hydrology

MDPI

Article

Exploring the Effects of Runoff Generation–Routing Structural Combinations in Conceptual Hydrological Models

Runhua Qiu [1], Lu Chen [1,2,*], Yang Song [1], Pengcheng Li [1] and Sihao Cao [1]

[1] College of Water Conservancy and Civil Engineering, Xizang Agricultural and Animal Husbandry University, Linzhi 860000, China; q975648591@163.com (R.Q.); jzzjbsdsg@163.com (Y.S.); 15181705405@163.com (P.L.); caosihao_1@163.com (S.C.)

[2] School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

* Correspondence: chen_lu@hust.edu.cn

Abstract

Structural uncertainty limits the reliability of conceptual hydrological models, yet the effect of pairing runoff-generation and routing structures remains poorly understood. We evaluated 46 models from the Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) in the Xunhe and Hanzhong catchments of the upper Han River Basin and recombined the modules of five representative models into 20 off-diagonal configurations. Validation KGE values for the 46 base models ranged from 0.166 to 0.806 in Xunhe and from 0.043 to 0.830 in Hanzhong. Only M58 and M60 improved KGE in both catchments: M58 increased KGE by 1.68% in Xunhe and 5.98% in Hanzhong, whereas M60 increased it by 5.05% and 3.72%, respectively. Most other recombinations preserved or reduced performance. The improved combinations retained contributions from runoff pathways with different response timescales and showed greater sensitivity to routing parameters; degraded combinations tended to be dominated by groundwater runoff and showed weak routing-parameter sensitivity. These convergent patterns provide indirect evidence that correspondence between runoff composition and routing response may influence recombined-model performance. Because structural compatibility was not measured directly and only two catchments were examined, the proposed relationship requires validation across broader hydroclimatic and structural settings.

Keywords: runoff-generation–routing coupling; modular hydrological modeling; model structural uncertainty; Sobol sensitivity analysis

Check for updates

Academic Editor: Andrea Petroselli

Received: 10 July 2026
Revised: 17 August 2026
Accepted: 18 August 2026
Published: 22 August 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.

1. Introduction

Hydrological models extend process understanding beyond observation networks and underpin catchment water-balance analysis, climate-impact assessment, and prediction in ungauged basins [1–4]. Yet building models that are both scientifically defensible and operationally reliable remains a central challenge [5]. A major reason is structural uncertainty: conceptual models encode runoff generation, storage, evapotranspiration, and routing differently [6–8]. Model choice can consequently reflect historical legacy as much as demonstrated adequacy [9], while a defensible assessment requires evidence that the adopted structure is appropriate for its intended behavior [10].

Equifinality makes structural assessment difficult because different models can reproduce similar outlet hydrographs while simulating contrasting internal states and fluxes [11–13]. Calibration to discharge alone can therefore reward 'right answers for the wrong reasons' [14]. Stepwise hypothesis testing, multiple working hypotheses, and process-based diagnostics help reveal which assumptions—not parameter compensation alone—control performance [15–17]. The problem is thus not only which whole model performs best, but why a structural hypothesis succeeds or fails.

Flexible modeling frameworks make controlled structural tests possible. FUSE recombines alternative formulations of soil moisture, evapotranspiration, runoff generation, and baseflow [18]; SUPERFLEX represents models as configurations of storage, transfer, and connection elements [19]; and SUMMA separates conservation equations from process parameterizations [20]. MARRMoT supplies standardized state-space implementations of established conceptual rainfall–runoff models [21,22]. By representing model structures as combinations of explicit process hypotheses, these frameworks support reproducible comparisons that are difficult when models are treated as indivisible systems.

Evidence from global applications nevertheless shows that the value of a process representation is conditional. Evapotranspiration formulations differ in effectiveness across climates, and remotely sensed evapotranspiration constraints interact with internal model organization [23,24]. Snow and cold-region formulations are similarly conditioned by environmental setting [25,26], while large-sample comparisons show that structural uncertainty persists across diverse catchments [27]. These findings imply that module performance cannot always be inferred from isolated evaluation; interactions between connected process representations also require testing.

Runoff generation and routing provide a particularly interpretable pair for such a test. Runoff-generation schemes partition precipitation into components with different response timescales, whereas routing schemes translate, attenuate, and delay those components at the outlet. A routing mechanism that is effective for one runoff composition may therefore be poorly matched to another. This dependency is broadly relevant across rainfall-driven catchments and can be examined without relying on processes restricted to specific settings, such as snow accumulation and melt.

Regional research in the upper Han River Basin reinforces the need to distinguish process representation from whole-model performance. Vegetation greening has been associated with reduced water yield and greater hydrological-drought risk [28]. Spatially explicit VIC parameterization combined with multi-gauge calibration improved streamflow simulation and parameter identifiability [29], and hybrid Xinanjiang–random-forest–wavelet methods improved daily streamflow simulation at Shiquan station [30]. These studies demonstrate the regional importance of process assumptions, observational constraints, and model integration, but none factorially exchanged runoff-generation and routing modules across a common set of conceptual models.

Two gaps remain. Globally, modular frameworks enable controlled comparison, but little work has tested runoff-generation outputs with routing mechanisms from other models. Regionally, upper Han River studies have modified whole models without isolating this interaction. We define structural compatibility as correspondence between runoff composition and routing response. Our novelty is a standardized factorial exchange diagnosed through performance, flood hydrographs, runoff composition, parameter sensitivity, and a supplementary complexity adjustment. The transferable workflow evaluates interaction indirectly; it neither quantifies compatibility nor identifies a universally optimal pair.

We apply this design in the Xunhe and Hanzhong catchments and ask: (1) how do the 46 base structures differ in validation performance; (2) which cross-model pairings improve, preserve, or degrade performance; and (3) do hydrographs, runoff composition, and parameter sensitivity provide consistent but indirect explanations for those changes?

2. Methodology

2.1. Study Area and Data

Two catchments in the upper Han River Basin were examined: Xunhe and Hanzhong (Figure 1). Xunhe is a left-bank tributary upstream of the Danjiangkou Reservoir and drains 6703 km^2. The 218.1 km river has a reported mean channel gradient of 2.9%; the forested catchment spans the climatic transition along the southern Qinling Mountains and receives strongly seasonal precipitation. The Hanzhong catchment is the 9329 km^2 drainage area upstream of the Hanzhong hydrometric station. Its elevation ranges from 489 to 3502 m, its terrain and river network are complex, and precipitation is concentrated mainly from July to September. These contrasts provide two regional settings for evaluating whether structural rankings and recombination effects persist across catchments.

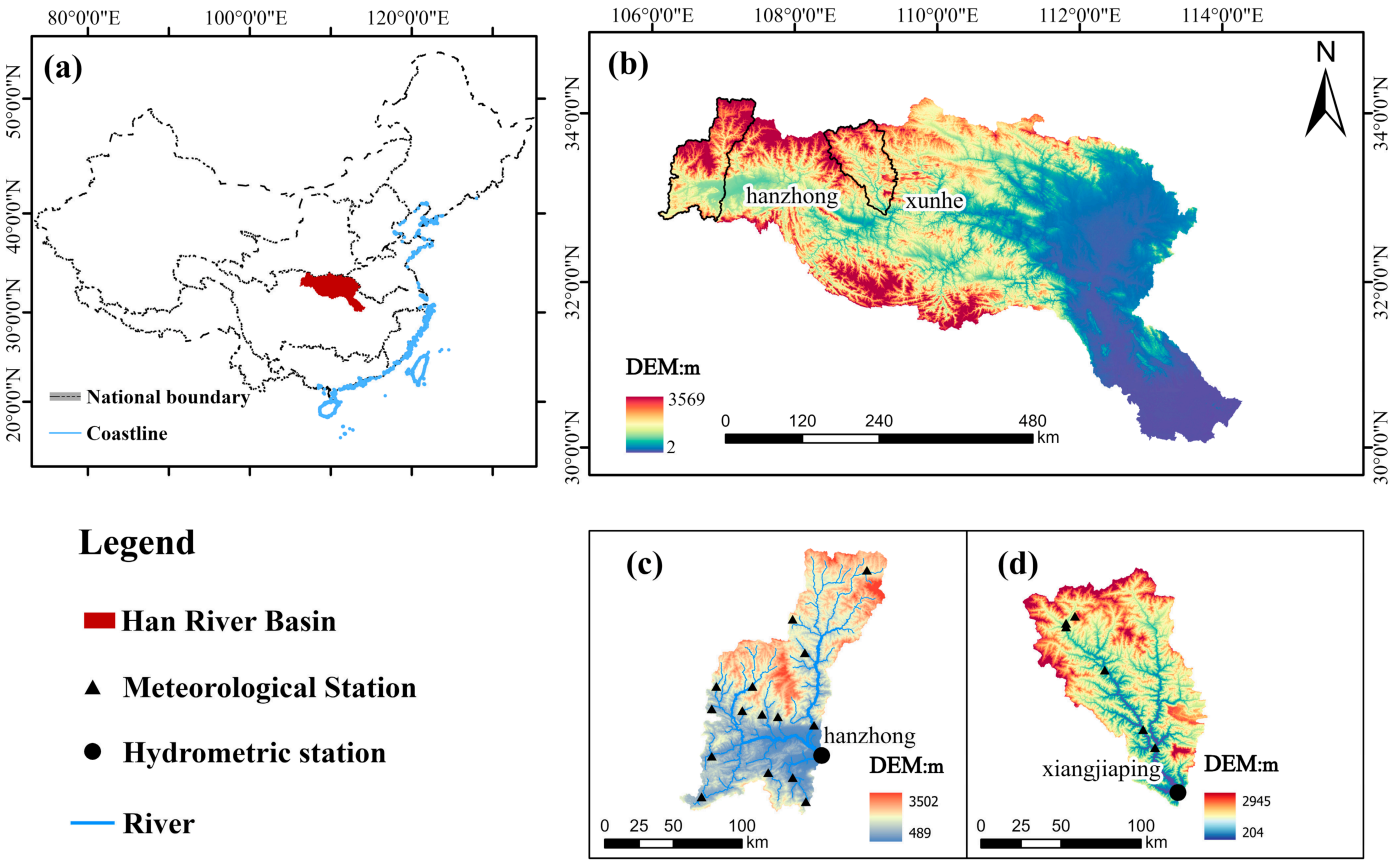

Figure 1. Study setting and monitoring network. (**a**) Location of the Han River Basin in China; (**b**) elevation and locations of the Hanzhong and Xunhe study catchments in the upper Han River Basin; (**c**) river network, 15 meteorological stations, and the Hanzhong hydrometric station in the Hanzhong catchment; and (**d**) river network, six meteorological stations, and the Xiangjiaping hydrometric station in the Xunhe catchment. Triangles and circles denote meteorological and hydrometric stations, respectively; the panels highlight the contrasting terrain and drainage structures of the two catchments.

Daily hydrometeorological records covered 1 January 2000–31 December 2020 for Xunhe and 1 January 2001–31 December 2022 for Hanzhong. Observed daily streamflow for Xunhe and Hanzhong was obtained from the Xiangjiaping and Hanzhong hydrometric stations, respectively (Figure 1). The station discharge records had been preprocessed to catchment-average runoff depth (mm d^{-1}) before model application.

Daily precipitation and station-based observed potential evapotranspiration (PET) for Xunhe were obtained from six meteorological stations: Shagou, Dongjiangfang, Chongjiagou, Tongmu, Dongjiangkou, and Liangheguan. The corresponding inputs for Hanzhong were obtained from 15 stations: Changgouhe, Liuhou, Baohe, Lianshui, Zhangjiaba, Yudongzi, Liuba (柳坝), Changba, Ningqiang, Xishenba, Taibai, Liuba (留坝), Zhangjiahe, Da'an, and Daheba. Catchment-average precipitation and PET were calculated by Thiessen-polygon weighting. Daily mean air temperature was taken from the national daily observational dataset compiled by the National Meteorological Information Center from more than 2400 national basic, reference, and ordinary stations. Model inputs were expressed as precipitation (mm d^{-1}), PET (mm d^{-1}), mean air temperature (°C), and streamflow depth (mm d^{-1}).

The preprocessed catchment-scale model-input series were screened for missing and physically impossible values before modeling. The supplied precipitation, PET, temperature, and streamflow series contained no missing values; precipitation, PET, and streamflow were non-negative, while physically plausible subzero air temperatures were retained. The same daily records and preprocessing conventions were used for every model within each catchment.

A fixed chronological indexing protocol was used so that all base and recombined models were calibrated and evaluated on identical samples within each catchment. For Xunhe, 1 January–30 December 2000 was used for model warm-up, 31 December 2000–13 September 2015 for calibration, and 14 September 2015–1 January 2020 for validation; the final 365 days were outside the scored periods. For Hanzhong, 1 January–31 December 2001 was used for warm-up, 1 January 2002–25 May 2017 for calibration, and 26 May 2017–31 December 2021 for validation; the final 365 days were outside the scored periods. The chronological design avoided random interleaving of calibration and validation observations and retained a later, non-overlapping interval for out-of-sample evaluation.

2.2. Flexible Architecture for Runoff-Generation–Routing Recombination

Conceptual hydrological models differ substantially in their representations of runoff generation and routing. To test how these representations interact, we decomposed each model into a runoff-generation module that transforms meteorological inputs into runoff components and a routing module that propagates those components to the outlet. Standardized flux interfaces preserved mass balance while allowing modules from different models to be paired. The public MARRMoT v2.1 code base is implemented in MATLAB-R2022a [21,22]. We benchmarked the 46 established structures M1–M46; the subsequently added IHM19 structure (M47) was not included so that the benchmark set remained consistent with the original 46-model MARRMoT experiment. This modular formulation provides a common environment for comparing the separate and joint effects of runoff-generation and routing structures.

Figure 2 summarizes the controlled experimental workflow used to benchmark the base models, construct the full-factorial module combinations, and diagnose changes in validation performance.

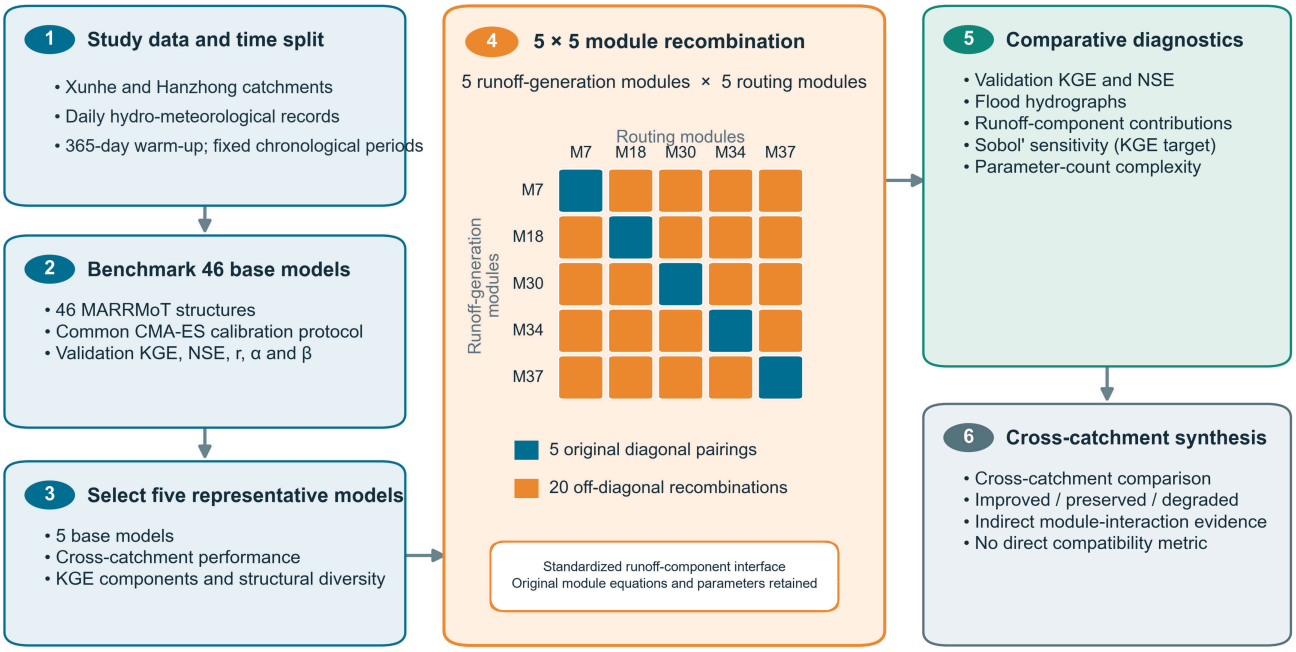

Figure 2. Methodological workflow for evaluating runoff-generation–routing structural combinations. Daily hydrometeorological records from the Xunhe and Hanzhong catchments were subjected to a 365-day warm-up and fixed chronological calibration and validation periods. Forty-six MARRMoT base models were calibrated with CMA-ES and evaluated using KGE, NSE, and the KGE components. Five representative models were selected using cross-catchment performance, KGE-component behavior, and structural-diversity considerations, then crossed in a 5 × 5 factorial design containing five original diagonal configurations and 20 off-diagonal recombinations. Validation metrics, hydrographs, runoff-component contributions, Sobol sensitivity, and a parameter-count complexity check were interpreted jointly across catchments. This synthesis provides indirect evidence about runoff-generation–routing correspondence rather than a direct measure of structural compatibility.

2.2.1. Runoff-Generation Process

Conceptual hydrological models differ substantially in their representations of runoff-generation processes, including whether and how they account for snowmelt, vegetation interception, surface depression storage, soil-layer partitioning, soil moisture deficit, saturation-excess runoff, and infiltration-excess runoff. Under a discrete time-step framework, the runoff-generation module satisfies the following water balance relationship:

$$R_e(t) = P(t) - E_a(t) - \Delta S_p(t) \quad (1)$$

where $P(t)$ denotes the precipitation input during time step t, $E_a(t)$ represents actual evapotranspiration, and $\Delta S_p(t)$ denotes the change in the total water storage within the runoff-generation module during the same time step.

$$\Delta S_p(t) = S_p(t) - S_p(t-1) \quad (2)$$

where $\Delta S_p(t) = S_p(t) - S_p(t-1)$ denotes the total water storage of the runoff generation module at the end of the time step. Based on the underlying runoff generation mechanisms, the water outputs from the runoff generation modules within the MARRMoT library are standardized into three categories of fluxes: surface runoff (R_{SRUN}), interflow (R_{INT}), and groundwater flow (R_{GW}). These fluxes satisfy the following mass conservation equation:

$$R_e(t) = R_{SRUN} + R_{INT} + R_{GW} \quad (3)$$

The runoff components generated by different runoff-generation schemes and their corresponding formation mechanisms are determined by the original model structure. Differences exist not only in the classification of runoff components among models, but also in the underlying process representations. Some models explicitly distinguish among surface runoff, interflow, and groundwater runoff, whereas others represent runoff using only fast and slow flow components. Even when the same runoff component is simulated, its generation mechanism may differ across models. For example, fast runoff may originate from either infiltration-excess runoff or saturation-excess runoff processes. When a runoff-generation module contains only two runoff components, namely surface runoff and groundwater runoff, the following relationship applies:

$$R_{INT} = 0 \quad (4)$$

2.2.2. Routing Process

The primary function of the flow-routing module is to propagate, delay, and regulate the runoff components generated by the runoff-generation process, thereby representing catchment response processes such as hillslope routing, interflow concentration, and groundwater recession. Different routing structures describe runoff propagation in markedly different ways, including mechanisms such as linear reservoir routing, unit hydrograph routing, and multi-path flow routing. For the same set of runoff-generation outputs, different routing structures may produce substantially different catchment response behaviors. Consequently, when the runoff components generated by the runoff-generation process and their characteristic response timescales are inconsistent with the propagation mechanisms represented by the routing structure, the resulting model combination may distort the runoff-routing process. For example, combining a runoff-generation structure dominated by fast-flow components with a routing structure characterized primarily by slow groundwater storage and release may lead to excessive smoothing and delayed flood peaks, thereby reducing the model's ability to reproduce the observed dynamics of streamflow response.

(1) The dual-path routing scheme partitions runoff into fast-flow and slow-flow components, which are routed through independent fast-response and slow-response storage reservoirs, respectively. The fast pathway primarily represents surface runoff and interflow processes, whereas the slow reservoir is intended to represent delayed flow components such as groundwater baseflow and other slow recession processes.

$$Q_T = F_f(R_{SRUN} + R_{INT}) + F_s(R_{GW}) \quad (5)$$

where F_f and F_s denote the routing functions for fast-flow and slow-flow components, respectively. The two runoff pathways are routed independently through their corresponding routing schemes, and the total catchment outflow, Q_T, is obtained by summing the routed outputs from both pathways.

(2) In the lumped routing scheme, all runoff components are first aggregated into a single total inflow and are subsequently propagated and regulated through a common routing unit, without distinguishing independent routing pathways for different runoff components. This structure emphasizes the overall catchment response, requires fewer parameters, and is relatively simple in formulation.

$$Q_T = F(R_{GW} + R_{SRUN} + R_{INT}, \theta) \quad (6)$$

where F denotes the routing function. The aggregated runoff is routed and regulated through a shared routing structure, such as a unit hydrograph or a linear reservoir, to generate the catchment outlet discharge Q_T.

(3) In the component-specific routing scheme, different runoff components are routed and regulated through separate routing pathways, thereby preserving the response characteristics and characteristic timescales of each runoff component. The total catchment outlet discharge is obtained by summing the outflows from all routing pathways:

$$Q_T = F_{SR}(R_{SRUN}) + F_I(R_{INT}) + F_G(R_{GW}) \quad (7)$$

where F_{SR} represents the routing function for surface runoff, F_I denotes the routing function for interflow, and F_G denotes the routing function for groundwater runoff. This routing structure explicitly accounts for the distinct propagation behaviors of different runoff components and enables their individual contributions to the total streamflow response to be retained.

2.2.3. Classification of Runoff-Generation and Routing Processes

The MARRMoT framework enables structural comparison and module recombination across conceptual hydrological models through a unified state-space representation and standardized flux interfaces. Although individual models differ in their process representations and parameterization schemes, their fundamental functions can generally be described as two sequential components: the transformation of precipitation into runoff and the subsequent routing of runoff through the catchment. Accordingly, this study partitions model structures into two functional components: a runoff generation process and a routing process.

The runoff generation process describes the formation of runoff from precipitation through interception, evapotranspiration, soil water storage, and water partitioning. The associated structural elements primarily control precipitation inputs, water storage dynamics, and runoff production. In contrast, the routing process describes the propagation, attenuation, and aggregation of runoff components within the catchment. The corresponding structural elements mainly govern flow travel times and the characteristics of the outlet response. Following this functional classification, storage elements and associated flux functions involved in the transformation of precipitation into runoff are assigned to the runoff generation process, whereas storage elements, routing functions, and transfer structures responsible for runoff propagation and flow organization are assigned to the routing process.

Based on how runoff components are organized after entering the routing module, the routing structures implemented in MARRMoT can be broadly classified into three categories: dual-path routing, lumped routing, and component-specific routing (Table 1). This classification preserves the original structural characteristics of each model while providing a consistent basis for cross-model comparisons of runoff generation and routing modules, as well as for subsequent module recombination experiments.

Table 1. Runoff-component and routing classifications for the 46 MARRMoT models used in this study. A check mark indicates that the corresponding runoff component or routing category is represented in the model.

Model ID	Model Name	Surface	Interflow	Groundwater	Lumped	Dual-Path	Component-Specific
M1	collie1	✓			✓		
M2	wetland	✓		✓	✓		
M3	collie2	✓	✓		✓		

Table 1. *Cont.*

Model ID	Model Name	Surface	Interflow	Groundwater	Lumped	Dual-Path	Component-Specific
M4	newzealand1	✓	✓	✓	✓		
M5	ihacres	✓		✓		✓	
M6	alpine1	✓	✓		✓		
M7	gr4j	✓		✓		✓	
M8	us1	✓	✓		✓		
M9	susannah1	✓	✓	✓			✓
M10	susannah2	✓	✓		✓		
M11	collie3	✓	✓	✓			✓
M12	alpine2	✓	✓	✓	✓		
M13	hillslope		✓	✓	✓		
M14	topmodel	✓	✓	✓			✓
M15	plateau	✓		✓		✓	
M16	newzealand2	✓	✓	✓	✓		
M17	penman	✓		✓	✓		
M18	simhyd	✓	✓	✓			✓
M19	australia	✓	✓	✓			✓
M20	gsfb	✓		✓	✓		
M21	flexb	✓		✓		✓	
M22	vic	✓	✓	✓	✓		
M23	lascam	✓	✓	✓	✓		
M24	mopex1	✓		✓		✓	
M25	tcm			✓	✓		
M26	flexi	✓		✓		✓	
M27	tank	✓	✓	✓			✓
M28	xinanjiang	✓	✓	✓			✓
M29	hymod	✓		✓		✓	
M30	mopex2	✓		✓		✓	
M31	mopex3	✓	✓	✓		✓	
M32	mopex4	✓	✓	✓		✓	
M33	sacramento	✓	✓	✓			✓
M34	flexis	✓		✓		✓	
M35	mopex5	✓	✓	✓		✓	
M36	modhydrolog	✓	✓	✓			✓
M37	hbv	✓		✓	✓		
M38	tank2	✓	✓	✓			✓
M39	mcrm	✓	✓	✓	✓		
M40	smar	✓		✓		✓	
M41	nam	✓	✓	✓			✓
M42	hycymodel	✓	✓	✓			✓
M43	gsmsocont	✓	✓	✓			✓
M44	echo	✓		✓		✓	
M45	prms	✓	✓	✓			✓
M46	classic	✓	✓	✓			

2.3. Parameter Calibration and Model Evaluation

All model configurations were calibrated with the Covariance Matrix Adaptation Evolution Strategy (CMA-ES), a derivative-free stochastic optimizer suited to nonlinear, multimodal objective functions [31,32]. Applying one optimization framework across structures reduced procedural differences in the comparison.

Validation performance was evaluated with Nash–Sutcliffe efficiency (NSE) and Kling–Gupta efficiency (KGE). NSE emphasizes squared streamflow errors, whereas KGE separates correlation, variability, and bias contributions [33,34]. The two metrics were

therefore reported together rather than interpreted as interchangeable scores. NSE is defined as follows:

$$NSE = 1 - \frac{\sum_{t=1}^{T} (Q_{sim,t} - Q_{obs,t})^2}{\sum_{t=1}^{T} (Q_{obs,t} - \overline{Q}_{obs,t})^2} \tag{8}$$

where $Q_{sim,t}$ and $Q_{obs,t}$ denote the simulated and observed streamflow at time step t, respectively, and $\overline{Q}_{obs,t}$ represents the mean observed streamflow. T denotes the total number of time steps within the validation period.

The KGE is defined as follows:

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{9}$$

where r is the linear correlation coefficient between simulated and observed streamflow, α is the variability ratio, and β is the bias ratio. These components are defined as follows:

$$\alpha = \frac{\sigma_{sim}}{\sigma_{obs}}, \beta = \frac{\mu_{sim}}{\mu_{obs}} \tag{10}$$

where σ_{sim} and σ_{obs} are the standard deviations of simulated and observed streamflow, respectively, and μ_{sim} and μ_{obs} are their corresponding mean values. Compared to NSE, KGE explicitly decomposes model performance into three components: temporal dynamics, variability, and overall bias. This decomposition facilitates a more comprehensive assessment of model performance in terms of hydrograph shape, flow variability, and water balance representation. Therefore, both NSE and KGE were employed in this study to reduce the potential bias associated with relying on a single performance metric. The combined evaluation results were further used to identify benchmark models with strong overall simulation capabilities, providing a basis for subsequent modular decomposition and structural recombination analyses.

Considering that the number of parameters in the recombined models is generally greater than that in the original models, evaluating model performance solely based on KGE is insufficient to determine whether performance improvements arise from structural optimization or merely from the increase in model complexity. Therefore, a complexity-control approach based on the Structural Risk Minimization (SRM) principle was adopted to assess model complexity [35–37]. In general, the SRM estimator can be expressed as follows:

$$\hat{R}_e = \bar{r} R_e \tag{11}$$

where $\bar{r}$ is the penalty factor, with $\bar{r} \geq 1$. The penalty factor increases as model complexity grows relative to the size of the available dataset. R_e denotes the empirical error, and $\hat{R}_e$ represents the empirical error adjusted for model complexity. Structural Risk Minimization (SRM) is rooted in statistical learning theory and adopts the following formulation for the penalty factor $\bar{r}$:

$$r_{SRM} = \left[1 - \sqrt{\frac{d}{n} - \frac{d}{n}\ln\left(\frac{d}{n}\right) + \frac{\ln n}{2n}}\right]^{-1} \tag{12}$$

where d is the number of calibrated parameters used here as a proxy for model capacity, and n is the sample size.

2.4. Method for Parameter Sensitivity Analysis

Sensitivity analysis quantifies how uncertainty in model parameters contributes to uncertainty in a selected output. Local or one-at-a-time methods may miss parameter interactions in nonlinear conceptual models [38,39]. We therefore used variance-based Sobol

global sensitivity analysis, which partitions output variance into individual and interaction contributions [40]. KGE was used as the target response for the analysis.

The Sobol sensitivity analysis is a variance-based global sensitivity analysis technique that decomposes the variance of model outputs into contributions from individual parameters and their interactions. The first-order sensitivity index (S_i) quantifies the effect of varying a single parameter i on model output, whereas the total-order sensitivity index (S_{T_i}) quantifies both the direct effect of parameter i and the effects arising from its interactions with all other parameters.

$$S_i = \frac{V_K}{V(y)} \tag{13}$$

$$S_{T_i} = \frac{V_{N,K}}{V(y)} \tag{14}$$

where $V(y)$ denotes the total variance of the model output y; V_K represents the variance contribution attributable to parameter K, and $V_{N,K}$ denotes the variance contribution from all factors other than parameter K.

The target variable in Sobol sensitivity analysis can be either simulated streamflow or a model performance metric. In this study, KGE was selected as the target variable for sensitivity analysis.

3. Results and Analysis

3.1. Overall Performance of the Base Models

This section presents the performance of the 46 conceptual hydrological models included in the MARRMoT framework. Unless otherwise stated, all figures and analyses are based on the validation period, and model abbreviations follow those listed in Table 1. Overall, the 46 models demonstrated satisfactory simulation performance in both the Hanzhong and Xunhe catchments (Figures 3 and 4). For the Hanzhong catchment, KGE values ranged from 0.043 to 0.830. Among the 46 models, 41 (89.1%) achieved KGE values greater than 0.5, while only seven models (15.2%) reached a relatively high level of performance (KGE > 0.75). Models M7, M32, M36, M45, and M34 ranked among the best-performing structures, with M7 exhibiting the highest overall performance, achieving a KGE of 0.830 and an NSE of 0.687. For the Xunhe catchment, KGE values ranged from 0.166 to 0.806. A total of 44 models (96%) achieved KGE values above 0.5, and nine models (approximately 20%) exceeded 0.75. Models M30, M40, M7, M18, and M12 showed comparatively strong performance. Among them, M30 achieved the highest KGE (0.806) and NSE (0.652), indicating strong overall predictive skill and a good representation of catchment hydrological dynamics.

Overall, most models achieved satisfactory baseline performance in both catchments, although the proportion of high-performing models remained relatively limited, highlighting the substantial influence of model structure on simulation performance. At the same time, several models exhibited relatively high correlation coefficients (r), yet their KGE values were considerably reduced because either the bias ratio (β) or the variability ratio (α) deviated substantially from the ideal value of 1. For example, in the Xunhe catchment, both M19 and M11 achieved correlation coefficients greater than 0.80. However, the KGE values of these models were only 0.532 and 0.482, respectively, owing to a pronounced bias in β (1.385) for M19 and in α (1.460) for M11. These results indicate that correlation alone is insufficient to fully characterize model performance. The ability of a model to reproduce the observed water balance and flow variability, as reflected by β and α, is equally important for achieving good overall performance [34]. Because KGE has no universal zero-value or satisfactory benchmark, the thresholds above are used descriptively rather than as evidence of structural adequacy [34].

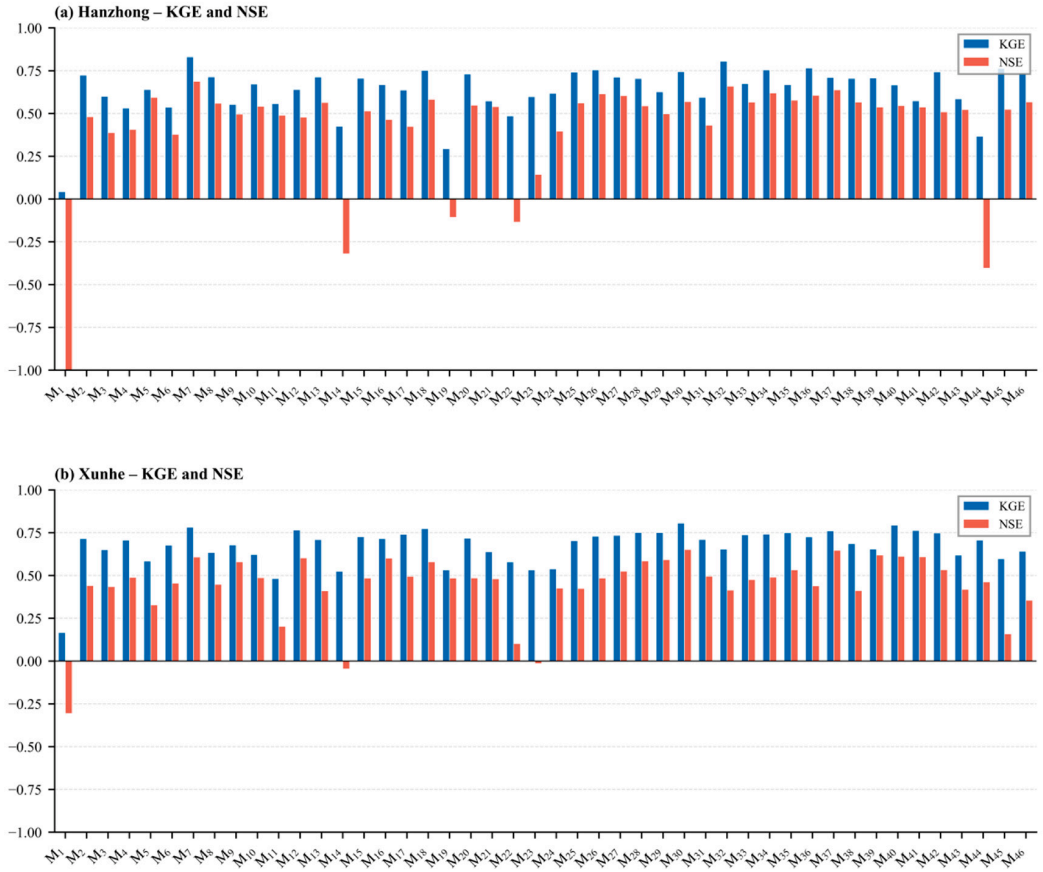

Figure 3. Validation-period KGE and NSE for the 46 MARRMoT models in Hanzhong and Xunhe. KGE and NSE are dimensionless; model identifiers are defined in Table 1.

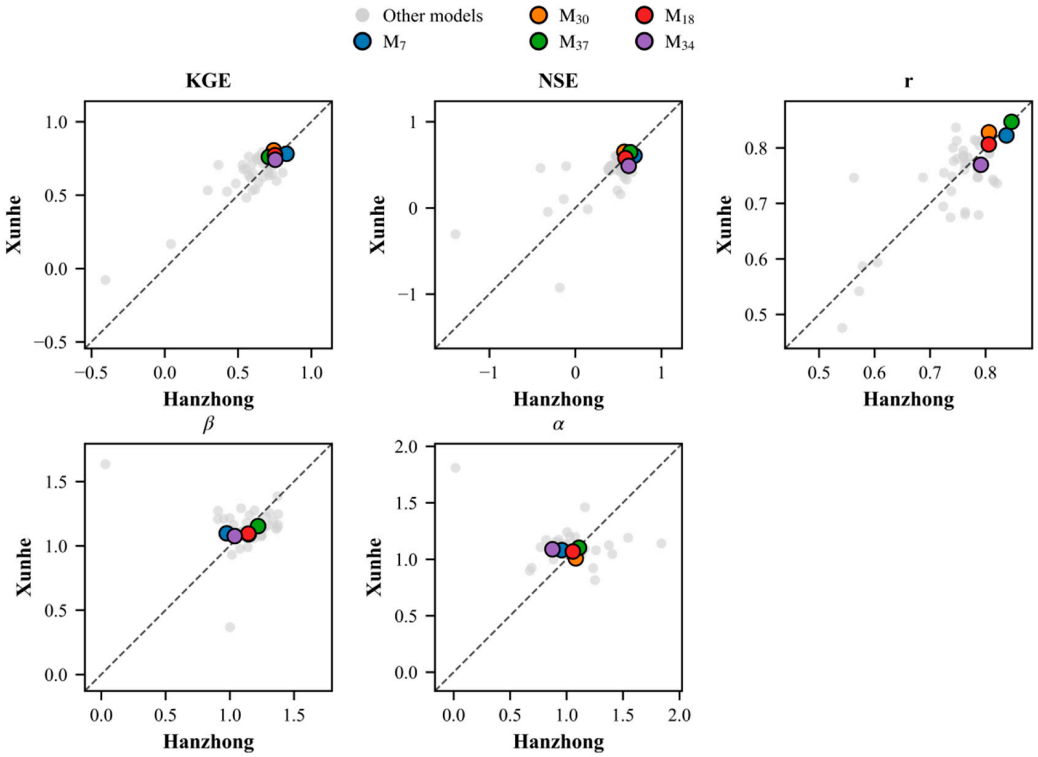

Figure 4. Cross-catchment comparison of validation KGE, NSE, correlation (r), variability ratio (α), and bias ratio (β) for the 46 base models. Dashed reference lines denote the ideal values for the KGE components.

3.2. Selection of Representative Models and Recombination Design

Five base models were selected by jointly considering four features: KGE and NSE in both catchments, consistency of performance across catchments, the KGE components r, α, and β, and diversity in runoff-generation and routing structures. These features were assessed qualitatively without numerical weights or a composite ranking, and the selection was completed before the off-diagonal recombination results were interpreted. The resulting set comprised M7 (GR4J), M18 (SIMHYD), M30 (MOPEX-2), M34 (FLEX-IS), and M37 (HBV). This multi-criterion selection was intended to retain both competitive performance and structural contrast, which is important when model hypotheses rather than only whole-model rankings are compared [9,10,19].

The five selected models also represent distinct structural hypotheses (Table 2). SIMHYD generates three runoff components—surface runoff, interflow, and groundwater runoff—but only groundwater runoff is routed through a linear reservoir, while the other components contribute directly to catchment outflow. MOPEX-2 and FLEX-IS primarily generate surface runoff and groundwater recharge and use fast and slow routing pathways to represent rapid responses and recession behavior; FLEX-IS additionally uses a unit hydrograph to organize response timescales. HBV does not explicitly distinguish runoff sources; runoff generation and release are controlled jointly by soil-moisture, upper, and lower reservoirs, with a lumped unit hydrograph representing catchment-scale routing. GR4J generates effective rainfall through a production store, partitions it into fast and slow components using a fixed ratio, and routes them through two unit-hydrograph pathways with an additional groundwater-exchange process. The selected set therefore spans three-source, two-source, and soil-moisture-based runoff-generation schemes together with component-specific, dual-path, lumped, and unit-hydrograph routing structures.

Table 2. MARRMoT identifiers and structural characteristics of the five selected base models.

Model (ID)	Runoff Generation	Routing
GR4J (M7)	Soil-moisture-based generation	Dual unit hydrographs
SIMHYD (M18)	Three-source	Component-specific routing
MOPEX-2 (M30)	Two-source	Dual-path routing
FLEX-IS (M34)	Two-source	Dual-path routing with unit hydrograph routing
HBV (M37)	Soil-moisture-based generation	Lumped routing

We paired each of the five runoff-generation modules with each of the five routing modules, yielding a 5 × 5 factorial design (Figure 5). The five diagonal pairings were the original models; the 20 off-diagonal pairings were the recombined models analyzed here. Original module equations and parameter definitions were retained, and the modules were linked through the standardized runoff-component interface. Each off-diagonal structure was independently recalibrated using the same CMA-ES objective, parameter-bound convention, warm-up, calibration period, and validation period used for the base structures. This design isolates changes associated with replacing the runoff-generation module, the routing module, or their interaction.

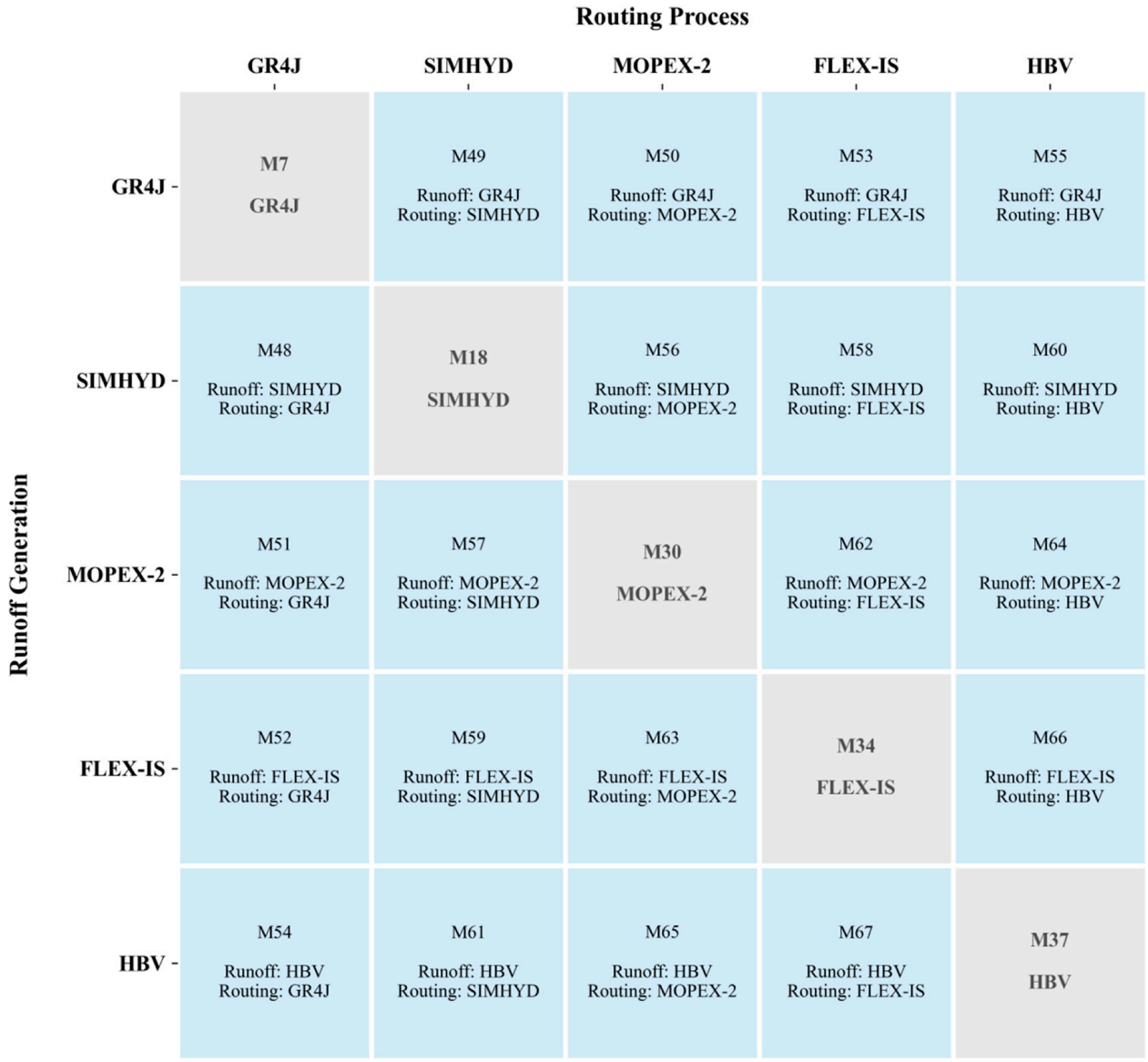

Figure 5. Full-factorial pairing of runoff-generation modules (rows) and routing modules (columns) from the five selected base models. Diagonal cells are the original models (M7, M18, M30, M34, and M37); the 20 off-diagonal cells are recombined models (M48–M67). Gray diagonal cells denote the five original models, whereas light-blue off-diagonal cells denote the 20 recombined configurations.

3.3. Performance Comparison of Recombined Models

Validation KGE and NSE varied substantially among the 20 off-diagonal combinations (Figure 6). The terms improved, approximately preserved, and degraded are used descriptively: improvement or degradation denotes a consistent direction of change in the relevant KGE and NSE comparison across both catchments, whereas mixed cross-catchment responses are described as approximately preserved. No additional tolerance threshold was imposed.

Only M58 (SIMHYD runoff generation with FLEX-IS routing) and M60 (SIMHYD runoff generation with HBV routing) improved KGE in both catchments. Relative to M18, M58 increased KGE by 5.98% in Hanzhong and 1.68% in Xunhe; M60 increased it by 3.72% and 5.05%, respectively. M60 showed the strongest cross-catchment validation performance among the recombined models.

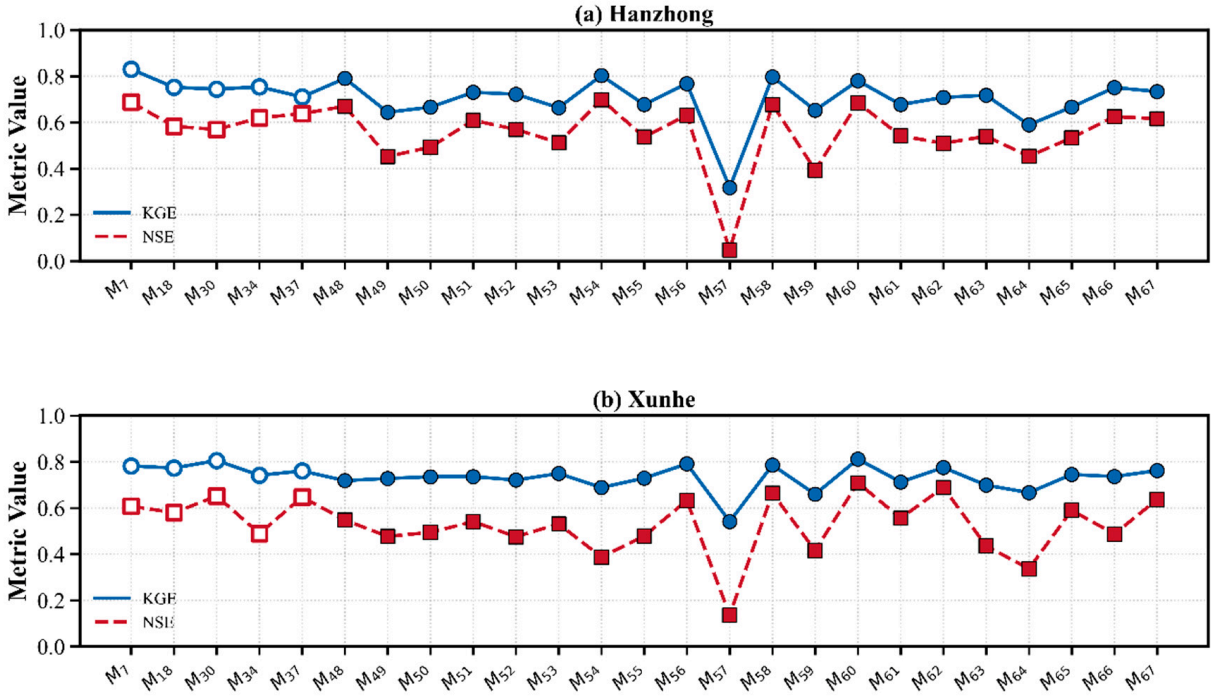

Figure 6. Validation-period KGE and NSE for the five selected base models (open markers) and 20 recombined models (filled markers) in (**a**) Hanzhong and (**b**) Xunhe. Key combinations are M57 (MOPEX-2 runoff generation + SIMHYD routing), M58 (SIMHYD runoff generation + FLEX-IS routing), M59 (FLEX-IS runoff generation + SIMHYD routing), and M60 (SIMHYD runoff generation + HBV routing). Both metrics are dimensionless.

M48, M56, M62, M66, and M67 remained close to their parent-model performance, whereas M57 (MOPEX-2 runoff generation with SIMHYD routing), M59 (FLEX-IS runoff generation with SIMHYD routing), and most combinations using GR4J runoff generation deteriorated. Among the degraded combinations, M57 showed the most pronounced decline, with KGE markedly lower than the corresponding parent-model values in both catchments. These outcomes show that modules that perform effectively in their original models do not necessarily retain that behavior after transfer.

Overall, improvement was limited to a small subset of the 20 recombined models, while most combinations approximately preserved or degraded performance. The following diagnostics therefore compare hydrographs, runoff composition, and parameter sensitivity to determine whether the performance classes also differ in their internal response patterns.

3.4. Diagnostic Differences Among Recombined Models

3.4.1. Flood Hydrographs

To further evaluate the ability of the recombined models to reproduce catchment response dynamics, M58 (SIMHYD runoff generation + FLEX-IS routing) and M60 (SIMHYD runoff generation + HBV routing), which exhibited improved simulation performance, were selected as representative cases, with the original SIMHYD model (M18) serving as a benchmark. Figure 7 presents the simulated continuous streamflow series and selected flood events during the validation period. Overall, all three models were able to reproduce the general patterns of streamflow variability, but clear differences were observed in their representation of flood peaks and recession behavior. Compared to M18, the recombined models M58 and M60 exhibited more stable performance in simulating both peak flows and recession processes, resulting in an improved representation of catchment hydrological dynamics.

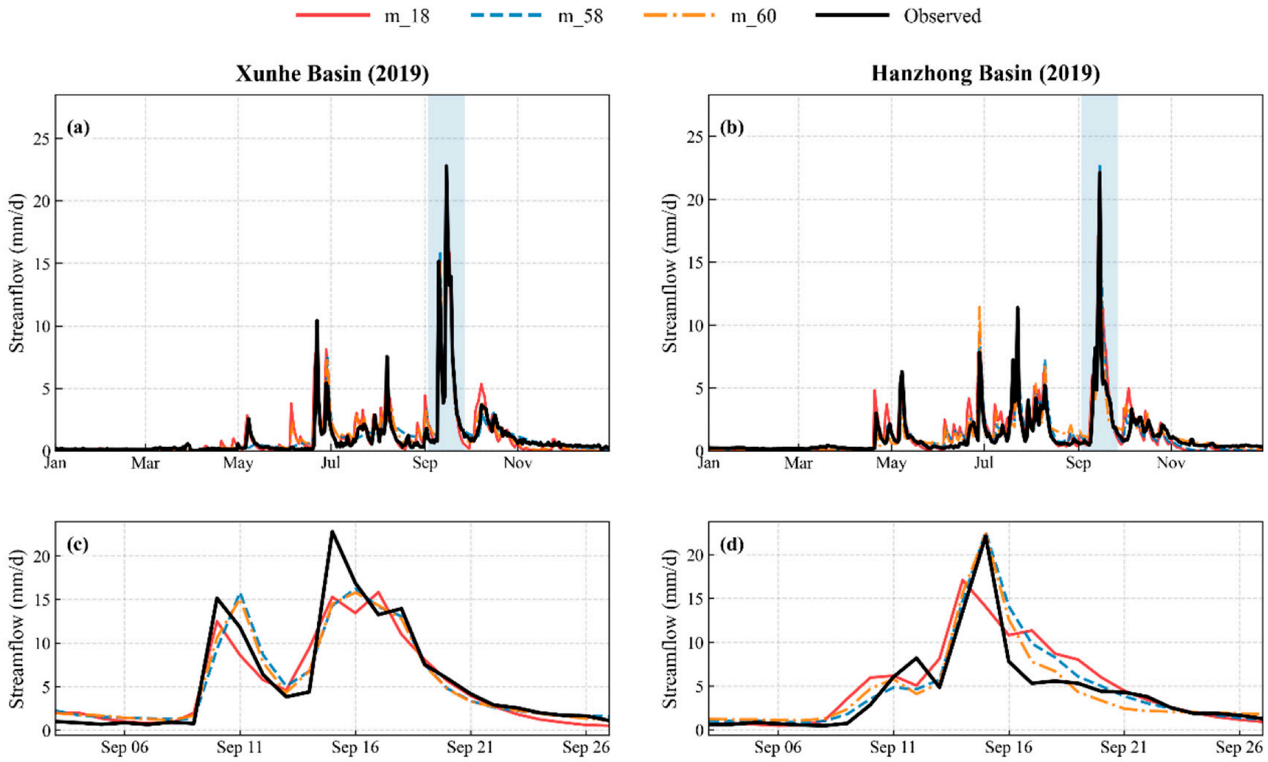

Figure 7. Observed and simulated daily streamflow for the original SIMHYD model (M18) and the improved recombinations M58 (SIMHYD runoff generation + FLEX-IS routing) and M60 (SIMHYD runoff generation + HBV routing). (**a**) Xunhe continuous hydrograph for 2019; (**b**) Hanzhong continuous hydrograph for 2019; (**c**) enlarged Xunhe flood event; and (**d**) enlarged Hanzhong flood event; streamflow is in mm d^{-1}.

The continuous streamflow simulations (Figure 7a,b) show that M18 tended to overestimate peak discharge during several flood events. In contrast, M58 and M60 reduced the occurrence of unrealistically sharp peaks, producing flood responses that were more consistent with the observations while retaining the ability to capture high-flow events. Among the three models, M60 showed the closest agreement with the observed hydrographs during high-flow periods. For the selected flood events (Figure 7c,d), all three models accurately reproduced the timing of peak occurrence. However, M18 exhibited noticeable biases in peak magnitude and showed discrepancies during the recession phase. M58 improved the simulation of peak flows to some extent, although local deviations remained. By comparison, M60 provided a more faithful representation of peak magnitude, peak timing, and recession characteristics, yielding the best overall agreement with the observed flood hydrographs.

To further investigate the causes of performance deterioration following structural recombination, M57 (MOPEX-2 runoff generation + SIMHYD routing) and M59 (FLEX-IS runoff generation + SIMHYD routing), both of which exhibited reduced simulation performance, were selected as representative cases. The original SIMHYD model (M18) was used as a reference for comparison. Figure 8 presents the simulated continuous streamflow series and selected flood events during the validation period. Overall, M57 and M59 exhibited two contrasting response patterns, characterized by overly rapid and overly damped responses, respectively. Although both models were constructed by combining individually well-performing runoff-generation and routing structures, their ability to reproduce flood peaks and recession dynamics declined after recombination.

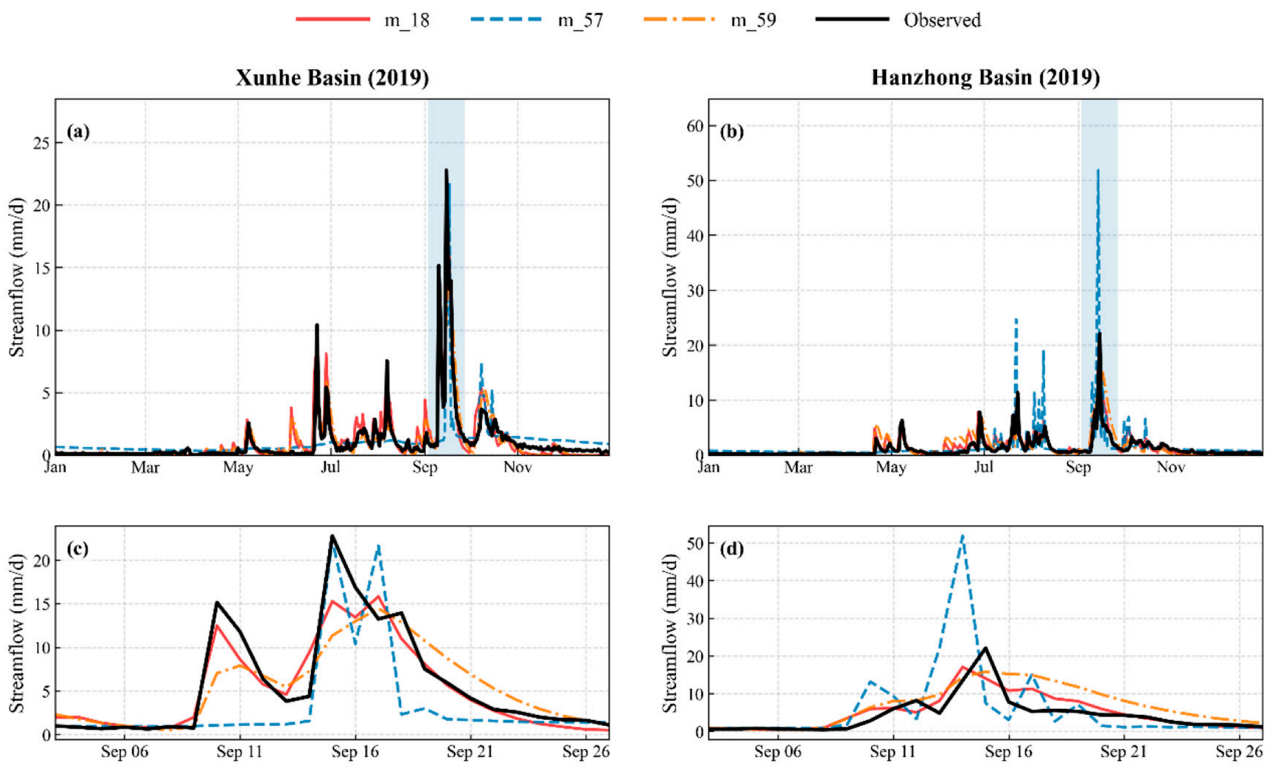

Figure 8. Observed and simulated daily streamflow for M18 and the degraded recombinations M57 (MOPEX-2 runoff generation + SIMHYD routing) and M59 (FLEX-IS runoff generation + SIMHYD routing). (**a**) Xunhe continuous hydrograph for 2019; (**b**) Hanzhong continuous hydrograph for 2019; (**c**) enlarged Xunhe flood event; and (**d**) enlarged Hanzhong flood event; streamflow is in mm d^{-1}.

The continuous streamflow simulations (Figure 8a,b) reveal distinct deficiencies in the two models. M57 frequently produced sharp spikes during high-flow periods, resulting in substantial overestimation of some flood peaks and increased variability in the simulated hydrograph. In contrast, M59 exhibited excessive smoothing, leading to systematic underestimation of many flood peaks. This tendency was particularly evident during periods of rapid flow rise and recession, where the model struggled to reproduce the observed streamflow dynamics. For the selected flood events (Figure 8c,d), both models generally captured the timing of peak occurrence, but substantial errors remained in the simulation of peak magnitude and recession behavior. In both catchments, M57 displayed an excessively rapid response, producing exaggerated peak flows and unrealistically sharp flood peaks. Moreover, the simulated recession limbs deviated markedly from the observations. By contrast, M59 showed pronounced smoothing effects, which reduced its ability to reproduce double-peaked flood events. As a result, peak magnitudes were consistently underestimated and recession processes tended to be delayed. In double-peaked events, M59 failed to adequately distinguish between successive peaks, whereas M57 struggled to maintain stable hydrograph behavior, resulting in substantial distortion of flood-peak characteristics.

3.4.2. Runoff-Component Analysis

To further investigate the sources of performance differences among the recombined models, the average contribution of each runoff component to total streamflow during the validation period was calculated. The results are presented in Figure 9. Overall, the recombined models exhibited markedly different runoff-composition patterns, indicating that the way runoff-generation and routing structures are combined can substantially influence the relative contributions of different flow pathways to the catchment response.

However, no simple relationship was found between runoff composition and model performance. In some models, multiple runoff components contributed to the routing process, yet this did not lead to a noticeable improvement in simulation performance. Conversely, other models achieved satisfactory performance despite being dominated by only a limited number of flow pathways. These findings suggest that the number of runoff components alone is not a primary determinant of model performance.

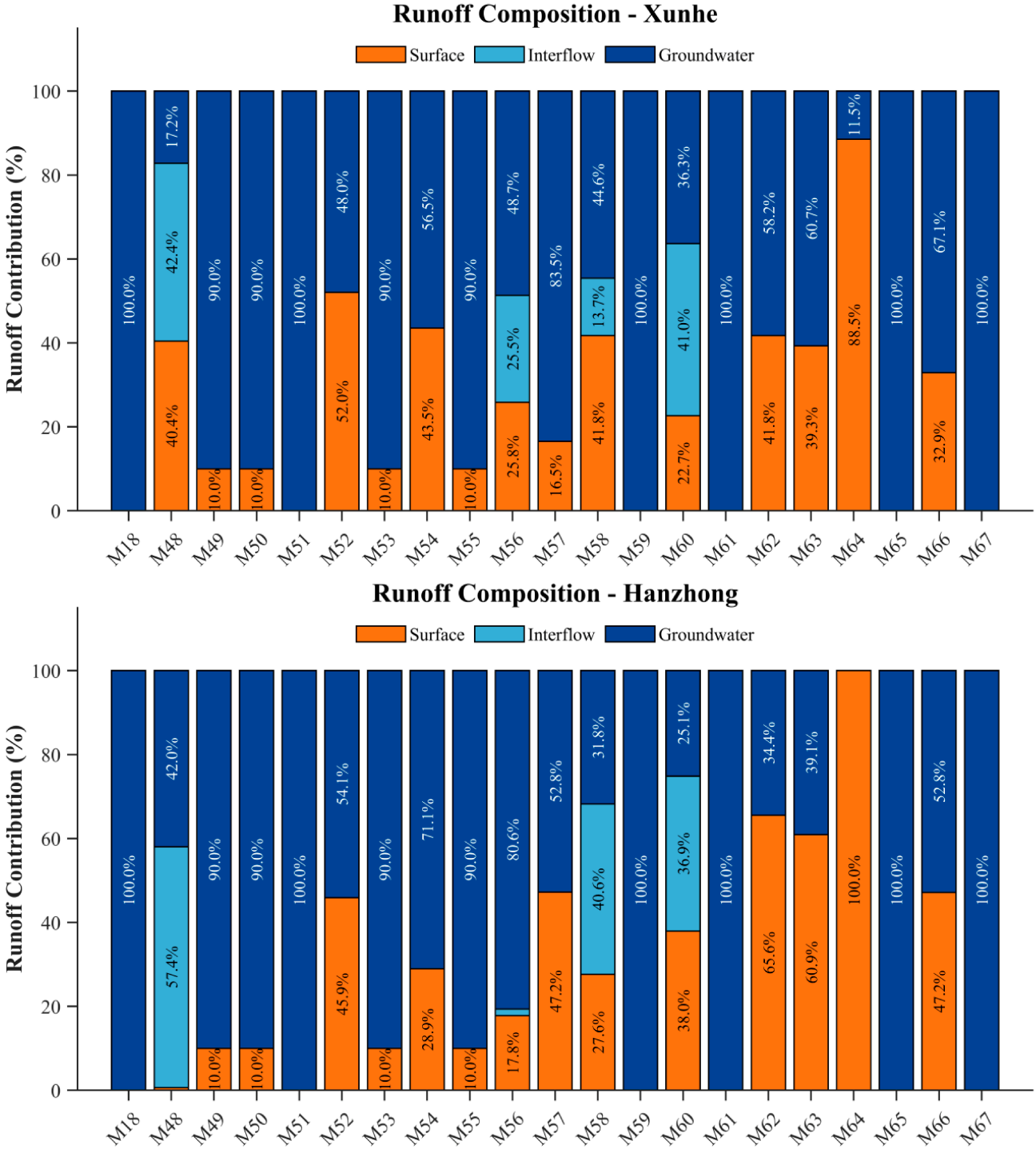

Figure 9. Mean validation-period contributions of surface runoff, interflow, and groundwater runoff in Xunhe and Hanzhong. Key identifiers are M18 (SIMHYD), M57 (MOPEX-2 runoff generation + SIMHYD routing), M58 (SIMHYD runoff generation + FLEX-IS routing), M59 (FLEX-IS runoff generation + SIMHYD routing), and M60 (SIMHYD runoff generation + HBV routing).

Using the SIMHYD runoff-generation structure as an example, the original model M18 exhibited a clear dominance of groundwater runoff in both catchments, with groundwater contributions reaching 100% of total streamflow. Although the SIMHYD structure is capable of generating three runoff components—surface runoff, interflow, and groundwater runoff—the calibration results indicate that the routing response was ultimately dominated by the groundwater pathway. Consequently, only a limited number of runoff pathways actively contributed to the simulated catchment response, while the influence of fast-response components on total streamflow remained relatively weak. In contrast, the two recombined models that showed the greatest performance improvements, M58 (SIMHYD runoff generation + FLEX-IS routing) and M60 (SIMHYD runoff generation + HBV routing), exhibited a much more balanced runoff composition. In the Xunhe catchment, the contributions of surface runoff, interflow, and groundwater runoff in M58 were 41.8%, 13.7%, and 44.6%, respectively, whereas the corresponding contributions in M60 were 22.7%, 41.0%, and 36.3%. Similar patterns were observed in the Hanzhong catchment. Compared with M18, both recombined models were characterized by the simultaneous contribution of multiple runoff pathways to the routing response. At the same time, both models achieved improvements in flood-peak simulation and overall KGE performance, suggesting that the routing structures altered the way runoff components associated with different response timescales were expressed in total streamflow. By contrast, the models exhibiting the most pronounced performance deterioration, M57 (MOPEX-2 runoff generation + SIMHYD routing) and M59 (FLEX-IS runoff generation + SIMHYD routing), showed a strong dominance of groundwater runoff. For M59, groundwater runoff accounted for nearly 100% of total streamflow in both catchments, whereas the corresponding contributions in M57 reached 83.5% and 52.8%. Although the original MOPEX-2 and FLEX-IS structures both include fast- and slow-response flow pathways, recombination caused a substantial redistribution of runoff toward the groundwater pathway. As a result, the overall routing response became increasingly controlled by slow-response processes.

These results suggest that model performance is related to runoff composition, but this relationship cannot be explained simply by the number of runoff components involved. Rather, it appears to depend on the extent to which different runoff components continue to contribute to the routing response. In this study, the better-performing recombined models were generally characterized by the joint contribution of runoff components operating across multiple response timescales, whereas the poorer-performing models tended to exhibit a concentration of runoff contributions within a single dominant flow pathway. Because runoff composition reflects only the model output and does not directly reveal the underlying control mechanisms, it is insufficient for explaining the processes responsible for the observed performance differences. Therefore, the following section employs Sobol sensitivity analysis to further investigate the relative roles of runoff-generation and routing parameters across different model structures.

3.4.3. Parameter Sensitivity Analysis

Figure 10 presents the first-order and total-order sensitivity indices for the parameters of different model structures. Based on the parameter classifications shown in Table 3, substantial differences in sensitivity patterns can be observed among the models, providing additional insight into the mechanisms underlying the performance changes in the recombined models.

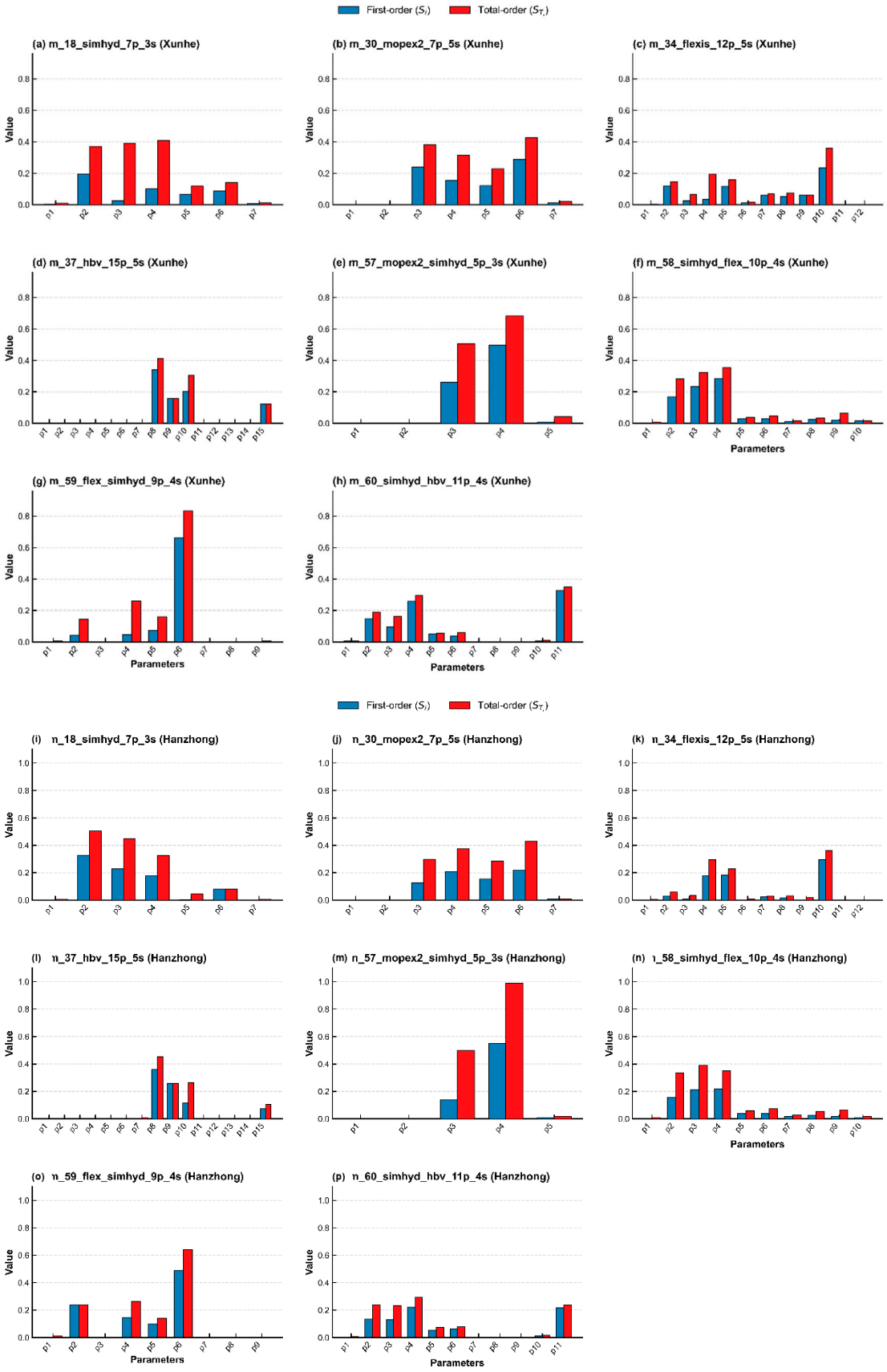

Figure 10. First-order (Si) and total-order (STi) Sobol sensitivity indices for model parameters in Xunhe (**a**–**h**) and Hanzhong (**i**–**p**). Parameter-to-process assignments are given in Table 3; the dominant bars indicate which runoff-generation or routing parameters most strongly controlled KGE in each structure.

Table 3. Assignment of calibrated parameters to runoff-generation and routing processes in the selected original and recombined models.

Model	Parameter	Parameter Name	Process Type
M18	P1–P6 P7	insc, coeff, sq, smsc, sub, crak k	Runoff Generation Routing
M30	P1–P4, P6 P5, P7	tcrit, ddf, s2max, tw, se tu, tc	Runoff Generation Routing
M34	P1–P5, P10–P12 P6–P9	smax, beta, d, percmax, lp, imax, tt, ddf nlagf, nlags, kf, ks	Runoff Generation Routing
M37	P1–P10 P11–P15	tt, tti, ttm, cfr, cfmax, whc, cflux, fc, lp, beta k0, alpha, perc, k1, maxbas	Runoff Generation Routing
M57	P1–P4 P5	tcrit, ddf, s2max, tw k	Runoff Generation Routing
M58	P1–P6 P7–P10	INSC, COEFF, SQ, SMSC, SUB, CRAK nlagf, nlags, kf, ks	Runoff Generation Routing
M59	P1–P8 P9	smax, beta, d, percmax, lp, imax, tt, ddf k	Runoff Generation Routing
M60	P1–P6 P7–P11	INSC, COEFF, SQ, SMSC, SUB, CRAK K0, ALPHA, PERC, K1, MAXBAS	Runoff Generation Routing

For the original model M18 and the degraded recombinations M57 and M59, parameter sensitivity was concentrated primarily in a small number of runoff-generation parameters, whereas routing parameters exhibited consistently low sensitivity. This indicates that variations in model output were controlled mainly by water-partitioning processes during runoff generation, while the routing structures exerted only a limited influence on simulated streamflow. In these models, changes in routing contributed little to the overall runoff response, and model behavior was therefore dominated largely by the runoff-generation module. In contrast, the better-performing recombinations M58 and M60 retained relatively high sensitivity to runoff-generation parameters while also showing increased sensitivity to several routing parameters. In particular, the HBV routing parameters in M60 exhibited relatively high first-order and total-order indices, indicating that routing played a more influential role in shaping model behavior.

Sensitivity analysis does not measure structural compatibility directly and is interpreted here only as supporting evidence. When considered together with the runoff-composition and hydrograph analyses, however, a consistent pattern emerges. M58 and M60 maintained contributions from runoff pathways operating at different response timescales and reproduced flood-peak propagation and recession dynamics more effectively. By contrast, M57 and M59 concentrated a larger share of runoff in the groundwater pathway and reproduced flood peaks less successfully. The sensitivity results are consistent with these observations: the improved models were jointly influenced by runoff-generation and routing processes, whereas the degraded models were controlled predominantly by water partitioning during runoff generation. The greater influence of routing parameters in M58 and M60 therefore provides a plausible process-based explanation for their improved performance. When routing participates effectively in propagating and attenuating multiple runoff components, the model is better able to represent catchment response dynamics.

4. Discussion

4.1. Structural Compatibility Between Runoff Generation and Routing Structures and Its Potential Influence on Model Performance

The results demonstrate that module recombination does not necessarily lead to improved model performance through the simple combination of individually well-performing components. Among the 20 recombined models, only a small number of combinations achieved consistent performance improvements, whereas most either maintained performance levels comparable to their parent models or exhibited varying degrees of deterioration. This suggests that runoff generation and routing processes in conceptual hydrological models are not independent of one another, but are linked through a degree of structural coupling. Consequently, model behavior depends not only on the capability of individual modules to represent specific processes, but also on how effectively different modules function together. It should be emphasized that the concept of structural compatibility discussed in this study is not defined through a direct quantitative metric. Instead, it is inferred from multiple lines of evidence, including changes in model performance, runoff composition characteristics, and parameter sensitivity patterns.

This interpretation agrees with broader calls to evaluate models through internal process diagnostics rather than outlet performance alone [10,12,13,17]. It also provides a module-scale analog to the interaction reported by Zheng et al. [29], who found that spatially explicit parameterization and multi-gauge calibration produced joint benefits and trade-offs in the upper Han River Basin. Their experiment concerns a distributed VIC configuration, not runoff-generation–routing exchange, so the comparison is conceptual rather than direct. Our factorial design adds evidence that the effect of one design element can depend on the structural and informational context supplied by another.

During the model recombination process, incorporating runoff generation or routing modules with more comprehensive process representations inevitably increases the number of model parameters. In hydrological modeling, an increase in parameterization complexity can often lead to improvements in simulation performance. To objectively evaluate the benefits of model recombination, it is therefore necessary to assess model performance while accounting for differences in model complexity. To this end, the Structural Risk Minimization (SRM) criterion was adopted as a complexity-aware evaluation metric. SRM incorporates a complexity penalty into the mean squared error (MSE), thereby balancing model fit against model complexity, where n denotes the sample size. As shown in Table 4, after accounting for model complexity, the recombined model M60 achieved the lowest structural risk values in both the Xunhe and Hanzhong catchments, indicating the best overall performance, followed by M58. These results suggest that the performance improvements observed in the recombined models cannot be attributed solely to the increase in parameter numbers. Instead, they are more likely associated with changes in structural representation and process organization introduced through model recombination, rather than with increased parameterization alone.

Table 4. Complexity-penalized mean squared error (MSE) for M18, M58, and M60 in the two catchments. n is the sample size; MSE is reported in (mm d^{-1})^2.

Basin	Model	n	$\hat{R}e$ (mm d^{-1})^2
Xunhe	M_{18}	7306	1.342839526
	M_{58}		0.912815927
	M_{60}		0.864759452
Hanzhong	M_{18}	7670	1.619772011
	M_{58}		1.219919362
	M_{60}		1.114722319

4.2. Implications and Limitations of Flexible Modeling Frameworks

Traditional conceptual hydrological models are typically calibrated and evaluated as integrated systems, making it difficult to determine which specific process representations are responsible for differences in model performance. By decomposing and recombining runoff generation and routing modules within the MARRMoT framework, this study enables individual structural components to be compared under a common modeling environment, thereby providing a clearer understanding of how different process representations contribute to overall model behavior.

The results demonstrate that model performance alone is insufficient for identifying the structural mechanisms underlying model behavior. In contrast, a modular analysis can more effectively reveal the processes that control flood peak dynamics, water balance, and catchment response characteristics. While overall model performance, as measured by KGE, was found to be strongly influenced by water-partitioning parameters within the runoff generation module (e.g., soil water storage capacity, interception, and nonlinear runoff generation coefficients), the sensitivity of routing parameters was often relatively low. Nevertheless, the routing structure itself played a critical role in determining whether flood dynamics could be reproduced successfully. This finding highlights the importance of considering not only parameter sensitivity but also structural organization when evaluating hydrological models.

Although the effectiveness of the recombination approach was demonstrated in two representative catchments, several limitations remain. First, the analysis was conducted in only two catchments, and the applicability of the identified structural relationships across different climatic and physiographic settings requires further investigation. Second, this study focused primarily on runoff generation and routing processes, whereas other important processes, such as evapotranspiration, snowmelt, and vegetation dynamics, were not included in the recombination analysis. Third, model evaluation was based mainly on streamflow observations. Future studies could incorporate multiple sources of information, including soil moisture, evapotranspiration, and groundwater observations, within a multi-objective evaluation framework to provide stronger constraints on internal model processes. Finally, the concept of structural compatibility proposed in this study is currently supported primarily by indirect evidence derived from model performance, runoff composition, and parameter sensitivity analyses. A dedicated quantitative metric has not yet been developed. Consequently, the present results should be viewed as empirical evidence supporting the existence of structural compatibility rather than as a definitive demonstration of the concept. Future research should focus on developing quantitative measures of structural compatibility for modular model recombination, thereby providing a stronger theoretical basis for evaluating and interpreting compatibility relationships among hydrological process representations.

5. Conclusions

Using MARRMoT, we evaluated 46 conceptual hydrological models in the Xunhe and Hanzhong catchments and tested 20 off-diagonal runoff-generation–routing combinations derived from five representative structures. The main conclusions are:

(1) The base models showed substantial structural and performance heterogeneity. GR4J, SIMHYD, MOPEX-2, FLEX-IS, and HBV were selected to retain cross-catchment performance and structural contrast.

(2) Recombination did not generally improve performance. Only M58 (SIMHYD–FLEX-IS) and M60 (SIMHYD–HBV) improved KGE in both catchments, whereas M57 (MOPEX-2–SIMHYD), M59 (FLEX-IS–SIMHYD), and several other pairings deteriorated.

(3) Improved combinations retained runoff pathways with different response timescales and showed greater routing-parameter sensitivity; degraded combinations tended to be dominated by one pathway and by runoff-generation parameters. Thus, runoff-component count alone did not explain performance.

(4) M58 and M60 retained lower complexity-adjusted MSE than M18, so their gains were not attributable solely to additional parameters. Together, the diagnostics provide indirect—not definitive—evidence that correspondence between runoff composition and routing response may influence performance; a direct compatibility metric and tests in more catchments are still needed.

Author Contributions: Conceptualization, R.Q. and L.C.; methodology, R.Q.; software, R.Q.; validation, R.Q.; formal analysis, R.Q.; investigation, R.Q.; data curation, R.Q.; writing—original draft preparation, R.Q.; writing—review and editing, L.C., Y.S., P.L. and S.C.; visualization, R.Q.; supervision, L.C.; project administration, L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Xizang Natural Science Foundation, grant number XZ202401ZR0044, and the High-Level Talent Program of the Tibet Autonomous Region, grant number XZNMXYRCDWJS-2024-001.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Berghuijs, W.R.; Sivapalan, M.; Woods, R.A.; Savenije, H.H.G. Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales. *Water Resour. Res.* **2014**, *50*, 5638–5661. [CrossRef]
2. Guo, D.; Westra, S.; Maier, H.R. Impact of evapotranspiration process representation on runoff projections from conceptual rainfall–runoff models. *Water Resour. Res.* **2017**, *53*, 435–454. [CrossRef]
3. Zhang, Y.; Chiew, F.H.S.; Liu, C.; Tang, Q.; Xia, J.; Tian, J.; Kong, D.; Li, C. Can remotely sensed actual evapotranspiration facilitate hydrological prediction in ungauged regions without runoff calibration? *Water Resour. Res.* **2020**, *56*, e2019WR026236. [CrossRef]
4. Singh, V.P.; Woolhiser, D.A. Mathematical modeling of watershed hydrology. *J. Hydrol. Eng.* **2002**, *7*, 270–292. [CrossRef]
5. Sivapalan, M. The secret to 'doing better hydrological science': Change the question! *Hydrol. Process.* **2009**, *23*, 1391–1396. [CrossRef]
6. Beven, K. A manifesto for the equifinality thesis. *J. Hydrol.* **2006**, *320*, 18–36. [CrossRef]
7. Refsgaard, J.C.; van der Sluijs, J.P.; Brown, J.; van der Keur, P. A framework for dealing with uncertainty due to model structure error. *Adv. Water Resour.* **2006**, *29*, 1586–1597. [CrossRef]
8. Beven, K.J. *Rainfall–Runoff Modelling: The Primer*, 2nd ed.; Wiley-Blackwell: Chichester, UK, 2012.
9. Addor, N.; Melsen, L.A. Legacy, rather than adequacy, drives the selection of hydrological models. *Water Resour. Res.* **2019**, *55*, 378–390. [CrossRef]
10. Gupta, H.V.; Clark, M.P.; Vrugt, J.A.; Abramowitz, G.; Ye, M. Towards a comprehensive assessment of model structural adequacy. *Water Resour. Res.* **2012**, *48*, W08301. [CrossRef]
11. Beven, K.; Freer, J. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J. Hydrol.* **2001**, *249*, 11–29. [CrossRef]
12. Gupta, H.V.; Wagener, T.; Liu, Y. Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrol. Process.* **2008**, *22*, 3802–3813. [CrossRef]
13. Bouaziz, L.J.E.; Fenicia, F.; Thirel, G.; de Boer-Euser, T.; Buitink, J.; Brauer, C.C.; De Niel, J.; Dewals, B.J.; Drogue, G.; Grelier, B.; et al. Behind the scenes of streamflow model performance. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 1069–1095. [CrossRef]
14. Kirchner, J.W. Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resour. Res.* **2006**, *42*, W03S04. [CrossRef]
15. Fenicia, F.; Savenije, H.H.G.; Matgen, P.; Pfister, L. Understanding catchment behavior through stepwise model concept improvement. *Water Resour. Res.* **2008**, *44*, W01402. [CrossRef]
16. Clark, M.P.; Kavetski, D.; Fenicia, F. Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.* **2011**, *47*, W09301. [CrossRef]

17. Clark, M.P.; McMillan, H.K.; Collins, D.B.G.; Kavetski, D.; Woods, R.A. Hydrological field data from a modeller's perspective: Part 2: Process-based evaluation of model hypotheses. *Hydrol. Process.* **2011**, *25*, 523–543. [CrossRef]
18. Clark, M.P.; Slater, A.G.; Rupp, D.E.; Woods, R.A.; Vrugt, J.A.; Gupta, H.V.; Wagener, T.; Hay, L.E. Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resour. Res.* **2008**, *44*, W00B02. [CrossRef]
19. Fenicia, F.; Kavetski, D.; Savenije, H.H.G. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resour. Res.* **2011**, *47*, W11510. [CrossRef]
20. Clark, M.P.; Nijssen, B.; Lundquist, J.D.; Kavetski, D.; Rupp, D.E.; Woods, R.A.; Freer, J.E.; Gutmann, E.D.; Wood, A.W.; Brekke, L.D.; et al. A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resour. Res.* **2015**, *51*, 2498–2514. [CrossRef]
21. Knoben, W.J.M.; Freer, J.E.; Fowler, K.J.A.; Peel, M.C.; Woods, R.A. Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v1.2: An open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. *Geosci. Model Dev.* **2019**, *12*, 2463–2480. [CrossRef]
22. Trotter, L.; Knoben, W.J.M.; Fowler, K.J.A.; Saft, M.; Peel, M.C. Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v2.1: An object-oriented implementation of 47 established hydrological models for improved speed and readability. *Geosci. Model Dev.* **2022**, *15*, 6359–6369. [CrossRef]
23. Wu, S.; Yang, Y.; Li, C.; Yang, W.; Zhao, J. Evaluating evapotranspiration simulation performance in 30 conceptual hydrological models: Insights into ET representation across diverse climates. *Water Resour. Res.* **2026**, *62*, e2025WR040017. [CrossRef]
24. Gui, Z.; Zhang, F.; Chang, D.; Xie, A.; Yue, K.; Wang, H. A general method to improve runoff prediction in ungauged basins based on remotely sensed actual evapotranspiration data. *Water* **2023**, *15*, 3307. [CrossRef]
25. Pomeroy, J.W.; Brown, T.; Fang, X.; Shook, K.R.; Pradhananga, D.; Armstrong, R.; Harder, P.; Marsh, C.; Costa, D.; Krogh, S.A.; et al. The cold regions hydrological modelling platform for hydrological diagnosis and prediction based on process understanding. *J. Hydrol.* **2022**, *615*, 128711. [CrossRef]
26. Essery, R.; Morin, S.; Lejeune, Y.; Ménard, C.B. A comparison of 1701 snow models using observations from an alpine site. *Adv. Water Resour.* **2013**, *55*, 131–148. [CrossRef]
27. Knoben, W.J.M.; Freer, J.E.; Peel, M.C.; Fowler, K.J.A.; Woods, R.A. A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. *Water Resour. Res.* **2020**, *56*, e2019WR025975. [CrossRef]
28. Zhang, J.; Zhang, Y.; Sun, G.; Song, C.; Dannenberg, M.P.; Li, J.; Liu, N.; Zhang, K.; Zhang, Q.; Hao, L. Vegetation greening weakened the capacity of water supply to China's South-to-North Water Diversion Project. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 5623–5640. [CrossRef]
29. Zheng, X.; Liu, D.; Wang, H.; Ma, C.; Liu, H.; Ming, G.; Li, Q.; Khan, M.Y.A.; Hussain, F. What can hydrological modelling gain from spatially explicit parameterization and multi-gauge calibration? *Hydrol. Earth Syst. Sci.* **2026**, *30*, 2493–2521. [CrossRef]
30. Wang, J.; Bao, W.; Gao, Q.; Si, W.; Sun, Y. Coupling the Xinanjiang model and wavelet-based random forests method for improved daily streamflow simulation. *J. Hydroinform.* **2021**, *23*, 589–604. [CrossRef]
31. Arsenault, R.; Poulin, A.; Côté, P.; Brissette, F. Comparison of stochastic optimization algorithms in hydrological model calibration. *J. Hydrol. Eng.* **2014**, *19*, 1374–1384. [CrossRef]
32. Hansen, N. The CMA Evolution Strategy: A Tutorial. *arXiv* **2016**, arXiv:1604.00772. [CrossRef]
33. 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]
34. 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]
35. Vapnik, V.N. *Statistical Learning Theory*; Wiley-Interscience: New York, NY, USA, 1998.
36. Schoups, G.; van de Giesen, N.C.; Savenije, H.H.G. Model complexity control for hydrologic prediction. *Water Resour. Res.* **2008**, *44*, W00B03. [CrossRef]
37. Cherkassky, V.; Mulier, F.M. *Learning from Data: Concepts, Theory, and Methods*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2007. [CrossRef]
38. Tang, Y.; Reed, P.; Wagener, T.; van Werkhoven, K. Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. *Hydrol. Earth Syst. Sci.* **2007**, *11*, 793–817. [CrossRef]
39. Pianosi, F.; Beven, K.; Freer, J.; Hall, J.W.; Rougier, J.; Stephenson, D.B.; Wagener, T. Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environ. Model. Softw.* **2016**, *79*, 214–232. [CrossRef]
40. Saltelli, A.; Annoni, P.; Azzini, I.; Campolongo, F.; Ratto, M.; Tarantola, S. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Comput. Phys. Commun.* **2010**, *181*, 259–270. [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.