

Comparing multi-model mosaic and multi-model combination methods to simulate streamflow across the contiguous USA

Cyril Thébault^{1*}, Wouter J. M. Knoben¹, Nans Addor^{2,3}, Andrew J. Newman⁴, Diana Spieler¹, Nicolás A. Vásquez¹, Yalan Song⁵, Gaby J. Gründemann¹, Shaun Carney⁶, Mukesh Kumar⁷, Katie van Werkhoven⁶,
5 Chaopeng Shen⁵, Andrew W. Wood^{8,9}, and Martyn P. Clark¹

¹Schulich School of Engineering, University of Calgary, Calgary, Alberta, Canada.

²Fathom, Bristol, United Kingdom.

³Geography, University of Exeter, Exeter, United Kingdom.

⁴RAL, NSF National Center for Atmospheric Research, Boulder, Colorado, United States of America.

10 ⁵Civil and Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania, United States of America.

⁶Research Triangle Institute, Research Triangle Park, North Carolina, United States of America.

⁷Civil, Construction, and Environmental Engineering, University of Alabama, Tuscaloosa, Alabama, United States of America.

⁸CGD, NSF National Center for Atmospheric Research, Boulder, Colorado, United States of America.

15 ⁹Civil and Environmental Engineering, Colorado School of Mines, Golden, Colorado, United States of America.

Correspondence to: Cyril Thébault (cyril.thebault@ucalgary.ca)

Abstract. The ability to accurately predict streamflow underpins decisions in water management, flood prevention, and sectoral planning. Traditional approaches for streamflow prediction often rely on ~~a one~~ single model, thereby overlooking
20 potential benefits from using multiple models. To address this limitation, this study explores alternative methods that select and combine multiple models to enhance streamflow simulations. Specifically, we assess the performance of multi-model mosaic methods that assign a single model to each catchment, and multi-model combination methods that merge multiple models using static or dynamic weighting schemes. The Framework for Understanding Structural Errors (FUSE) is used to
25 create an ensemble of 78 hydrological models, which were applied to ~~559544~~ catchments from the CAMELS dataset across the contiguous United States. Each of the 78 models is calibrated utilizing a composite objective function, calculated as the average of a high-flow and a low-flow performance metric, to cover a wide range of streamflow conditions. Based on our selection of lumped FUSE models, ~~T~~he results show that a carefully chosen single model from a larger ensemble can closely approach the performance of more complex multi-model strategies. Among the multi-model approaches, the combination and mosaic methods show broadly similar overall skill, although the combination approaches deliver slightly higher performance
30 and lower sampling uncertainty. However, per-catchment differences persist, indicating that no single multi-model strategy dominates everywhere. This heterogeneity in performance makes it difficult to determine a priori which multi-model method will best represent streamflow in a given catchment.

1 Introduction

In hydrology, applications such as water resource management, climate impact assessment, and flood prediction usually rely on hydrologic modelling. Yet, this task remains difficult given the complexity of most catchments, where numerous hydrologic processes are intertwined. Consequently, representing these interactions within a single modelling framework that covers large spatial domains remains a central challenge ~~in hydrology~~. Over the years, numerous hydrological models have emerged, most of which were developed to meet a specific need in a specific area (Singh and Woolhiser, 2002; Clark et al., 2011; Sidle, 2021; Horton et al., 2022). The substantial differences among hydrological models in their data requirements, process representations, and parameterizations make it unlikely that any single model will outperform all others across all catchments; in other words, identifying a “one-size-fits-all” model is difficult (Andréassian et al., 2009; Savenije, 2009). Nevertheless, despite these challenges, the ambition of developing broadly applicable hydrological models has long been present in the discipline (e.g., Perrin et al., 2003; Beck et al., 2016; Arheimer et al., 2020; Mathevet et al., 2020). (Linsley, (1982) already argued that general modelling frameworks would make it unnecessary for each hydrologist to develop his or her own model for each catchment, highlighting that repeated application of the same model enhances cumulative learning. ~~Moreover~~In addition, there are often only weak relationships between model performance and catchment geological, topographical, soil, and vegetation characteristics (Addor et al., 2018; Knoben et al., 2020; David et al., 2022). Hence, it remains challenging to determine which model is most appropriate for a given modelling project. The difficulty — and the potential lack of hydrologic realism — of selecting a single model for individual catchments or across multiple catchments has motivated the use of multi-model approaches, where the choice of models may vary across space (Georgakakos et al., 2004; Knoben et al., 2020; Wan et al., 2021; Thébault et al., 2024; Spieler and Schütze, 2024).

While traditional hydrologic modelling often relies on a single model structure, multi-model approaches are increasingly explored for their potential to enhance predictive reliability and inform decision-making (Refsgaard et al., 2007; Ramos et al., 2013; Dion et al., 2021; Ogden et al., 2021; Caillouet et al., 2022; Johnson et al., 2023). Multi-model approaches can be implemented either as ensembles, which retain multiple simulations to characterize uncertainty, or as deterministic selections/combinations that aim to produce a single best estimate. Although we acknowledge the benefits of ensemble multi-model methods — namely, their ability to explicitly characterize structural uncertainty and provide probabilistic predictions (see e.g., Thébault et al., 2025b) — the use of ensemble model configurations is outside the scope of our study.

Here, we focus instead on two complementary multi-model paradigms: (1) multi-model *mosaics*, and (2) multi-model *combinations*. For a given selection of catchments, multi-model mosaics aim to assign a single suitable model from a larger ensemble of available models to each catchment. Note that this term has been increasingly used in recent hydrological literature to describe spatial model selection approaches across large domains (e.g., Ogden et al., 2021; Johnson et al., 2023; Knoben et al., 2025; Thébault et al., 2025b; Ogden et al., 2026). For such an approach, the main question to answer is: how should this model selection be made? In contrast, multi-model combinations aim to take advantage of the potential complementarity of hydrological models within a given catchment and seek to create suitable multi-model combinations for each catchment. The

question to answer here is: how can multiple models be effectively combined? There is currently limited guidance on the implementation of multi-model mosaic and multi-model combination methods.

70 For multi-model mosaics, ~~(Ogden et al., 2021; Johnson et al., 2023; Knoben et al., 2025; Thébault et al., 2025b; Ogden et al., 2026)~~ one possible implementation strategy is to select models in specific catchments based on aggregate measures of model performance (e.g., ~~based on~~ the Nash-Sutcliffe or Kling-Gupta efficiency metrics). Early large-sample model intercomparison studies already demonstrated that no single model structure consistently outperforms others across all catchments (e.g., Perrin et al., 2001). Subsequent applications explored spatial model selection strategies (e.g., van Esse et al., 2013), illustrating both the potential and the challenges associated with mosaic-type approaches. Performance-based selection generally leads to higher overall performance across large domains compared to a one-size-fits-all model approach (Knoben et al., 2020; Mai et al., 2022; Thébault et al., 2024). However, this approach introduces several challenges. First, as highlighted by multiple authors (Perrin et al., 2001; Thébault et al., 2024; Knoben et al., 2025), the multi-model mosaic based on performance exhibits a highly heterogeneous pattern, with many different models identified as locally optimal across the domain. Second, many previous studies have demonstrated that there is a high degree of equifinality between models (Beven, 2006) — while a single model may perform slightly better than others for a specific catchment, the performance difference between the best and the next-
80 best models is often very small (Knoben et al., 2020; Spieler and Schütze, 2024; Knoben et al., 2025). This non-uniqueness of models highlights the uncertainty surrounding the choice of specific model structures and is compounded by the fact that the performance scores used for model selection can themselves be highly uncertain and overly sensitive to model errors on individual time steps (see e.g. Brigode et al., 2015; Newman et al., 2015; Clark et al., 2021; Klotz et al., 2024). The dual problems of lack of spatial coherence and non-uniqueness of models make it difficult to reliably identify a single best model
85 for any given catchment or to link model performance to the physical characteristics of the catchments. An alternative selection strategy is landscape-based, where models are chosen to match dominant processes. Although this approach can be promising, it typically requires substantial expert judgement and detailed catchment information that are rarely available consistently at large scales (McMillan et al., 2023) and is therefore not considered further here.

90 For multi-model combinations, the most common approach is based on weighted average multi-model combination algorithms such as Bayesian Model Averaging (Raftery et al., 2005; Vrugt et al., 2008). These methods have shown improvements in performance compared with one-size-fits-all models over large domains (Shamseldin et al., 1997; Georgakakos et al., 2004; Seiller et al., 2012; Thébault et al., 2024). However, assigning weights to the different members of the hydrological ensemble is not trivial and is still widely debated in the scientific community. For example, Arsenault et al. (2015), Wan et al. (2021), and Todorović et al. (2024) compare different algorithms for streamflow combinations and highlight that it is difficult to
95 establish the benefits of one method over another. Furthermore, combination methods also suffer from issues of spatial coherence and model equifinality over large domains, which limits their ability to consistently represent regional hydrologic behaviour or to provide physically interpretable relationships between model performance and catchment characteristics.

Despite these challenges, some general principles for building an effective multi-model ensemble can be found in the literature. Winter and Nychka (2010) showed that the key to multi-model approaches does not lie in the number of models but in their

100 differences: the effectiveness of an ensemble increases when one model's strengths can compensate for another's weaknesses. Seiller et al. (2012) highlighted that an individually 'poor' (low-performing) model may occasionally provide useful information to the ensemble. Several studies have emphasised the importance of accounting for model diversity, fidelity, and sensitivity when constructing ensembles (e.g. Evans et al., 2013; Knutti et al., 2013; Clark et al., 2016).

105 More generally, previous work has shown that multi-model approaches (whether mosaics or combinations) have benefits over traditional one-size-fits-all models, particularly because of their flexibility in space. However, the studies carried out in this area of research are generally temporally static. In other words, once the model or the combination of models has been chosen for a catchment, it remains fixed throughout the length of the simulation period. This represents a major limitation, given that hydrological systems exhibit significant temporal variability, from low-flow conditions to flood events, and that different models tend to perform better under different objectives (e.g., reproducing baseflow versus peak flow), which usually
110 correspond to distinct periods within the streamflow regime (Kollat et al., 2012). The initial progress on dynamic combination approaches by Oudin et al. (2006) or more recently by Thébault et al. (2025) shows promising results, but such dynamic combination approaches have not yet been systematically compared to other multi-model methods.

From the literature review, we identify a large diversity of multi-model approaches, ranging from multi-model combinations to spatial mosaics. Previous studies have primarily focused on comparing averaging or weighting methods within ensemble
115 combinations (e.g. Arsenault et al., 2015; Wan et al., 2021; Todorović et al., 2024). However, to our knowledge, no systematic assessment has been made between the mosaic and combination strategies themselves. To address this gap, we conduct such a comparison to address the following question: *which multi-model approach is best suited for streamflow simulation across a large sample of catchments?* The remainder of the paper is organized as follows: first, the catchments, the hydrometeorological data, and the hydrological models are presented (Sections 2.1 and 2.2). Next, the multi-model mosaic
120 and multi-model combination methodologies are detailed (Sections 2.3 and 2.4). Then, the results are presented and analysed (Section 3). This section is followed by a broader discussion of the benefits and limitations of the multi-model approaches (Section 4). Finally, we summarize the main conclusions and expand on the potential perspectives for this work (Section 5).

2 Materials and methods

2.1 Catchments and hydrometeorological data

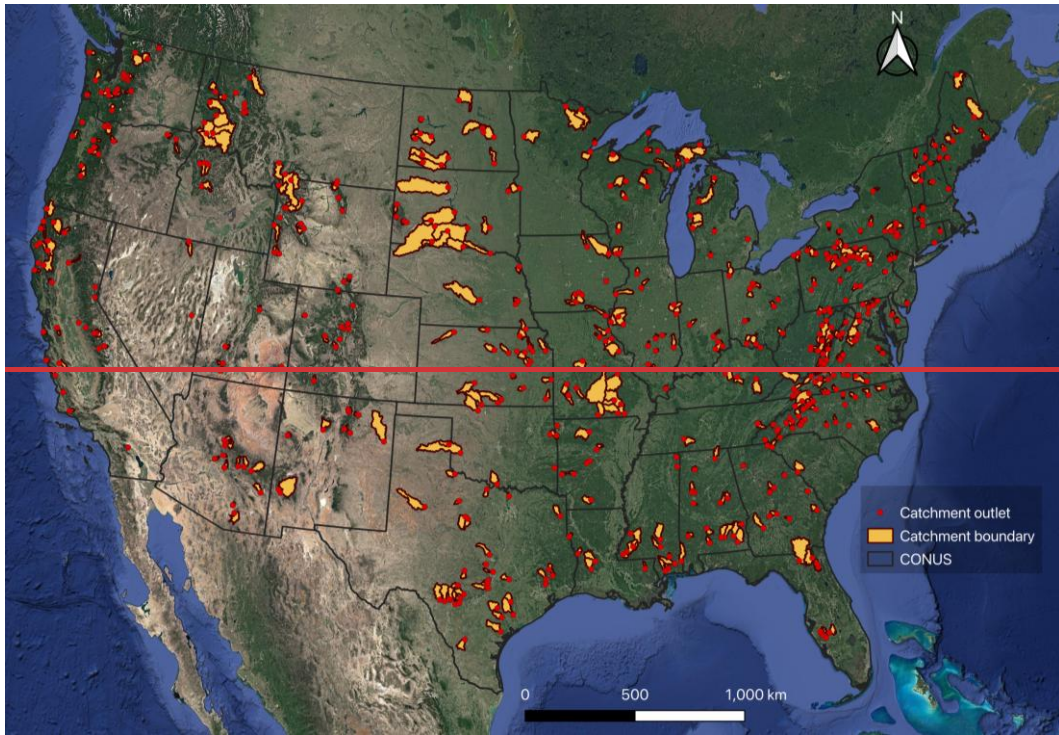
125 This study is based on the same catchments and hydrometeorological data as Thébault et al. (2025). In particular, we used the CAMELS data set (Newman et al., 2015; Addor et al., 2017), which provides daily meteorological and streamflow time series for 671 catchments with limited human influence across the contiguous United States (CONUS). CAMELS includes several meteorological forcing products: Daymet (Thornton et al., 2012), Maurer (Livneh et al., 2013), and NLDAS (Xia et al., 2012). In this work, we used only the Daymet product because it provides the required variables (precipitation and temperature) at
130 the highest spatial resolution (1 km x 1 km) compared with the other datasets (12 km x 12 km), and has shown better performance in past large-sample hydrologic studies (e.g. Kratzert et al., 2021; Sawadekar et al., 2025). Potential

evapotranspiration is calculated using the formulation proposed by Oudin et al. (2005). This equation was chosen for its simplicity, as it requires only daily air temperature and extraterrestrial radiation (a function of Julian day and latitude), and because it was developed and tested with conceptual models, which aligns with the modelling framework used here (see Section 2.2). Following Knoben et al. (2020), the number of catchments was reduced to 559 (Figure 1) by excluding catchments with large area discrepancies between the provided polygons and reference values, as well as catchments that fall outside the water and energy limits on a Budyko plot. From these 559 catchments, a final subset of 544 catchments was retained for the comparative analyses presented in this study. The remaining catchments were excluded due to model failures or because sampling uncertainties could not be computed consistently across all modelling approaches (see Appendix B for details). Figure 1 provides a map of the 544 catchments retained for analysis, and Table 1 summarizes some of their key attributes available in from the CAMELS dataset. ~~set for the 559 catchments selected for this study. For the remainder of this document, the results are presented for a subset of 544 catchments out of the 559 available. This selection ensures a fair comparison among all approaches, as some catchments were excluded due to model failures or because sampling uncertainties could not be computed. Further details are provided in Appendix B.~~

135

140

145



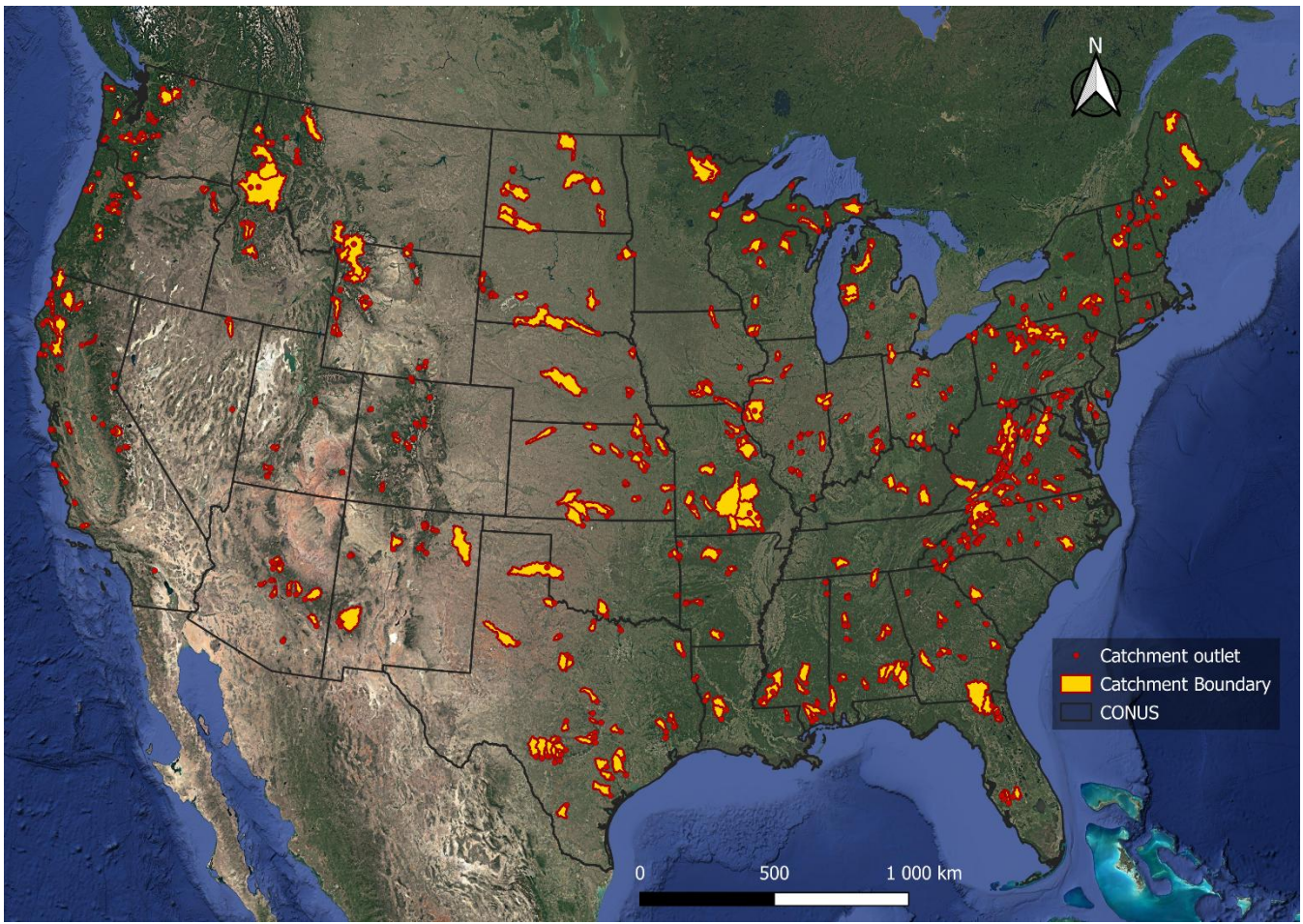


Figure 1: Location of the 54459 catchments selected for this study, using state boundaries from the “North American Atlas - Political Boundaries” (Commission for Environmental Cooperation, 2022) and the basemap derived from “Google Satellite”, available through QGIS via the QuickMapServices plugin. Figure reproduced based on from Thébault et al. (2025).

150

Table 1: Statistical summary of some of the key attributes available in the CAMELS data over the 54459 catchments selected for this study.

Main characteristics of the catchments	Min.	Median	Max.
Catchment area [km ²]	4	<u>36584</u>	<u>1425,257818</u>
Mean elevation [m]	14	<u>43743</u>	3,457
Mean slope [m/km]	1	22	227
<u>Mean temperature [°C]</u>	<u>-1</u>	<u>10</u>	<u>22</u>
Mean annual precipitation (P) [mm/year]	372	1,1 <u>6559</u>	3,2 <u>21563</u>
Mean annual potential evapotranspiration (PET), estimated with the Oudin formula [mm/year]	320	<u>7775</u>	1,407
Mean annual streamflow (Q) [mm/year]	2	<u>4096</u>	2,904
Aridity (P/PET) [-]	0.35	1.3 <u>87</u>	5.73
Runoff ratio (Q/P) [-]	0.01	0.35	0.99
Snowfall fraction [-]	0	0.10	0.91
Baseflow index [-]	0.01	0.50	0.89

2.2 Hydrological model framework: FUSE

155 For this study, we use the Framework for Understanding Structural Errors (FUSE) developed by Clark et al. (2008). FUSE is a modular framework that enables the construction of new hydrological model structures from different modules. FUSE is based on four existing conceptual models (Variable Infiltration Capacity, VIC; Precipitation Runoff Modelling System, PRMS; Sacramento Soil Moisture Accounting, SAC-SMA; and TOPMODEL), which were decomposed into individual model components that can be used interchangeably in a general model template.

160 FUSE is designed to enable the investigation of the structural uncertainties and errors inherent in hydrological models by allowing modellers to test different model decisions. In total, more than 3,000 structures can be generated using this modular modelling framework. For this study, we focus on the 78 original configurations proposed by Clark et al. (2008). These model structures were created by combining different formulations of the state equations for the upper layer, lower layer and baseflow, percolation, and surface runoff. All other decisions are set to their default values to reduce computational costs and limit the size of the hydrological ensemble to a manageable number. As described in Henn et al. (2015), FUSE was also coupled with
165 a temperature-index snow model based on the Snow-17 formulation (Anderson, 2006). Table 2 summarizes the different decisions in FUSE. Considering the options outlined in each row of Table 2, a total of $3 \times 4 \times 3 \times 3 \times 1 \times 1 = 108$ model structures is possible. However, as certain combinations of options are not functional, a total of 78 model structures was considered in this analysis. It should be noted that the four parent FUSE configurations (VIC, PRMS, SAC-SMA and TOPMODEL) are included within the 78 structures. Although the 78 model structures provide a broad exploration of conceptual structural decisions within FUSE, they share several common characteristics: all models are lumped, operate at a daily time step, use identical forcing data, and are bucket-style. While this design enables a controlled comparison of structural choices, it does not span fundamentally different modelling paradigms (e.g., fully distributed or data-driven approaches), which may limit the overall structural diversity of the ensemble and could limit the benefits of multi-model approaches (Winter and Nychka, 2010).

175 **Table 2: Model decisions available in FUSE. The options in light grey correspond to options that are available in FUSE but not considered in this study.**

Decision	Options			
Rainfall correction	Additive	Multiplicative		
Upper-layer architecture	Single state variable	Separate state variables (tension storage and free storage)	Separate state variables (tension storage with two zones and free storage)	
Lower-layer architecture + baseflow	Single linear reservoir with no evaporation	Single power-law reservoir with no evaporation	Single nonlinear reservoir with evaporation	Two parallel linear reservoirs
Surface runoff	PRMS	ARNO/VIC	TOPMODEL	
Percolation	Linear with drainage below field capacity limited	Linear with drainage below field capacity not allowed	Nonlinear	
Evaporation	Sequential	Root weighting		
Interflow	No	Yes		
Routing	No	Gamma distribution		
Snow model	No	Yes (Henn et al., 2015)		

2.3 Model calibration

The parameters of each of these 78 structures (between 15 and 20 depending on the selected decisions) are calibrated for each catchment using up to 10,000 iterations (e.g., Duan et al., 1994; Tolson and Shoemaker, 2007; Feyen et al., 2008; Lan et al., 2020) of the shuffled complex evolution approach (Duan et al., 1993) to maximize a composite criterion (Garcia et al., 2017):

$$KGE_{comp} = \frac{KGE(Q) + KGE(1/Q)}{2} \quad (1)$$

where KGE, the Kling-Gupta efficiency (Gupta et al., 2009), is given by:

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

with r the correlation coefficient, α the ratio of standard deviations, and β the ratio of the means (i.e. the bias) between simulated and observed time-series.

In hydrology, the KGE is typically calculated on streamflow time series without transformation, i.e. as KGE(Q). When used as an objective function for calibration, it is particularly sensitive to high-flow values, often at the expense of low flows (Pushpalatha et al., 2012; Garcia et al., 2017). A common alternative to reduce the influence of high-flow values is to apply a transformation of the time series (e.g., square root, box-cox, inverse). However, Thirel et al. (2024) showed that such an approach only targets a different part of the hydrograph, without successfully capturing the full range of streamflow. This is a known limitation of using a single metric (e.g., Booij and Krol, 2010; Clark et al., 2021). To address this limitation, multi-objective algorithms can be used to better account for multiple facets of streamflow behaviour (e.g. Gupta et al., 1998;

195 Efstratiadis and Koutsoyiannis, 2010; Kollat et al., 2012; Zhang et al., 2018). Such approaches aim to identify trade-offs among competing objectives by approximating a Pareto front rather than a single optimum. An alternative approach is to combine several metrics into a single objective function — here referred to as a composite metric (e.g. Garcia et al., 2017; Hallouin et al., 2020; Thébault et al., 2024). Although this method presents some limitations, such as the loss of some information due to dimensionality reduction or the subjectivity through the weights applied to the different metrics, it typically requires fewer model evaluations because it only needs to converge to a single optimum instead of determining trade-offs with the Pareto front (Efstratiadis and Koutsoyiannis, 2010; Mai, 2023).

In this work, we therefore use a composite metric for the calibration and evaluation of hydrological models, as it provides a good compromise between multi-objectivity and computation time. It aims to provide a balance between high-flow and low-flow (Eq. 1). This metric has already been applied in the literature and has shown benefits compared to traditional single-objective approaches or other combinations of criteria (Garcia et al., 2017).

205 Each model is calibrated for each catchment over the period 1989-1998 (calendar year), including both relatively wet and dry years, with a preliminary two-year warm-up period ~~of two years~~(1987-1988) preceding calibration. These periods were selected to maximize the number of catchments with sufficiently complete streamflow time series while maintaining consistency with previous large-sample CAMELS-based studies (e.g. Knoben et al., 2020).

2.4 **Model evaluation and performance**

The models and ~~the~~ multi-model approaches are evaluated over the period 1999-2009 (calendar year), including both relatively wet and dry years. ~~Our~~ The evaluation framework is divided in ~~three~~3 parts: (1) a comparison of performance, (2) an analysis of sampling uncertainty, and (3) an assessment of test on model equivalence.

210 2.4.1 **Performance**

The performance is assessed with the composite metric KGE_{comp} (Eq. 1).

2.4.2 **Sampling uncertainty**

215 We evaluate the robustness of the performance score — defined here as the consistency of model skill across different temporal evaluation regimes — by accounting for sampling uncertainty (Clark et al., 2021; Lamontagne et al., 2020). Specifically, sampling uncertainty refers to the dependency of performance score values, such as NSE and KGE, on the data used for their calculation. This effect is especially pronounced in catchments without strong seasonality in their streamflow regimes, where much of the model error may be concentrated in just a few time steps. Sampling uncertainty can be addressed using bootstrapping methods to provide an estimate of the uncertainty range of a model's performance for a given catchment (see, e.g., Clark et al., 2021). For this study, sampling uncertainty in performance metrics is quantified using the *gumboot* R package 220 (Clark and Shook, 2021), following the default bootstrap procedure described by the authors. Specifically, the method resamples non-overlapping blocks of complete water years with replacement to generate 1000 synthetic hydrographs,

preserving within-year autocorrelation and seasonality. Sampling uncertainty is expressed as the interval between the 5th and 95th percentiles of the resulting bootstrap distribution. Note that the KGE_{comp} is not implemented in the current version of *gumboot*; we therefore extended the package to include this metric for the present analysis.

225 2.4.3 Performance equivalence

Knoben et al. (2025) showed that many models often exhibit performances falling within the sampling uncertainty range of the model that initially achieves the highest KGE score in a given catchment (i.e., within the range of the 5th and 95th percentile range of the bootstrap distribution of the top-performing model). In such cases, the models are considered equivalent. Here, we apply this method to test the equivalence between multi-model approaches.

230 2.5 Multi-model approaches

2.5.1 Performance benchmark

We establish our benchmark by selecting, from the ensemble of 78 models, a single model across all catchments. Specifically, the benchmark is defined as the model with the highest median KGE_{comp} value over the calibration period and across the 559544 catchments. It should be noted that while we select a single model for all catchments as our benchmark, the process of
235 how it is selected differs from traditional “one-size-fits-all” approaches, which are typically based on legacy or convenience selections (Addor and Melsen, 2019). Our approach instead selects a single model based on proven performance superiority compared to the 77 considered alternative FUSE models and can therefore be viewed as a multi-model approach. The benchmark serves as a stringent baseline against which the value of more complex multi-model strategies can be assessed. Therefore, the benchmark is not intended to represent a typical legacy model choice, but rather to provide either a high
240 boundary for single-model modelling and a reference for multi-model experiments. Additional details on the benchmark structure and its performance distribution are provided in Appendix A (Figure A1 and Table A1).

2.5.2 Multi-model mosaics

Multi-model mosaics aim to assign a single suitable model taken from a larger model ensemble to each individual catchment. Here, we explore two approaches for building this spatial model mosaic: one based on performance only (Section 2.5.2.1), and
245 the other on performance-equivalence (Section 2.5.2.2).

2.5.2.1 Mosaic based on performance

One possible implementation of the multi-model mosaic approach is to select a single model per catchment based on aggregated performance metrics such as KGE or NSE (Knoben et al., 2020; Spieler et al., 2020; Mai et al., 2022; Thébault et al., 2024). In this study, we replicate such a performance-based mosaic approach by selecting, for each catchment, the model
250 with the highest KGE_{comp} value during the calibration period. Further details on the spatial distribution and frequency of selected models are provided in Appendix A (Figure A2).

2.5.2.2 Mosaic based on performance-equivalence

Another possible implementation of the multi-model mosaic approach uses the principle of performance-equivalence introduced by Knoben et al. (2025) to account for sampling uncertainty in performance scores. By assessing which models are equivalent to the top-performing one (i.e., whose scores fall within the sampling uncertainty of the best-top-performing model) in each catchment, it is possible to minimise the number of models required across the domain. This task is carried out using a linear programming algorithm from the *lpSolve* R package (Csárdi and Berkelaar, 2024) by solving a set cover type problem. In this study, the mosaic based on performance-equivalence is derived from the ensemble of 78 models, where sampling uncertainty and model selection are guided by the KGE_{comp} metric over the calibration period. The purpose of minimizing the number of models is to promote parsimony and spatial coherence. While many structures may be statistically equivalent, in term of performance metric value, in a given catchment (Figure A3), selecting a different model in each location may lead to a highly fragmented mosaic with limited interpretability. Across the domain, this procedure reduces the 78 candidate structures to only eight distinct models required to cover all catchments (Figures A4). Additional methodological details and supporting results are provided in Appendix A (Figure A3-A4).

2.5.3 Multi-model combinations

Multi-model combinations aim to leverage the potential complementarity of hydrological models within a given catchment by averaging streamflow outputs from an ensemble of models, thereby improving approximation of the expected hydrologic response and providing insight into uncertainty arising from differences in model structure. In this study, we explore three approaches for constructing such combinations: the first combination uses static weights across time and space (Section 2.5.3.1), the second combination uses weights that are static in time but variable in space (Section 2.5.3.2), and the last combination uses weights that change dynamically in space and in time (Section 2.5.3.3).

2.5.3.1 Spatially and temporally static combination

A static combination in time and space ~~means-refers to that~~ a single combination of several models ~~is-~~used for streamflow predictions in all catchments, with weights that do not change. There are numerous ways to derive the weights for such a setup (see, e.g., Arsenault et al., 2015; Wan et al., 2021; Todorović et al., 2024), with no particular approach appearing to be consistently better than the others. We adopt an approach similar ~~approach-ofo~~ that of Thébault et al. (2024), who demonstrated the benefits of a simple static combination approach on a large sample of catchments: using a simple average of up to four models for a combination of n models among an ensemble of m models, each selected model receives a weight of $1/n$, and the weight of the other $m-n$ models is set to 0. Thébault et al. (2024) also showed that the performance gain from combining four models is limited compared to the performance that can be achieved by combining three models. In this work, considering all combinations of up to four models (among an ensemble of 78 models) would result in a substantially larger number of possibilities compared to three (${}^78_2C + {}^78_3C + {}^78_4C = 1,505,504$ vs ${}^78_2C + {}^78_3C = 79,079$ combinations per catchment), which would drastically increase computational cost. To achieve a better balance between computational ~~ing~~ cost and performance,

285 this study considers only combinations of ~~up to~~ ~~two or~~ ~~three~~ models (see also Figure A5, ~~that which~~ shows that imposing a ~~three~~-model limit is acceptable in our case). Note that combinations of size one (i.e. equivalent to a single model) are not included in the search space, although they may yield marginally higher performance in a subset of catchments (see Appendix C). The spatially and temporally static combination is defined as the combination that yields the highest median KGE_{comp} score over the calibration period across the ~~559~~544 catchments. The distribution of performance across all tested combinations and identification of the top-performing combination are shown in Appendix A (Figure A5).

290 2.5.3.2 Spatially variable and temporally static combination

This approach recognizes that not every model might be equally appropriate for every catchment and instead selects an optimal combination for each individual catchment. This approach is similar to the previous combination (Section 2.5.3.1), except that the models selected in the combination differ for each individual catchment. In other words, for each of the ~~559~~544 catchments, we identify the combination of ~~up to three~~two or three models that yields the highest KGE_{comp} scores over the calibration period. The diversity of selected combinations and model frequencies are detailed in Appendix A (Figure A6-A7).

295

2.5.3.3 Spatially and temporally variable combination

The aim of this approach, hereafter called the “dynamic combination”, is to dynamically combine model simulations across time and space — in this case selecting from all 78 models. Therefore, the weight given to each of the 78 simulations varies at each time step and for each catchment. The general principle of the dynamic combination method is to identify past periods with similar hydrological conditions to the current one and define model weights for the current time step based on model performance during those analogous periods. In this study, we adopt the specific implementation used in Thébault et al. (2025) which involves two chained components run for each time step: (1) a search for past time periods (typically in the calibration period) similar to current conditions (typically in the evaluation period), implemented as a k-nearest-neighbour (k-NN) search based on Mean Absolute Error (MAE) scores between the streamflow simulation representing current conditions and past observations; and (2) a weighting of simulations based on mean MAE scores between each model’s simulations and past observations across the identified neighbours. Further methodological details on the dynamic combination framework can be found in Thébault et al. (2025).

300

305

We deviate slightly from the implementation in Thébault et al. (2025), by allowing the parameters of the dynamic combination — namely, the length of the time window (τ , ranging from 4 to 28 days), the number of nearest neighbours (k , ranging from 1 to 19), and the number of models to combine (m , ranging from 1 to 19) — to vary across catchments (see Figure A8). These parameters control how the method identifies similar past conditions and determines the weights assigned to individual models. Here, their values are optimized for each catchment to account for spatial heterogeneity. In addition, Thébault et al. (2025) highlighted that ~~part of~~ the strength of the dynamic combination comes from using an ensemble calibrated on different objective functions. Here, we deliberately limit the optimization to a single metric (KGE_{comp}) to ensure that observed

310

315 differences among approaches reflect their intrinsic combination mechanisms rather than differences in calibration effort or metric alignment.

It should be noted that a major difference between the dynamic combination and the two previous multi-model combination approaches (Sections 2.5.3.1 and 2.5.3.2) is that, in the dynamic combination, models are selected based on their individual performance (78 simulations evaluated) rather than on their combined performance (${}^78_2C + {}^78_3C = 79,079$ simulations evaluated). While adapting the method to evaluate combined performances (e.g., by assessing all possible combinations of ~~up to three~~two or three models across each neighbour) would be straightforward in principle, it would drastically increase computation time and complexity. Distributions of optimized parameters and model usage patterns are presented in Appendix A (Figure A8-A10).

2.6 Summary

325 We test six different multi-model approaches based on an ensemble of 78 models calibrated with KGE_{comp} , across ~~559544~~ CONUS catchments. Further details and analyses on each multi-model approach are provided in Appendix A. Table 3 summarizes these experiments, Figure 2 provides a schematic illustration of the different approaches, and the following points briefly summarize their implementation:

- 330 • The benchmark corresponds to a single model (the highest median performance across all catchments) applied everywhere.
- The performance-based mosaic assigns one model per catchment (the top-performing model across each catchment).
- The performance-equivalence mosaic extends this idea by explicitly accounting for sampling uncertainty in performance scores. A linear-programming algorithm identifies the minimum set of equivalent models needed to maintain performance within these uncertainty bounds, yielding a spatially parsimonious mosaic.
- 335 • The spatially and temporally static combination selects a single combination of models that is applied to all catchments. All possible combinations of ~~up to three~~two or three models (${}^78_2C + {}^78_3C = 79,079$ simulations per catchment) are evaluated, and the combination with the highest median performance across catchments is selected. The combination is computed by simple averaging, i.e., assigning equal weights to the selected models.
- The spatially variable and temporally static combination follows the same principle but determines the optimal combination (of ~~up to three~~two or three models) independently for each catchment.
- 340 • The dynamic (i.e., spatially and temporally variable) combination adapts model weights (up to 19 models selected based on individual performance) across space and time based on similarities during past conditions.

Table 3: Overview of the multi-model approaches evaluated in this study. Although the dynamic combination involves up to 19 models, the model selection is based on individual rather than combined performance, reducing its apparent complexity (see Section 2.5.3.3).

	Models per catchment	Variable in space	Variable in time
Benchmark (top-performing single model for everywhere)	1	No	No
Mosaic performance based	1	Yes	No
Mosaic performance-equivalence based	1	Yes	No
Spatially and temporally static combination	Up to 3	No	No
Spatially variable and temporally static combination	Up to 3	Yes	No
Dynamic combination	Up to 19	Yes	Yes

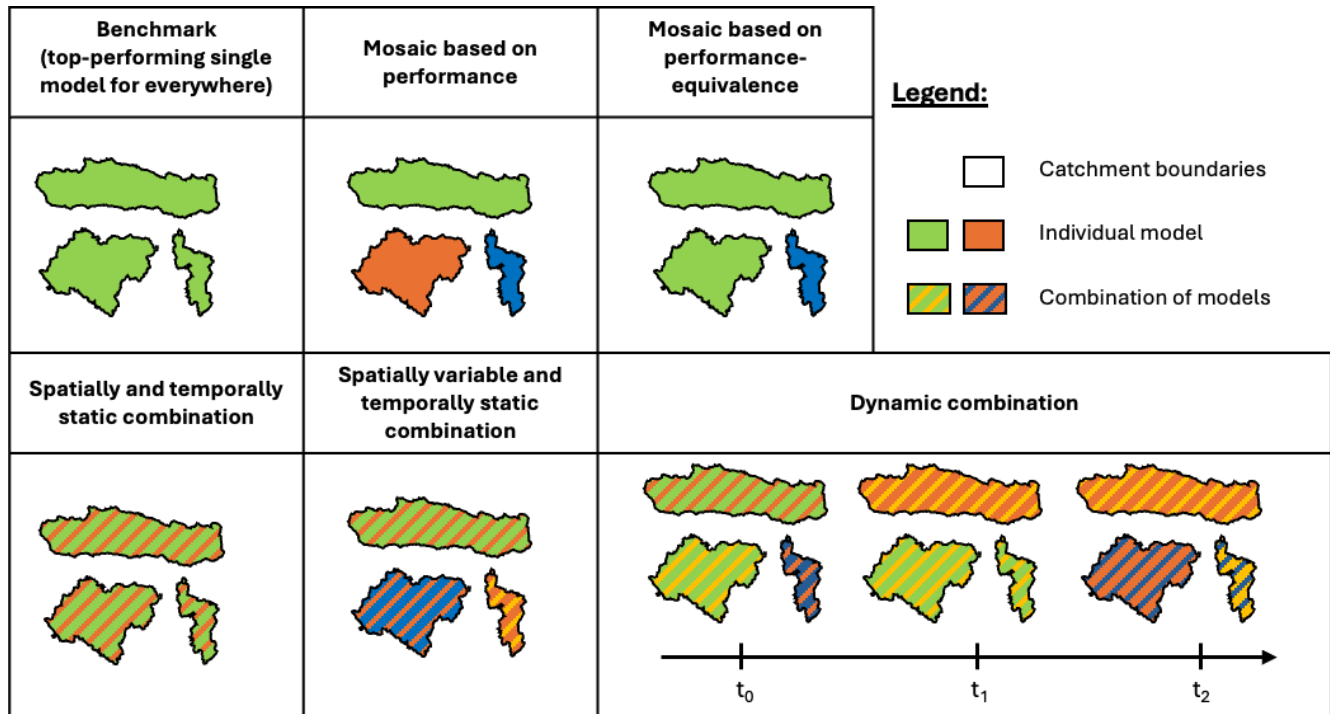


Figure 2: Schematic illustration of the six multi-model approaches tested. Each colour represents a model, e.g. the benchmark uses the same model everywhere, whereas in the mosaic approaches the selected model can be different for each catchment. Note that when we compare the two mosaic approaches (top row), the orange catchment becomes green, meaning that green model is equivalent to the orange one in that specific catchment. Hatched areas (bottom row) indicate combinations of models. The x-axis in the bottom right plot represents time, showing that the combination of models can vary on a per-timestep basis.

3 Results

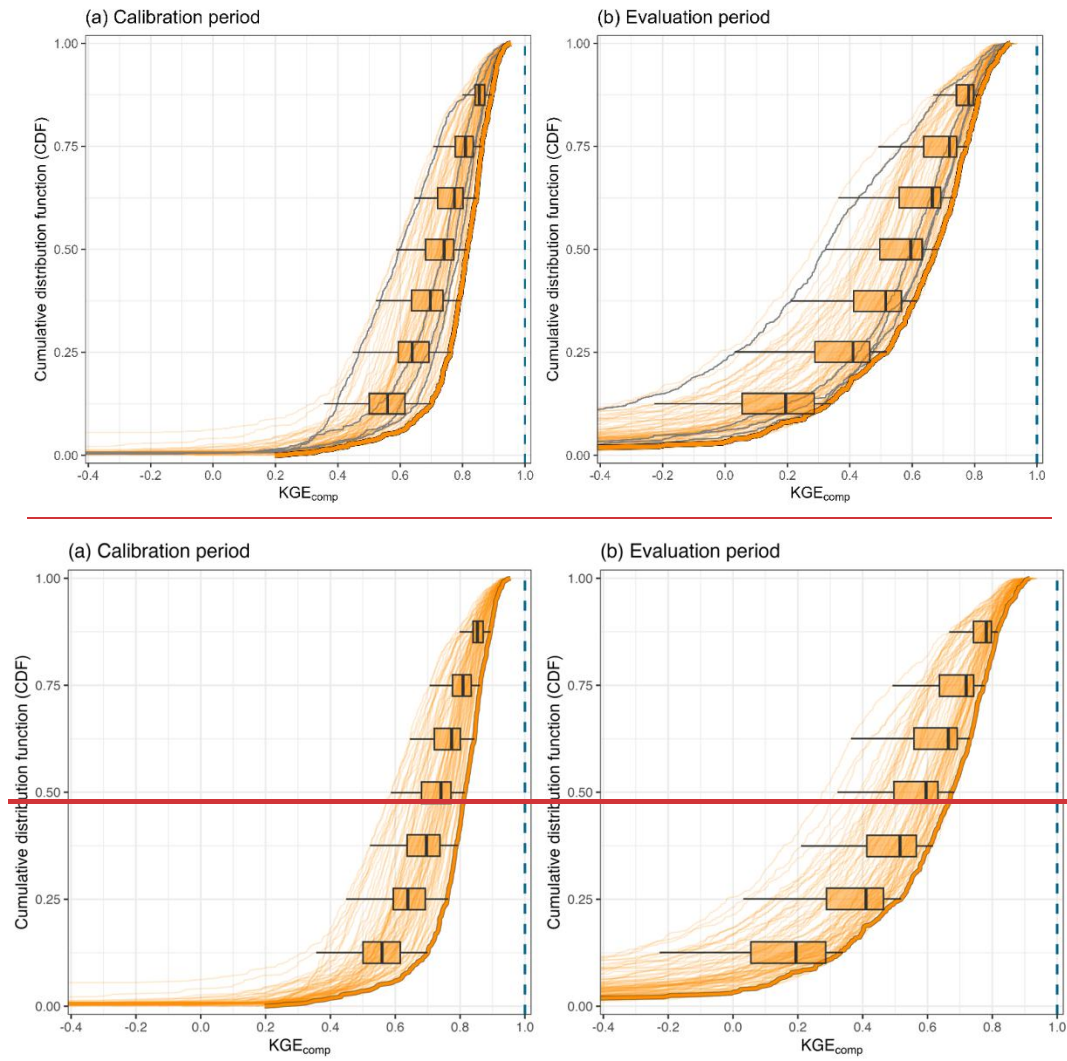
~~For the remainder of this document, the results are presented for a subset of 544 catchments out of the 559 available. This selection ensures a fair comparison among all approaches, as some catchments were excluded due to model failures or because sampling uncertainties could not be computed. Further details are provided in Appendix B.~~

In the following sections, we first examine the overall behaviour of the FUSE ensemble and the single-model benchmark (Section 3.1). We then compare the different multi-model approaches in terms of their predictive performance (Section 3.2.1) and the associated sampling uncertainty (Section 3.2.2). Finally, we assess whether the multi-model approaches are equivalent across the catchments (Section 3.3). This analysis aims to answer three main questions:

- (i) How does each multi-model approach perform relative to the single-model benchmark across a large sample of catchments?
- (ii) How robust are these performance differences given sampling uncertainty?
- (iii) To what extent do different multi-model approaches provide equivalent performance across space?

3.1 FUSE model ensemble and benchmark

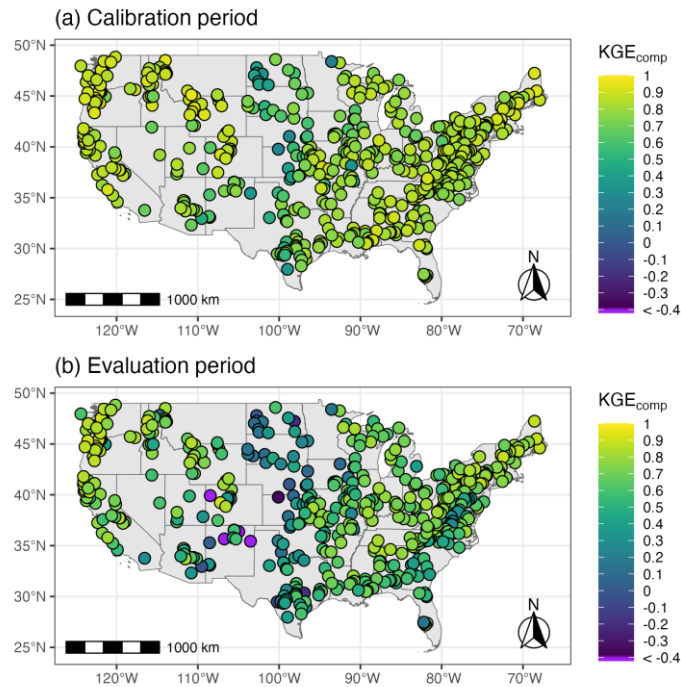
Figure 3 shows the KGE_{comp} metric of the individual FUSE models over both calibration and evaluation periods. The spread demonstrates substantial performance differences attributable to model structure, with median values (i.e. CDF at 0.5) varying from 0.59 to 0.82 during the calibration period and from 0.29 to 0.69 during the evaluation period, depending solely on the structure employed. The benchmark, defined as the model achieving the highest median KGE_{comp} across catchments during the calibration period, is highlighted in bold. This model also maintains one of the highest performance distributions during the evaluation period. However, this does not mean that it systematically outperforms all other models in every catchment (result not shown here for simplicity, but valid for both the calibration and evaluation period). The per-catchment performances of the benchmark model, presented in ~~Figure 4~~**Figure 4**, exhibit a spatial pattern typical of streamflow modelling across CONUS (e.g., Newman et al., 2015; Knoben et al., 2020), with lower performance scores in the drier central area. While it is well documented that KGE (and related composite metrics) are influenced by streamflow variability and therefore are not directly comparable across locations (Schaepli and Gupta, 2007; Knoben et al., 2019; Williams, 2025), these maps provide a general sense of model skill and enable per-catchment comparison between the calibration and evaluation periods. Details about the model selected as a benchmark are given in Table A1. As a reminder, this benchmark aims to provide either a high boundary for single-model modelling and a reference for multi-model experiments, rather than a simple baseline based on legacy choices. However, for transparency, the four parent FUSE configurations corresponding to VIC, PRMS, SAC-SMA, and TOPMODEL are explicitly highlighted in grey in Figure 3, allowing readers to assess their relative performance within the full ensemble.



385

Figure 3: Distribution of the performance score KGE_{comp} of the individual FUSE models (78 structures) across the 544 catchments during (a) the calibration and (b) the evaluation period. A slight transparency has been applied to avoid problems of misinterpretation due to line overlaps. The bold line highlights the benchmark, i.e., the top-performing model based on median KGE_{comp} value across all catchments for the calibration period. The grey lines show the four parent FUSE configurations (VIC, PRMS, SAC-SMA, and TOPMODEL) The boxplots illustrate the ensemble spread (differences linked to model structure) for different values of the cumulative distribution function (CDF): 0.125, 0.25, 0.375, 0.5, 0.625, 0.75 and 0.875. The dashed blue line indicates the optimum value of the KGE_{comp} score, i.e. 1.

390



395 **Figure 4: Spatial distribution of the performance score KGE_{comp} of the benchmark model across 544 catchments during (a) the calibration and (b) the evaluation period. The benchmark corresponds to the top-performing model based on the median KGE_{comp} value across all catchments during the calibration period.**

3.2 Comparison of the multi-model approaches

3.2.1 Performance

400 Figure 5 presents a comparison of performance scores of all multi-model approaches with the benchmark model for the composite criterion KGE_{comp} over the evaluation period. Overall, all various modelling approaches tested here yield similar performance scores.

Surprisingly, the **benchmark** (orange) — a single model applied across all catchments — achieves a score distribution that is very close to those of multi-model approaches (slightly lower) with a median value of 0.68. However, unlike typical “one-size-fits-all” strategies, the model selected here was not chosen based on legacy or convenience. Although the outcome is the same (one model for all catchments), it is chosen on performance across a broad spatial domain (~~559~~544 catchments) from a large ensemble of models (78 structures). In this sense, the benchmark stems from an initial multi-model experiment, where
 405 all structures were candidates to become the benchmark. Its comparison with purely multi-model methods points to a promising pathway toward developing a robust single-model solution if the selection method is carefully considered.

The **multi-model mosaic** approaches do not exhibit a substantial improvement in performance compared to the single-model
 410 benchmark across the sample of catchments. The mosaic based purely on performance (light pink) provides an increase in the

median performance of 0.02 (to 0.70) and also benefits the other quantiles considered. The mosaic approach based on performance-equivalence (dark pink) shows scores more similar to the benchmark model (with a median of 0.69 and similar quantiles). This result is consistent with Figure A4, where the model structure that was selected as the benchmark model is used by nearly 80% of the catchments for the mosaic based on performance-equivalence.

415 The **multi-model combination** approaches also do not exhibit performance that substantially improves upon the single-model benchmark. The spatially and temporally static combination (light green) provides a distribution of performance scores very close to those obtained with a performance-based mosaic approach (median of 0.70 and similar quantile distributions), even though the model choice does not vary spatially. This result thus highlights that there are some benefits of combining several models — i.e., using an ensemble of models rather than a single one. As expected, enabling a degree of freedom in space, with
 420 the spatially variable and temporally static combination (dark green), leads to slightly better results over the evaluation period (median of 0.72). This approach achieved the highest performance among all the multi-model approaches tested, although the differences remain small. Although more complex, the dynamic combination (red) does not provide the highest scores (median of 0.70) and leads to a performance distribution similar to what can be obtained with a static combination in time and space (light green) or a mosaic based on performance (light pink). This result is discussed in Section 4.3, which examines why the
 425 dynamic combination does not outperform other multi-model approaches.

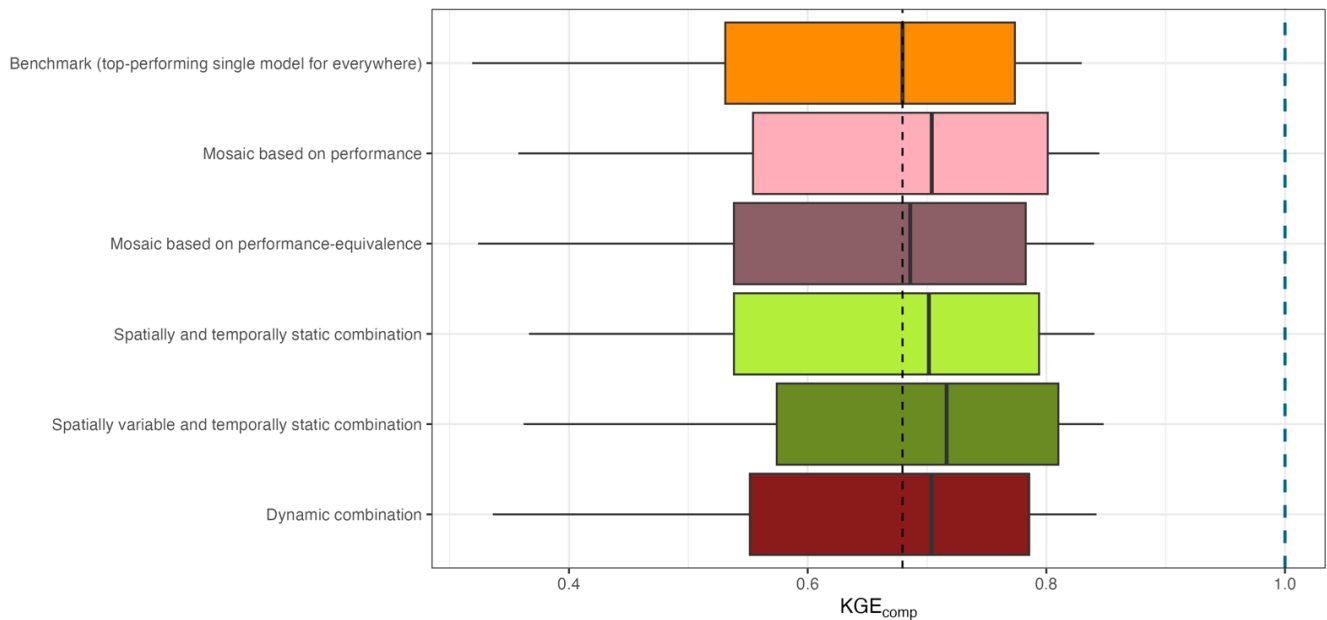


Figure 5: Boxplot of performance scores, KGE_{comp} , over the evaluation period for the different multi-model approaches. The boxplots represent the 10 %, 25 %, 50 %, 75 % and 90 % quantiles. The dashed black line indicates the median value of the benchmark. The dashed blue line highlights the optimal value of KGE_{comp} .

430 To complement the domain-wide performance summaries, [Figure 6](#) illustrates the spatial distribution of KGE_{comp} scores for each multi-model approach and provides a comparison with the benchmark model. The maps reveal that all multi-

model approaches produce broadly consistent spatial patterns, similar to those found in the literature for streamflow modelling across CONUS (e.g, Newman et al., 2015; Knoben et al., 2020). Differences to the benchmark are mostly limited in magnitude, with predominantly light blue tones overall, indicating a slight overall improvement (consistent with the previous results).

435 However, a few outliers can be observed, showing large improvements or deteriorations for specific catchments. These strong variations mostly occur in catchments where modelling is initially challenging (i.e., where performances are low in [Figure 4](#)), likely reflecting some degree of overfitting during calibration in poorly constrained catchments. This interpretation is supported by [Figure 6](#): the performance-based mosaic cannot perform worse than the benchmark during calibration by construction, yet during evaluation it sometimes yields substantially poorer performance, illustrating how calibration-time gains can fail to generalize.

440

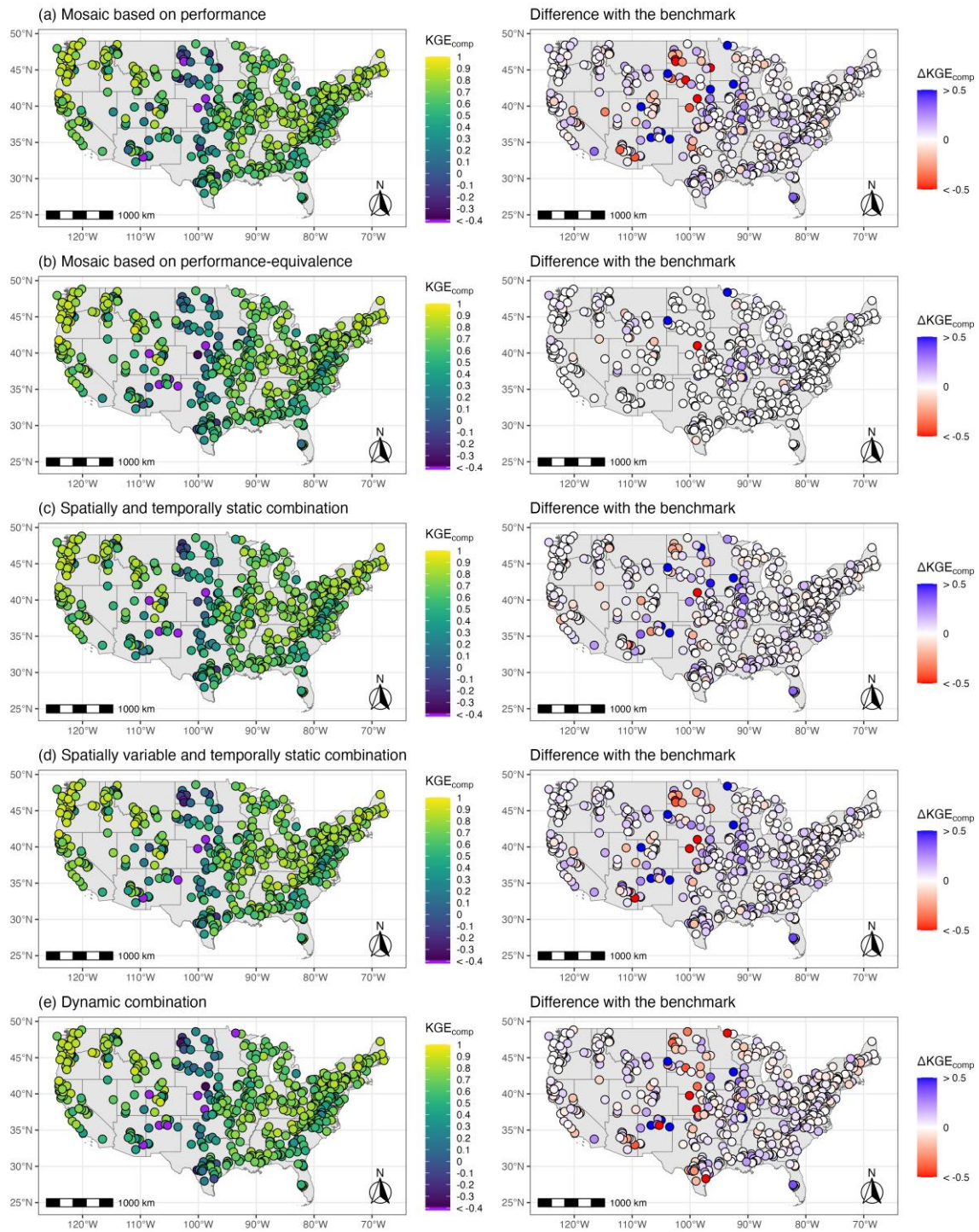


Figure 6: Spatial distribution of the performance score KGE_{comp} of the different multi-model approaches across 544 catchments over the evaluation period. The difference with the benchmark is also shown in the right column; blue colours indicate better performance with the multi-model approach, while red colours indicate worse performance.

445 3.2.2 Sampling uncertainty

The purpose of sampling uncertainty is to assess the sensitivity of performance scores (here, KGE_{comp}) to the specific period over which they are calculated (here, the evaluation period). Figure 7 presents a comparison of the sampling uncertainty associated with KGE_{comp} for the various multi-model approaches, calculated using the *gumboot* package. For each multi-model approach and each catchment, we computed the 5th-95th percentile uncertainty interval around the KGE_{comp} scores using
450 bootstrap resampling. In practical terms, narrower intervals (i.e., values closer to zero) indicate more robust scores, meaning that the KGE_{comp} values vary little with the choice of the time period used for their calculation. Such comparisons are useful to assess whether the differences in KGE_{comp} among multi-model approaches exceed the uncertainty in the scores themselves. Differences in sampling uncertainty across the approaches are minor. Nevertheless, Figure 7 shows slightly lower (i.e. better) sampling uncertainty for combination approaches (light green, dark green and red) compared with the mosaic approaches (light
455 and dark pink) or the single-model benchmark (orange). This tendency likely reflects that selecting a single model for each catchment (or across all catchments) increases the risk of choosing a model that appears to perform well at first glance (i.e., during calibration) but is highly sensitive to the evaluation period (i.e., lacks robustness). In contrast, combination approaches tend to dampen this effect because they aggregate multiple models. Note that the pattern in Figure 7 (sampling-uncertainty analysis) differs from that in Figure 5 (performance analysis): the top-performing approach is not the least uncertain. In other
460 words, although the spatially variable and temporally static combination approach achieves the highest evaluation performance, its scores do not exhibit the lowest uncertainty among all approaches, indicating that its apparent superiority may be less robust across time. By contrast, the dynamic combination approach yields the lowest sampling uncertainty; this likely occurs because the time-varying weights in the dynamic combination adapt to shifts in hydro-climatic conditions, and, as such, the dynamic combination method is more resilient to differences in the temporal samples that are used to quantify sampling
465 uncertainty.

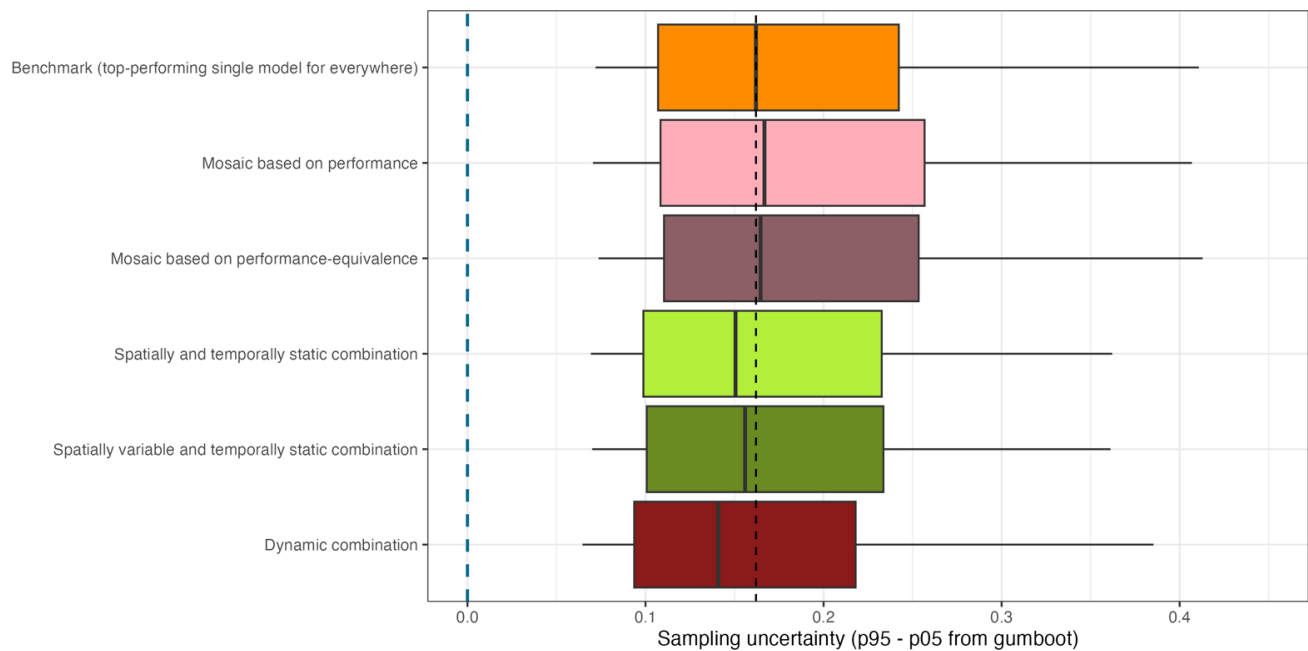
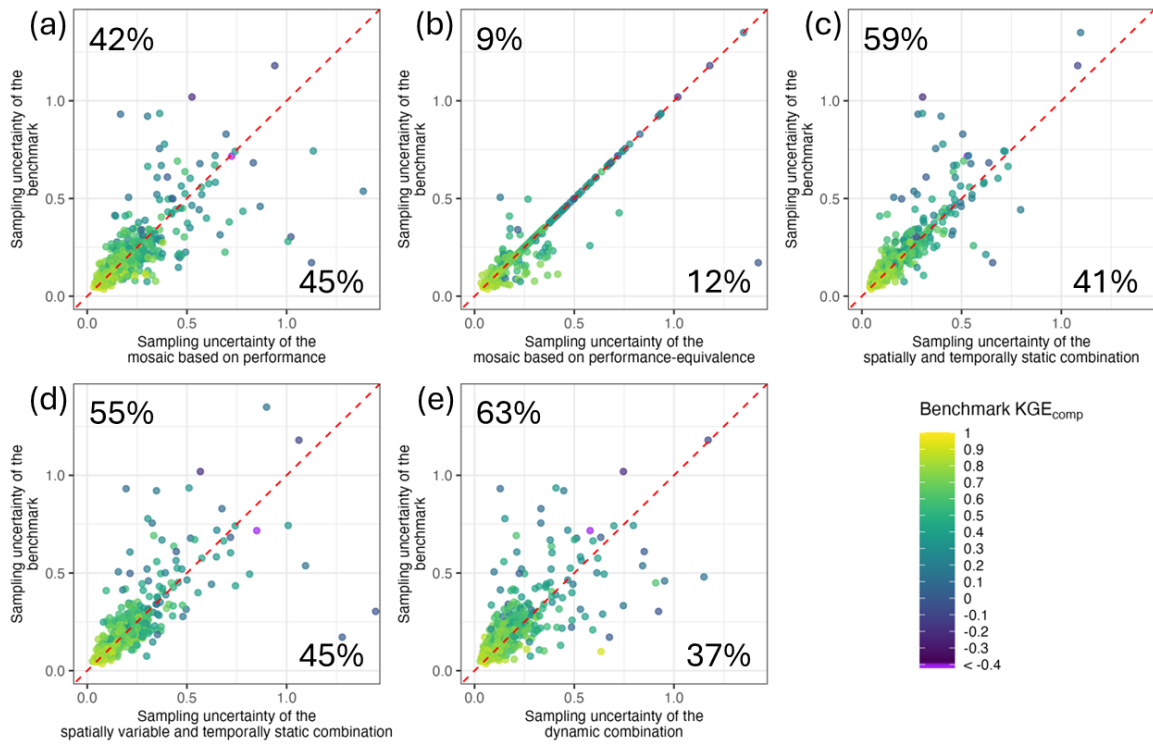


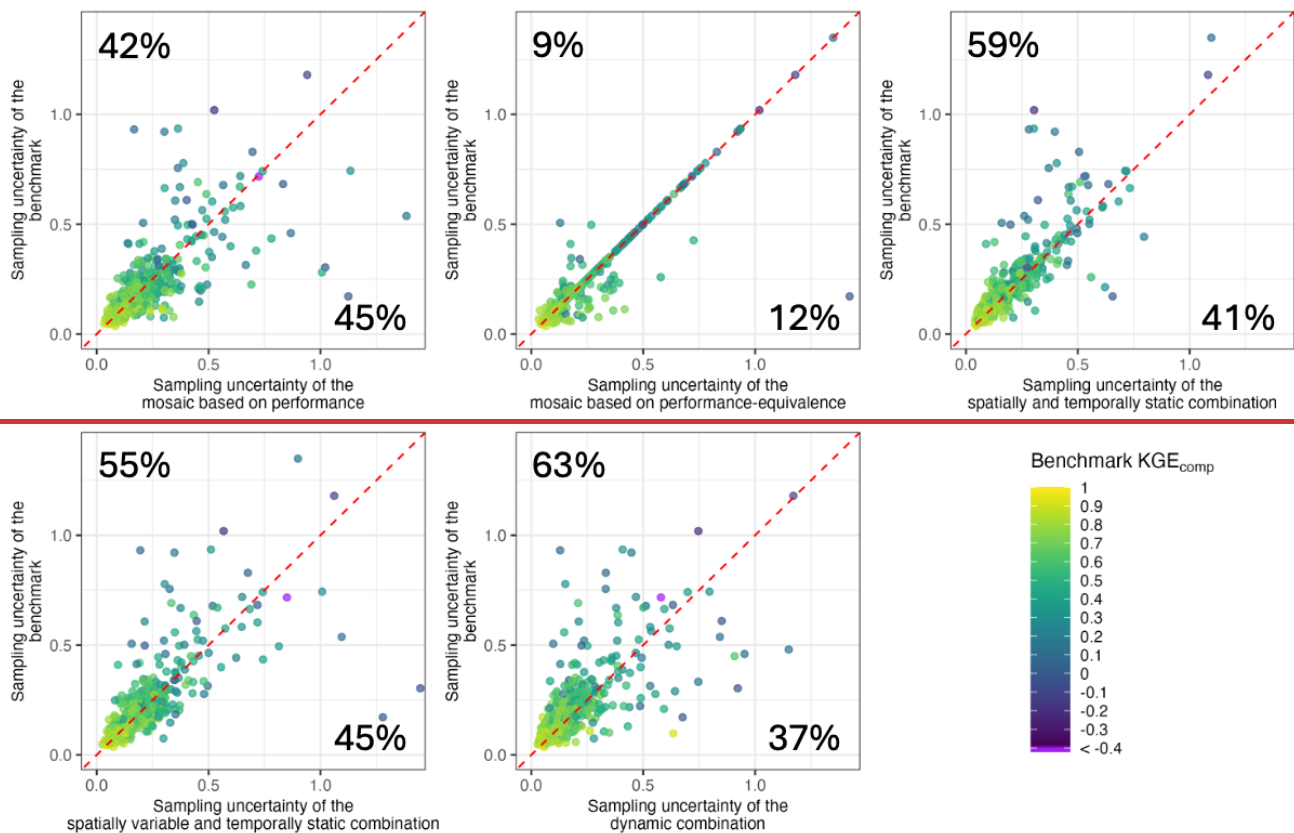
Figure 7: Boxplot of sampling uncertainty surrounding the performance score KGE_{comp} over the evaluation period for the various multi-model approaches. The boxplots represent the 10 %, 25 %, 50 %, 75 % and 90 % quantiles. The dashed black line indicates the median value of the benchmark. The dashed blue line highlights the optimal value of sampling uncertainty range.

470 Although the multi-model combination methods tend to exhibit lower sampling uncertainty when aggregated across all catchments, this does not necessarily hold on a per-catchment basis. Figure 8 illustrates this point by comparing the sampling uncertainty of each multi-model approach against that of the benchmark model for each catchment. The percentages shown in each panel indicate the proportion of catchments where the multi-model approach has lower (bottom-right value) or higher (top-left value) sampling uncertainty than the benchmark. For example, the dynamic combination reduced sampling uncertainty compared to the benchmark for about 63 % of catchments, but increased it for 37 % (Figure 8e), indicating that

475 improvement is not absolute. This result highlights that while multi-model combinations can improve overall sampling uncertainty (Figure 7), they do not systematically yield more robust scores across all catchments. Note that for multi-model mosaic approaches, the percentages do not sum to 100%. This reflects the fact that, in multi-model mosaic approaches, the model used as the benchmark can also be selected to simulate a given catchment, resulting in identical sampling uncertainty

480 for those cases (3% of the catchments for the mosaic based on performance, and 79% for the mosaic based on performance-equivalence). Most of the larger degradations occur in initially poor-performing catchments (blue and purple colours).





485 **Figure 8: Scatter plot comparing sampling uncertainty (defined as the difference between the 5th and 95th percentiles of the KGE_{comp} bootstrap distribution) between the benchmark and the multi-model approaches: (a) mosaic based on performance, (b) mosaic based on performance-equivalence, (c) spatially and temporally static combination, (d) spatially variable and temporally static combination, and (e) dynamic combination.** The dashed red line indicates equality. Dots below the 1:1 line indicate an increase in sampling uncertainty (i.e. a degradation), whereas dots above the line indicate a decrease (i.e. an improvement) resulting from the multi-model approach relative to the benchmark. The numbers indicate the percentage of catchments that fall on each side of the 1:1 line. The colour scale indicates the initial KGE_{comp} performance of the benchmark.

490

3.3 Equivalence

In this section, we assess whether the different multi-model methods provide added value that makes them distinguishable from one another. Using the notion of performance-equivalence introduced in Section 2.4.3, we determine whether two approaches are effectively indistinguishable for a given catchment — i.e., whether the KGE_{comp} score of the lower-performing approach lies within the sampling uncertainty interval of the other, here over the evaluation period.

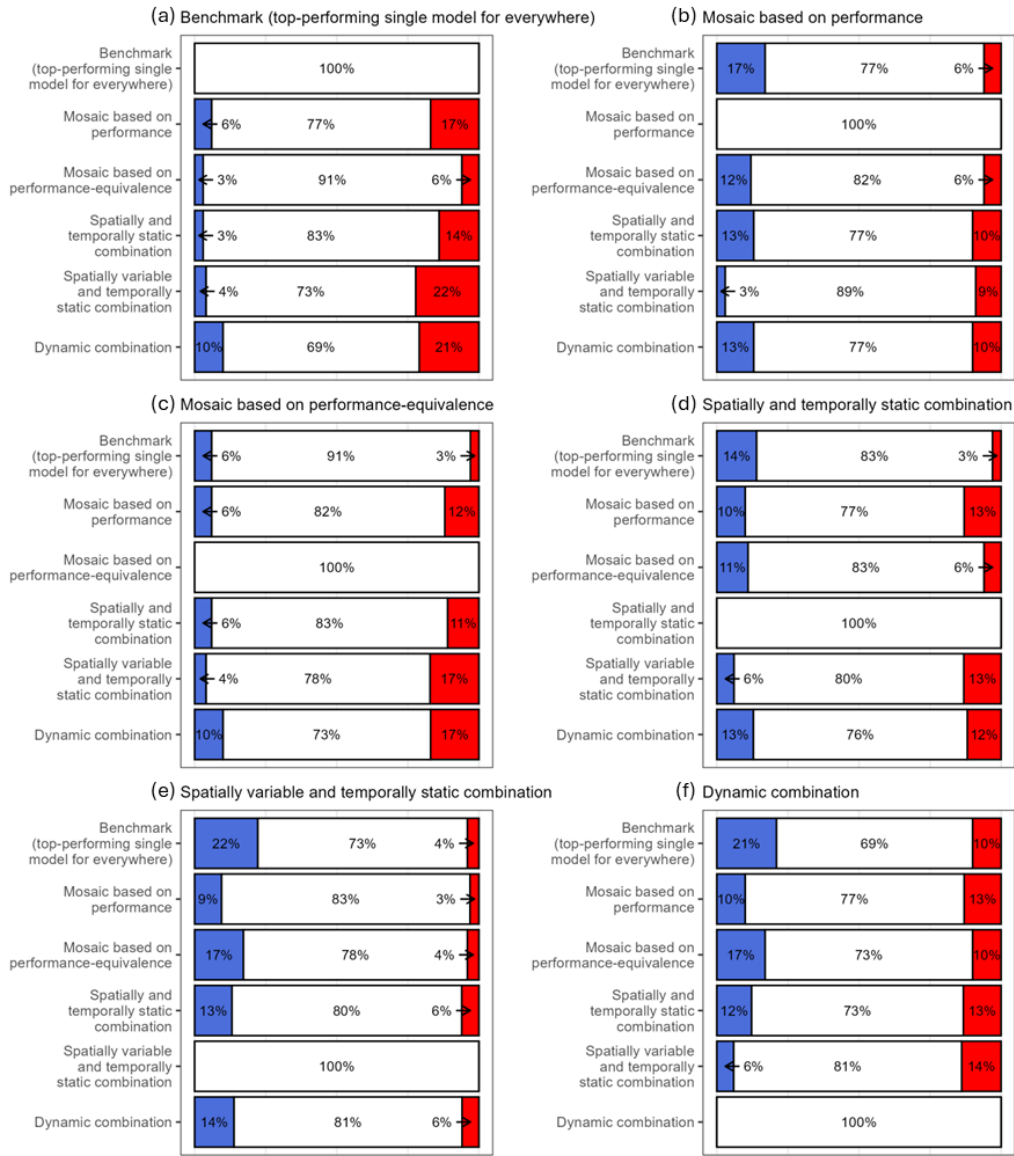
495

Figure 9 summarizes these equivalence analyses. Overall, the various multi-model approaches are indistinguishable for at least 6970% of the catchments (white dots and black numbers bars), regardless of the method considered. In other words, in most catchments, changing the multi-model strategy does not yield statistically meaningful differences in performance. Among the remaining 301% of catchments, no approach consistently outperforms the others. Each method shows a mixture

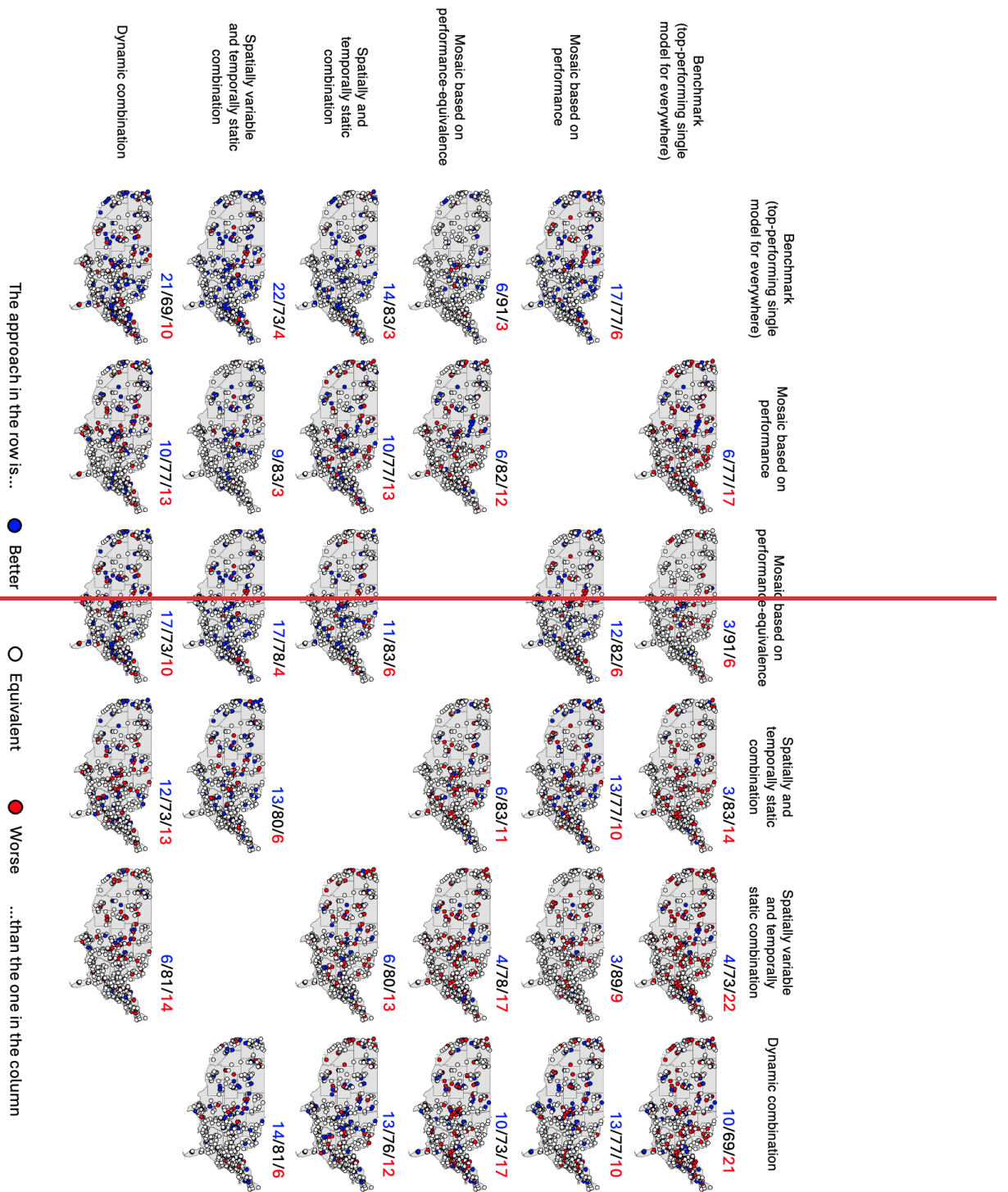
500 of improvements (blue ~~dots~~bars) and deteriorations (red ~~dots~~bars), with no strategy achieving systematic superiority. The benchmark (i.e., the top-performing single model applied everywhere) exhibits the least improvement and the most deterioration in pairwise comparisons, with up to 22~~0~~% of catchments showing worse performance compared to the spatially variable and temporally static combination, and 21% compared~~er~~ to the dynamic combination (Figure 9a). Consistent with the results on performance (Section 3.2.1), the spatially variable and temporally static combination stands out modestly: across
505 catchments, it exhibits the highest proportion of improvements and the lowest proportion of deteriorations relative to other multi-model approaches (Figure 9e). The spatial distribution of the equivalence analysis is provided in Appendix D for completeness.

~~The spatial organization shows that multi-model approaches—particularly the combination-based methods—tend to improve performance in catchments that already exhibit relatively high KGE_{comp} values and low uncertainty. In contrast, deteriorations relative to the benchmark occur mainly in poor performing catchments, where uncertainty is very large. This pattern likely reflects overfitting during the calibration period in catchments where key hydrological processes are either poorly or not at all represented in the models (e.g., prairie potholes, highly arid regions).~~

510



The approach in the title is... ● Better ○ Equivalent ● Worse ...than the one in the y-axis



515 **Figure 9: Maps-Bar plots** comparing the equivalence between each the multi-model approach against all others: (a) benchmark, (b) mosaic based on performance, (c) mosaic based on performance-equivalence, (d) spatially and temporally static combination, (e) spatially variable and temporally static combination, and (f) dynamic combination. The numbers (e.g. 0/100/0) represent

the ~~percentage-percentage~~ of catchments falling in each category (colour) for each pairwise comparison: better (~~blue numbers and dots~~), equivalent (~~black numbers, white dots~~) or worse (~~red numbers and dots~~).

520 4 Discussion

4.1 Does equivalent performance mean equivalent behaviour?

Although the different approaches tested here yield broadly similar performance scores, this does not necessarily imply that they reproduce streamflow dynamics in the same way. The composite KGE metric used for calibration and evaluation was designed to balance high- and low-flow conditions (Garcia et al., 2017), yet it remains a scalar summary of model behaviour.

525 As such, it cannot fully capture important aspects of hydrograph dynamics — such as timing of peak flows, representation of recession limbs, or the sequencing of extreme events — which are known limitations of aggregated performance metrics (Gupta et al., 2009). Two approaches may therefore achieve comparable KGE values while differing substantially in how they represent underlying hydrological processes (e.g., Beven, 2006; Kirchner, 2006; Khatami et al., 2019; Bouaziz et al., 2021; Cinkus et al., 2023). This issue is particularly relevant for multi-model combinations, where averaging effects can mask
530 deficiencies in individual models. Therefore, equifinality remains an important challenge: different model structures or combination schemes can lead to similar scores, but not necessarily to similar hydrological realism (e.g. Butts et al., 2004; Wagener and Gupta, 2005; Renard et al., 2010; Gupta and Govindaraju, 2019). Future analyses should therefore focus on process-oriented diagnostics to better assess to what extent “equivalent performance” implies “equivalent behaviour”. These findings also contribute to the motivation for using ensembles in a probabilistic framework, which aim to represent the range
535 of plausible model behaviours rather than relying on a single simulation.

4.2 Can we link model structure, model performance and catchment attributes?

Establishing systematic links between model structures, their performance, and catchment attributes remains a fundamental challenge in hydrology (e.g., Knoben et al., 2020; David et al., 2022; Kiraz et al., 2023; Spieler and Schütze, 2024). However, the evidence so far suggests that these relationships are often weak or inconsistent across domains. For example, Knoben et al.
540 (2020) showed that model performance aligns more strongly with climate and hydrological signatures than with geomorphological ones (e.g., geology, soils, vegetation). Knoben et al. also ranked the model structures according to structural similarity but found no clear relationship between model structure and catchment attributes. David et al. (2022) demonstrated that although aridity and baseflow index influence the behaviour of different structures, anticipating where a specific structure will perform well remains difficult. Kiraz et al. (2023) hence suggest selecting hydrological model structures a priori based on
545 explicit perceptual models of the region and looking beyond statistical performance alone. This call to incorporate more process-based metrics is echoed by many other studies (e.g., Yilmaz et al., 2008; Althoff and Rodrigues, 2021; Todorović et al., 2022).

Our results point in the same direction. Across the CONUS, several single models exhibit performance comparable to the benchmark (Figure A1) despite substantial structural differences (e.g., differences in lower architecture, baseflow computation,

550 and percolation equations between the two top-performing models). This makes it difficult to establish clear links between model structure and performance. Although we do not conduct a dedicated analysis of catchment attributes, the spatial distribution maps of the models selected within the multi-model mosaic approaches (Figure A2 and Figure A4) do not reveal patterns that correspond to catchment characteristics.

In this study, we also explored multi-model combinations. Within such a modelling framework, the structure–performance–
555 attributes relationship becomes even more complex: the goal is no longer to determine whether a single model is appropriate for a catchment, but to evaluate whether the interactions among multiple models are appropriate for that catchment. Although previous work has shown that diverse ensembles can improve streamflow simulations (e.g., Winter and Nychka, 2010; Seiller et al., 2012; Thébault et al., 2024), the direct connections between ensemble composition and catchment characteristics remain poorly constrained.

560 A promising avenue for addressing these challenges is the use of large-sample emulators that predict model parameters as a function of catchment characteristics (Tang et al., 2025; Farahani et al., 2025) or differentiable models that build a direct connection between catchment attributes and parameters (Feng et al., 2022; Song et al., 2024). Another research perspective is to enrich multi-model mosaic approaches with perceptual-model constraints or additional process-based diagnostics to help reduce equifinality and improve realism.

565 4.3 Why does the dynamic combination not outperform other multi-model approaches?

Although the dynamic combination represents the most sophisticated of the tested approaches, its performance did not surpass that of the simpler methods. Compared to more traditional multi-model approaches such as mosaics (Knoben et al., 2020; Mai et al., 2022; Spieler and Schütze, 2024; Knoben et al., 2025) or static combinations (Shamseldin et al., 1997; Georgakakos et al., 2004; Seiller et al., 2012; Thébault et al., 2024), the dynamic combination approach (Thébault et al., 2025a) was developed
570 recently and deliberately kept simple in this initial development. Thébault et al. (2025) highlighted possible extensions to improve performance, such as incorporating machine learning to increase flexibility at different stages of the method. Such improvements have not yet been made in this study.

There are several further potential explanations for why the dynamic combination method does not perform as well as other multi-model methods in this paper. First, the dynamic combination was optimized using the Mean Absolute Error (MAE), a
575 criterion that tends to emphasize the higher errors more commonly found at higher flows, while our evaluation was based on a composite KGE metric aiming to target both high- and low-flow dynamics. This mismatch may limit the apparent benefits of the method. Second, and as noted by Winter and Nychka (2010), ensemble diversity is key to enhancing the performance of multi-model approaches. Although 78 conceptual models were used, they were all calibrated using only one composite criterion, resulting in limited heterogeneity in how the models target different parts of the hydrograph. This limitation is
580 particularly relevant for the dynamic approach, which is able to select a different model at each timestep and would therefore benefit from having parameter sets trained for specific hydrological conditions.

While the development of the dynamic combination approach is still at an early stage, performance differences remain minimal here (Section 3.2.1), and the results on sampling uncertainty (Section 3.2.2) showed a lower sensitivity to the evaluation period compared to other multi-model approaches. This reduced sensitivity is expected to some extent, because the dynamic combination is the only approach whose weights can adjust over time and thus partially compensate for differences introduced by temporal sampling. In other words, the dynamic combination appears to yield simulations with more robust performance under unseen (in time) data, which is a helpful characteristic when modelling systems undergoing change. It is also worth noting that the dynamic approach uses a different rule for forming the combination than the two other approaches evaluated here. Specifically, the current implementation of the dynamic combination selects the model weights at each time step by assuming that the m models with the highest performance over the k past windows of length τ — chosen to be similar to the current hydrological conditions — will also form the best-performing combination. In practice, this means that the combination is derived directly from individual model performance rankings, rather than by evaluating all possible model combinations. While this is a reasonable heuristic, it is nevertheless not always optimal (see e.g., Ajami et al., 2006; Seiller et al., 2012). Comparing every combination of ~~up to three~~two or three models (as done for the other combination approaches) across the different neighbours at each time step could provide better results but requires much more computational resources (${}^78_2C + {}^78_3C = 79,079$ simulations evaluated instead of the current 78). Future developments could therefore explore a more consistent or adaptive selection of the number of models within the dynamic framework to ensure a fairer comparison with static approaches.

4.4 What are the implications for hydrological prediction and practice?

From an operational perspective, our results highlight a promising pathway for using a single model everywhere. Although the benchmark — i.e., a single model for everywhere — was identified from a comprehensive multi-model evaluation (selected among a large ensemble), once this best-performing structure is known, it can be applied operationally with very limited computational cost. In other words, the initial ensemble exploration represents a research investment, whereas the resulting selected model offers a pragmatic option for agencies or practitioners looking for minimal ongoing complexity. This strategy is therefore very different from usual practice, where the model is selected based on legacy choices or convenience of use (Addor and Melsen, 2019). Yet, realizing this pathway in practice requires operational support systems (e.g. staff training, backup capabilities, product-generation workflows, and input engine) that are flexible enough to accommodate alternative model structures; the absence of such infrastructure is a key reason why operational systems often default to using a single model based on legacy. However, it is important to note that all models and multi-model approaches analysed here originate from FUSE and thus share structural similarities, which may partly explain why the top-performing single model performs comparably to the multi-model approaches. Consequently, the conclusion that a single model can perform well across diverse catchments should be interpreted with caution, as it may depend on the diversity of model structures and hydroclimatic conditions considered. Expanding the analysis to include a wider range of model architectures or climate regimes would help to further test the robustness of this finding.

Starting from an ensemble of models can also be useful to further explore the structural uncertainties of different models.

615 Following the advances highlighted by the HEPEX community (Schaake et al., 2007; Ramos et al., 2018), there is a growing shift toward probabilistic modelling frameworks. In this regard, the FUSE model ensemble offers a particularly valuable tool, enabling the exploration of a wide range of model structures while maintaining low computational costs due to its conceptual simplicity.

620 At the same time, the sampling uncertainty analysis underscores that combination approaches provide greater robustness by being less sensitive to the evaluation period. For operational hydrology, this robustness may be more important than marginal gains in performance, especially when forecasting floods or low flows, or generating forecasts under changing climatic conditions. Dynamic combinations, although not yet fully mature, remain particularly promising in this regard, as they explicitly allow model weights to vary through time in response to shifting hydrological conditions.

5 Conclusions

625 Our study aims to compare various multi-model approaches across a large sample of catchments for streamflow simulations. To this end, ~~559~~544 catchments from the CAMELS dataset (Addor et al., 2017) are used, and an ensemble of 78 lumped structures is built within FUSE (Clark et al., 2008). Six different multi-model approaches — a single model selected from a larger ensemble, a mosaic based on performance, a mosaic based on performance-equivalence, a spatially and temporally static combination, a spatially variable and temporally static combination, and a dynamic combination — are tested and compared.

630 In conclusion, our analysis highlights the following key points:

- Multi-model approaches, as tested here and commonly found in the literature, show similar distributions ~~for~~ of performance score and associated sampling uncertainty across a large sample of catchments. Yet, differences remain in a per-catchment comparison. Indeed, multi-model approaches achieve equivalent performance for more than 70% of the catchments, but no approach consistently outperforms the other approaches in the remaining 30% of catchments.

- Compared to the other methods, the spatially variable and temporally static combination shows a slightly better performance score distribution, together with ~~among one of~~ the lowest (i.e. best) sampling uncertainty distributions, leading to the highest (i.e. best) improvement/deterioration ratio among the various multi-model approaches tested. The dynamic combination approach shows the lowest sampling uncertainty across the sample of catchments, ~~which~~ highlightings its relative robustness when applied to unseen data.

- The availability of a large structural ensemble creates scope to identify a single model that performs well across all catchments, ~~where the whose~~ performance and sampling uncertainty ~~of a single model~~ are comparable to ~~that those~~ of more complex multi-model approaches.

645 Looking forward, several avenues for research emerge from this study. Given the strong performance of the one-size-fits-all benchmark identified in this study, a promising extension would be to test adaptive calibration strategies within a single-

structure framework. For instance, the multi-calibration approach proposed by (Oudin et al., (2006), which combines parameter sets optimized for contrasting flow regimes, could offer a parsimonious alternative to multi-model strategies while retaining structural simplicity. Exploring whether such regime-dependent calibration could further enhance the robustness of a single well-chosen structure across diverse hydro-climatic conditions would constitute a valuable direction for future research. First,

650 ~~w~~While the dynamic combination approach shows no clear benefits compared to simpler multi-model strategies, it remains in its infancy; future work should explore its potential using ensembles with greater structural diversity (e.g. physically-based or machine learning models), alternative calibration strategies (e.g. various objective functions) and a larger evaluation framework (e.g. several metrics, hydrological signatures). Importantly, this study accounts for sampling uncertainty in model performance metrics, demonstrating that apparent differences in skill can often be statistically indistinguishable. This

655 highlights that performance scores alone may be insufficient to identify or select appropriate model structures. Moreover, an important next step is to test multi-model approaches in ungauged catchments, where model choice or weights must be transferred in space; in this context, linking multi-model performance more explicitly to catchment attributes could help improve understanding when a specific model is needed and which one. In addition, given the computational burden of large ensembles, future studies should also assess the trade-offs between complexity and practicality, to determine when

660 sophisticated multi-model schemes are justified compared to simpler benchmark strategies. Finally, evaluating the robustness of single- and multi-model approaches under non-stationary conditions — for example using dedicated frameworks such as the Robustness Assessment Test (RAT; (Nicolle et al., 2021)) or structured climate-sensitivity experiments (e.g., Wood et al., 2004) — such as climate change, land-use shifts, or extreme events remains an open challenge. This issue is particularly critical in the context of climate change, land-use shifts, or increasing hydro-climatic extremes, where model transferability becomes a central concern for operational forecasting and long-term water resource planning, yet is critical for operational forecasting and long-term water resource planning. In such contexts, ~~shifting moving beyond from~~ a deterministic paradigm based on a single prediction to a probabilistic framework using ~~an ensemble~~ offers may provide a practical means to better represent parameter and ~~model~~-structural uncertainties.

Benchmark:

As a reminder, the benchmark is defined as the model that attains the highest median KGE_{comp} value over the calibration period and across all the catchments. Figure A1 provides an alternative representation of Figure 3a, showing the distribution of model performance across catchments for each individual structure. Based on this criterion, model no. 126 is selected as the benchmark. Note that this structure is mixed, i.e. it does not correspond to any of the original models (VIC, PRMS, SAC-SMA, and TOPMODEL). The details of its structural configuration are provided in Table A1.

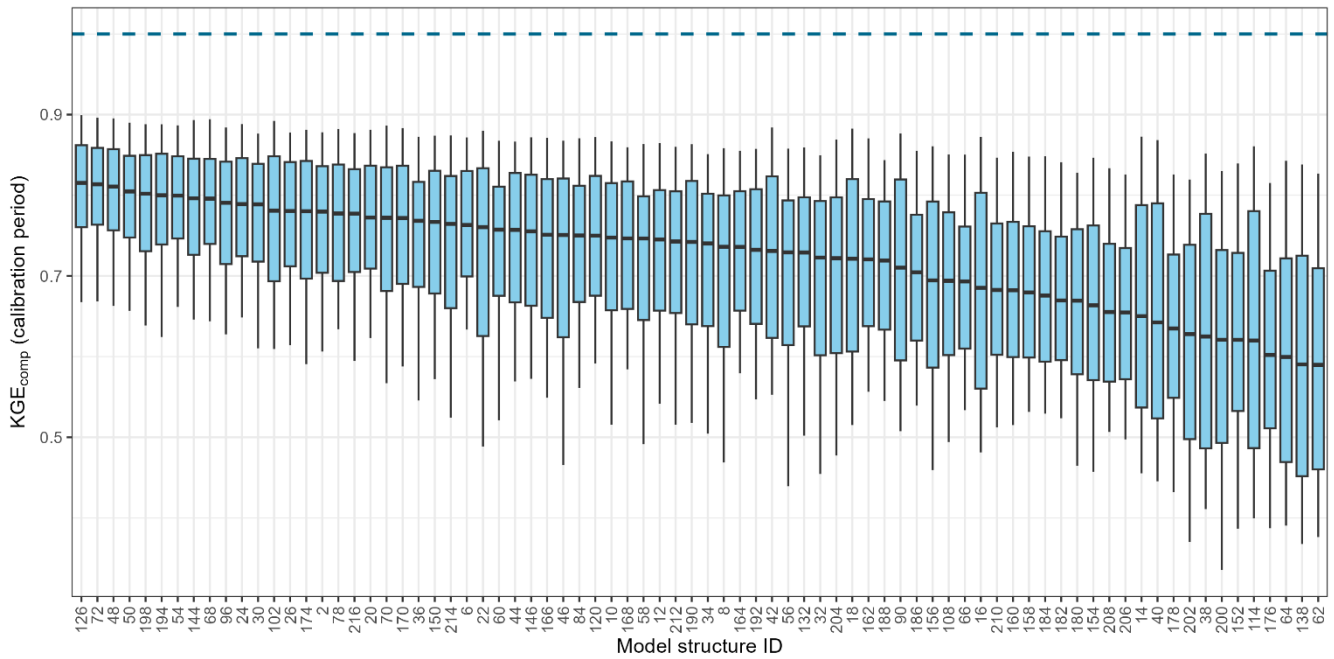
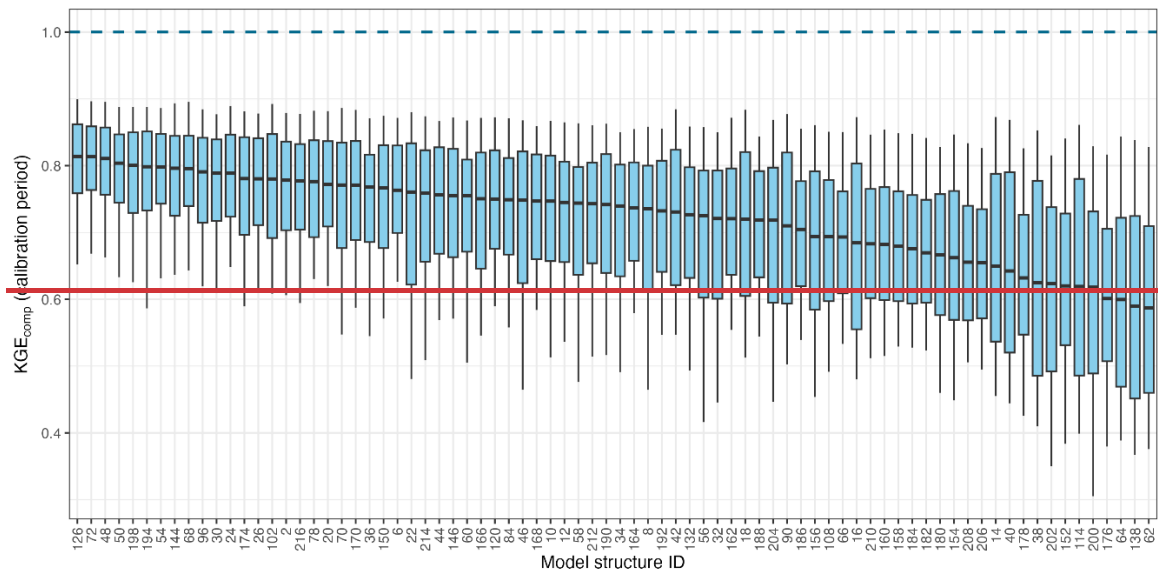


Figure A1: Boxplot of performance scores, KGE_{comp} , over the calibration period for each of the 78 individual FUSE structures across all catchments. The boxplots represent the 10 %, 25 %, 50 %, 75 % and 90 % quantiles. The dashed blue line highlights the optimal value of KGE_{comp} .

Table A1: Description of the structure selected as benchmark, structure no. 126. Here, we keep the notation as in Notations are derived from Clark et al. (2008), with the same notations.

Structure component	Description of the selected decision	Equation of the selected decision
Rainfall error	Multiplicative correction coefficient	$p_{eff} = p_{ini}\epsilon_{mlt}$
Upper-layer architecture	Single state variable	$\frac{dS_1}{dt} = (p - q_{sx}) - e_1 - q_{12} - q_{if} - q_{ufof}$
Lower-layer architecture	Tension reservoir combined with two parallel linear reservoirs	$\frac{dS_2^T}{dt} = \kappa q_{12} - e_2 - q_{stof}$ $\frac{dS_2^{FA}}{dt} = \frac{(1 - \kappa)q_{12}}{2} - \frac{q_{stof}}{2} - q_b^A - q_{sfofa}$ $\frac{dS_2^{FB}}{dt} = \frac{(1 - \kappa)q_{12}}{2} - \frac{q_{stof}}{2} - q_b^B - q_{sfofb}$
Baseflow		$q_b = v_A S_2^{FA} + v_B S_2^{FB}$
Surface runoff	Saturation-excess mechanism using TOPMODEL parametrization	$A_c = \int_{\zeta_{crit}}^{\infty} f(\zeta) d\zeta$
Percolation	Continuous gravity drainage	$q_{12} = k_u \left(\frac{S_1}{S_{1,max}} \right)^c$
Evaporation	Sequential	$e_1 = pet \frac{\min(S_1^T, S_{1,max}^T)}{S_{1,max}^T}$ $e_2 = (pet - e_1) \frac{\min(S_2^T, S_{2,max}^T)}{S_{2,max}^T}$
Interflow	No interflow	$q_{if} = 0$
Routing	Gamma distribution	$P(a, x) = \frac{\gamma(a, x)}{\Gamma(a)}$
Snow model	Temperature index snow model running on elevation-bands (Henn et al., 2015)	$p_{snow} = \begin{cases} p_{eff}, & T \leq T_{snow} \\ 0, & T > T_{snow} \end{cases}$ $s_{melt} = \max(\min[f\{T - T_{melt}\}, SWE], 0)$ $\frac{dSWE}{dt} = p_{snow} - s_{melt}$ $p = p_{eff} - p_{snow} + s_{melt}$

685 Mosaic based on performance:

The mosaic based on performance selects a single model for each catchment according to the highest KGE_{comp} value during the calibration period. Figure A2 shows that 57 of the 78 models are selected at least once as the top-performing model for a catchment. Structure no. 126 (Table A1) is the most dominant, being selected in 13% of the catchments (73 out of 54452).

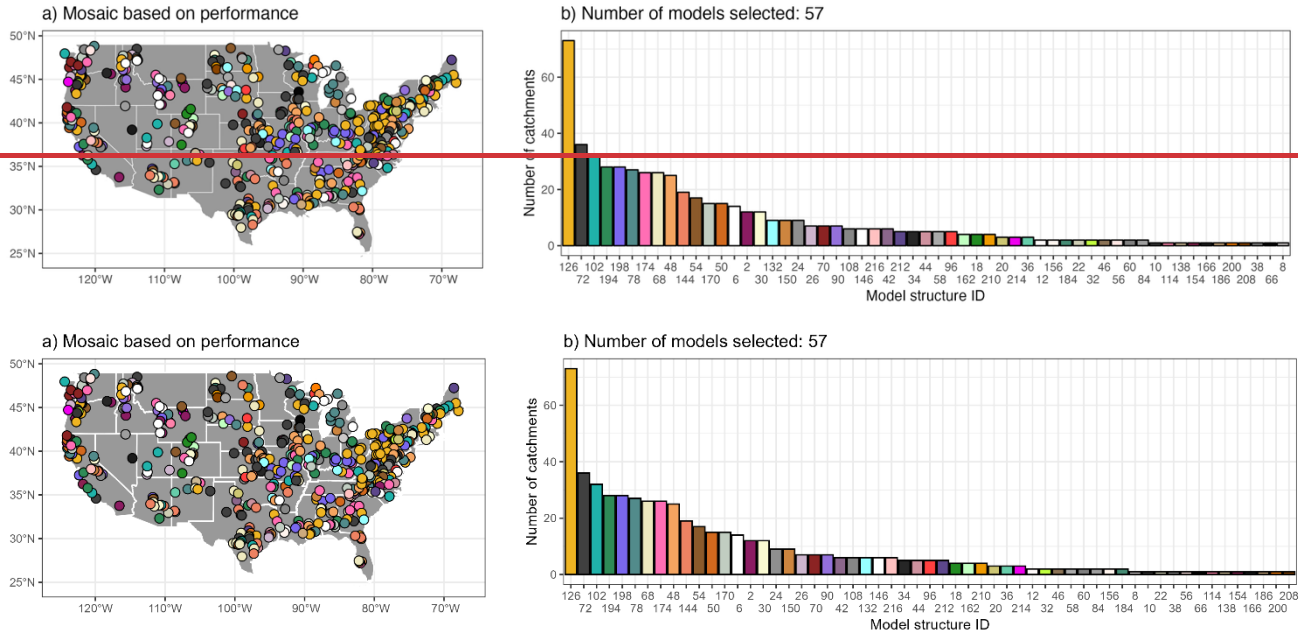
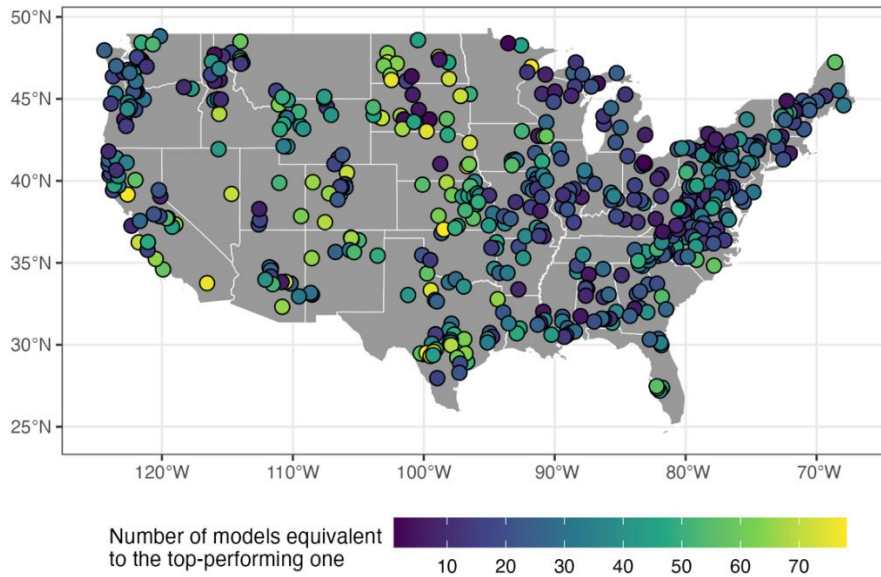


Figure A2: (a) spatial distribution of the models selected within a mosaic based on performance and (b) histogram illustrating the number of catchments where a specific model is selected, ranked by their frequency of selection.

Mosaic based on performance-equivalence:

695 The performance-equivalence mosaic also selects a single model for each catchment. In this case, model equivalence is defined by considering both performance and the associated sampling uncertainty, resulting in multiple candidate models for a given catchment (Figure A3). Figure A3 highlights that a large number of structures can be considered equivalent to the top-performing model in most catchments: at least 10 structures are equivalents in more than 90% of the catchments, and the median number of equivalent structures is 28.



700

Figure A3: Spatial distribution of the number of models that are performance-equivalent to the top-performing model for each catchment.

The final model assignment is then determined by minimizing the number of distinct models used across all catchments. Figure A4 shows that only 8 of the 78 models are required to cover all the catchments. Notably, structure no. 126 (Table A1) alone accounts for 80% of the catchments (430 out of 54452). Interestingly, structure no. 72 is not used at all within this framework despite being the second top-performing single model in Figure A1. This suggests similarities between different models. It is important to note that equivalence in this context refers solely to statistical indistinguishability with respect to KGE_{comp} within sampling-uncertainty bounds. Such equivalence does not imply identical hydrograph dynamics or process representations.

705

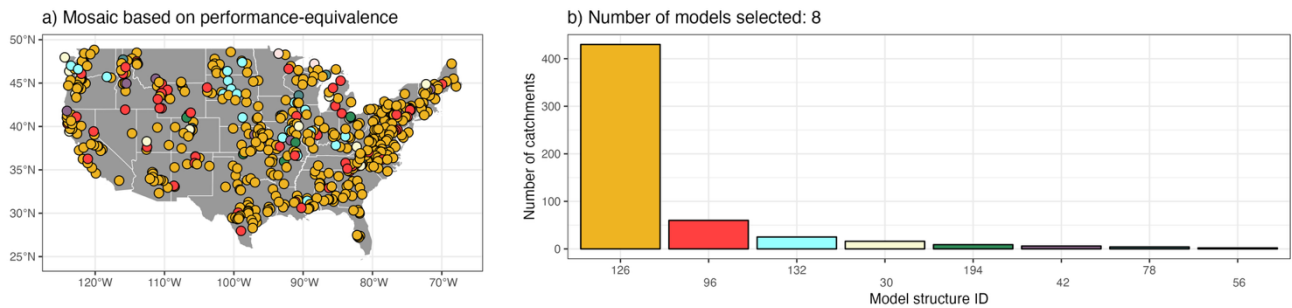


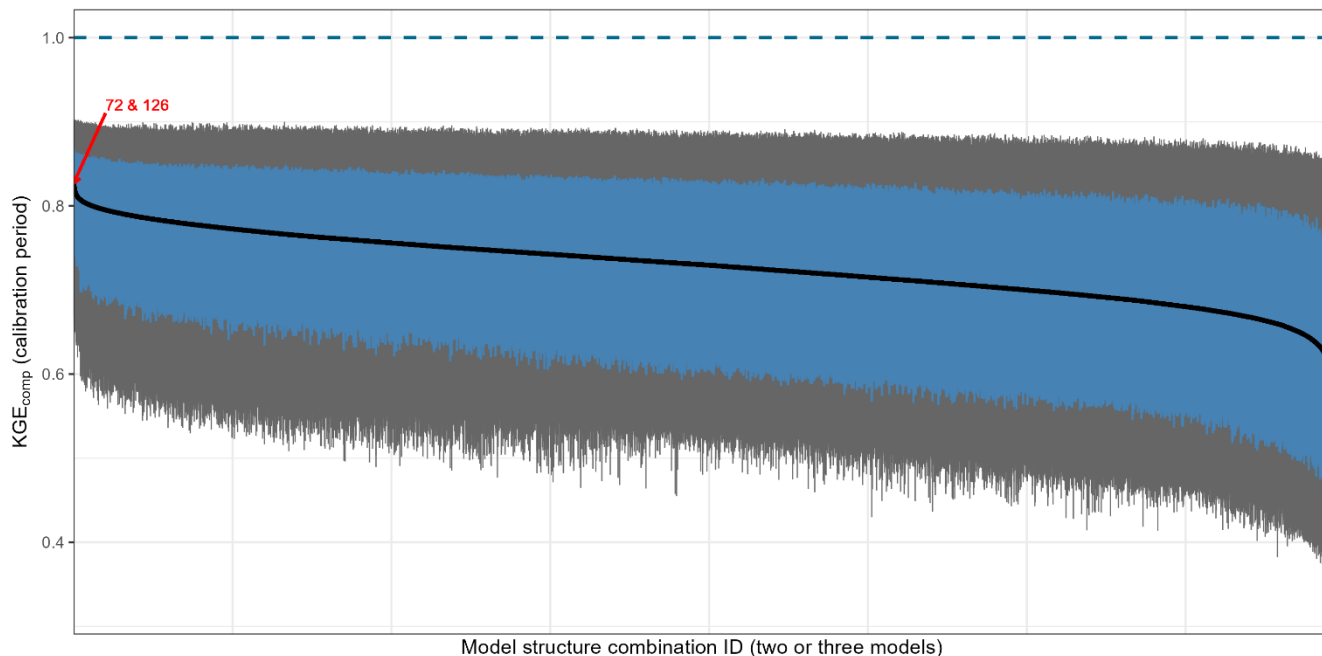
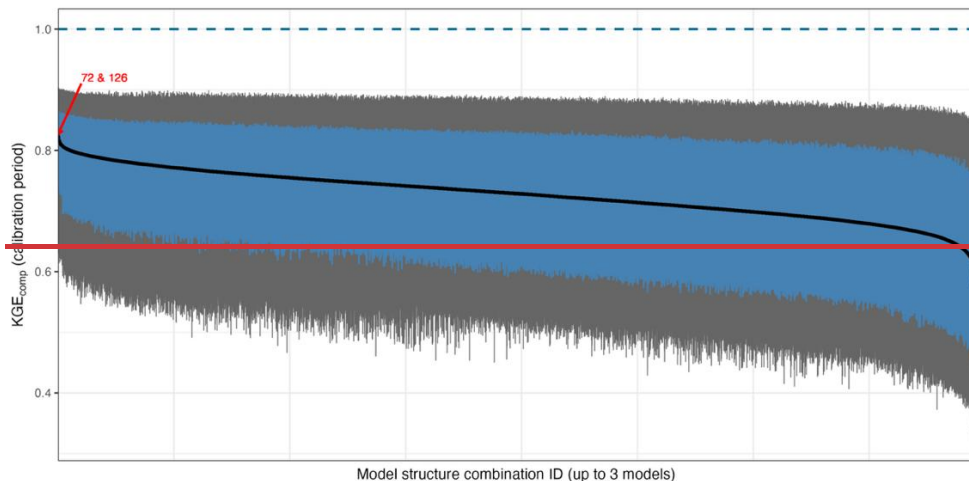
Figure A4: (a) spatial distribution of the models selected within a mosaic based on performance-equivalence and (b) histogram illustrating the number of catchments where a specific model is selected, ranked by their frequency of selection.

Spatially and temporally static combination:

The spatially and temporally static combination selects a single combination of models (here, up to threetwo or three models) that attains the highest median KGE_{comp} value over the calibration period and across all the catchments. Model combinations are computed by simple averaging, i.e., each selected model receives an equal weight. Figure A5 shows the distribution of

715

720 model performance across all catchments for each possible combination. The top-performing combination is ~~the average of~~ based on models no. 72 and no. 126, which ~~is therefore selected to represent all catchments under the spatially and temporally static combination framework~~ are also the two top-performing models selected in the mosaic based on performance (Figure A2). Note that although combinations of up to three ~~three~~ models were permitted, the top-performing option over the entire domain consists of only two structures. ~~Interestingly, these are the two top-performing models in Figure A2, but model 72 is not selected in Figure A4, suggesting a large degree of similarity (i.e., equivalence) between both models.~~

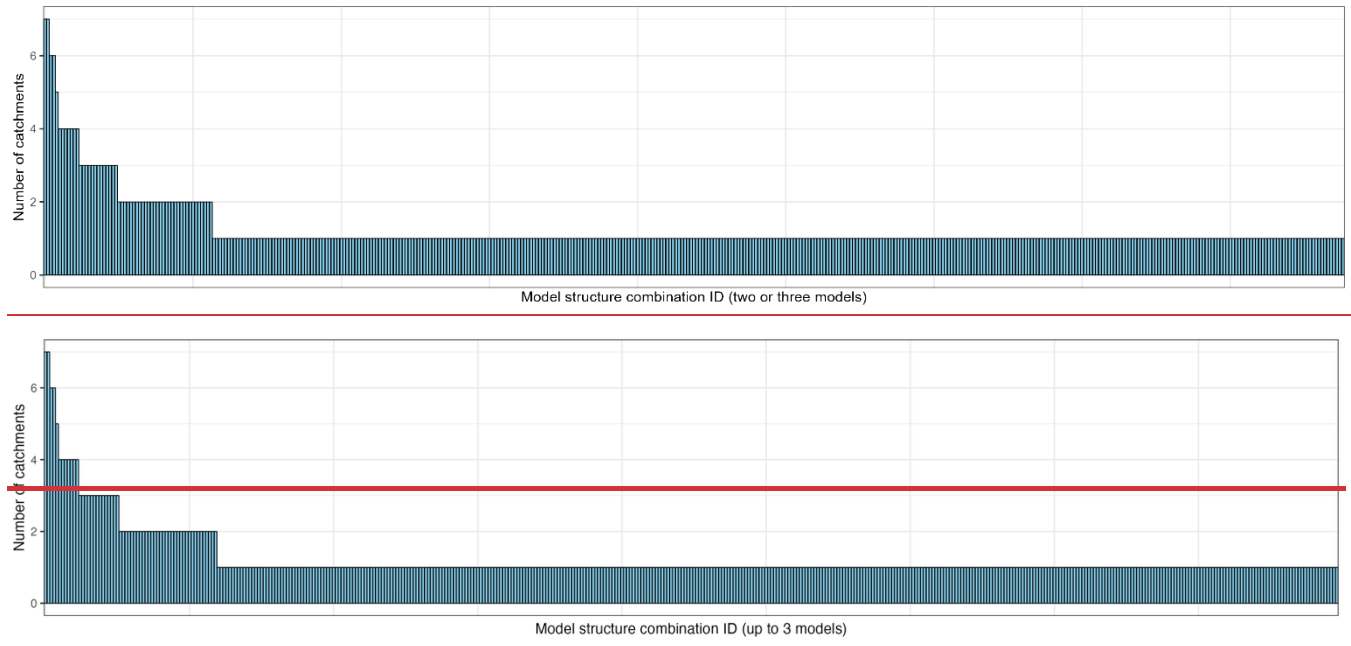


725 **Figure A5: Distribution of performance scores, KGE_{comp} , over the calibration period for each of the 79,079 combinations across all catchments. The boxplots represent the 10 %, 25 %, 50 %, 75 % and 90 % quantiles. The dashed blue line highlights the optimal value of KGE_{comp} . The number in red highlights the top-performing combination.**

Spatially variable and temporally static combination:

730

The spatially variable and temporally static combination selects a single combination of models (here, ~~up to three~~ two or three models) independently for each catchment based on performance over the calibration period. Figure A6 shows that almost every catchment required a different combination to achieve the highest performance: 4439 distinct combinations are selected across the ~~54452~~ catchments. Note that ~~all-most of the~~ combinations selected for more than four catchments include structure no. 126 (Table A1) and they all consists of only two structures. This result supports that testing combinations of four models was unnecessary in our case.

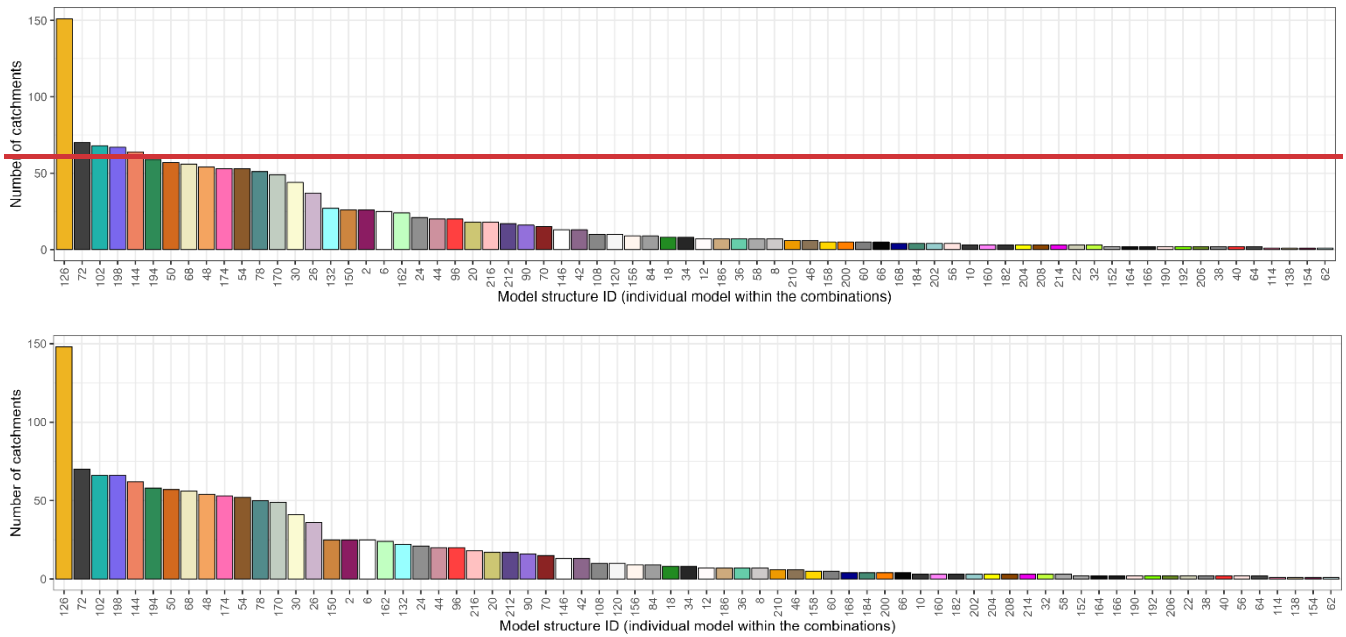


735

Figure A6: Histogram illustrating the number of catchments where a specific combination of models is selected, ranked by their frequency of selection, for the spatially variable and temporally static combination approach.

740

This analysis can be extended to the model level, i.e. identify which individual models are most frequently selected within the combinations. Figure A7 shows that all 78 structures are used at least once in a top-performing combination. In addition, structure no. 126 stands out once again, being selected in the top-performing combinations for 27% of the catchments (14850 out of ~~54452~~), with structure 72 being a distant (and barely) second.



745 **Figure A7: Histogram illustrating the number of catchments where a specific model is selected (within the combinations), ranked by their frequency of selection, for the spatially variable and temporally static combination approach.**

Dynamic combination:

750 The dynamic combination adapts the models included in the combination across both space and time based on similarities in past hydrological conditions. This method requires three parameters: the length of the time window (τ , ranging from 4 to 28 days), the number of nearest neighbours (k , ranging from 1 to 19), and the number of models to combine (m , ranging from 1 to 19). Figure A8 provides the distribution of these parameters across catchments. The parameter controlling the length of the time window (τ) reveals a clear divide between catchments that require short periods and those that require long periods to characterize current conditions and similar past windows. This difference likely reflects differences between catchments dominated by fast, event-driven dynamics and those governed by slower, longer-term processes. The distribution of the number of models to be combined is strongly skewed toward low values, with $m \leq 5$ in nearly 70% of the catchments.

755

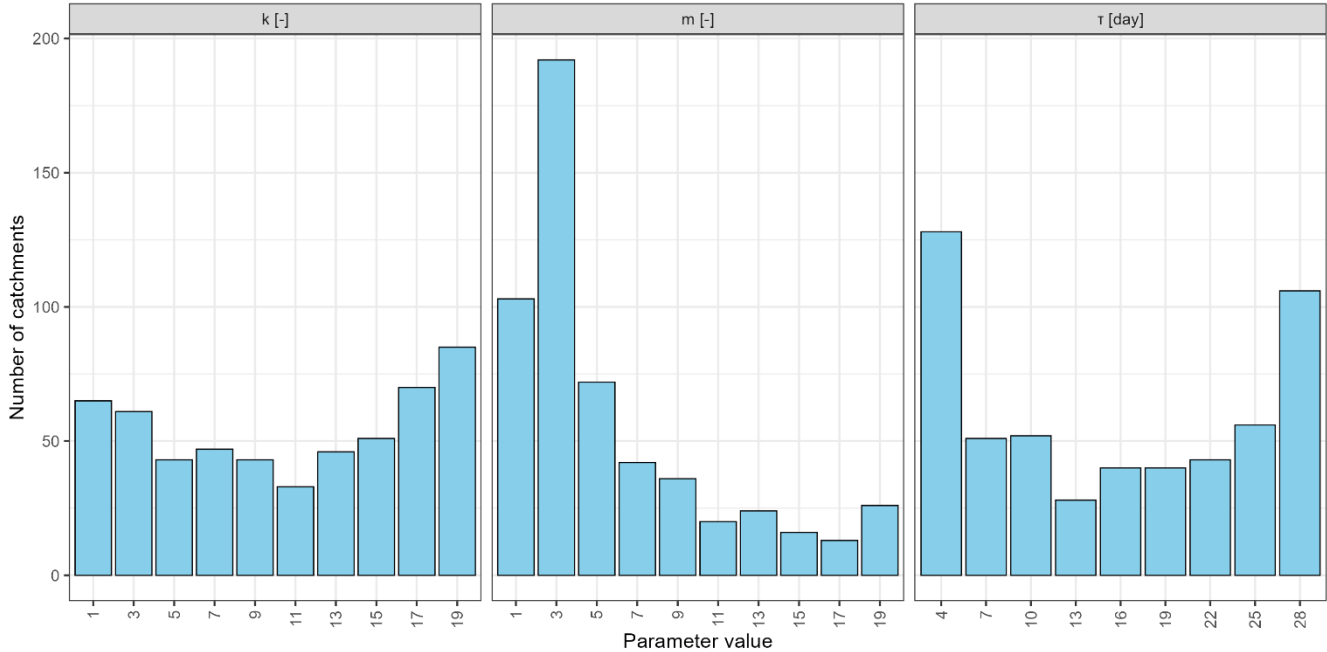
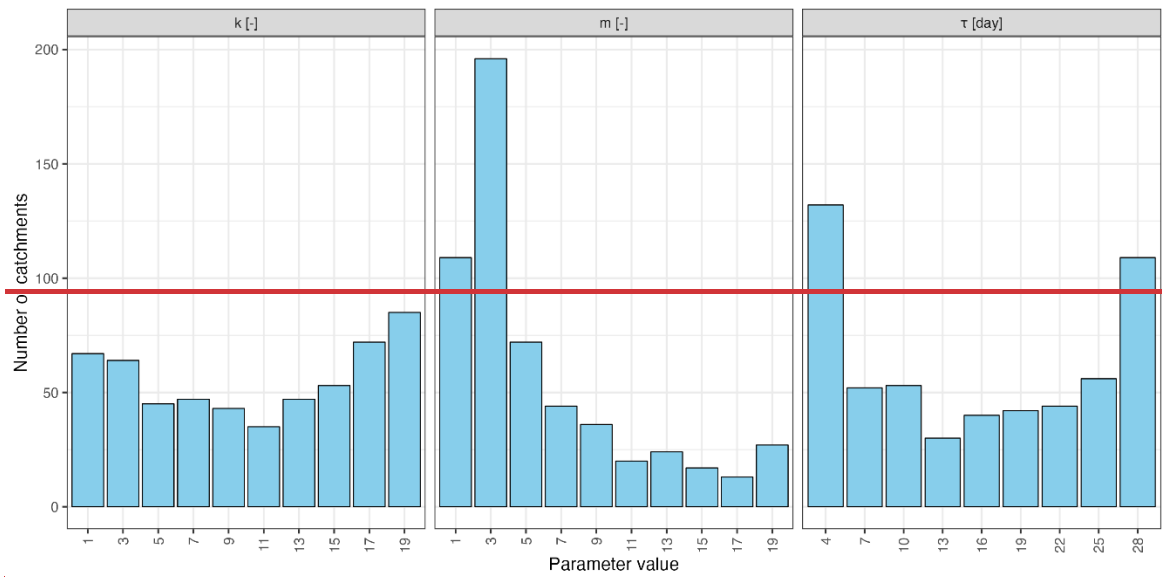


Figure A8: Distribution of the dynamic combination parameters (k , m and τ) across all the catchments.

760

Because the combinations vary over time and may include up to 19 different models, analysing the individual combinations is impractical. However, an analysis at the structure level — i.e. examining how often each individual model is included in the combinations — remains feasible. Figure A9 shows the percentage of use (both temporal and spatial) for each structure. All 78 structures are used at least once in at least one catchment. In contrast to the results from the previous multi-model approaches, structure no. 126 does not stand out markedly from the others. The small range in usage percentages (from 0.5 %

to 23.85 %) indicates that all models are used relatively frequently in the dynamic combination, suggesting that temporal dynamics must play a key role in model selection.

765

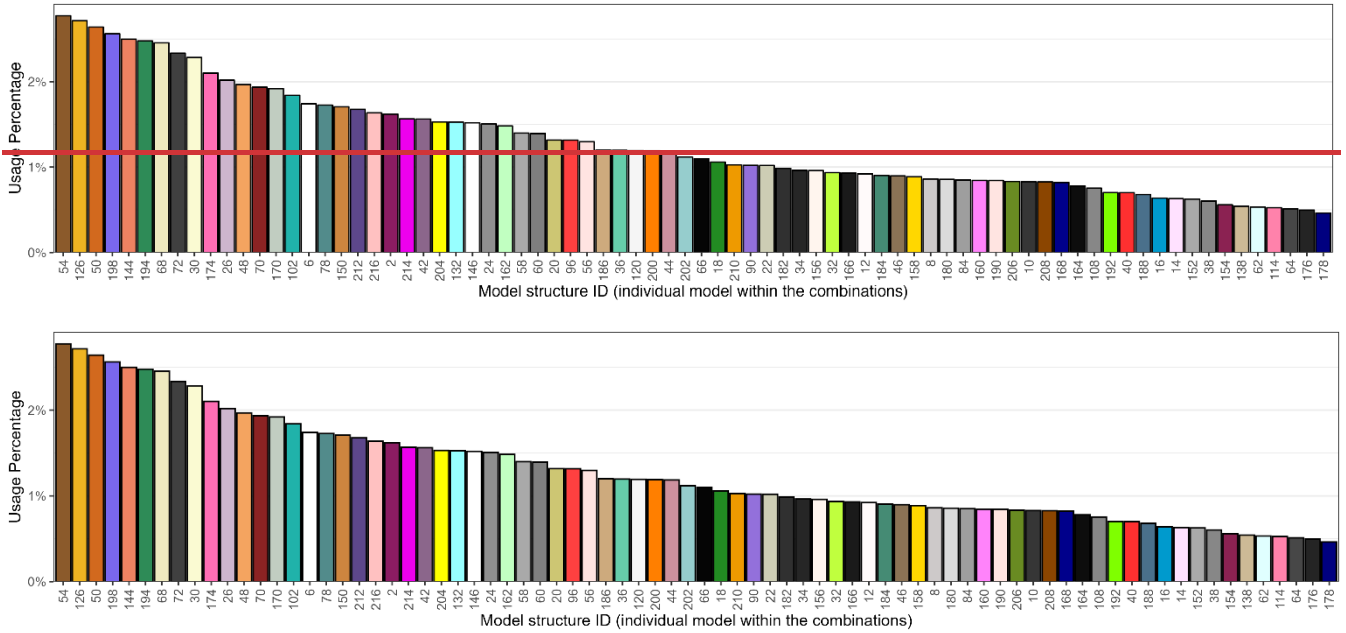
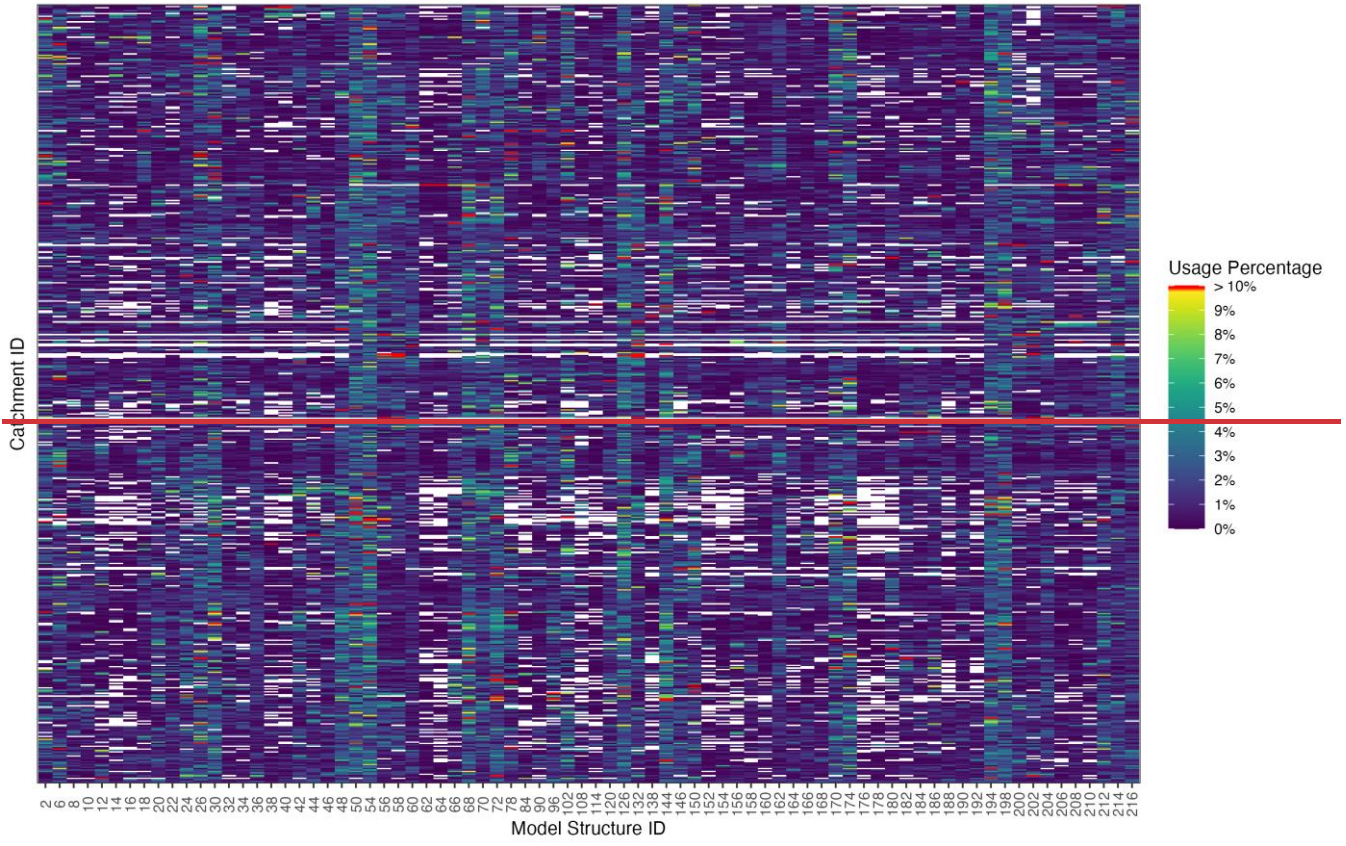
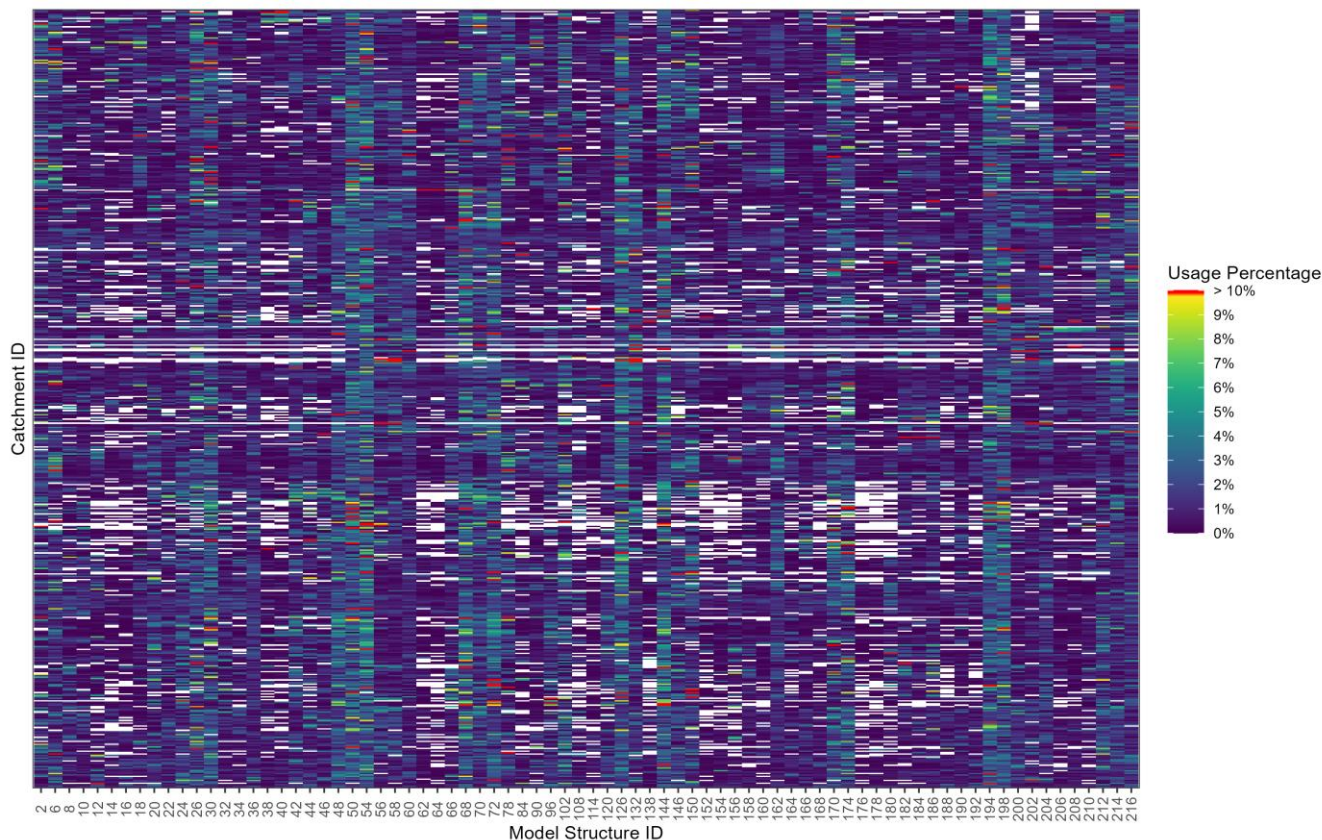


Figure A9: Histogram illustrating the number of catchments where a specific model is selected (within the combinations), ranked by their frequency of selection, for the dynamic combination approach.

770

This analysis can be further refined to examine the temporal dynamics of model selection at the scale of each catchment. Figure A10 shows the percentage of use (through time) of each structure (within the combinations) for every catchment. Notably, several vertical bands of lighter colours indicate specific models that are frequently selected across many catchments. Conversely, some horizontal lines exhibit gaps, showing that only a limited subset of models is selected for those specific catchments. It is worth noting that a temporal-scale analysis would also be possible. However, given the catchment-specific dynamics, it is difficult to identify general trends without first grouping the catchments into clusters.





780 **Figure A10: Percentage of use of each model structure within the dynamic combinations for every catchment. Each row corresponds to a catchment and each column to a model structure. Colours indicate the proportion of time steps in which a given structure is included in the optimal combination for that catchment. White colour indicates 0%. Catchments are ordered according to their gauged station identifiers; it does not reflect performance or catchment attributes, although station numbering broadly follows a regional logic.**

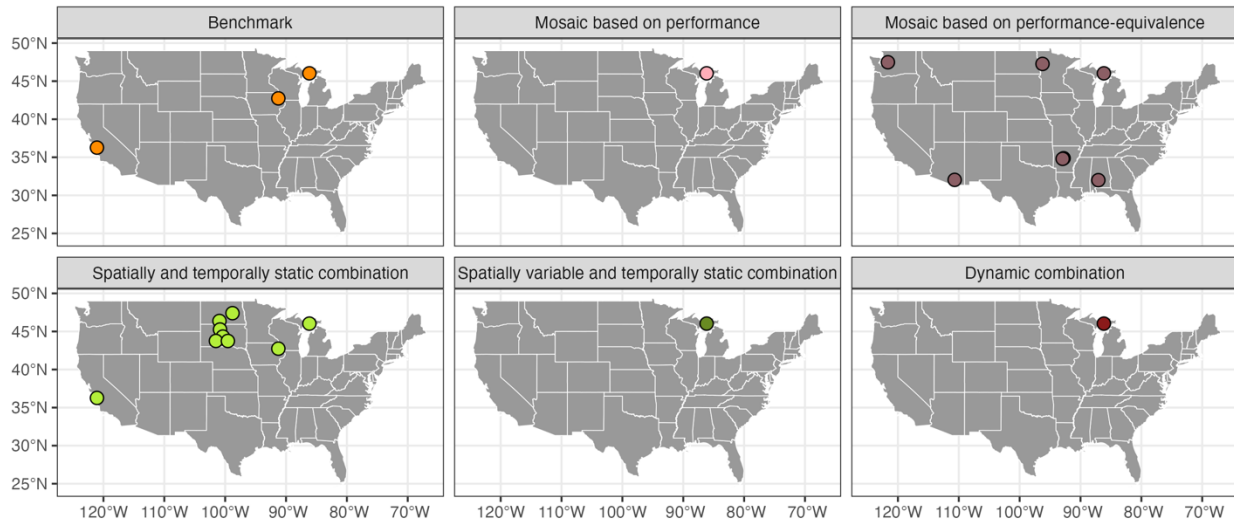
Appendix B Appendix B: Individual models and sampling uncertainty failures.

785 Among the initial 559 catchments, 15 were removed due to model failures or because sampling uncertainties could not be computed. Figure B1 shows the location of the catchments experiencing such failures for each approach. The following bullet points explain the failure for each catchment (USGS gauge ID):

- 02427250: Failed on the mosaic based on performance-equivalence due to a criterion in the sampling-uncertainty calculation: only nine years contained more than 100 valid days over the calculation period (here the calibration), whereas ten years are required.
- 04056500: All approaches failed because no model calibrated successfully for this catchment. The snow module failed due to a negative area for elevation band 2 in the CAMELS file.

- 05062500: Failed on the mosaic based on performance-equivalence due to a criterion in the sampling-uncertainty calculation: only nine years contained more than 100 valid days over the calculation period (here the calibration), whereas ten years are required.
- 795 • 05412500: Failed on the benchmark and the spatially and temporally static combination because both structure no. 72 and no. 126 did not calibrate successfully. Notably, 71 out of 78 structures did not calibrate due to a negative area for elevation band 3 in the CAMELS file.
- 06354000: Failed on the spatially and temporally static combination because structure no. 72 did not calibrate successfully. Notably, 56 out of 78 structures did not calibrate due to a negative area for elevation bands 2 & 4 in the CAMELS file.
- 800 • 06360500: Failed on the spatially and temporally static combination because structure no. 72 did not calibrate successfully. Notably, 66 out of 78 structures did not calibrate due to a negative area for elevation bands 3, 4 & 5 in the CAMELS file.
- 06441500: Failed on the spatially and temporally static combination because structure no. 72 did not calibrate successfully. Notably, 64 out of 78 structures did not calibrate due to a negative area for elevation bands 3 & 4 in the CAMELS file.
- 805 • 06447000: Failed on the spatially and temporally static combination because structure no. 72 did not calibrate successfully. Notably, 59 out of 78 structures did not calibrate due to a negative area for elevation bands 3, 4 & 5 in the CAMELS file.
- 06452000: Failed on the spatially and temporally static combination because structure no. 72 did not calibrate successfully. Notably, 64 out of 78 structures did not calibrate due to a negative area for elevation bands 2, 3, 4, 5 & 7 in the CAMELS file.
- 810 • 06468250: Failed on the spatially and temporally static combination because structure no. 72 did not calibrate successfully. Notably, 64 out of 78 structures did not calibrate due to a negative area for elevation band 1 in the CAMELS file.
- 815 • 07263295: Failed on the mosaic based on performance-equivalence due to a criterion in the sampling-uncertainty calculation: only nine years contained more than 100 valid days over the calculation period (here the calibration), whereas ten years are required.
- 07362587: Failed on the mosaic based on performance-equivalence due to a criterion in the sampling-uncertainty calculation: only nine years contained more than 100 valid days over the calculation period (here the calibration), whereas ten years are required.
- 820 • 09484600: Failed on the mosaic based on performance-equivalence due to a criterion in the sampling-uncertainty calculation: only nine years contained more than 100 valid days over the calculation period (here the calibration), whereas ten years are required.

- 11151300: Failed on the benchmark and the spatially and temporally static combination because structure no. 126 did not calibrate successfully.
- 12141300: Failed on the mosaic based on performance-equivalence due to a criterion in the sampling-uncertainty calculation: only nine years contained more than 100 valid days over the calculation period (here the calibration), whereas ten years are required.



830 **Figure B1: Spatial distribution of the catchments that were excluded due to a ~~fail~~failure during model calibration or sampling uncertainty calculation for each multi-model approach.**

Appendix C: Comparison between single-model and multi-model combinations.

In the multi-model combination framework presented in Section 2.5.3, combinations were restricted to sets of two or three models. Single-model combinations (i.e., combinations of size one) were not included in the search space. This choice was made to explicitly distinguish combinations from mosaics. To assess the potential impact of this restriction, we compared, for each catchment, the top-performing single model (i.e. mosaic based on performance) with the top-performing two- or three-model combination (i.e. spatially variable and temporally static combination) over the calibration period.

Figure C1 shows that across all catchments, the top-performing single model achieved a higher KGE_{comp} score than any two- or three-model combination in 73 catchments (13% of the sample). In the remaining 471 catchments, the top-performing multi-model combination outperformed the single structure. This indicates that although model combinations tend to improve performance overall, they do not systematically outperform individual models in every catchment. Figure C1 also highlights that performance difference between the top-performing single model and the top-performing multi-model combination remain small, suggesting that the exclusion of size-one combinations has a limited influence on the overall performance distributions reported in the main manuscript. Allowing combinations of size one would lead to minor improvements in a limited subset of catchments but would not alter the overall ranking of modelling approaches or the main conclusions of this study.

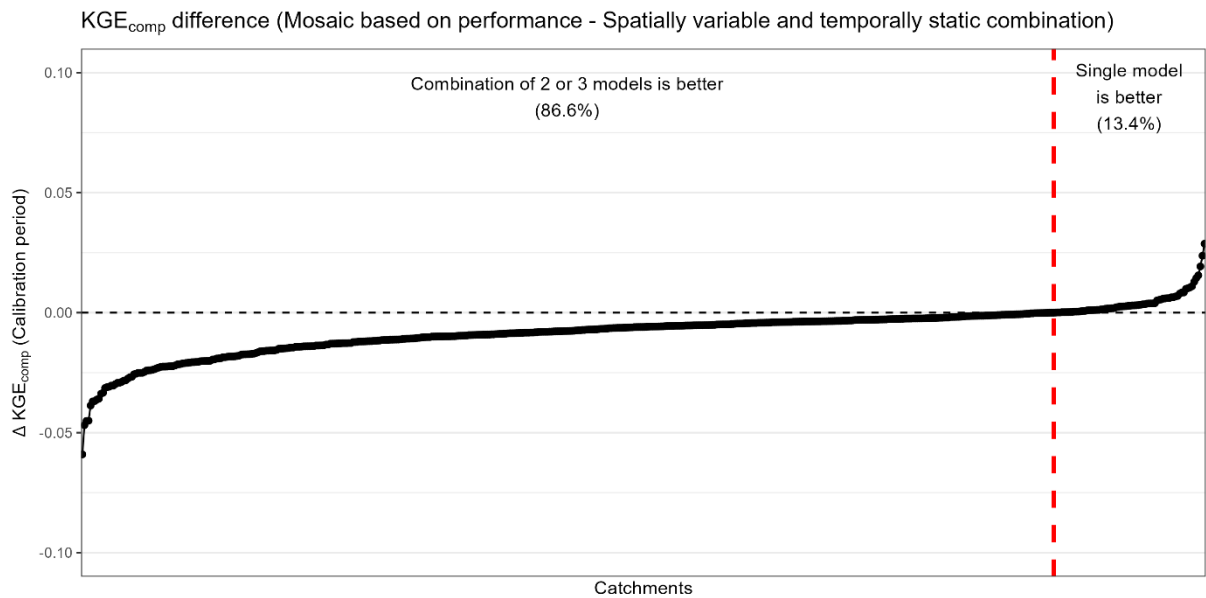


Figure C1: Difference of performance KGE_{comp} between the top-performing single model and the top-performing combination (two or three models) for each catchment over the calibration period. Each dot represents a catchment, and they are ordered by difference of performance. The dashed black line indicates equality; dots under show that the top-performing combination is better than the top-performing single model and vice versa. The dashed red line highlights the tipping point.

Appendix D: Spatial pattern of equivalence between multi-model approaches.

Figure D1 shows the spatial distribution of the equivalence analysis between each multi-model approach. The spatial organization indicates that multi-model approaches — particularly the combination-based methods — tend to improve performance in catchments that already exhibit relatively high KGE_{comp} values and comparatively low sampling uncertainty. In contrast, deteriorations relative to the benchmark are primarily observed in catchments with lower performance and larger uncertainty.

This pattern may reflect the greater difficulty of improving simulations in catchments where dominant hydrological processes are either weakly represented or structurally absent in the model ensemble (e.g., prairie pothole regions or highly arid areas).

In such cases, calibration may lead to overfitting during the calibration period without improving generalization.

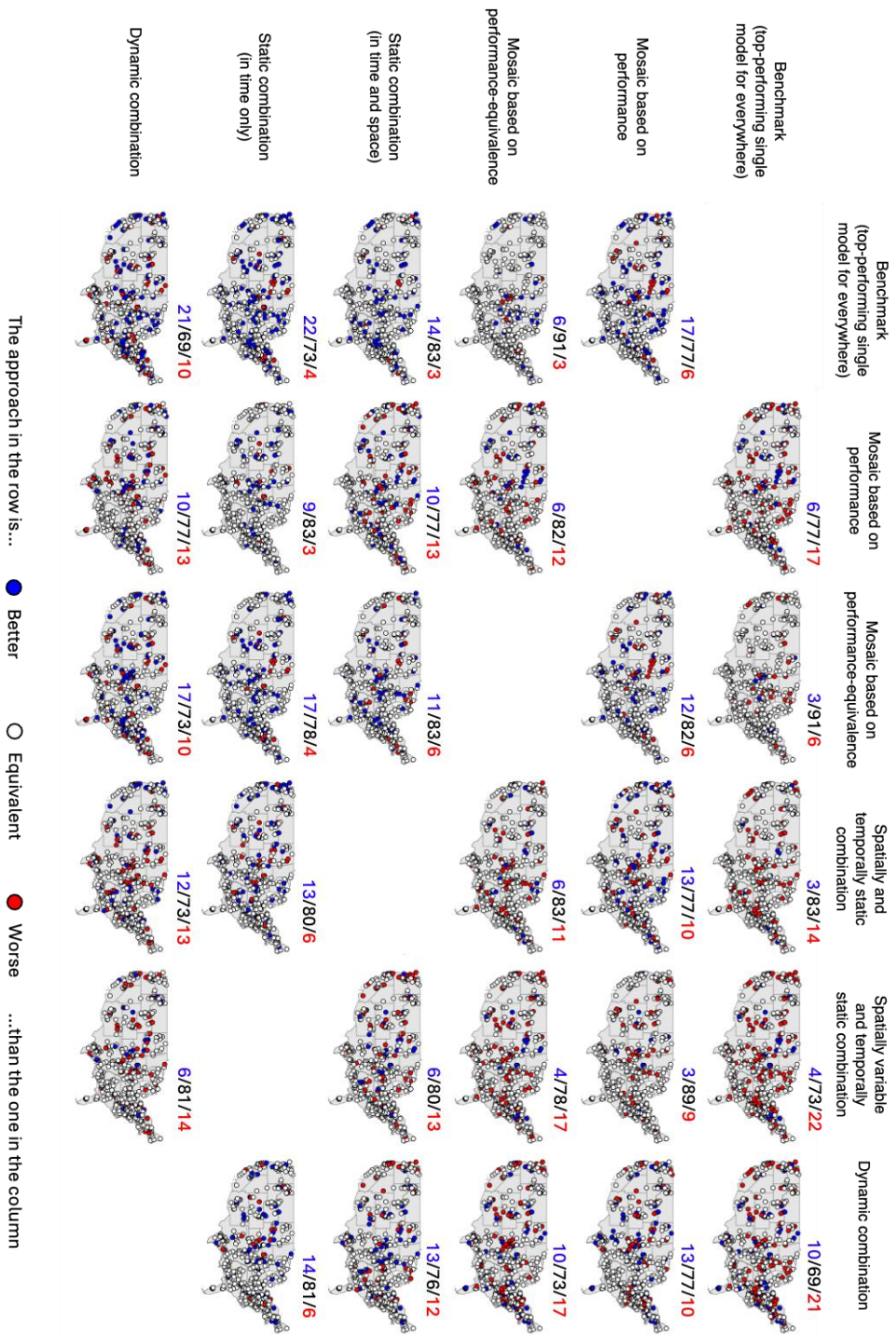


Figure D1: Maps comparing the equivalence between the multi-model approaches. The numbers (e.g. 0/100/0) represent the percentage of catchments falling in each category (colour): better (blue numbers and dots), equivalent (black numbers, white dots) or worse (red numbers and dots).

865 **Code and data availability**

The CAMELS dataset (Newman et al., 2015; Addor et al., 2017) used for this work can be download at <https://ral.ucar.edu/solutions/products/camels>. The FUSE (Clark et al., 2008) version is accessible through https://github.com/CyrilThebault/fuse/tree/FUSE_KGECOMP. Sampling uncertainty was calculated with the R package *gumboot* (Clark and Shook, 2021) accessible on the CRAN at <https://cran.r-project.org/web/packages/gumboot/index.html>.

870 **Author contributions**

MPC wrote the original project proposal, with contributions from AWW, CS, KvW, and MK. The original idea of this work was first discussed between AWW, KvW, MPC, and then with the broader team of AWW, CS, CT, DS, GJG, MPC, SC, KvW, MK, NAV, YS and WJMK. Then, CT, WJMK, NA, and MPC conceptualized the study and developed the methodological framework. CT carried out the simulations and analyses. The manuscript was prepared by CT, with contributions from WJMK,

875 NA, AJN, DS, NAV, YS, MK, and MPC.

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

880 The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the opinions of NOAA.

Acknowledgements

The authors acknowledge the Research Computing Services group at the University of Calgary for the use of their high-performance computer.

Financial support

885 This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) with funding under award NA22NWS4320003 from the NOAA Cooperative Institute Program. This material is based upon work supported by the NSF National Center for Atmospheric Research, which is a major facility sponsored by the U.S. National Science Foundation under Cooperative Agreement No. 1852977.

References

- 890 Addor, N. and Melsen, L. A.: Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models, *Water Resources Research*, 55, 378–390, <https://doi.org/10.1029/2018WR022958>, 2019.
- Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, *Hydrology and Earth System Sciences*, 21, 5293–5313, <https://doi.org/10.5194/hess-21-5293-2017>, 2017.
- 895 Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., and Clark, M. P.: A Ranking of Hydrological Signatures Based on Their Predictability in Space, *Water Resources Research*, 54, 8792–8812, <https://doi.org/10.1029/2018WR022606>, 2018.
- Ajami, N. K., Duan, Q., Gao, X., and Sorooshian, S.: Multimodel Combination Techniques for Analysis of Hydrological Simulations: Application to Distributed Model Intercomparison Project Results, *Journal of Hydrometeorology*, 7, 755–768, <https://doi.org/10.1175/JHM519.1>, 2006.
- 900 Althoff, D. and Rodrigues, L. N.: Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment, *Journal of Hydrology*, 600, 126674, <https://doi.org/10.1016/j.jhydrol.2021.126674>, 2021.
- Anderson, E.: Snow accumulation and ablation model—SNOW-17, 2006.
- Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.-H., and Valéry, A.: HESS Opinions “Crash tests for a standardized evaluation of hydrological models,” *Hydrology and Earth System Sciences*, 13, 1757–1764, <https://doi.org/10.5194/hess-13-1757-2009>, 2009.
- 905 Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J. C. M., Hasan, A., and Pineda, L.: Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation, *Hydrology and Earth System Sciences*, 24, 535–559, <https://doi.org/10.5194/hess-24-535-2020>, 2020.
- Arsenault, R., Gatién, P., Renaud, B., Brissette, F., and Martel, J.-L.: A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation, *Journal of Hydrology*, 529, 754–767, <https://doi.org/10.1016/j.jhydrol.2015.09.001>, 2015.
- Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, L. A.: Global-scale regionalization of hydrologic model parameters, *Water Resources Research*, 52, 3599–3622, <https://doi.org/10.1002/2015WR018247>, 2016.
- 915 Beven, K.: A manifesto for the equifinality thesis, *Journal of Hydrology*, 320, 18–36, <https://doi.org/10.1016/j.jhydrol.2005.07.007>, 2006.
- Booij, M. J. and Krol, M. S.: Balance between calibration objectives in a conceptual hydrological model, *Hydrological Sciences Journal*, 55, 1017–1032, <https://doi.org/10.1080/02626667.2010.505892>, 2010.
- 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., Melsen, L. A., Moustakas, S., Nossent, J., Pereira, F., Sprokkereef, E., Stam, J., Weerts, A. H., Willems, P., Savenije, H. H. G., and Hrachowitz, M.: Behind the scenes of streamflow model performance, *Hydrology and Earth System Sciences*, 25, 1069–1095, <https://doi.org/10.5194/hess-25-1069-2021>, 2021.
- 920 Brigode, P., Paquet, E., Bernardara, P., Gailhard, J., Garavaglia, F., Ribstein, P., Bourgin, F., Perrin, C., and Andréassian, V.: Dependence of model-based extreme flood estimation on the calibration period: case study of the Kamp River (Austria), *Hydrological Sciences Journal*, 60, 1424–1437, <https://doi.org/10.1080/02626667.2015.1006632>, 2015.
- 925 Butts, M. B., Payne, J. T., Kristensen, M., and Madsen, H.: An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, *Journal of Hydrology*, 298, 242–266, <https://doi.org/10.1016/j.jhydrol.2004.03.042>, 2004.

- 930 Caillouet, L., Celie, S., Vannier, O., Bontron, G., and Legrand, S.: Operational hydrometeorological forecasting on the Rhône River in France: moving toward a seamless probabilistic approach, *LHB*, 108, 2061312, <https://doi.org/10.1080/27678490.2022.2061312>, 2022.
- Cinkus, G., Mazzilli, N., Jourde, H., Wunsch, A., Liesch, T., Ravbar, N., Chen, Z., and Goldscheider, N.: When best is the enemy of good – critical evaluation of performance criteria in hydrological models, *Hydrology and Earth System Sciences*, 27, 2397–2411, <https://doi.org/10.5194/hess-27-2397-2023>, 2023.
- 935 Clark, M. and Shook, K.: *gumboot: Bootstrap Analyses of Sampling Uncertainty in Goodness-of-Fit statistics*, 2021.
- Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, *Water Resources Research*, 44, <https://doi.org/10.1029/2007WR006735>, 2008.
- 940 Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resources Research*, 47, <https://doi.org/10.1029/2010WR009827>, 2011.
- Clark, M. P., Wilby, R. L., Gutmann, E. D., Vano, J. A., Gangopadhyay, S., Wood, A. W., Fowler, H. J., Prudhomme, C., Arnold, J. R., and Brekke, L. D.: Characterizing Uncertainty of the Hydrologic Impacts of Climate Change, *Curr Clim Change Rep*, 2, 55–64, <https://doi.org/10.1007/s40641-016-0034-x>, 2016.
- 945 Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R., and Papalexiou, S. M.: The Abuse of Popular Performance Metrics in Hydrologic Modeling, *Water Resources Research*, 57, e2020WR029001, <https://doi.org/10.1029/2020WR029001>, 2021.
- Csárdi, G. and Berkelaar, M.: *lpSolve: Interface to “Lp_solve” v. 5.5 to Solve Linear/Integer Programs*, 2024.
- 950 David, P. C., Chaffe, P. L. B., Chagas, V. B. P., Dal Molin, M., Oliveira, D. Y., Klein, A. H. F., and Fenicia, F.: Correspondence Between Model Structures and Hydrological Signatures: A Large-Sample Case Study Using 508 Brazilian Catchments, *Water Resources Research*, 58, e2021WR030619, <https://doi.org/10.1029/2021WR030619>, 2022.
- Dion, P., Martel, J.-L., and Arsenault, R.: Hydrological ensemble forecasting using a multi-model framework, *Journal of Hydrology*, 600, 126537, <https://doi.org/10.1016/j.jhydrol.2021.126537>, 2021.
- Duan, Q., Sorooshian, S., and Gupta, V. K.: Optimal use of the SCE-UA global optimization method for calibrating watershed models, *Journal of Hydrology*, 158, 265–284, [https://doi.org/10.1016/0022-1694\(94\)90057-4](https://doi.org/10.1016/0022-1694(94)90057-4), 1994.
- 955 Duan, Q. Y., Gupta, V. K., and Sorooshian, S.: Shuffled complex evolution approach for effective and efficient global minimization, *J Optim Theory Appl*, 76, 501–521, <https://doi.org/10.1007/BF00939380>, 1993.
- Efstratiadis, A. and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a review, *Hydrological Sciences Journal*, 55, 58–78, <https://doi.org/10.1080/02626660903526292>, 2010.
- 960 van Esse, W. R., Perrin, C., Booij, M. J., Augustijn, D. C. M., Fenicia, F., Kavetski, D., and Lobligois, F.: The influence of conceptual model structure on model performance: a comparative study for 237 French catchments, *Hydrology and Earth System Sciences*, 17, 4227–4239, <https://doi.org/10.5194/hess-17-4227-2013>, 2013.
- Evans, J. P., Ji, F., Abramowitz, G., and Ekström, M.: Optimally choosing small ensemble members to produce robust climate simulations, *Environ. Res. Lett.*, 8, 044050, <https://doi.org/10.1088/1748-9326/8/4/044050>, 2013.
- 965 Farahani, M. A., Wood, A. W., Tang, G., and Mizukami, N.: Calibrating a large-domain land/hydrology process model in the age of AI: the SUMMA CAMELS experiments, *Hydrology and Earth System Sciences*, 29, 4515–4537, <https://doi.org/10.5194/hess-29-4515-2025>, 2025.
- Feng, D., Liu, J., Lawson, K., and Shen, C.: Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy, *Water Resources Research*, 58, e2022WR032404, <https://doi.org/10.1029/2022WR032404>, 2022.

- 970 Feyen, L., Kalas, M., and Vrugt, J. A.: Semi-distributed parameter optimization and uncertainty assessment for large-scale streamflow simulation using global optimization / Optimisation de paramètres semi-distribués et évaluation de l'incertitude pour la simulation de débits à grande échelle par l'utilisation d'une optimisation globale, *Hydrological Sciences Journal*, 53, 293–308, <https://doi.org/10.1623/hysj.53.2.293>, 2008.
- Garcia, F., Folton, N., and Oudin, L.: Which objective function to calibrate rainfall–runoff models for low-flow index simulations?, *Hydrological Sciences Journal*, 62, 1149–1166, <https://doi.org/10.1080/02626667.2017.1308511>, 2017.
- 975 Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J., and Butts, M. B.: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, *Journal of Hydrology*, 298, 222–241, <https://doi.org/10.1016/j.jhydrol.2004.03.037>, 2004.
- Gupta, A. and Govindaraju, R. S.: Propagation of structural uncertainty in watershed hydrologic models, *Journal of Hydrology*, 980 575, 66–81, <https://doi.org/10.1016/j.jhydrol.2019.05.026>, 2019.
- Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resources Research*, 34, 751–763, <https://doi.org/10.1029/97WR03495>, 1998.
- 985 Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *Journal of Hydrology*, 377, 80–91, <https://doi.org/10.1016/j.jhydrol.2009.08.003>, 2009.
- Hallouin, T., Bruen, M., and O'Loughlin, F. E.: Calibration of hydrological models for ecologically relevant streamflow predictions: a trade-off between fitting well to data and estimating consistent parameter sets?, *Hydrology and Earth System Sciences*, 24, 1031–1054, <https://doi.org/10.5194/hess-24-1031-2020>, 2020.
- 990 Henn, B., Clark, M. P., Kavetski, D., and Lundquist, J. D.: Estimating mountain basin-mean precipitation from streamflow using Bayesian inference, *Water Resources Research*, 51, 8012–8033, <https://doi.org/10.1002/2014WR016736>, 2015.
- Horton, P., Schaefli, B., and Kauzlaric, M.: Why do we have so many different hydrological models? A review based on the case of Switzerland, *WIREs Water*, 9, e1574, <https://doi.org/10.1002/wat2.1574>, 2022.
- 995 Johnson, J. M., Fang, S., Sankarasubramanian, A., Rad, A. M., Kindl da Cunha, L., Jennings, K. S., Clarke, K. C., Mazrooei, A., and Yeghiazarian, L.: Comprehensive Analysis of the NOAA National Water Model: A Call for Heterogeneous Formulations and Diagnostic Model Selection, *Journal of Geophysical Research: Atmospheres*, 128, e2023JD038534, <https://doi.org/10.1029/2023JD038534>, 2023.
- 1000 Khatami, S., Peel, M. C., Peterson, T. J., and Western, A. W.: Equifinality and Flux Mapping: A New Approach to Model Evaluation and Process Representation Under Uncertainty, *Water Resources Research*, 55, 8922–8941, <https://doi.org/10.1029/2018WR023750>, 2019.
- Kiraz, M., Coxon, G., and Wagener, T.: A priori selection of hydrological model structures in modular modelling frameworks: application to Great Britain, *Hydrological Sciences Journal*, 68, 2042–2056, <https://doi.org/10.1080/02626667.2023.2251968>, 2023.
- 1005 Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resources Research*, 42, <https://doi.org/10.1029/2005WR004362>, 2006.
- Klotz, D., Gauch, M., Kratzert, F., Nearing, G., and Zscheischler, J.: Technical Note: The divide and measure nonconformity – how metrics can mislead when we evaluate on different data partitions, *Hydrology and Earth System Sciences*, 28, 3665–3673, <https://doi.org/10.5194/hess-28-3665-2024>, 2024.
- 1010 Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, *Hydrology and Earth System Sciences*, 23, 4323–4331, <https://doi.org/10.5194/hess-23-4323-2019>, 2019.

- Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., and Woods, R. A.: A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments, *Water Resources Research*, 56, e2019WR025975, <https://doi.org/10.1029/2019WR025975>, 2020.
- 1015 Knoben, W. J. M., Raman, A., Gründemann, G. J., Kumar, M., Pietroniro, A., Shen, C., Song, Y., Thébault, C., van Werkhoven, K., Wood, A. W., and Clark, M. P.: Technical note: How many models do we need to simulate hydrologic processes across large geographical domains?, *Hydrology and Earth System Sciences*, 29, 2361–2375, <https://doi.org/10.5194/hess-29-2361-2025>, 2025.
- Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, *Geophysical Research Letters*, 40, 1194–1199, <https://doi.org/10.1002/grl.50256>, 2013.
- 1020 Kollat, J. B., Reed, P. M., and Wagener, T.: When are multiobjective calibration trade-offs in hydrologic models meaningful?, *Water Resources Research*, 48, <https://doi.org/10.1029/2011WR011534>, 2012.
- Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, *Hydrology and Earth System Sciences*, 25, 2685–2703, <https://doi.org/10.5194/hess-25-2685-2021>, 2021.
- 1025 Lamontagne, J. R., Barber, C. A., and Vogel, R. M.: Improved Estimators of Model Performance Efficiency for Skewed Hydrologic Data, *Water Resources Research*, 56, e2020WR027101, <https://doi.org/10.1029/2020WR027101>, 2020.
- Lan, T., Lin, K., Xu, C.-Y., Liu, Z., and Cai, H.: A framework for seasonal variations of hydrological model parameters: impact on model results and response to dynamic catchment characteristics, *Hydrology and Earth System Sciences*, 24, 5859–5874, <https://doi.org/10.5194/hess-24-5859-2020>, 2020.
- 1030 Linsley, R. K.: Rainfall-runoff models-an overview, in: *Rainfall-Runoff Relationship: Proceedings of the International Symposium on Rainfall-Runoff Modeling*, Water Resources Publications, Littleton, CO, 3–22, 1982.
- Livneh, B., Rosenberg, E. A., Lin, C., Nijssen, B., Mishra, V., Andreadis, K. M., Maurer, E. P., and Lettenmaier, D. P.: A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States: Update and Extensions, <https://doi.org/10.1175/JCLI-D-12-00508.1>, 2013.
- 1035 Mai, J.: Ten strategies towards successful calibration of environmental models, *Journal of Hydrology*, 620, 129414, <https://doi.org/10.1016/j.jhydrol.2023.129414>, 2023.
- Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O’Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., and Waddell, J. W.: The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL), *Hydrology and Earth System Sciences*, 26, 3537–3572, <https://doi.org/10.5194/hess-26-3537-2022>, 2022.
- 1040 Mathevet, T., Gupta, H., Perrin, C., Andréassian, V., and Le Moine, N.: Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds, *Journal of Hydrology*, 585, 124698, <https://doi.org/10.1016/j.jhydrol.2020.124698>, 2020.
- 1045 McMillan, H., Araki, R., Gnann, S., Woods, R., and Wagener, T.: How do hydrologists perceive watersheds? A survey and analysis of perceptual model figures for experimental watersheds, *Hydrological Processes*, 37, e14845, <https://doi.org/10.1002/hyp.14845>, 2023.
- Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, *Hydrology and Earth System Sciences*, 19, 209–223, <https://doi.org/10.5194/hess-19-209-2015>, 2015.
- 1050 Nicolle, P., Andréassian, V., Royer-Gaspard, P., Perrin, C., Thirel, G., Coron, L., and Santos, L.: Technical note: RAT – a robustness assessment test for calibrated and uncalibrated hydrological models, *Hydrology and Earth System Sciences*, 25, 5013–5027, <https://doi.org/10.5194/hess-25-5013-2021>, 2021.

- 1055 Ogden, F., Avant, B., Bartel, R., Blodgett, D., Clark, E., Coon, E., Cosgrove, B., Cui, S., Kindl da Cunha, L., Farthing, M., Flowers, T., Frame, J., Frazier, N., Graziano, T., Gutenson, J., Johnson, D., McDaniel, R., Moulton, J., Loney, D., Peckham, S., Mattern, D., Jennings, K., Williamson, M., Savant, G., Tubbs, C., Garrett, J., Wood, A., and Johnson, J.: The Next Generation Water Resources Modeling Framework: Open Source, Standards Based, Community Accessible, Model Interoperability for Large Scale Water Prediction, 2021, H43D-01, 2021.
- 1060 Ogden, F. L., Jennings, K., Clark, E. P., Coon, E., Cosgrove, B., da Cunha, L. K., Farthing, M. W., Flowers, T., Frame, J. M., Frazier, N. J., Garrett, J. L., Graziano, T. M., Hughes, J. D., Johnson, J. M., McDaniel, R., Moulton, J. D., Peckham, S. D., Salas, F. R., Savant, G., Viger, R., and Wood, A.: The NextGen Water Resources Modeling Framework: Community Innovation at the Intersection of Hydrologic, Data and Computer Sciences, *JAWRA Journal of the American Water Resources Association*, 62, e70089, <https://doi.org/10.1111/1752-1688.70089>, 2026.
- 1065 Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and 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, *Journal of Hydrology*, 303, 290–306, <https://doi.org/10.1016/j.jhydrol.2004.08.026>, 2005.
- Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., and Michel, C.: Dynamic Averaging of Rainfall-Runoff Model Simulations from Complementary Model Parameterizations, *Water Resources Research*, 42, <https://doi.org/10.1029/2005WR004636>, 2006.
- 1070 Perrin, C., Michel, C., and Andréassian, V.: Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, *Journal of Hydrology*, 242, 275–301, [https://doi.org/10.1016/S0022-1694\(00\)00393-0](https://doi.org/10.1016/S0022-1694(00)00393-0), 2001.
- 1075 Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, *Journal of Hydrology*, 279, 275–289, [https://doi.org/10.1016/S0022-1694\(03\)00225-7](https://doi.org/10.1016/S0022-1694(03)00225-7), 2003.
- Pushpalatha, R., Perrin, C., Moine, N. L., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, *Journal of Hydrology*, 420–421, 171–182, <https://doi.org/10.1016/j.jhydrol.2011.11.055>, 2012.
- 1080 Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, *Monthly Weather Review*, 133, 1155–1174, <https://doi.org/10.1175/MWR2906.1>, 2005.
- Ramos, M. H., van Andel, S. J., and Pappenberger, F.: Do probabilistic forecasts lead to better decisions?, *Hydrology and Earth System Sciences*, 17, 2219–2232, <https://doi.org/10.5194/hess-17-2219-2013>, 2013.
- Ramos, M.-H., Pappenberger, F., Wood, A., Wetterhall, F., Wang, Q., Verkade, J., Pechlivanidis, I., Thielen-del Pozo, J., Buizza, R., and Schaake, J.: The history of HEPEX - a community of practice in hydrologic prediction, 10137, 2018.
- 1085 Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L., and Vanrolleghem, P. A.: Uncertainty in the environmental modelling process – A framework and guidance, *Environmental Modelling & Software*, 22, 1543–1556, <https://doi.org/10.1016/j.envsoft.2007.02.004>, 2007.
- Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resources Research*, 46, <https://doi.org/10.1029/2009WR008328>, 2010.
- 1090 Savenije, H. H. G.: HESS Opinions “The art of hydrology”*, *Hydrology and Earth System Sciences*, 13, 157–161, <https://doi.org/10.5194/hess-13-157-2009>, 2009.
- Sawadekar, K., Song, Y., Pan, M., Beck, H., McCrary, R., Ullrich, P., Lawson, K., and Shen, C.: Improving differentiable hydrologic modeling with interpretable forcing fusion, *Journal of Hydrology*, 659, 133320, <https://doi.org/10.1016/j.jhydrol.2025.133320>, 2025.
- 1095 Schaake, J. C., Hamill, T. M., Buizza, R., and Clark, M.: HEPEX: The Hydrological Ensemble Prediction Experiment, *Bulletin of the American Meteorological Society*, 88, 1541–1548, <https://doi.org/10.1175/BAMS-88-10-1541>, 2007.

- Schaeffli, B. and Gupta, H. V.: Do Nash values have value?, *Hydrological Processes*, 21, 2075–2080, <https://doi.org/10.1002/hyp.6825>, 2007.
- 1100 Seiller, G., Anctil, F., and Perrin, C.: Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, *Hydrology and Earth System Sciences*, 16, 1171–1189, <https://doi.org/10.5194/hess-16-1171-2012>, 2012.
- Shamseldin, A. Y., O’Connor, K. M., and Liang, G. C.: Methods for combining the outputs of different rainfall runoff models, *Journal of Hydrology*, 197, 203–229, [https://doi.org/10.1016/S0022-1694\(96\)03259-3](https://doi.org/10.1016/S0022-1694(96)03259-3), 1997.
- 1105 Sidle, R. C.: Strategies for smarter catchment hydrology models: incorporating scaling and better process representation, *Geoscience Letters*, 8, 24, <https://doi.org/10.1186/s40562-021-00193-9>, 2021.
- Singh, V. P. and Woolhiser, D. A.: Mathematical Modeling of Watershed Hydrology, *Journal of Hydrologic Engineering*, 7, 270–292, [https://doi.org/10.1061/\(ASCE\)1084-0699\(2002\)7:4\(270\)](https://doi.org/10.1061/(ASCE)1084-0699(2002)7:4(270)), 2002.
- Song, Y., Knoben, W. J. M., Clark, M. P., Feng, D., Lawson, K., Sawadekar, K., and Shen, C.: When ancient numerical demons meet physics-informed machine learning: adjoint-based gradients for implicit differentiable modeling, *Hydrology and Earth System Sciences*, 28, 3051–3077, <https://doi.org/10.5194/hess-28-3051-2024>, 2024.
- 1110 Spieler, D. and Schütze, N.: Investigating the Model Hypothesis Space: Benchmarking Automatic Model Structure Identification With a Large Model Ensemble, *Water Resources Research*, 60, e2023WR036199, <https://doi.org/10.1029/2023WR036199>, 2024.
- 1115 Spieler, D., Mai, J., Craig, J. R., Tolson, B. A., and Schütze, N.: Automatic Model Structure Identification for Conceptual Hydrologic Models, *Water Resources Research*, 56, e2019WR027009, <https://doi.org/10.1029/2019WR027009>, 2020.
- Tang, G., Wood, A. W., and Swenson, S.: On Using AI-Based Large-Sample Emulators for Land/Hydrology Model Calibration and Regionalization, *Water Resources Research*, 61, e2024WR039525, <https://doi.org/10.1029/2024WR039525>, 2025.
- 1120 Thébault, C., Perrin, C., Andréassian, V., Thirel, G., Legrand, S., and Delaigue, O.: Multi-model approach in a variable spatial framework for streamflow simulation, *Hydrology and Earth System Sciences*, 28, 1539–1566, <https://doi.org/10.5194/hess-28-1539-2024>, 2024.
- Thébault, C., Knoben, W. J. M., Addor, N., Newman, A. J., and Clark, M. P.: Varying the Combination of Hydrological Models in Time and Space: Towards a More Accurate Representation of Streamflow Across Large Domains, <https://doi.org/10.22541/essoar.175855455.54894703/v1>, 2025a.
- 1125 Thébault, C., Perrin, C., Legrand, S., Andréassian, V., Thirel, G., and Delaigue, O.: What can be expected from a semi-distributed multi-model approach for streamflow forecasting? Tailoring the structure and size of a super-ensemble on the Rhône basin, *Journal of Hydrology*, 661, 133589, <https://doi.org/10.1016/j.jhydrol.2025.133589>, 2025b.
- Thirel, G., Santos, L., Delaigue, O., and Perrin, C.: On the use of streamflow transformations for hydrological model calibration, *Hydrology and Earth System Sciences*, 28, 4837–4860, <https://doi.org/10.5194/hess-28-4837-2024>, 2024.
- 1130 Thornton, P. E., Thornton, M. M., Mayer, B. W., Wilhelmi, N., Wei, Y., and Cook, R. B.: Daymet: Daily surface weather on a 1km grid for North America, 2012.
- Todorović, A., Grabs, T., and Teutschbein, C.: Advancing traditional strategies for testing hydrological model fitness in a changing climate, *Hydrological Sciences Journal*, 67, 1790–1811, <https://doi.org/10.1080/02626667.2022.2104646>, 2022.
- 1135 Todorović, A., Grabs, T., and Teutschbein, C.: Improving performance of bucket-type hydrological models in high latitudes with multi-model combination methods: Can we wring water from a stone?, *Journal of Hydrology*, 632, 130829, <https://doi.org/10.1016/j.jhydrol.2024.130829>, 2024.
- Tolson, B. A. and Shoemaker, C. A.: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, *Water Resources Research*, 43, <https://doi.org/10.1029/2005WR004723>, 2007.

- 140 Vrugt, J. A., Diks, C. G. H., and Clark, M. P.: Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling, *Environ Fluid Mech*, 8, 579–595, <https://doi.org/10.1007/s10652-008-9106-3>, 2008.
- Wagener, T. and Gupta, H. V.: Model identification for hydrological forecasting under uncertainty, *Stoch Environ Res Ris Assess*, 19, 378–387, <https://doi.org/10.1007/s00477-005-0006-5>, 2005.
- 145 Wan, Y., Chen, J., Xu, C.-Y., Xie, P., Qi, W., Li, D., and Zhang, S.: Performance dependence of multi-model combination methods on hydrological model calibration strategy and ensemble size, *Journal of Hydrology*, 603, 127065, <https://doi.org/10.1016/j.jhydrol.2021.127065>, 2021.
- Williams, G. P.: Friends don't let friends use Nash-Sutcliffe Efficiency (NSE) or KGE for hydrologic model accuracy evaluation: A rant with data and suggestions for better practice, *Environmental Modelling & Software*, 194, 106665, <https://doi.org/10.1016/j.envsoft.2025.106665>, 2025.
- 150 Winter, C. L. and Nychka, D.: Forecasting skill of model averages, *Stochastic Environmental Research and Risk Assessment*, 24, 633–638, <https://doi.org/10.1007/s00477-009-0350-y>, 2010.
- Wood, A. W., Leung, L. R., Sridhar, V., and Lettenmaier, D. P.: Hydrologic Implications of Dynamical and Statistical Approaches to Downscaling Climate Model Outputs, *Climatic Change*, 62, 189–216, <https://doi.org/10.1023/B:CLIM.0000013685.99609.9e>, 2004.
- 155 Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., and Mocko, D.: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products, *Journal of Geophysical Research: Atmospheres*, 117, <https://doi.org/10.1029/2011JD016048>, 2012.
- 160 Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resources Research*, 44, <https://doi.org/10.1029/2007WR006716>, 2008.
- Zhang, R., Liu, J., Gao, H., and Mao, G.: Can multi-objective calibration of streamflow guarantee better hydrological model accuracy?, *Journal of Hydroinformatics*, 20, 687–698, <https://doi.org/10.2166/hydro.2018.131>, 2018.