagnostics

Predictive performance was evaluated with NSE, the Kling–Gupta efficiency (KGE), root mean squared error (RMSE), percent bias (PBIAS), and the Pearson correlation coefficient [34–37]. The metrics were calculated for each held-out June–October period and for all test years combined. Pooled NSE uses the mean discharge of the combined daily record, whereas mean annual NSE is the arithmetic mean of six NSE values, each calculated relative to its own annual mean discharge. Paired differences in annual NSE were tested with an exact two-sided Wilcoxon signed-rank test.

Uncertainty in the pooled performance differences was estimated using 5000 paired block bootstrap samples, resampling consecutive days within the flood season of each test year. The primary block length was 14 days, with 7- and 30-day blocks used for sensitivity analysis.

Spatial heterogeneity was summarized from the cell weights of each test year model. For each model weight, we calculated the range and standard deviation across cells of the test year mean field, the standard deviation among test year models at each cell, and the pairwise Pearson correlations between the weight fields of different test year models. The normalized entropy of the three weights was also calculated for each cell, with a value of 1 indicating equal weights. For repeated initializations, weight fields from the same held-out year were compared using Spearman rank correlation and mean absolute differences.

The dependence of the gating network on each attribute was evaluated by randomly permuting that attribute among cells while keeping the other inputs unchanged. The resulting changes in model weights were normalized across attributes and summarized by their median across test years. Spearman correlations between attributes and weights were calculated as descriptive measures of spatial association.

## 3. Results

### 3.1. Out-of-Sample Discharge Prediction

Spatial ENS achieved the best flood season performance on all metrics when the test years were pooled (Table 2). Its NSE was 0.78, compared with 0.77 for GR4J, 0.70 for VMM, and 0.74 for HBV. It also had the highest KGE (0.80) and correlation (0.89), the lowest RMSE (33.25 $m^3 s^{-1}$), and the smallest absolute PBIAS (4.40%). Among the individual models, GR4J performed best, followed by HBV and VMM. All four configurations underestimated flood season discharge, with PBIAS ranging from −8.57% to −4.40%.

**Table 2.** Flood season performance of the three individual models and Spatial ENS, with all test years pooled.

| Method | NSE | KGE | RMSE ($m^3 s^{-1}$) | PBIAS (%) | Correlation |
|---|---|---|---|---|---|
| GR4J | 0.77 | 0.77 | 34.35 | −4.57 | 0.88 |
| VMM | 0.70 | 0.70 | 39.19 | −8.57 | 0.85 |
| HBV | 0.74 | 0.72 | 36.59 | −7.94 | 0.87 |
| Spatial ENS | 0.78 | 0.80 | 33.25 | −4.40 | 0.89 |

Spatial ENS had a higher annual NSE than each individual model in all six main experiments (Table 3; Figures 3 and 4). The NSE gain over GR4J was below 0.01 in 2018 and reached 0.03 in 2019, averaging 0.02; the mean gains over VMM and HBV were 0.09 and 0.05, respectively. For each of the three paired comparisons, the exact two-sided Wilcoxon signed-rank test gave $p = 0.03$, the smallest value attainable with six paired years. Absolute NSE varied considerably between years, and all models performed worst in 2019, but the ranking of the four configurations was the same in every year.

**Table 3.** NSE in each held-out year.

| Method | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | Mean |
|---|---|---|---|---|---|---|---|
| GR4J | 0.82 | 0.74 | 0.48 | 0.84 | 0.69 | 0.82 | 0.73 |
| VMM | 0.76 | 0.68 | 0.42 | 0.74 | 0.61 | 0.75 | 0.66 |
| HBV | 0.81 | 0.71 | 0.47 | 0.77 | 0.67 | 0.79 | 0.70 |
| Spatial ENS | 0.84 | 0.74 | 0.51 | 0.85 | 0.70 | 0.84 | 0.75 |

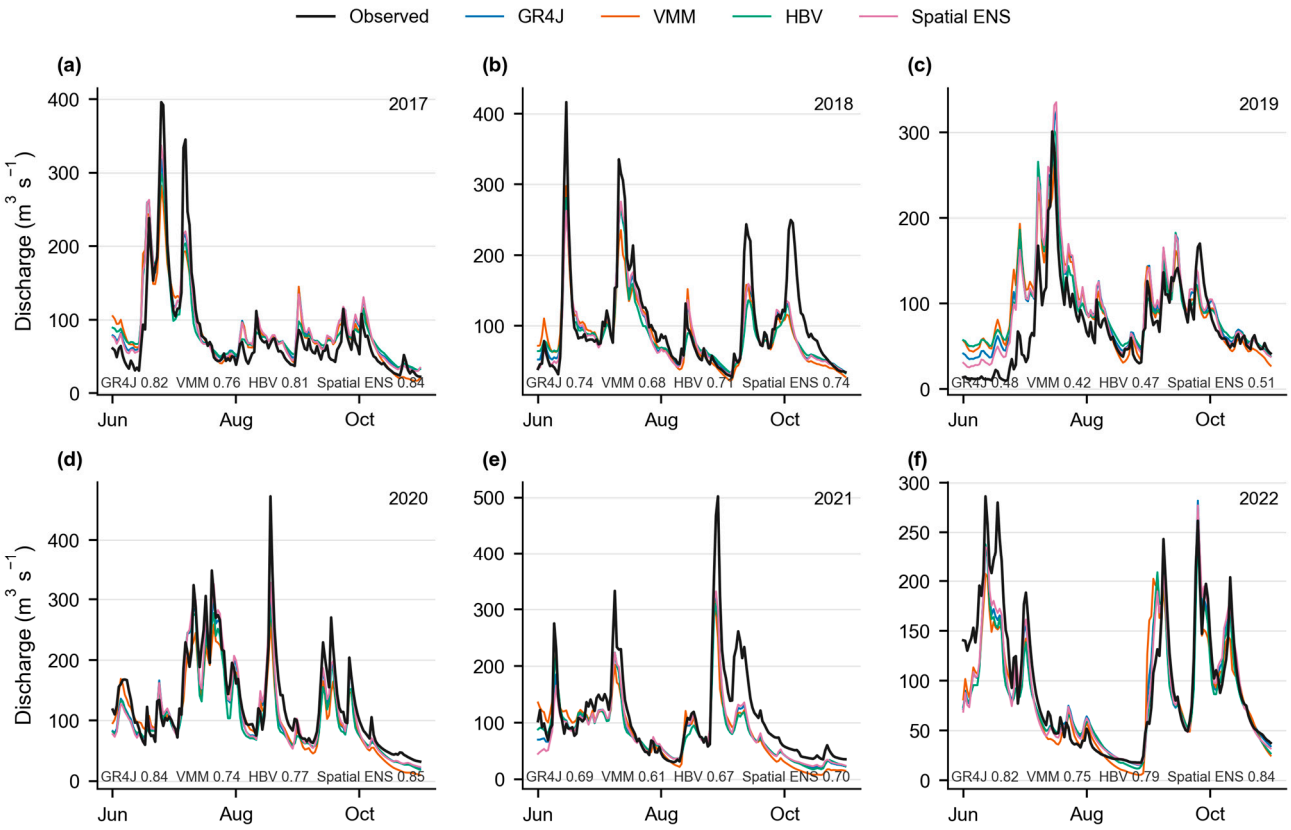



**Figure 3.** Observed and simulated daily discharge from June to October in each held-out year. (**a**–**f**) was produced by a separate fit that excluded discharge from the displayed year. Values at the bottom of (**a**–**f**) are the NSE of GR4J, VMM, HBV, and Spatial ENS.

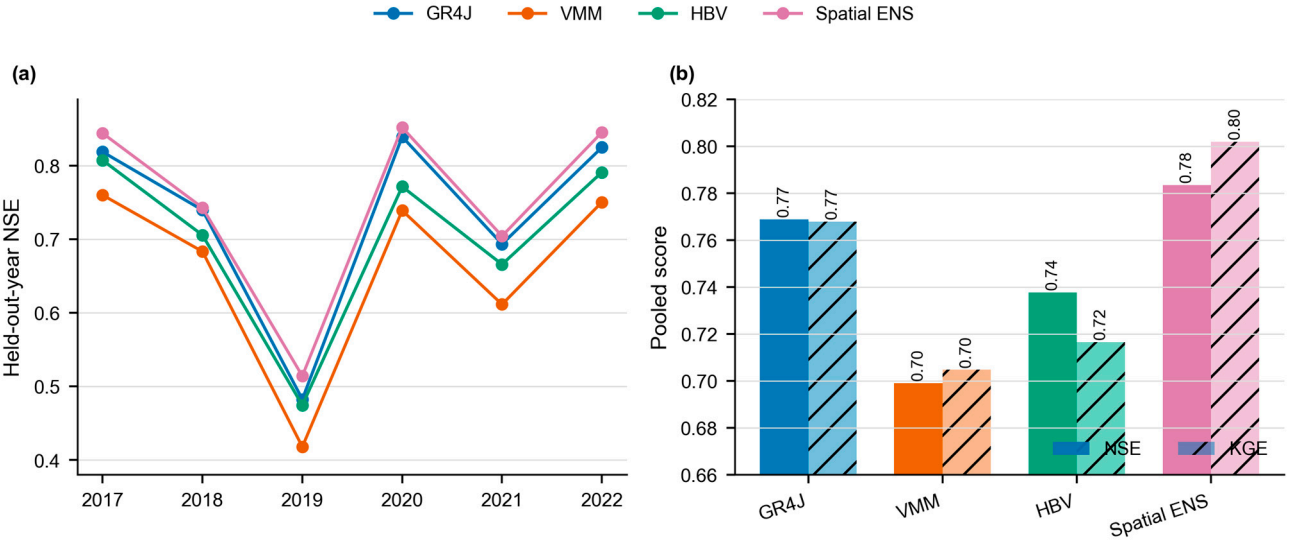



**Figure 4.** Predictive performance of the three individual models and Spatial ENS. (**a**) NSE in each held-out year. (**b**) NSE and KGE with all test years pooled.

The pooled NSE gain over GR4J was 0.01, with a 95% block bootstrap interval of 0.01–0.03. Mean annual NSE was 0.75, with each year evaluated relative to its own observed mean, compared with a pooled NSE of 0.78. The ensemble retained its pooled performance advantage across three initializations and had a higher annual NSE in 17 of the 18 comparisons. GR4J was marginally better in the 2020 test under the second initialization.

### 3.2. Spatial Pattern and Stability of the Learned Weights

The gating network produced weights that varied gradually across the catchment (Figure 5). Averaged over all cells and test year models, the weights were 0.465 for GR4J, 0.206 for VMM, and 0.329 for HBV. In the test year mean fields, GR4J weights ranged from 0.410 to 0.515, VMM weights from 0.167 to 0.255, and HBV weights from 0.318 to 0.335, with spatial standard deviations of 0.0175, 0.0145, and 0.0030, respectively. Spatial variation therefore consisted mainly of a trade-off between GR4J and VMM, while the HBV weight was nearly uniform. The mean normalized entropy was 0.951, and all three models retained substantial weight throughout the catchment.

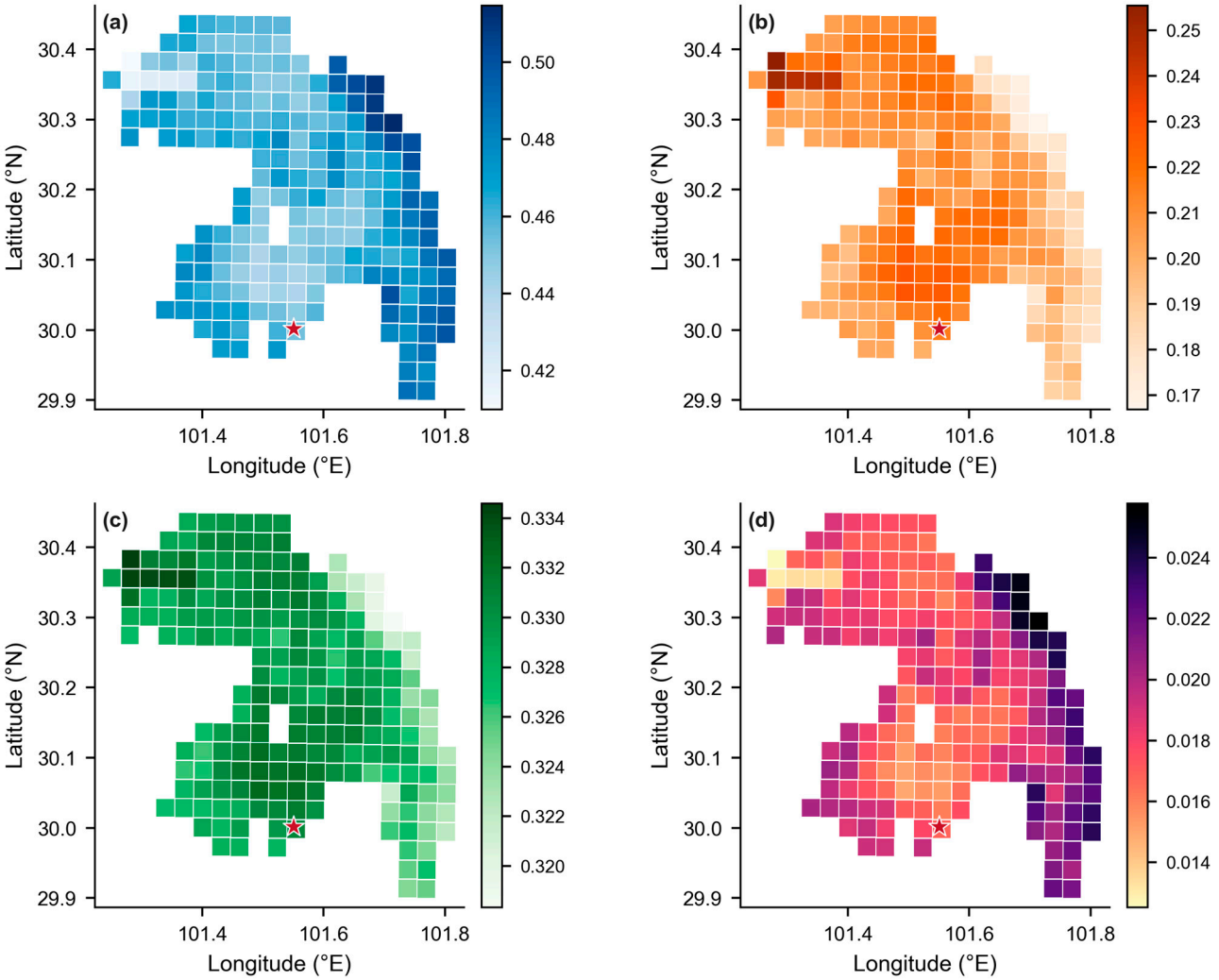



**Figure 5.** Mean Spatial ENS weight fields across the test year models: (**a**) GR4J, (**b**) VMM, and (**c**) HBV. (**d**) Mean standard deviation among the test year models. Colour scales are independent across panels, and the HBV weights span a narrower range than the GR4J and VMM weights. The star marks the Jiagenba outlet.

The spatial ordering of the weights was highly reproducible when different years were withheld. The median pairwise Pearson spatial correlation exceeded 0.99 for each of the three model weights. One HBV pair had a lower correlation, consistent with its nearly flat weight field. GR4J received the largest weight throughout the catchment in every test year. The weight fields mainly showed gradual shifts in the relative support for the three models.

Across initializations, median pairwise Spearman correlations for the same held-out year were 0.89, 0.89, and 0.87 for the GR4J, VMM, and HBV weights, respectively. Their

broad spatial ordering was similar, while absolute weights varied more, particularly for GR4J and VMM.

### *3.3. Attributes Associated with the Learned Weights*

Permutation diagnostics showed that the gating network relied mainly on clay fraction, terrain slope, and PET (Figure 6a). Clay fraction accounted for median normalized effects of 51.1%, 52.7%, and 43.2% on the GR4J, VMM, and HBV weights, respectively; slope accounted for 31.5%, 30.6%, and 35.6%, and PET for 11.0%, 10.6%, and 12.7%. No other attribute exceeded 4% for any weight. Together, these three attributes accounted for more than 90% of the normalized permutation effect on each model weight.

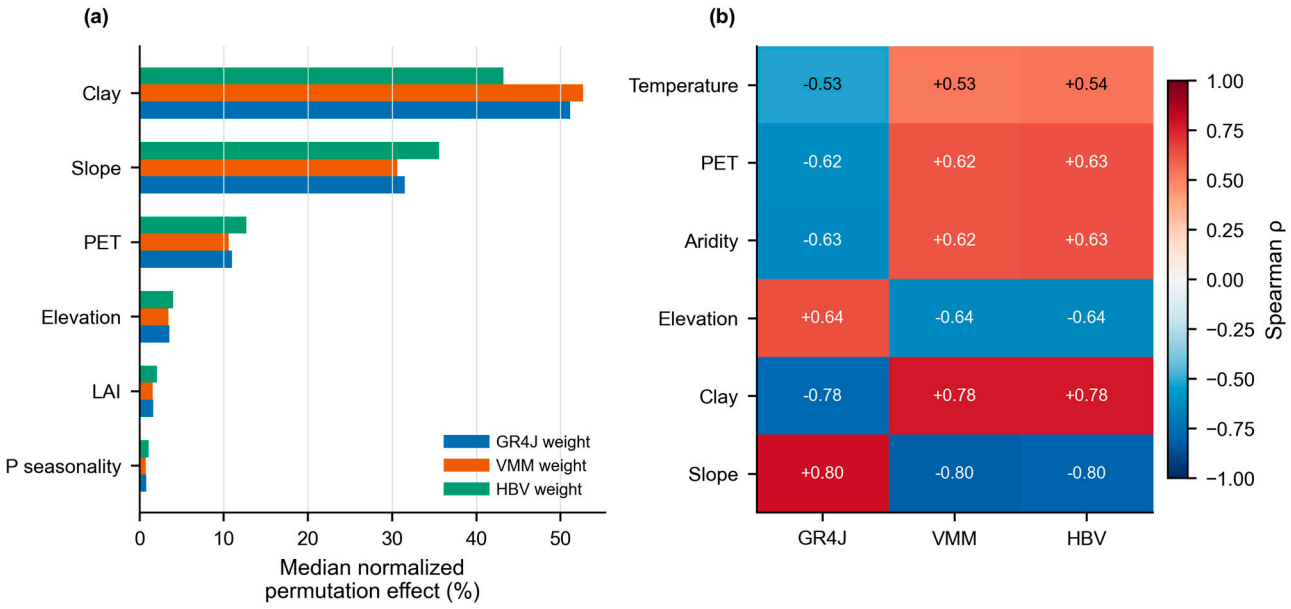



**Figure 6.** Attributes associated with the learned model weights. (**a**) Median normalized permutation effects across test years. (**b**) Spearman correlations between selected attributes and the test year mean weight fields.

The spatial correlations revealed a coherent terrain–soil gradient (Figure 6b). The GR4J weight was positively correlated with slope ($\rho = 0.80$) and elevation ($\rho = 0.64$) and negatively correlated with clay fraction ($\rho = -0.78$), aridity ($\rho = -0.63$), and PET ($\rho = -0.62$). The VMM and HBV weights showed largely opposite associations, as expected from the sum-to-one constraint of the softmax weights. GR4J was thus favored on steeper, higher, and less clay-rich cells with lower evaporative demand, whereas the combined weight of VMM and HBV was higher on flatter, clay-rich, and drier cells. The permutation effects, spatial correlations, and stability across test years consistently indicate that the weight pattern is organized mainly along clay and slope gradients.

## 4. Discussion

### *4.1. Learning Spatial Model Weights from Outlet Discharge*

The main finding is that outlet discharge alone constrained a spatial pattern of model weights that was stable across test years and accompanied improved prediction in held-out years. The weights were not uniform, their spatial ordering was nearly identical among test year models, and their variation followed soil, terrain, and evaporative demand attributes. Within a single gauged catchment, Spatial ENS therefore indicates where each runoff structure receives more or less support from the data.

The gating network expresses the outlet discharge signal as a single function of ten static attributes that is applied to every runoff cell. This shared mapping links the weights to landscape properties and reduces the number of quantities estimated from outlet discharge. The same principle underlies multiscale parameter regionalization [10,11] and recent differentiable models that couple process simulation with neural parameterization [20–22,38–43]. Here the attribute-based mappings generate both model parameters and the relative weights used to mix the three runoff estimates for each cell before river routing.

The spatial pattern is also consistent with the model structures. VMM is the only structure with an explicit infiltration-excess component, and its weight was higher on clay-rich cells, where lower infiltration capacity would favor infiltration-excess runoff. Different combinations of parameters and weights can nevertheless reproduce a similar outlet hydrograph, including through compensation for errors in forcing, storage, and routing [7–9]. The weights express relative model support under the present configuration. Internal soil moisture, snow, or tributary flow observations would help test their process interpretation.

### *4.2. Sources of the Performance Gain*

GR4J was the strongest individual model, yet Spatial ENS increased its pooled flood season NSE by 0.01 and its KGE by 0.03. Because mixing occurs in each runoff cell before the common routing, the ensemble can combine the structures in different proportions in different landscape settings. The consistent gain suggests that VMM and HBV supply responses that complement GR4J, even though both performed worse on their own.

All four configurations share the same forcing, snow and evapotranspiration components, river network, routing operator, and optimization budget. The difference between Spatial ENS and the individual models therefore lies in how runoff generation is represented within the catchment: a spatially weighted combination of three structures instead of a single structure applied everywhere. The ensemble also improved on GR4J in 2019, when all models had their lowest skill.

The weights were also fixed in time. They describe differences among landscape settings, whereas the runoff of each model still responds to meteorological forcing and model state. Time-varying weights could represent event-scale shifts in the dominant runoff process, but with a single outlet gauge they would need stronger constraints to remain identifiable.

Finally, the three structures differ in complexity. The reduced GR4J is parsimonious, VMM includes an equivalent-duration parameter for sub-daily rainfall variability, and HBV separates upper- and lower-zone responses. These differences are part of the controlled comparison: each structure enters the ensemble with its own parameterization, and the gating network weighs them under identical forcing and routing.

### *4.3. Scope and Limitations*

This study covers one catchment, one outlet gauge, and six evaluation years, and the primary target is June–October discharge. The short evaluation record and overlapping training sets limit statistical power. Catchments with different aridity, snow regimes, soils, or terrain may favour different model contributions. Evaluation in additional catchments and at internal gauges would help assess these differences.

The ensemble's pooled performance advantage persisted across three initializations, while the absolute weights were more sensitive to initialization. Further runs with longer training would help assess convergence uncertainty.

The discharge record involves the interpolation of short gaps, and gridded precipitation remains uncertain in mountainous terrain, where gauges are sparse. The shared snow and routing components rely on temperature-index and Muskingum approximations whose parameters can vary regionally [44–47], and the choice of evapotranspiration formulation can also affect rainfall–runoff simulation [48,49]. Multiple forcing products and internal hydrological observations would strengthen the evaluation. Recent work on large-sample machine learning [50,51], multi-forcing ensembles and forcing fusion [23,52], and gradient computation for implicit differentiable models [53] offers directions for these extensions.

For water resources assessment, the framework provides continuous streamflow simulations together with runoff at the cell scale and the contributions of the individual models. Spatial differences in these contributions could help guide tributary flow or soil moisture monitoring. Flood forecasting applications would additionally require evaluation with forecast meteorological inputs and discharge errors at the relevant lead times.

## 5. Conclusions

This study developed a differentiable spatially varying ensemble of GR4J, VMM, and HBV and evaluated it with leave-one-year-out cross-validation in the Jiagenba catchment. In the main experiments, Spatial ENS outperformed the three individual models in all six test years. It increased the pooled flood season NSE from 0.77 for the best individual model (GR4J) to 0.78 and also improved KGE, RMSE, bias, and correlation.

The learned weights varied systematically within the catchment, and their spatial ordering was nearly identical among test year models. Clay fraction, terrain slope, and PET were the dominant inputs to the gating network; GR4J was favored on steeper and more humid cells, whereas VMM and HBV together received more weight on flatter, clay-rich, and drier cells. The repeated initializations supported similar spatial ordering, although the absolute weights varied among fits. Applying the approach to additional catchments and internal gauges will further test its transferability.

**Author Contributions:** Conceptualization, J.D., Z.H. and N.D.; methodology, Z.H. and N.D.; software, Z.H., N.D. and R.M.; validation, Y.D. and Z.H.; formal analysis, Y.D. and Z.H.; investigation, Y.D. and J.D.; resources, J.D.; data curation, Y.D. and Z.H.; writing—original draft preparation, Y.D. and Z.H.; writing—review and editing, J.D., N.D. and R.M.; visualization, Z.H.; supervision, J.D. and N.D.; project administration, J.D.; funding acquisition, N.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** Ningpeng Dong received funding for this study through the National Natural Science Foundation of China (42401053, 42671068), and the Young Elite Scientist Sponsorship Program by CSHE (CSHE-YESS-2026006). This study was also funded by the Research Project of the State Key Laboratory of Water Cycle and Water Security (SKL2025TDGG04), Science and Technology Project of China Huaneng Group (HNKJ25-H190), and the Yunnan Key Laboratory of Water Security (No. 20254916CE340051).

**Data Availability Statement:** The model configurations, trained checkpoints, out-of-sample simulations, learned weight fields, and figure-generation scripts are available from the corresponding authors upon reasonable request. The discharge, meteorological, and static-attribute data are subject to the redistribution conditions of their providers.

**Acknowledgments:** The authors thank the hydrological monitoring staff of Yalong River Hydropower Development Co., Ltd. for providing the discharge observations at the Jiagenba station.

**Conflicts of Interest:** Authors Yi Ding, Zhaohui He and Jintao Ding were employed by Yalong River Hydropower Development Co., Ltd. The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential

conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References


1. Beven, K. *Rainfall–Runoff Modelling: The Primer*; Wiley: Chichester, UK, 2001.
2. Clark, M.P.; Bierkens, M.F.P.; Samaniego, L.; Woods, R.A.; Uijlenhoet, R.; Bennett, K.E.; Pauwels, V.R.N.; Cai, X.; Wood, A.W.; Peters-Lidard, C.D. The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 3427–3440. [CrossRef] [PubMed]
3. Hrachowitz, M.; Savenije, H.H.G.; Blöschl, G.; McDonnell, J.J.; Sivapalan, M.; Pomeroy, J.W.; Arheimer, B.; Blume, T.; Clark, M.P.; Ehret, U.; et al. A decade of Predictions in Ungauged Basins (PUB)—A review. *Hydrol. Sci. J.* **2013**, *58*, 1198–1255. [CrossRef]
4. Perrin, C.; Michel, C.; Andréassian, V. Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.* **2003**, *279*, 275–289. [CrossRef]
5. Lindström, G.; Johansson, B.; Persson, M.; Gardelin, M.; Bergström, S. Development and test of the distributed HBV-96 hydrological model. *J. Hydrol.* **1997**, *201*, 272–288. [CrossRef]
6. Lin, K.; Lv, F.; Chen, L.; Singh, V.P.; Zhang, Q.; Chen, X. A statistical vertically mixed runoff model for regions featured by complex runoff generation process. *Water* **2020**, *12*, 2324. [CrossRef]
7. Beven, K.; Binley, A. The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Process.* **1992**, *6*, 279–298. [CrossRef]
8. Marshall, L.; Sharma, A.; Nott, D. Modeling the catchment via mixtures: Issues of model specification and validation. *Water Resour. Res.* **2006**, *42*, W11409. [CrossRef]
9. Jeremiah, E.; Marshall, L.; Sisson, S.A.; Sharma, A. Specifying a hierarchical mixture of experts for hydrologic modeling: Gating function variable selection. *Water Resour. Res.* **2013**, *49*, 2926–2939. [CrossRef]
10. Samaniego, L.; Kumar, R.; Attinger, S. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resour. Res.* **2010**, *46*, W05523. [CrossRef]
11. Feigl, M.; Thober, S.; Schweppe, R.; Herrnegger, M.; Samaniego, L.; Schulz, K. Automatic regionalization of model parameters for hydrological models. *Water Resour. Res.* **2022**, *58*, e2022WR031966. [CrossRef] [PubMed]
12. Merz, R.; Blöschl, G. Regionalisation of catchment model parameters. *J. Hydrol.* **2004**, *287*, 95–123. [CrossRef]
13. Oudin, L.; Andréassian, V.; Perrin, C.; Michel, C.; Le Moine, N. Spatial proximity, physical similarity, regression and ungauged catchments: A comparison of regionalization approaches. *Water Resour. Res.* **2008**, *44*, W03413. [CrossRef]
14. Razavi, T.; Coulibaly, P. Streamflow prediction in ungauged basins: Review of regionalization methods. *J. Hydrol. Eng.* **2013**, *18*, 958–975. [CrossRef]
15. Parajka, J.; Merz, R.; Blöschl, G. A comparison of regionalisation methods for catchment model parameters. *Hydrol. Earth Syst. Sci.* **2005**, *9*, 157–171. [CrossRef]
16. Wagener, T.; Wheater, H.S. Parameter estimation and regionalization for continuous rainfall–runoff models including uncertainty. *J. Hydrol.* **2006**, *320*, 132–154. [CrossRef]
17. Tsai, W.-P.; Feng, D.; Pan, M.; Beck, H.; Lawson, K.; Yang, Y.; Liu, J.; Shen, C. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nat. Commun.* **2021**, *12*, 5988. [CrossRef] [PubMed]
18. Feng, D.; Liu, J.; Lawson, K.; Shen, C. Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resour. Res.* **2022**, *58*, e2022WR032404. [CrossRef]
19. Shen, C.; Appling, A.P.; Gentine, P.; Bandai, T.; Gupta, H.; Tartakovsky, A.; Baity-Jesi, M.; Fenicia, F.; Kifer, D.; Li, L.; et al. Differentiable modelling to unify machine learning and physical models for geosciences. *Nat. Rev. Earth Environ.* **2023**, *4*, 552–567. [CrossRef]
20. Huynh, N.N.T.; Garambois, P.-A.; Colleoni, F.; Renard, B.; Roux, H.; Demargne, J.; Jay-Allemand, M.; Javelle, P. Learning regionalization using accurate spatial cost gradients within a differentiable high-resolution hydrological model: Application to the French Mediterranean region. *Water Resour. Res.* **2024**, *60*, e2024WR037544. [CrossRef]
21. Song, Y.; Bindas, T.; Shen, C.; Ji, H.; Knoben, W.J.M.; Lonzarich, L.; Clark, M.P.; Liu, J.; van Werkhoven, K.; Lamont, S.; et al. High-resolution national-scale water modeling is enhanced by multiscale differentiable physics-informed machine learning. *Water Resour. Res.* **2025**, *61*, e2024WR038928. [CrossRef]
22. Scholz, F.; Traub, M.; Zarfl, C.; Scholten, T.; Butz, M.V. Fully differentiable, fully distributed rainfall–runoff modeling. *Hydrol. Earth Syst. Sci.* **2025**, *29*, 6257–6283. [CrossRef]
23. Li, P.; Song, Y.; Pan, M.; Lawson, K.; Shen, C. Ensembling differentiable process-based and data-driven models with diverse meteorological forcing datasets to advance streamflow simulation. *Hydrol. Earth Syst. Sci.* **2025**, *29*, 6829–6861. [CrossRef]
24. Han, J.; Miao, C.; Gou, J.; Zheng, H.; Zhang, Q.; Guo, X. A new daily gridded precipitation dataset for the Chinese mainland based on gauge observations. *Earth Syst. Sci. Data* **2023**, *15*, 3147–3161. [CrossRef]

25. Hu, J.; Miao, C.; Su, J.; Zhang, Q.; Gou, J.; Sun, Q. An upgraded high-precision gridded precipitation dataset for the Chinese mainland considering spatial autocorrelation and covariates. *Earth Syst. Sci. Data* **2025**, *17*, 3987–4004. [CrossRef]
26. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2049. [CrossRef]
27. Allen, R.G.; Pereira, L.S.; Raes, D.; Smith, M. *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*; FAO Irrigation and Drainage Paper 56; Food and Agriculture Organization of the United Nations: Rome, Italy, 1998.
28. Addor, N.; Newman, A.J.; Mizukami, N.; Clark, M.P. The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 5293–5313. [CrossRef]
29. Yin, Z.; Lin, P.; Riggs, R.; Allen, G.H.; Lei, X.; Zheng, Z.; Cai, S. A synthesis of global streamflow characteristics, hydrometeorology, and catchment attributes for large-sample river-centric studies. *Earth Syst. Sci. Data* **2024**, *16*, 1559–1590. [CrossRef]
30. Dong, N.; Yang, M.; Wei, J.; Xu, S.; Zhao, Y.; Zhang, X.; Liu, B.; Wu, M.; Wang, H.; Kunstmann, H. Reconstructing China's natural streamflow at 1-km resolution. *Water Resour. Res.* **2026**, *62*, e2025WR042606. [CrossRef]
31. Li, L.; Bisht, G.; Hao, D.; Leung, L.R. Global 1 km land surface parameters for kilometer-scale Earth system modeling. *Earth Syst. Sci. Data* **2024**, *16*, 2007–2032. [CrossRef]
32. Cunge, J.A. On the subject of a flood propagation computation method (Muskingum method). *J. Hydraul. Res.* **1969**, *7*, 205–230. [CrossRef]
33. Zhao, R.-J. The Xinanjiang model applied in China. *J. Hydrol.* **1992**, *135*, 371–381. [CrossRef]
34. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models. Part I—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290. [CrossRef]
35. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **2009**, *377*, 80–91. [CrossRef]
36. Kling, H.; Fuchs, M.; Paulin, M. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *J. Hydrol.* **2012**, *424–425*, 264–277. [CrossRef]
37. Knoben, W.J.M.; Freer, J.E.; Woods, R.A. Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 4323–4331. [CrossRef]
38. Jiang, S.; Zheng, Y.; Solomatine, D. Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. *Geophys. Res. Lett.* **2020**, *47*, e2020GL088229. [CrossRef]
39. Krapu, C.; Borsuk, M. A differentiable hydrology approach for modeling with time-varying parameters. *Water Resour. Res.* **2022**, *58*, e2021WR031377. [CrossRef]
40. Wang, C.; Jiang, S.; Zheng, Y.; Han, F.; Kumar, R.; Rakovec, O.; Li, S. Distributed hydrological modeling with physics-encoded deep learning: A general framework and its application in the Amazon. *Water Resour. Res.* **2024**, *60*, e2023WR036170. [CrossRef]
41. Huynh, N.N.T.; Garambois, P.-A.; Renard, B.; Colleoni, F.; Monnier, J.; Roux, H. A distributed hybrid physics–AI framework for learning corrections of internal hydrological fluxes and enhancing high-resolution regionalized flood modeling. *Hydrol. Earth Syst. Sci.* **2025**, *29*, 3589–3613. [CrossRef]
42. Ouyang, W.; Ye, L.; Chai, Y.; Ma, H.; Chu, J.; Peng, Y.; Zhang, C. A differentiable, physics-based hydrological model and its evaluation for data-limited basins. *J. Hydrol.* **2025**, *649*, 132471. [CrossRef]
43. Zhang, C.; Li, H.; Hu, Y.; Shen, D.; Xu, B.; Chen, M.; Chu, W.; Li, R. A differentiability-based processes and parameters learning hydrologic model for advancing runoff prediction and process understanding. *J. Hydrol.* **2025**, *661*, 133594. [CrossRef]
44. Hock, R. Temperature index melt modelling in mountain areas. *J. Hydrol.* **2003**, *282*, 104–115. [CrossRef]
45. Jennings, K.S.; Winchell, T.S.; Livneh, B.; Molotch, N.P. Spatial variation of the rain–snow temperature threshold across the Northern Hemisphere. *Nat. Commun.* **2018**, *9*, 1148. [CrossRef] [PubMed]
46. Zhang, Y.; Liu, S.; Ding, Y. Observed degree-day factors and their spatial variation on glaciers in western China. *Ann. Glaciol.* **2006**, *43*, 301–306. [CrossRef]
47. Ponce, V.M.; Yevjevich, V. Muskingum–Cunge method with variable parameters. *J. Hydraul. Div.* **1978**, *104*, 1663–1667. [CrossRef]
48. Leuning, R.; Zhang, Y.Q.; Rajaud, A.; Cleugh, H.; Tu, K. A simple surface conductance model to estimate regional evaporation using MODIS leaf area index and the Penman–Monteith equation. *Water Resour. Res.* **2008**, *44*, W10419. [CrossRef]
49. Oudin, L.; Hervieu, F.; Michel, C.; Perrin, C.; Andréassian, V.; Anctil, F.; Loumagne, C. Which potential evapotranspiration input for a lumped rainfall–runoff model? Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. *J. Hydrol.* **2005**, *303*, 290–306. [CrossRef]
50. Kratzert, F.; Klotz, D.; Herrnegger, M.; Sampson, A.K.; Hochreiter, S.; Nearing, G.S. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resour. Res.* **2019**, *55*, 11344–11354. [CrossRef]
51. Nearing, G.S.; Kratzert, F.; Sampson, A.K.; Pelissier, C.S.; Klotz, D.; Frame, J.M.; Prieto, C.; Gupta, H.V. What role does hydrological science play in the age of machine learning? *Water Resour. Res.* **2021**, *57*, e2020WR028091. [CrossRef]

52. Sawadekar, K.; Song, Y.; Pan, M.; Beck, H.; McCrary, R.; Ullrich, P.; Lawson, K.; Shen, C. Improving differentiable hydrologic modeling with interpretable forcing fusion. *J. Hydrol.* **2025**, *659*, 133320. [CrossRef]
53. Song, Y.; Knoben, W.J.M.; Clark, M.P.; Feng, D.; Lawson, K.; Sawadekar, K.; Shen, C. When ancient numerical demons meet physics-informed machine learning: Adjoint-based gradients for implicit differentiable modeling. *Hydrol. Earth Syst. Sci.* **2024**, *28*, 3051–3077. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.