Multi-model approach in a variable spatial framework for streamflow simulation

Cyril Thébault[1*], Charles Perrin[1], Vazken Andréassian[1], Guillaume Thirel[1], Sébastien Legrand[2], Olivier Delaigue[1]

5   [1] Université Paris-Saclay, INRAE, HYCAR, Antony, France
    [2] Compagnie nationale du Rhône, Lyon, France

Correspondence to: cyril.thebault@inrae.fr

Abstract.

Accounting for the variability of hydrological processes and climate conditions between catchments and within catchments
10   remains a challenge in rainfall–runoff modelling. Among the many approaches developed over the past decades, multi-model approaches provide a way to take into account the uncertainty linked to the choice of model structure and its parameter estimates. Semi-distributed approaches make it possible to account explicitly for spatial variability while maintaining a limited level of complexity. However, these two approaches have rarely been used together. Such a combination would allow us to take advantage of both methods. The aim of this work is to answer the following question: What is the possible contribution
15   of a multi-model approach within a variable spatial framework compared to lumped single models for streamflow simulation?

To this end, a set of 121 catchments with limited influence in France was assembled, with precipitation, potential evapotranspiration and streamflow data at the hourly time step over the period 1998–2018. The semi-distribution set-up was kept simple by considering a single downstream catchment defined by an outlet, and one or more upstream sub-catchments. The multi-model approach was implemented with 13 rainfall–runoff model structures, three calibration options and two spatial
20   frameworks, for a total of 78 distinct modelling options. A simple average method was used to combine the various simulated streamflow at the outlet of the catchments and sub-catchments. The most efficient lumped model on a given catchment was taken as the benchmark for model evaluation.

Overall, the semi-distributed multi-model approach yields better performance than the different lumped models considered individually. The gain is mainly brought about by the multi-model set-up, with the spatial framework providing a benefit on a
25   more occasional basis. These results, based on a large catchment set, evince the benefits of using a multi-model in a variable spatial framework to simulate streamflow.

Keywords: Rainfall–runoff modelling, multi-model, semi-distribution

1   Introduction

1.1   Uncertainty in rainfall–runoff modelling

A rainfall–runoff model is a numerical tool based on a simplified representation of a real-world system, namely the catchment (Moradkhani and Sorooshian, 2008). It usually computes streamflow time series from climatic data, such as rainfall and potential evapotranspiration. Many rainfall–runoff models have been developed according to various assumptions in order to meet specific needs (e.g. water resources management, flood and low-flow forecasting, hydroelectricity), with choices and constraints concerning (Perrin, 2000):

- the temporal resolution, i.e. the way variables and processes are aggregated over time;
- the spatial resolution, i.e. the way spatial variability is taken into account more or less explicitly in the model;
- the description of dominant processes.

Different models will necessarily produce different streamflow simulations. Intuitively, one often expects that working at a finer spatio-temporal scale should allow for a better description of the processes (Atkinson *et al.*, 2002). However, this generally leads to additional complexity, i.e. a larger number of parameters, which requires more information to be estimated and often yields more uncertain results (Her and Chaubey, 2015).

Uncertainty in rainfall–runoff models depends on the assumptions made regarding the choice of the general structure and also on the parameter estimates. The variety of model structures and equations results in a large variability of streamflow simulations (Ajami *et al.*, 2007). The spatial and temporal resolutions also result in different streamflow simulations. Due to the complexity of the real system and the lack of information to parameterize the various equations over all of the catchment, parameter estimates must be set. Usually, these parameters are determined for each entity of interest by minimizing the error induced by the simulation compared to an observation. The choice of the optimization algorithm, the objective function and the streamflow transformation is therefore also a source of uncertainty. Since input data are used to drive model structures and parameters, the uncertainty associated with these data also contributes to the overall model uncertainty (Beven, 1993; Liu and Gupta, 2007; Pechlivanidis *et al.*, 2011; McMillan *et al.*, 2012).

Various approaches aim to improve models by taking into account uncertainties, among which are multi-model approaches that are the main topic of our research.

1.2   Multi-models

The multi-model approach consists in using several models in order to take advantage of the strengths of each one. This concept has been gaining momentum in hydrology since the end of the 20th century for simulation (e.g. Shamseldin *et al.*, 1997) and forecasting (e.g. Loumagne *et al.*, 1995). In this section, we distinguish between probabilistic and deterministic approaches.

A probabilistic multi-model seeks an explicit quantification of the uncertainty associated with simulations or forecasts through
60 statistical methods. The ensemble concept has commonly been applied in meteorology for several decades, and subsequently
has been widely used in hydrology to improve prediction (i.e. simulation or forecast). The international Hydrologic Ensemble
Prediction EXperiment initiative (Schaake *et al.*, 2007) fostered the work on this topic. The ensemble concept has also been
adapted to rainfall–runoff models in order to reduce modelling bias: Duan *et al.* (2007) used multiple predictions made by
several rainfall–runoff models using the same hydroclimatic forcing variables. An ensemble consisting of nine different models
65 (from three different structures and parameterizations) was constructed and applied to three catchments in the United States.
The predictions were then combined through a statistical procedure (Bayesian model averaging or BMA), which assigns larger
weight to a probabilistic likelihood measure. The authors showed that the probabilistic multi-model approach improves flow
prediction and quantifies model uncertainty compared to using a single rainfall–runoff model. Block *et al.* (2009) coupled both
multiple climate and multiple rainfall–runoff models, increasing the pool of streamflow forecast ensemble members and
70 accounting for cumulative sources of uncertainty. In their study, 10 scenarios were built for each of the three climatic models
and applied to two rainfall–runoff models, i.e. 60 different forecasts. This super-ensemble was applied to the Iguatu Catchment
in Brazil and showed better performance than the hydroclimatic or rainfall–runoff model ensembles studied separately. Note
that the authors tested three different combination methods: pooling, linear regression weighting, and a kernel density
estimator. They found that the last technique seems to perform better. Velázquez *et al.* (2011) showed that the combination of
75 different climatic scenarios with several models in a forecasting context leads to a reduction in uncertainty, particularly when
the forecast horizon increases. However, such methods generate a large number of scenarios and can therefore become time-
consuming and difficult to analyse. The probabilistic combination of simulations remains a major topic in the scientific
community (see Bogner *et al.*, 2017).

A deterministic multi-model seeks to define a single best streamflow time series, which often consists in a combination of the
80 simulations of individual models. Shamseldin *et al.* (1997) tested three methods in order to combine model outputs: a simple
average, a weighted average and a non-linear neural network procedure. Their study was conducted on a sample of 11
catchments mainly located in Southeast Asia using five different lumped models operating at the daily time step and showed
that multi-models perform better than models applied individually. Similar conclusions were reached in the Distributed Model
Intercomparison Project (DMIP) (Smith *et al.*, 2004) conducted by Georgakakos *et al.* (2004) in simulation or by Ajami *et al.*
85 (2006) for forecasting. Six to 10 rainfall–runoff models were applied at the hourly time step over a few catchments in the
United States. These studies showed that a model that performs poorly individually can contribute positively to the multi-
model set-up. Studies combining rainfall–runoff models by machine learning techniques led to the same conclusions (see, e.g.
Zounemat-Kermani *et al.*, 2021 for a review). Some authors have also been interested in combining the different modules of
rainfall–runoff models in order to create flexible structures. The underlying idea is to define an appropriate fit-for-purpose
90 structure. This is the case, for example with the FUSE (Clark *et al.*, 2008), SUPERFLEX (Fenicia *et al.*, 2011; Kavetski *et al.*,
2011) or SUMMA (Clark *et al.*, 2015a, 2015b) methods. In an even more complex approach, some authors tried to make

various models interact with each other in order to exchange different information to correct their internal states. This super model method was first developed in climatology (van den Berge *et al.*, 2011) and then adapted to hydrology (Santos, 2018) with two lumped models, but it did not show significant improvements.

95   All of the aforementioned multi-model approaches only focus on the structural aspect of rainfall–runoff models. Some authors have also combined streamflow generated from different parameterizations of the same rainfall–runoff model. Oudin *et al.* (2006) proposed combining two outputs obtained with a single model (GR4J) from two calibrations, one adapted to high flows and the other to low flows, by weighting each of the simulations on the basis of a seasonal index (filling rate of the production reservoir). Such a method makes it possible to provide good efficiency in both low and high flows, whereas usually an a priori

100   modelling choice must be made to focus on a specific streamflow range. More recently, Wan *et al.* (2021) used a multi-model approach based on four rainfall–runoff models calibrated with four objective functions on a large set of 383 Chinese catchments. The authors showed that methods based on weighted averaging outperform the ensemble members, except in low-flow simulation. They also highlighted the benefit of using several structures with different objective functions. The size of the ensemble was also studied and it was found that using more than nine ensemble members does not further improve

105   performance. Note that different results for optimal size can be found in the literature (Arsenault *et al.*, 2015; Kumar *et al.*, 2015).

The aforementioned studies were carried out within a fixed spatial framework (e.g. lumped, semi-distributed, distributed), i.e. considering that the model structures implemented are relevant over the whole modelling domain. Implicitly, the underlying assumption is that a fixed rainfall–runoff model is able to capture the main hydrological processes affecting streamflow on a

110   catchment (and its sub-catchments). However, this may not be true. Introducing a variable spatial modelling framework into the multi-model approach could help to overcome this issue.

1.3   Scope of the paper

This study intends to improve streamflow simulation through a multi-model approach. More precisely, we aim to take into account uncertainty stemming from (i) the spatial dimension (e.g. catchment division, aggregation of hydroclimatic forcing,

115   boundary conditions), (ii) the general structure of the model (e.g. formulation of water storages, filling/draining equations) and (iii) the parameter estimation (e.g. calibration algorithm, objective function, calibration period). Referring to the literature review (Section 1.2), we chose a deterministic averaging approach. The aim of this study is to answer the following question:

What is the possible contribution of a multi-model approach within a variable spatial framework compared to lumped single models for streamflow simulation?

120   The remainder of the paper is organized as follows: first, the catchment set, the hydroclimatic data, the spatial framework and the rainfall–runoff models used for this work are presented. The multi-model methodology and the calibration/evaluation

procedure are described. Then we present, analyse and discuss the results. Last, we summarize the main conclusions of this work and discuss its perspectives.

2 Material and methods

2.1 Catchments and hydroclimatic data

This study was conducted at an hourly time step using precipitation, potential evapotranspiration and streamflow time series over the period 1998–2018 (Delaigue *et al.*, 2020). Precipitation (P) was extracted from the radar-based COMEPHORE re-analysis produced by Météo-France (Tabary *et al.*, 2012), which provides information at a 1-km² resolution and which has already been extensively used in hydrological studies (Artigue *et al.*, 2012; van Esse *et al.*, 2013; Bourgin *et al.*, 2014; Lobligeois *et al.*, 2014; Saadi *et al.*, 2021).

Potential evapotranspiration (E_0) is calculated with the formula proposed by Oudin *et al.* (2005). This equation was chosen for its simplicity, as the only input required is daily air temperature (from the SAFRAN re-analysis of Météo-France, see Vidal *et al.,* 2010) and extra-terrestrial radiation (which depends only on the Julian day and the latitude). Once calculated, the daily potential evapotranspiration was disaggregated to the hourly time step using a simple parabola (Lobligeois, 2014). These steps for converting daily temperature data into hourly potential evapotranspiration are directly possible in the airGR software (Coron *et al.*, 2017, 2021; developed using the R programming language; R Core Team, 2020), which was used for this work. We did not use any gap-filling method since all climatic data were complete during the study period.

Streamflow time series (Q) were extracted from the national streamflow archive Hydroportail (Dufeu *et al.*, 2022), which makes available the data produced by hydrometric services in regional environmental agencies in charge of measuring flows in France as well as by other data producers (e.g. hydropower companies, dam managers, etc.). Before being archived, flow data undergo quality control procedures applied by data producers, with corrections when necessary. Quality codes are also available, although this information is not uniformly provided for all stations. These data are freely available on the Hydroportail website and are widely used in France for hydraulic and hydrological studies.

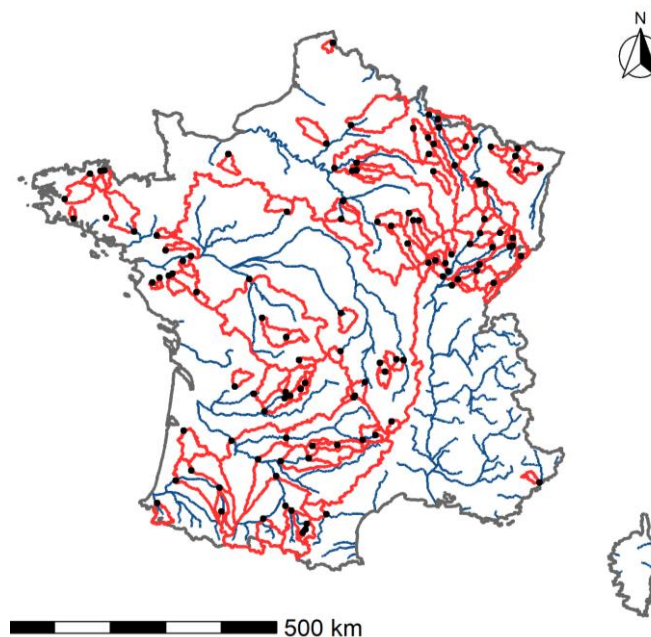
Here, we focus on simulating streamflow at the main catchment outlet, addressing the issue from a large-sample hydrology (LSH) perspective (Andréassian *et al.*, 2006; Gupta *et al.*, 2014), in which many catchments are used. After applying a simple criterion of streamflow availability (authorizing a maximum 10% gap per year with the possibility of exceeding it over 3 years), 121 catchments spread over mainland France with limited human influences were selected (Figure 1). Although snow-dominant or glacial regimes were rejected (due to lack of data or anthropogenic influence), the various catchments selected offer a wide hydroclimatic variability (Table 1).

150

Figure 1: Boundaries (in red) and outlets (black dots) of the 121 catchments selected for this study.

Table 1: Minimum, median and maximum values of some characteristics of the 121 catchments

Main characteristics of the catchment areas	Min.	Med.	Max.
Catchment area [km²]	15	880	110,188
Mean annual precipitation [mm y^{-1}]	661	960	1,699
Mean annual potential evapotranspiration [mm y^{-1}]	551	682	829
Mean annual flow [mm y^{-1}]	89	336	1,046
Humidity index (P/E$_0$) [-]	1.0	1.4	2.9
Runoff coefficient (Q/P) [-]	0.1	0.4	0.7

2.2 Principle of catchment spatial discretization

155 In this work, two spatial frameworks are used: lumped and semi-distributed. A lumped model considers the catchment as a single entity while the semi-distribution seeks to divide this catchment into several sub-catchments in order to partly take into account the spatial variability of hydroclimatic forcing and physical characteristics within the catchment.

Generally, the division of a catchment is defined on the basis of expertise and requires good knowledge of its characteristics (hydrological response units based on land use). From a large-sample hydrology perspective, an automatic definition of semi-

160 distribution was needed. To this end, we simplified the problem by looking at a first-order distribution, i.e. a single downstream

catchment defined by an outlet and one or more upstream sub-catchments. The underlying assumption is therefore that a second-order distribution (i.e. further dividing the upstream sub-catchments into a few smaller sub-catchments) will have a more limited impact on model behaviour than the first, when considering the main downstream outlet. Lobligeois *et al.* (2014) showed that a multitude of sub-basins of approximately 4 km² provide limited gain compared to a few sub-catchments of 64 km². Under these hypotheses, we used an automatic procedure to select semi-distributed configurations nested in each other, which we termed "Matryoshka doll" (Figure 2). This approach consists in creating different simple and distinct combinations of upstream–downstream gauged catchments starting from the main downstream station and progressively moving upstream.

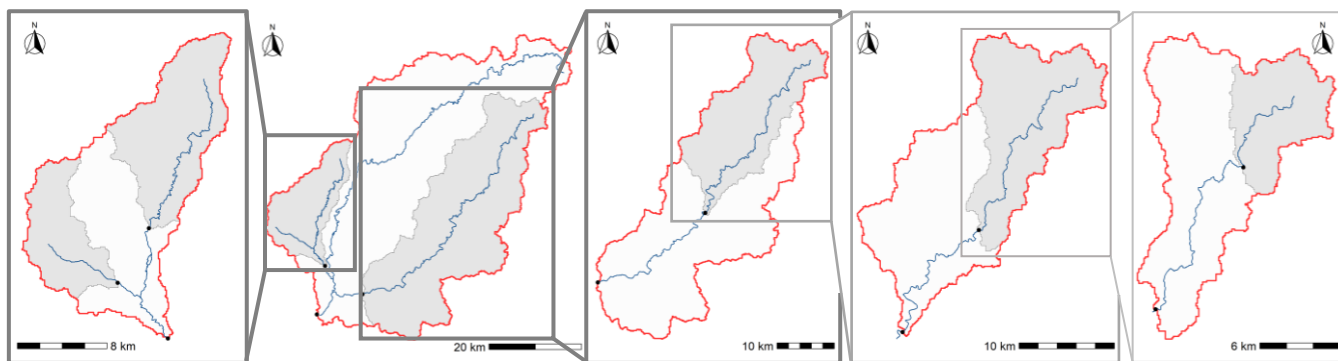
165

Specifically, the Matryoshka doll selection approach was implemented as follows:

1) Select a downstream station defining a catchment with one or more gauged internal points.

170 2) Restrict the upstream sub-catchment partitioning to a first-order split, i.e. going back only to the nearest upstream station(s) without going back to the stations further upstream, and respecting a size criterion to avoid sensitivity issues which may result from a too-small or too-large downstream catchment (in this study, we limited the area of the upstream sub-catchments to a value between 10% and 70% of the area of the total catchment). This step creates a combination of stations defining a single downstream catchment (which receives the upstream contributions).

175 3) If the upstream catchments have one or more internal gauged points, repeat step 1 and consider them as a downstream catchment.

The Matryoshka doll approach allows us to create distinct configurations (i.e. there cannot be two different semi-distributed configurations for the same downstream catchment) and therefore avoids over-sampling issues.

180 **Figure 2: Application of the Matryoshka dolls approach to the Vézère River at Larche. In this example, five semi-distributed configurations were obtained. The boundary of the main total catchment is in red, the first-order upstream catchments are filled in grey while the downstream catchment is white.**

The semi-distributed approach consists in performing lumped modelling on each sub-catchment by linking them through a routing scheme. For this study, a time lag between the upstream and downstream outlet was applied, as done by Lobligeois *et al.* (2014), and a sequential calibration was used, i.e. from upstream to downstream.

185

2.3 Models

In the context of this study, a model is defined as a configuration composed of a model structure and a method of parameter estimation (i.e. which may vary according to the objective function selected for calibration). These models will be applied independently in a lumped or a semi-distributed modelling framework.

190 For this study, the airGRplus software (Coron *et al.*, 2022), based on the works of Perrin (2000) and Mathevet (2005), was used. It includes various rainfall–runoff models structures running at the daily time step. airGRplus is an add-on to airGR (Coron *et al.*, 2017, 2021). An adaptation of the work by Perrin and Mathevet was carried out to use these structures at the hourly time step (mostly ensuring consistency of parameter ranges when changing simulation time steps and changing fixed time-dependent parameters). Finally, 13 structures were available for this study (Table 2). They are simplified versions of

195 original rainfall–runoff models taken from the literature (except GR5H, which corresponds to the original version). To avoid confusion with the original models, a four-letter abbreviation was used here. Since the various catchments used for this study are not prone to snowmelt issues, no snow module was implemented.

Table 2: List of rainfall–runoff models available in the airGRplus software at hourly time step and used for this work.

Original name	Number of free parameters	References	Name used in this study
Topmodel	7	Beven and Kirkby (1979)	TOPM
IHACRES	6	Jakeman *et al.* (1990)	IHAC
HBV	9	Bergström and Forsman (1973)	HBV0
Mohyse	8	Turcotte *et al.* (2010)	MOHY
Mordor	6	Garçon (1999)	MORD
Simhyd	8	Vaze *et al.* (2011)	SIMH
SMAR	9	O'Connell *et al.* (1970)	SMA0
TANK	10	Sugawara (1979)	TAN0
Gardenia	8	Thiéry (1982)	GARD
PDM	8	Moore and Clarke (1981)	PDM0
CREC	8	Cormary and Guilbot (1974)	CRE0
NAM	10	Nielsen and Hansen (1973)	NAM0
GR5H	5	Ficchì *et al.* (2019)	GR5H

200 The objective function used for parameter calibration is the Kling–Gupta efficiency (KGE) (Gupta *et al.*, 2009), defined by:

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

(1)

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

Three prior transformations on streamflow were applied in order to give more weight to some streamflow ranges. A KGE calculation on $Q^{+0.1}$ will give a relatively equivalent weight to the different streamflow ranges, $Q^{+0.5}$ will give more weight to
205 high flows, and $Q^{-0.5}$ more to low flows.

Overall, 13 structures and three calibration options were used, resulting in 39 models. Applied over two different spatial frameworks, a total of 78 distinct modelling options were available for this study.

2.4   Multi-model methodology

The multi-model approach consists in running various rainfall–runoff models. More specifically, here, we are interested in a
210 deterministic combination of the different streamflow simulations. Let us recall that for our study, a model corresponds to the association of a structure and a calibration option. By definition, a model is imperfect. Indeed, the different structures have been designed to meet different objectives (e.g. water resources management, forecasting, and climate change) in different geographical or geological contexts (e.g. high mountains, karstic zone, and alluvial plain). The calibration options (e.g. optimization algorithm, objective function, streamflow transformations), selected to optimize the parameters, are also choices
215 that will eventually impact the simulation *in fine*. The hypothesis made here is that the multi-model approach makes it possible to take advantage of the strengths of each model.

In the lumped framework, we consider every model on each catchment. In the semi-distributed framework, we consider every model on each sub-catchment. As the calibration is sequential, the various models are first applied to each upstream sub-catchment, and then their simulated streamflow is propagated to the downstream catchment to be modelled. However,
220 transferring every upstream possibility to the downstream catchment is excessively time consuming. Therefore, the simulated streamflow on each upstream sub-catchment was first set with an a priori choice, whatever the model used, and then transferred to the downstream catchment (this choice is discussed in Section 4.4).

The multi-model framework enables these different streamflow simulations to be combined on each catchment and sub-catchment in order to create multiple additional simulations. At the downstream outlet, we will consider mixed combinations,
225 using streamflow simulations from lumped and semi-distributed modelling (Figure 3). To this end, deterministic averaging methods were used. Here, we will focus on a simple average combination (SAC), i.e. giving an equal weight to all models combined, and defined by:

$$Q_{SAC} = \frac{\sum_{i=1}^{n} Q_i}{n} \qquad (2)$$

with Q_{SAC} the streamflow from a simple average combination, and Q_i the simulated streamflow with a model i selected among the n models.

230 Note that a weighted average combination (WAC) was also tested but did not significantly change the mean results and was therefore not used further (discussed in Section 4.3).

The number of possible combinations on a given outlet from the total number of available streamflow simulations increases exponentially and can be computed by:

$$n_c = \sum_{i=2}^{n_{sim}} \binom{n_{sim}}{i} \qquad (3)$$

with i the number of streamflow simulations to choose from the total number of available streamflow simulations n_{sim}.

235 As an indication, there are approximately 1,000 combinations for a streamflow ensemble simulated by 10 models but there are over 1,000,000 solutions for 20 models in a lumped framework. For this study, combinations will be set to a maximum of four different streamflow time-series among the total number of models available, i.e. approximately 1,500,000 different combinations (discussed in Section 4.2):

$$n_c = \sum_{i=2}^{4} \binom{78}{i} \approx 1{,}500{,}000 \qquad (4)$$

240 The objective of these combinations is to create a large set of simulations from which the best multi-model will be selected. We aim here to obtain simulations that are able to perform well over a wide range of streamflow and that can be applied to a large number of French catchments. Therefore, the best models (and multi-models) correspond to those which will achieve the highest performance on each catchment on average during the evaluation period.

2.5 Testing methodology

245 A split-sample test (Klemeš, 1986), commonly used in hydrology, was implemented. This practice consists in separating a streamflow time series into two distinct periods, the first for calibration and the second for evaluation, and then exchanging these two periods. The two periods chosen are 1999–2008 and 2009–2018. An initialization period of at least 2 years was used before each test period to avoid errors attributable to the wrong estimation of initial conditions within the rainfall–runoff model. For this study, results will be analysed only for the evaluation periods.

250 Model performance was evaluated on two levels:

- Over the whole time-series, model performance was evaluated with a composite criterion defined as the mean KGE calculated on $Q^{+0.5}$, $Q^{+0.1}$, and $Q^{-0.5}$.

- Based on events, specific criteria were defined for flood and low flows. In a context of high flows (5,447 events selected), the timing of the peak, the maximum flow and the flood volume were analysed, while the length and the severity of the low flows (1,332 events selected) were studied. Please refer to Appendix A for more details on the event selection method.

The best (i.e. giving the best performance over the evaluation period) lumped single model on a given catchment was taken as the benchmark for model evaluation.

3   Results

Results are presented from lumped (L) single models (SM), i.e. run individually, to more complex semi-distributed (SD) multi-models (MM) (see Figure 3). The mixed (M) multi-model allows for a variable spatial framework combining both lumped and semi-distributed approaches. The aim of this part is to present the results obtained with each modelling framework and their intercomparison.

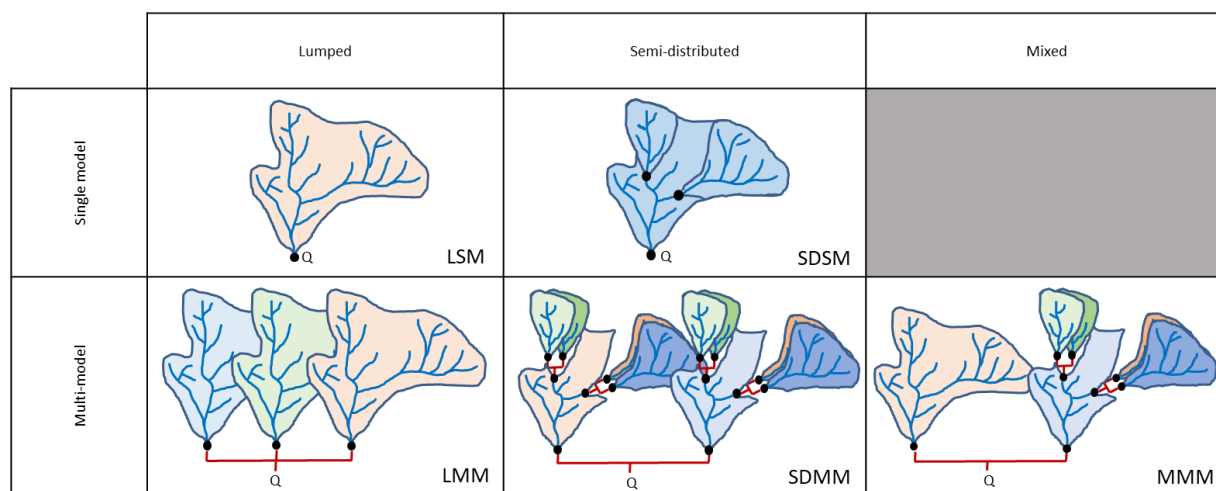



Figure 3: **Summary of the different approaches tested. Q is the target streamflow at the main catchment outlet; black dots show gauging stations used. The different colours represent different model structures. The variations of the same colour indicate different parameterizations. Red links represent the combination of streamflow.**

3.1   Lumped single models (LSM)

In this part, each model was run individually in a lumped mode (see Figure 3). Parameters of the 13 structures were calibrated successively with the three calibration options, resulting in 39 lumped models.

Figure 4 shows the distribution of the performance of lumped single models over the 121 downstream outlets over the whole time series. The KGE values obtained with different transformations applied over the evaluation period were first considered altogether to provide an overall picture of model performance (see Section 2.5). Overall, lumped single models give median KGE values between 0.70 and 0.88, which can be considered as slightly good to very good according to Crochemore *et al.*

275 (2015). The transformation applied during the calibration phase seems to have a variable impact on performance depending on the structure. For example, GR5H shows a similar performance regardless of the transformation applied, whereas TAN0 shows a large variation. The strong decrease in the 25% quantile of the latter is linked to the great difficulty for this structure to represent the low flows (KGE applied to $Q^{-0.5}$) when it is calibrated to enhance the high flows (KGE applied to $Q^{+0.5}$). Although the differences remain limited, the $Q^{+0.1}$ transformation applied to the streamflow for the objective function seems

280 to result in a better representation of the mean KGE (i.e. whatever the flow range considered) over the evaluation period.

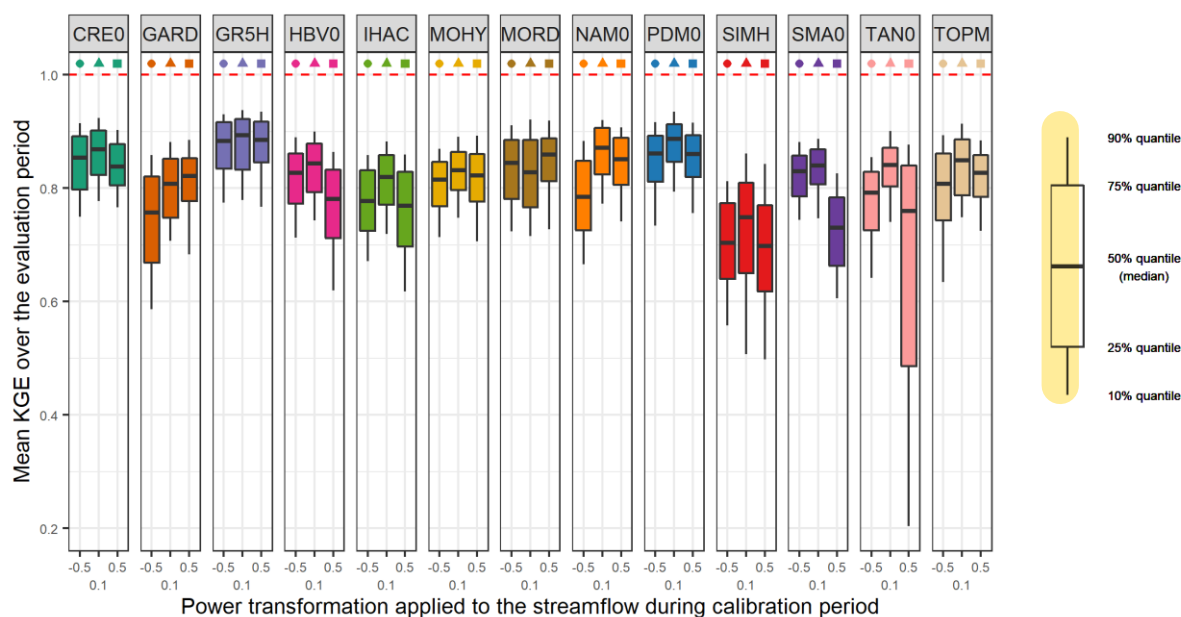



Figure 4: Distribution of the performance of the 39 lumped single models (average of the three evaluation criteria) over the 121 catchments and over the evaluation period. The boxplots represent the 10%, 25%, 50%, 75% and 90% quantiles. The dashed red line represents the optimal KGE value. Each colour represents a structure and each geometric pattern represents the power
285 **transformation applied to the streamflow during the calibration period.**

The left part of Figure 5 highlights the results obtained with the benchmark, i.e. by selecting the best lumped single model on each catchment. The benchmark presents high KGE values with a low variation (between 0.88 and 0.93 for the 25% and 75% quantiles, with a median of 0.91). The right part of Figure 5 indicates the number of catchments where each lumped single model is defined as the best. As expected, the models with high performance over the whole sample are selected more often

290 than the others. However, two thirds of the lumped single models have been selected at least once as the best on a catchment.

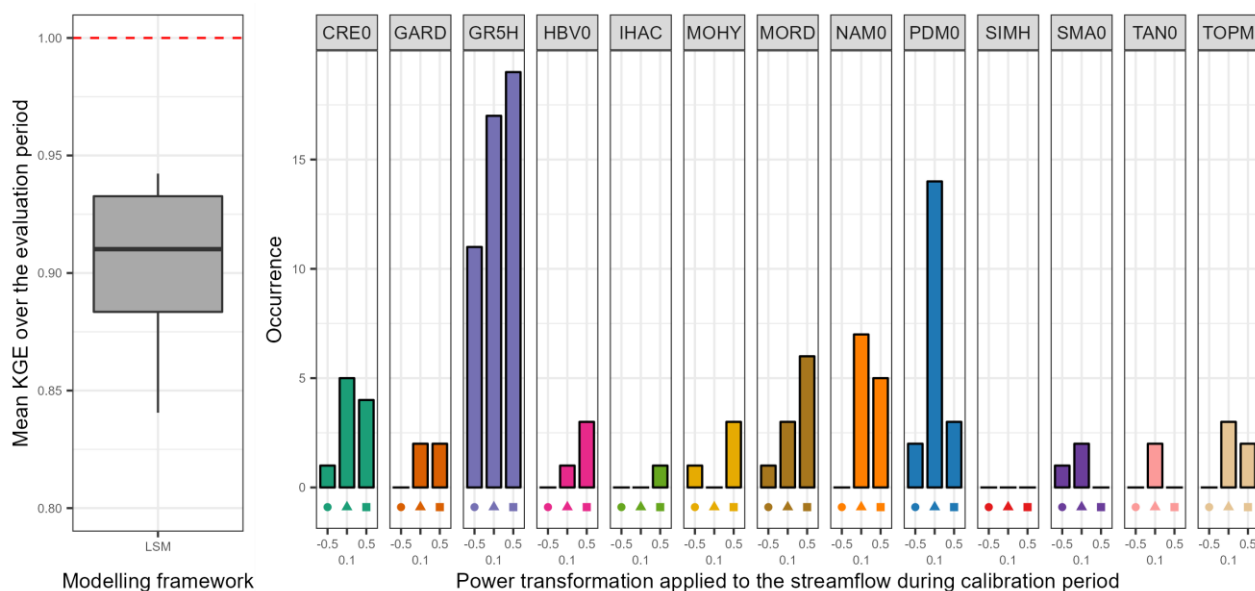



Figure 5: Distribution of the performance (average of the three evaluation criteria) of the best lumped single models (i.e. the benchmark) over the 121 catchments (left part) and model occurrence within this selection (right part). The boxplots represent the 10%, 25%, 50%, 75% and 90% quantiles. The dashed red line represents the optimal KGE value.

3.2   Semi-distributed single models (SDSM)

Remember that the semi-distribution with a single model (see Figure 3) is done sequentially, i.e. from upstream to downstream. Thus, each upstream sub-catchment is first modelled in a lumped mode with a single structure/calibration option pair. Then, the streamflow simulated upstream is propagated and the same model is calibrated and applied to the downstream catchment. This procedure is repeated for all 39 (13 structures and three calibration options) available models.

The left part of Figure 6 shows the comparison between mean KGE values obtained with lumped single models and semi-distributed single models. It shows a better performance for the semi-distributed models compared to their lumped versions in 65% of cases (and therefore worse or equal performance for the rest). However, about half of the KGE values show limited differences (lower than 0.02). The left part of Figure 6 aggregates these results in order to determine overall differences between the various models. Here, we only present the median value of the differences between mean KGEs for semi-distributed and lumped models over the entire catchment sample. The semi-distribution seems to have a variable impact on performance depending on the structure. For example, NAM0 shows a similar performance regardless of the spatial framework applied, whereas the performance of GARD improved with a spatial division. Although there is no clear trend in the impact of the semi-distribution in relation to the transformation applied during calibration, it seems that models calibrated on $Q^{+0.1}$ (i.e. giving "equal" weight on all flow ranges) show lower differences. Generally, the lumped models with the highest performance seem to benefit less from semi-distribution: the low-performing models may use the extra flexibility provided by the semi-distribution to compensate for some limitations of their structure.

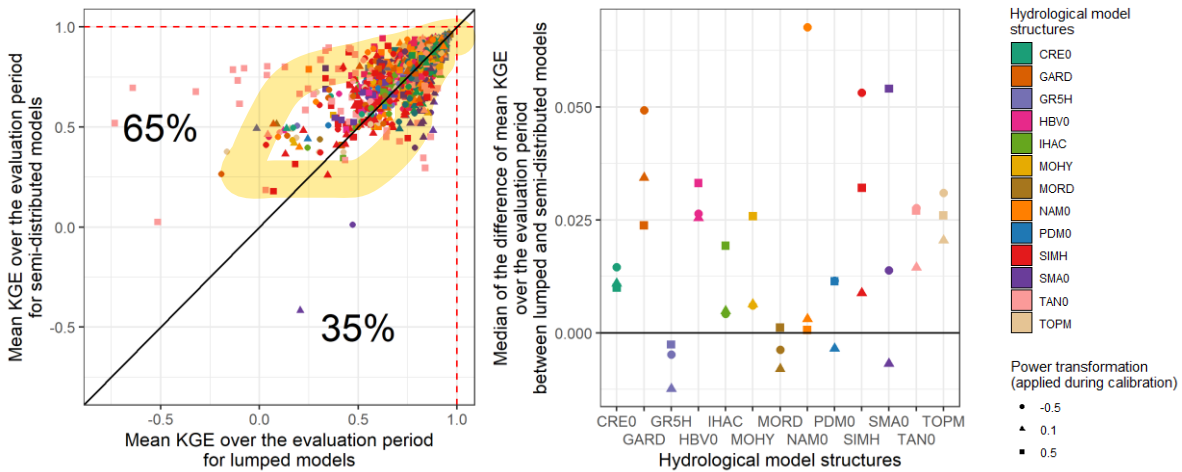



Figure 6: Comparison of the performance (average of the three evaluation criteria) obtained over the evaluation period by each model on each catchment in lumped mode vs in semi-distributed mode. The dashed red line represents the optimal KGE value. The black line indicates an equal performance between the lumped and semi-distributed approaches (dots above the line indicate better performance with the semi-distributed model, and vice versa).

Nevertheless, Figure 7 highlights that overall, if the focus is set on the best model on each catchment, the difference between semi-distributed and lumped single model (benchmark) remains small (no deviation for the quartiles and only 0.005 for the median). Once again, two thirds of the semi-distributed single models have been selected at least once as the best on a catchment.

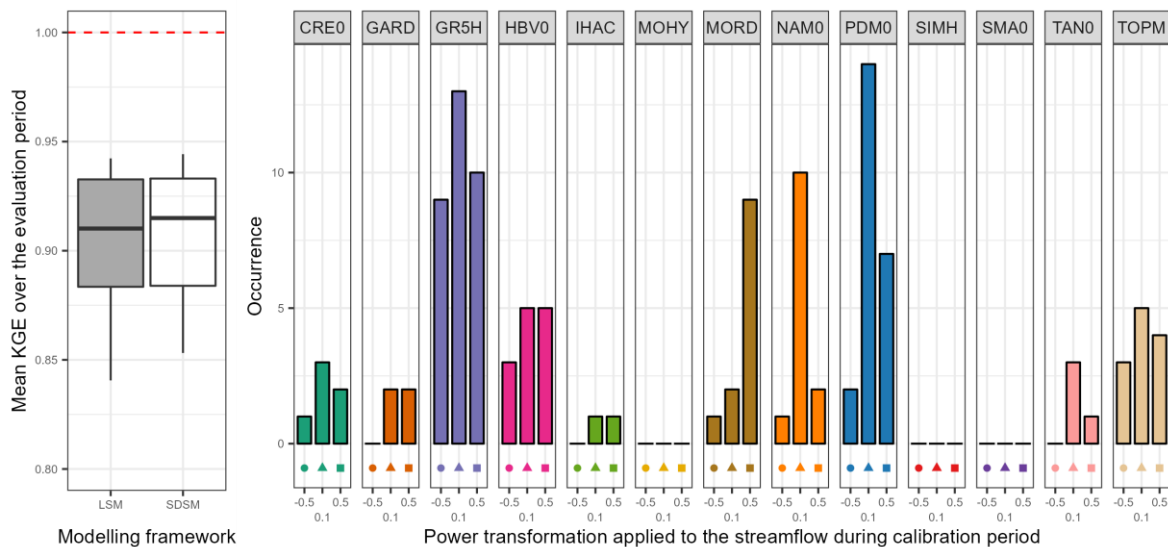



Figure 7: Distribution of the performance (average of the three evaluation criteria) of the best semi-distributed single models over the 121 catchments (left part) and model occurrence within this selection (right part). The boxplots represent the 10%, 25%, 50%, 75% and 90% quantiles. The dashed red line represents the optimal KGE value. The benchmark is the choice of the best lumped single model on each catchment and is represented by a grey box.

3.3 Lumped multi-models (LMM)

Here, each model was run in a lumped mode and model outputs were combined (see Figure 3). The multi-model approach used in this work is a deterministic combination with simple average (SAC) and will be limited to a combination of a maximum of four models among the available lumped models, i.e. approximately 92,000 different combinations.

330 The left part of Figure 8 shows the comparison between performance obtained with the benchmark and when the best lumped multi-model is selected on each catchment. As expected, the combination of lumped models enables an increase of 0.03 in the mean KGE value. While this gain may seem small at first glance, it is quite substantial since the performance obtained with the benchmark was already very high (median of 0.91), which makes improvements increasingly difficult. The right part of Figure 8 shows the number of times each model is selected within the best performing multi-model approach. It highlights the

335 benefits of a wide choice of models. Indeed, even if some of the models had never been used for the benchmark simulation (in a lumped single model framework), the multi-model approach shows that each of them can become a contributing factor to improve streamflow simulation in at least one catchment. Moreover, the models that are most often selected in the model combinations are not always the best models on their own. For example, TOPM, calibrated to favour high flows ($Q^{+0.5}$), was only used in the benchmark on 1.5% of the catchments but it is selected in the multi-model approach on 24% of the catchments.

340 However, the converse does not seem to be true since a model with good individual performance always seems to be a key element of the multi-model (e.g. GR5H, PDM0, MORD).

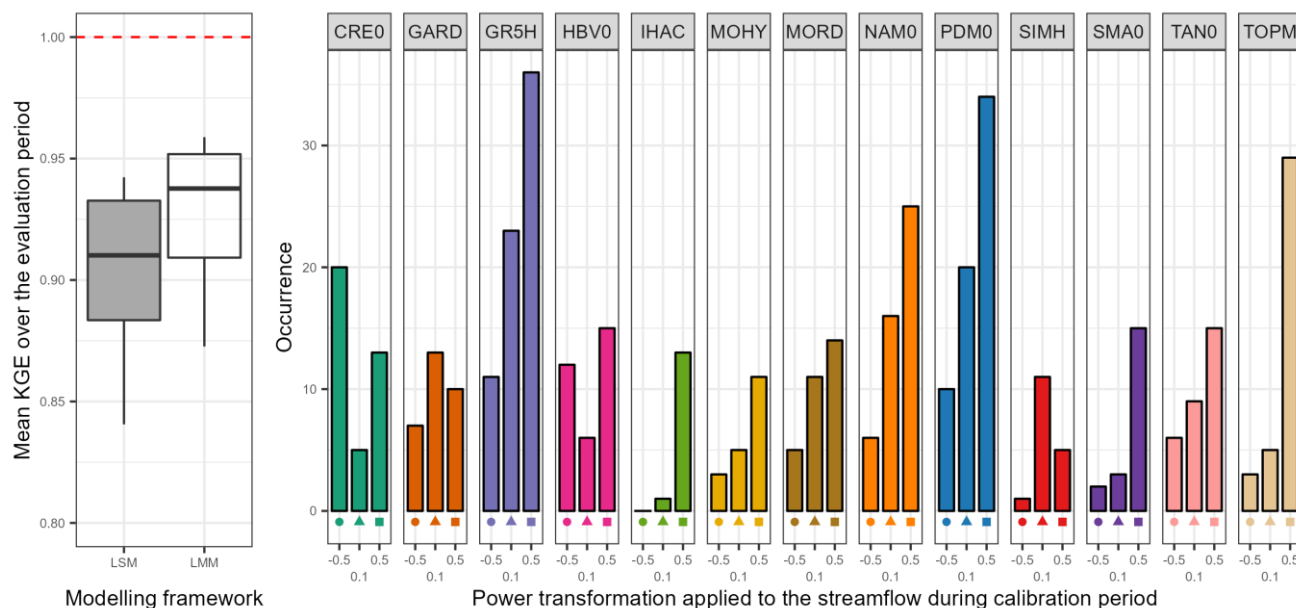



Figure 8: Distribution of the performance of the best lumped multi-models (average of the three evaluation criteria) over the 121 catchments (left part) and model occurrence within this selection (right part). The boxplots represent the 10%, 25%, 50%, 75% and
345 **90% quantiles. The dashed red line represents the optimal KGE value. The benchmark is the choice of the best lumped single model on each catchment and is represented by a grey box.**

3.4 Semi-distributed multi-models (SDMM)

For this study, the semi-distributed multi-model approach (see Figure 3) of the target catchment is performed in two steps.
First, the best multi-model (i.e. the combination of two to four models among the 39 available giving the highest performance
350 over the evaluation period) on each upstream sub-catchment is identified. In the second step, the simulated mean upstream
streamflow is propagated downstream and the different models are applied, and then combined (by 2, 3 or 4) on the downstream
catchment. Thus, approximately 92,000 different possible combinations of simulated streamflow are obtained at the outlet of
the total catchment.

Figure 9 shows very similar results to the LMM (Section 3.3). Indeed, we find again an improvement in the median of 0.03
355 compared to the benchmark and all the models are used on at least one downstream catchment. Moreover, the distribution of
the model count on the downstream catchment seems to be more homogeneous between the different members.

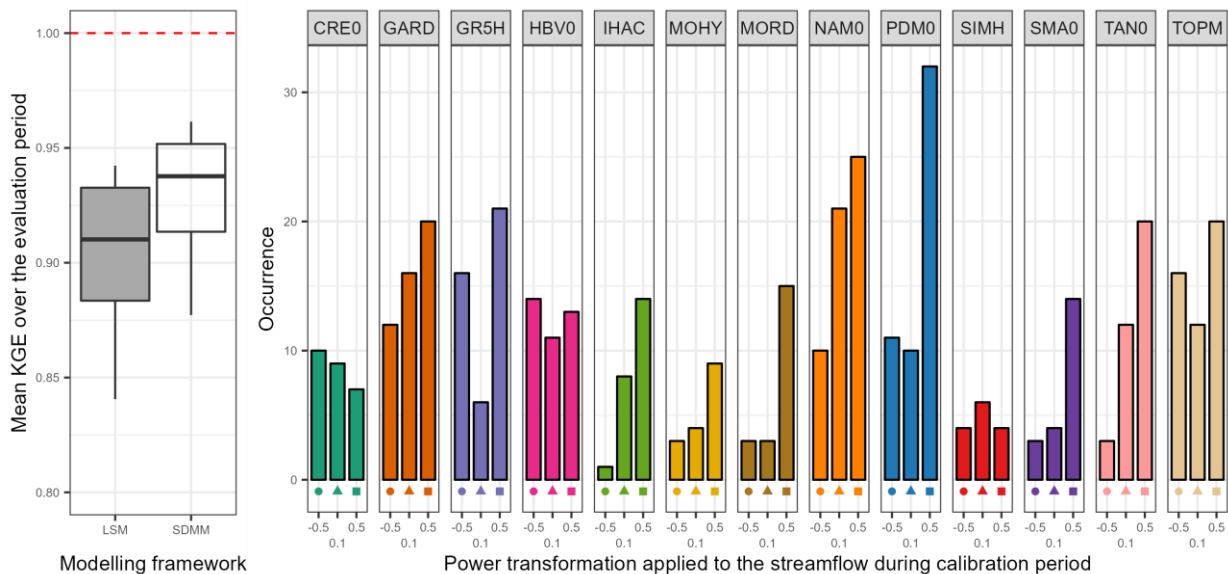



**Figure 9: Distribution of the performance of the best semi-distributed multi-models (average of the three evaluation criteria) over
the 121 catchments (left part) and model counts on the downstream catchment within this selection (right part). The boxplots
360 represent the 10%, 25%, 50%, 75% and 90% quantiles. The dashed red line represents the optimal KGE value. The benchmark is
the choice of the best lumped single model on each catchment and is represented by a grey box.**

3.5 Mixed multi-models (MMM)

Here, the mixed multi-model approach represents a combination of all the approaches tested above. This method allows for
each catchment a combination of models for a variable spatial framework (see Figure 3). In this context, the 39 models applied
365 to a lumped and a semi-distributed framework are used, resulting in 78 modelling options, each giving a different streamflow
at the outlet (Figure 10). These simulations can then be combined (by 2, 3 or 4) downstream in order to define the best mixed
multi-model on each catchment among more than 1,500,000 possibilities.

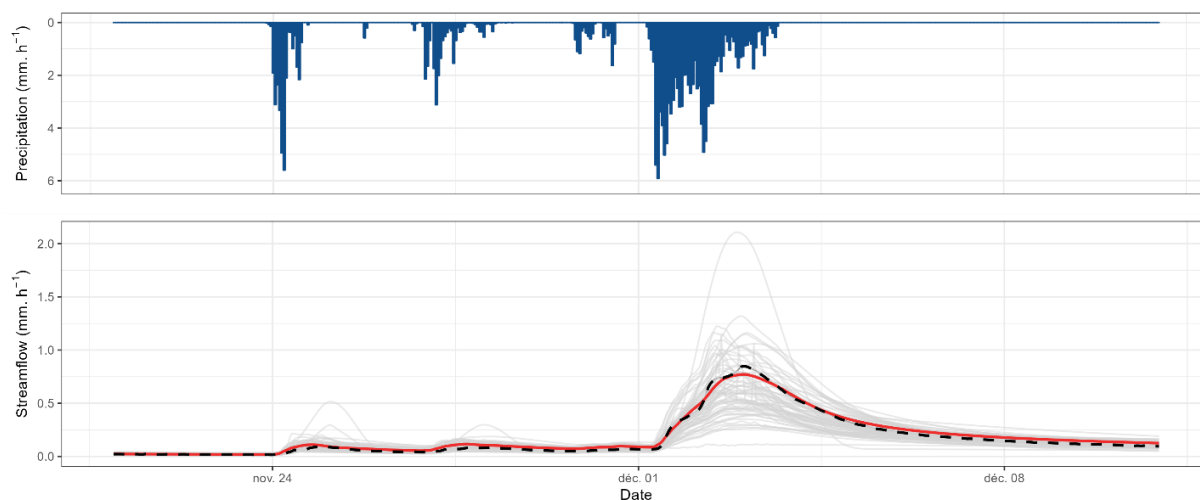



Figure 10: Hydrograph of November–December 2003 of the Dore River at Saint-Gervais-sous-Meymont (K287191001). The grey lines show the simulation obtained in an LSM and a SDSM framework. The red line highlights the best MMM combination. The dashed black line refers to the observed streamflow.

The left part of Figure 11 shows the performance obtained with the best mixed multi-model approach. The combination of lumped and semi-distributed models outperforms the benchmark. The right part of Figure 11 highlights the benefits gained from the wide choice of models and a variable spatial framework. Indeed, almost every lumped and semi-distributed models is used in order to improve the representation of streamflow with a multi-model approach in at least one catchment.

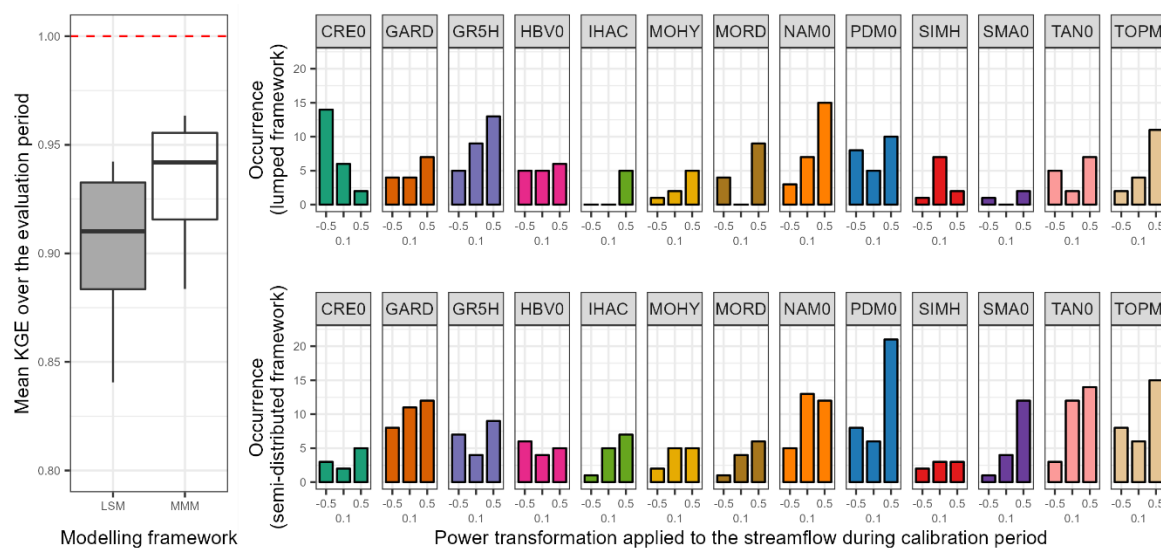



Figure 11: Distribution of the performance of the best mixed multi-models (average of the three evaluation criteria) over the 121 catchments (left part) and model counts within this selection (right part). The boxplots represent the 10%, 25%, 50%, 75% and 90% quantiles. The dashed red line represents the optimal KGE value. The benchmark is the choice of the best lumped single model on each catchment and is represented by a grey box.

3.6   Modelling framework comparison

Figure 12 compares the performance obtained when the best (multi-)model is selected on each catchment depending on the modelling approach used. The best LSM and SDSM distributions are almost identical and the same results are obtained with LMM and SDMM. This shows a limited gain of the semi-distribution approach compared to a lumped framework. However,

385   the LMM and SDMM outperformed the LSM and SDSM. Therefore, the increase in performance is mainly due to the multi-model aspect. Finally, the highest performance is obtained with the MMM but it remained close to the performance achieved by the LMM and SDMM.

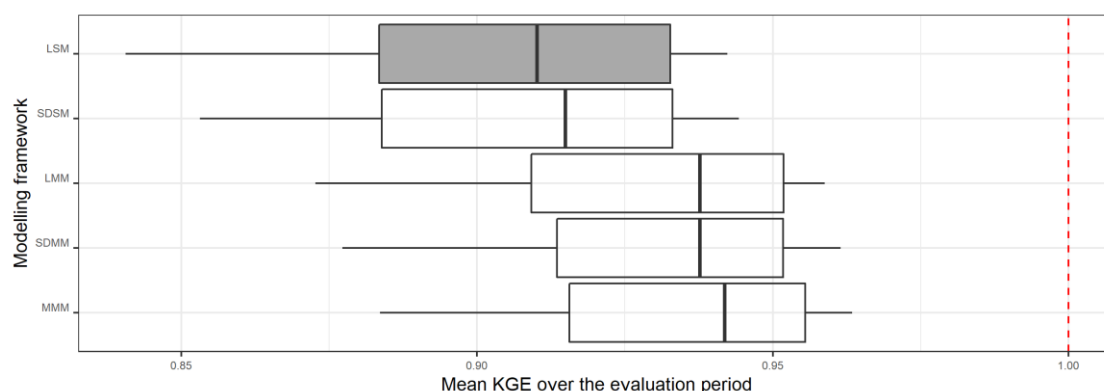



Figure 12: Distribution of the performance of the best (multi-)models (average of the three evaluation criteria) over the 121
390   **catchments according to their modelling framework (LSM: lumped single model; SDSM: semi-distributed single model; LMM: lumped multi-model; SDMM: semi-distributed multi-model; MMM: mixed multi-model). The dashed red line represents the optimal KGE value. The benchmark is the choice of the best lumped single model on each catchment and is represented by a grey box.**

Figure 13 shows the best performing modelling framework for each catchment. In general, the MMM approach seems to be

395   the most suitable for the majority of the catchments. However, if we accept to deviate by 0.005 (epsilon value arbitrarily set) from the optimal value, we notice that the lumped multi-model approach is sufficient on a large part of the catchments. There are no clear regional trends on which catchments require a more complex modelling framework.

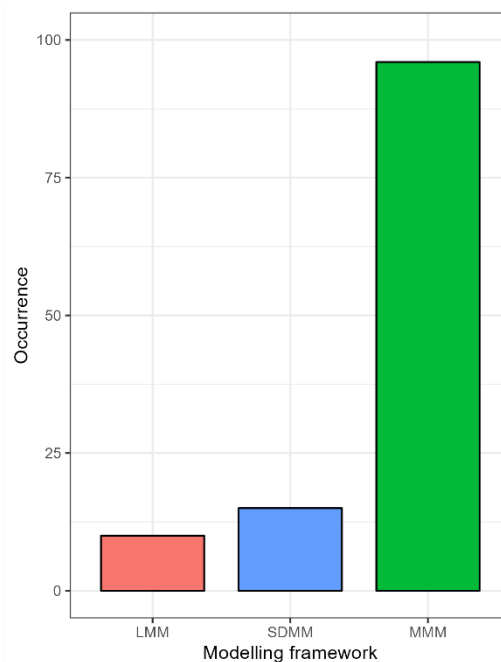
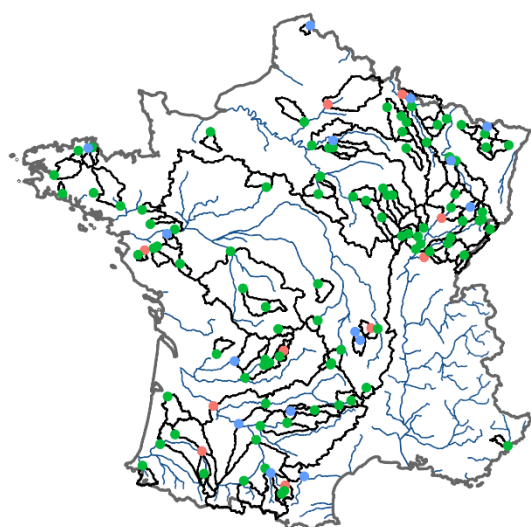
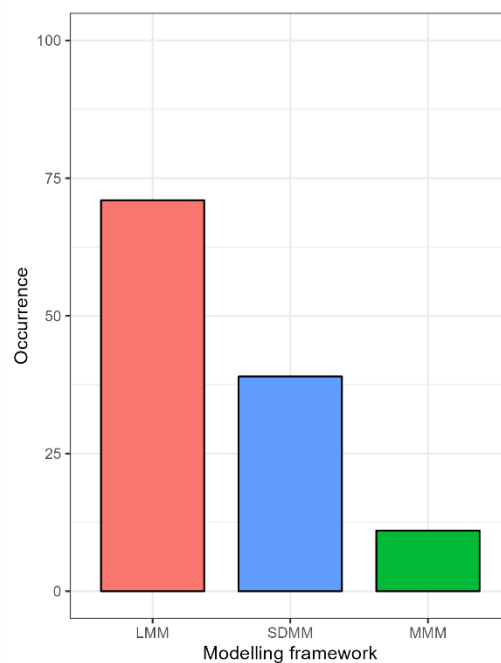
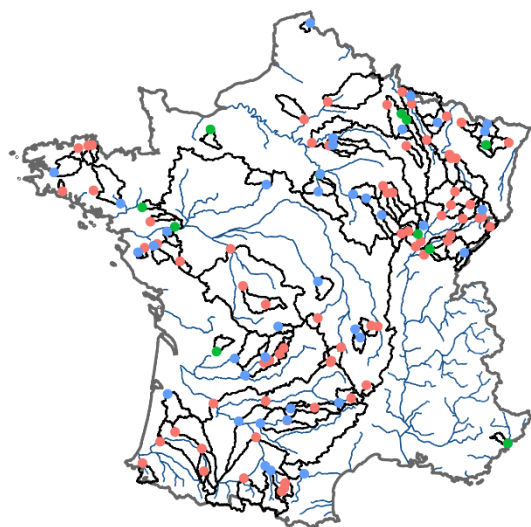



Figure 13: Maps of the best modelling framework on each catchment (top) and the simplest modelling framework among the best frameworks (bottom). The epsilon value corresponds to the performance deviation allowed from the best performance. If the drop between the evaluation performances is lower than or equal to epsilon, then the simplest modelling framework is selected.

Figure 14 shows the evaluation of the different modelling frameworks at the event scale. As a reminder, only the best model on each catchment is analysed for each approach tested. There does not seem to be a clear trend in the contribution of a complex mixed multi-model compared to a lumped single model (benchmark).

405   The duration of low-flow events is overestimated by about 7 days per year for all modelling frameworks. The maximum severity obtained each year during the events is also overestimated but remains close to the observations ($+ 0.18$ mm.d^{-1}). It should be noted that the performance of multi-model approaches is slightly less variable than for the single models alone during low-flow periods.

    The flood peak is late by about 1 h for the single-model approaches, whereas with a multi-model framework, it comes 1 h too

410   early. The most extreme values correspond to large catchments with slow responses and a strong base flow impact. In addition, multi-model approaches seem to have a lower variability for this criterion. The peak flow is slightly underestimated with a median of -0.05 mm.h^{-1}, like the flood flow, which shows a median deficit of 0.02 mm.h^{-1}.

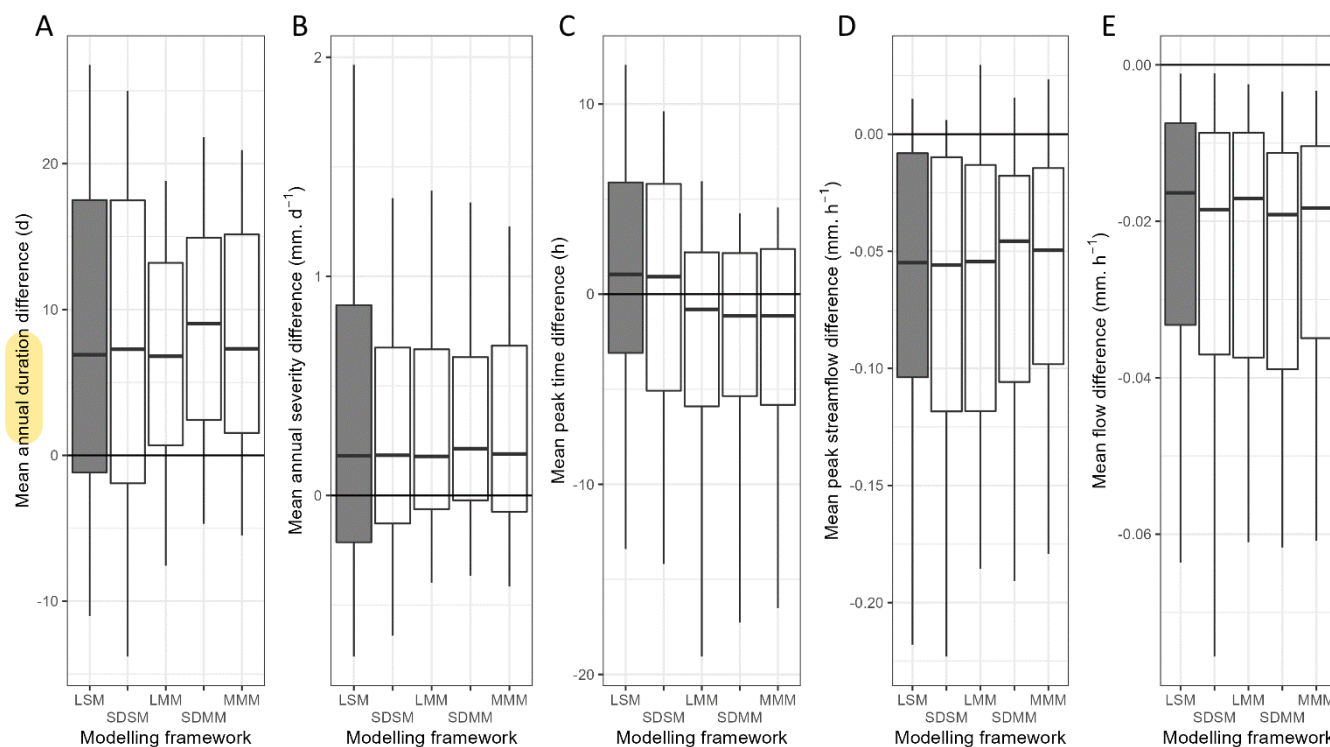



Figure 14: Comparison of various criteria between simulated and observed low-flow (A and B) or flood (C, D and E) events, according to the modelling approach used. The boxplots represent the 10%, 25%, 50%, 75% and 90% quantiles. The benchmark is the choice of the best lumped single model on each catchment and is represented by a grey box. The black line indicates the equivalence between the observed and simulated criteria. A positive value indicates an overestimation of the criterion by the simulation compared to observed streamflow.

4   Discussion

420   Here, the first objective is to discuss the results and answer the initial question: What is the possible contribution of a multi-model approach within a variable spatial framework for the simulation of streamflow over a large set of catchments? The second objective is to discuss the methodological choices by analysing them independently to determine to what extent they impact the results.

4.1   What is the possible contribution of a multi-model approach within a variable spatial framework?

425   First, our results confirmed the findings previously reported in the literature. Indeed, the multi-model approach outperformed the single models on a large sample of catchments (Shamseldin *et al.*, 1997; Georgakakos *et al.*, 2004; Ajami *et al.*, 2006; Velázquez *et al.*, 2011; Fenicia *et al.*, 2011; Santos, 2018; Wan *et al.*, 2021). Moreover, there is no clear benefit (on average) of using a semi-distributed framework because it degrades the streamflow simulation in some catchments and improves it in others (Khakbaz *et al.*, 2012; Lobligeois *et al.*, 2014; de Lavenne *et al.*, 2016).

430   The originality of our study is to combine these two approaches while providing a variable spatial framework. The mixed multi-model thus seems to benefit from the strengths of both methods. Most of the improvements compared to our benchmark come from the multi-model approach. On the other hand, although for a large part of the sample the differences are negligible, the variable spatial framework seems to generate an increase in the mean KGE values of up to 0.03 compared to a lumped multi-model approach (Figure 15). It should be noted that a similar difference is observed for the single models. By design,

435   the MMM does not deteriorate the performance when compared to what can be initially obtained with the LMM and SDMM approaches.

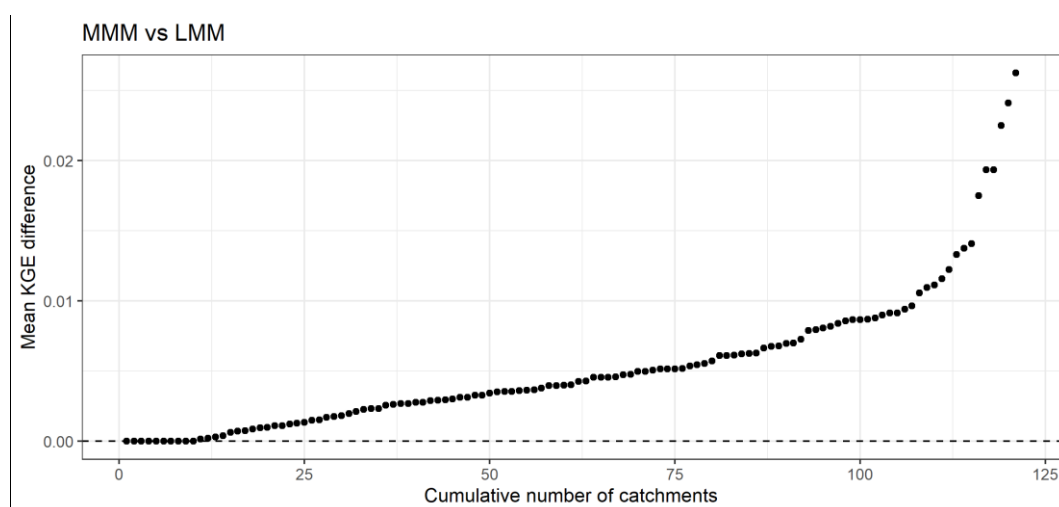



Figure 15: Ranked performance difference curve obtained for the evaluation period by the best lumped or mixed multi-model on each catchment. The black line indicates the equivalence between the two modelling frameworks. The higher the value, the greater
440   **the benefit of a variable spatial framework.**

Generally, this study has shown that a large number of models enables a better performance regardless of the streamflow range over a large sample of catchments. However, this methodology can be computationally expensive (due to the exponential number of combinations).

A mixed multi-model test was performed with a small pool of four models (i.e. structure/calibration option pairs) in order to
445   reduce this complexity (called "simplified MMM"; see Table 3 for more details). As a reminder, the procedure is the following:

-    In a semi-distributed framework, each of the four lumped models was applied and then combined on each upstream sub-catchment of the sample. Then, the best combination (i.e. giving the highest KGE value over the evaluation period) at each upstream outlet was propagated through the downstream catchment where the subsample of models was also used.

450   -   In a lumped framework, the modelling of each total catchment was performed with the different members of the simplified MMM.

-    Downstream simulations (four from the lumped approach and four from the semi-distributed approach) were then combined (resulting in 162 combinations) and the best multi-model at the outlet was selected.

Figure 16 shows that with the simplified MMM, the multi-model in a variable framework gives better KGE values than the
455   benchmark (which uses each lumped model independently and then selects the best one on each catchment). However, the performance obtained with the MMM approach is not reached, which again shows the added value of a wide choice of models.

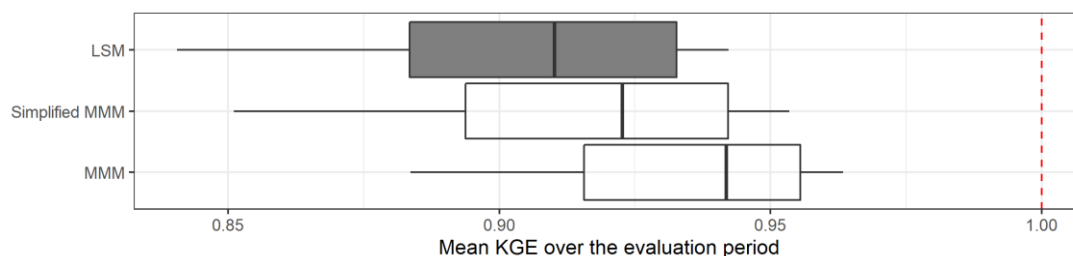



Figure 16: Distribution of the best simplified mixed multi-model, defined with a subset of four models, over the entire sample of catchments and over the evaluation period. The boxplots represent the 10%, 25%, 50%, 75% and 90% quantiles. The benchmark
460   **is the choice of the best lumped single model on each catchment and is represented by a grey box. The dashed red line represents the optimal KGE value.**

Table 3: Composition of the simplified MMM.

Structure	Parametrization
GR5H	KGE ($Q^{+0.5}$)
SMA0	KGE ($Q^{+0.5}$)
GARD	KGE ($Q^{+0.5}$)
NAM0	KGE ($Q^{+0.1}$)

4.2 What is the optimal number of models to combine in a multi-model framework?

The optimal number of models to combine in a multi-model framework varies between past studies. For example, Wan *et al.*
465   (2021) found that a limited improvement is achieved when more than nine models are combined; Arsenault *et al.* (2015)
concluded that seven models were sufficient; and Kumar *et al.* (2015) highlighted a combination of five members. This optimal
number therefore seems to vary with the catchment sample but also according to the number of models used and their qualities.
Figure 17 shows the results obtained by the best lumped (multi-)model on each catchment according to the number of members
combined. The largest improvement comes from a simple combination of two models and the performance gain becomes
470   limited from a combination of four different models (at least in our study).

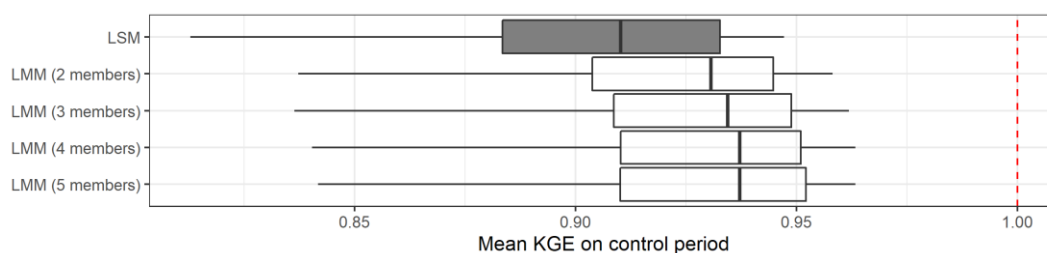



Figure 17: Distribution of the performance obtained with the best lumped multi-models on each catchment according to the number of models combined. The boxplots represent the 10%, 25%, 50%, 75% and 90% quantiles. The benchmark is the choice of the best lumped single model on each catchment and is represented by a grey box. The dashed red line represents the optimal KGE value.

475   ## 4.3 Is a weighted average combination always better than a simple average approach?

The weighted average combination (WAC) consists of assigning a weight that can be different for each model (Equation 3),
as opposed to the simple average combination (SAC), which considers each model in an identical manner (Equation 2).

$$Q_{WAC} = \frac{\sum_{i=1}^{n} \alpha_i \cdot Q_i}{\sum_{i=1}^{n} \alpha_i} \tag{5}$$

with Q_{WAC} the streamflow from a weighted average combination, Q_i the simulated streamflow with a model i selected among
the n models and α_i its attributed weight (between 0 and 1).

480   The complexity of the weighted average procedure lies in the estimation of the weights. Here, the weights have been optimized
according to the capacity of the combination to represent the observed streamflow by maximizing the KGE on the
transformations chosen during the calibration of the models. Thus, we obtain a set of weights for each calibration criterion
used by testing all possible values in steps of 0.1 between 0 and 1. The average of these different weight sets for the three
calibration options will then be taken as the final weighting.

485   This method becomes very expensive in terms of calculation time when the number of models increases. Thus, to answer this
question, a sub-sample of 13 models (corresponding to the 13 structures calculated on the square root of streamflow in a

lumped framework) was selected, and the tests were limited to a combination of three models (representing 364 distinct combinations in total).

Figure 18 shows the comparison of the SAC and WAC methods. Each point represents the mean KGE obtained over the evaluation period by each combination of models over the whole sample of catchments and over all the transformations evaluated. It highlights that the WAC and SAC methods provide similar mean results when dealing with a wide range of streamflow.

490

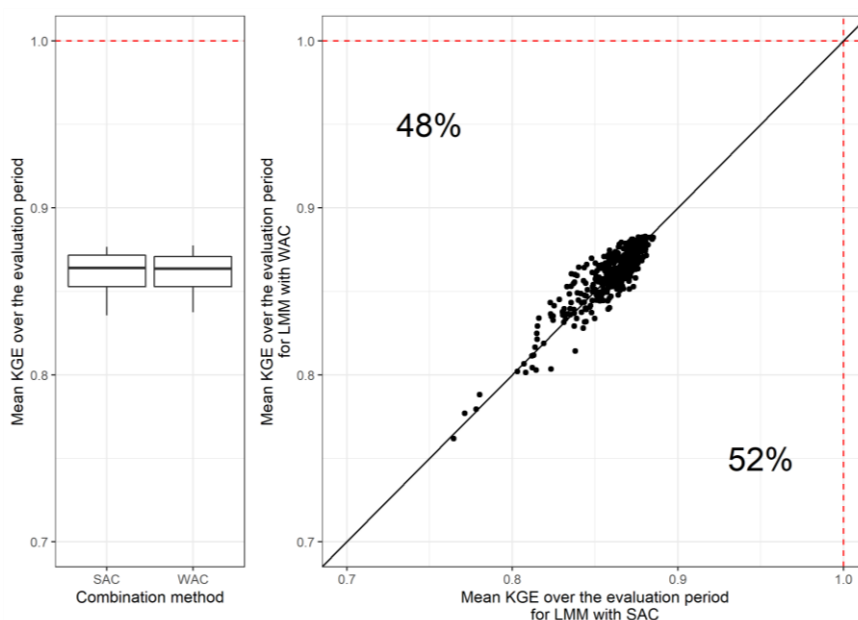



Figure 18: Comparison of the performance obtained over the evaluation period by each multi-model on the whole catchment sample according to the combination method used. The boxplots represent the 10%, 25%, 50%, 75% and 90% quantiles. The dashed red line represents the optimal KGE value. The black line indicates the equivalence between the two combination methods.

495

Another limit of the WAC procedure lies in the variability of the coefficients according to the calibration period. Moreover, this instability seems to increase with the number of models used in the combinations.

4.4 Is the a priori choice to use the best upstream multi-model always justified?

500 As a reminder, semi-distribution consists in dividing a catchment into several sub-catchments which can then be modelled individually with their own climate forcing and parameters, and then linked together by a propagation function. Generally, the number of possible downstream streamflow is set as:

$$n_{down} = n_{sim_{down}} + n_{c_{down}}(n_{sim_{down}}) \tag{6}$$

with $n_{sim_{down}}$ the number of simulations available downstream and $n_{c_{down}}$ the number of combinations from $n_{sim_{down}}$.

However, the number of simulations available downstream depends on the number of models available on this downstream
catchment and increases rapidly with the streamflow simulations injected from upstream catchments.

$$n_{sim_{down}} = n_{mod_{down}} \times \prod_{i=1}^{x} n_{sim_{up_i}} \qquad\qquad if\ x > 0$$

$$n_{sim_{down}} = n_{mod_{down}} \qquad\qquad if\ x = 0 \qquad\qquad (7)$$

with $n_{mod_{down}}$ the number of available models on the downstream catchment considered, x the number of direct upstream
sub-catchments and $n_{sim_{up_i}}$ the number of simulations available at the outlet of the upstream sub-catchment i.

It is therefore necessary to make an a priori choice on the different upstream sub-catchments in order to reduce the number of
possibilities downstream. The assumption made in this study is to propagate a single simulation, resulting from the best
combination of models, for each sub-catchment. Equation 6 then becomes:

$$n_{down} = n_{mod_{down}} + n_{c_{down}}(n_{mod_{down}}) \qquad\qquad (8)$$

This hypothesis ensures the same number of downstream streamflow simulations in a lumped or semi-distributed modelling
framework (as complex as it can be).

Simplified tests (semi-distributed configurations with one upstream catchment and only four distinct models - see Table 3 -
used without combination) were made in order to check the impact of this simplification on downstream performance. Figure
19 shows that for approximately 70% of the catchments, the use of an a priori choice on the injected upstream simulations has
a limited impact (< 0.01 difference in KGE values) on the downstream performance. However, a decrease of up to 0.05 can be
observed.

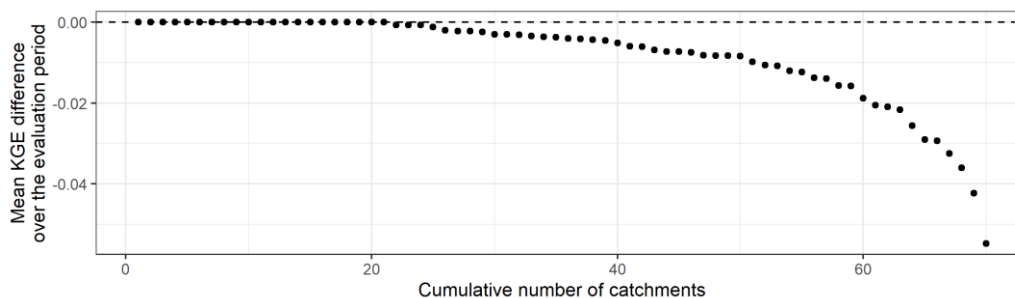



**Figure 19: Ranked performance difference curve obtained for the evaluation period on each catchment by the best downstream
model with an a priori choice on upstream simulation or not. The black line indicates the equivalence between both modelling
frameworks. A negative value indicates a decline in performance due to the a priori choice of upstream simulations injected.**

Although the assumption made here (i.e. to propagate the best upstream simulations) may occasionally lead to significant performance losses, we remain convinced that an a priori choice is necessary for a large gain in computation time even in simple semi-distributed configurations.

5   Conclusion

The main conclusions of this work are as follows:

1. The mixed multi-model approach provides higher KGE values than the different lumped single models. The gain is mainly due to the multi-model aspect, while the spatial framework brings a more limited added value.
2. At the event scale, the mixed multi-model approach does not show a large improvement on average, but seems to reduce the variability (i.e. inter-quantile deviation) compared to the benchmark.
3. Although some models are more often selected in multi-model combinations, almost all models have proven useful on at least one catchment. Moreover, the models that are most often selected in the model combinations are not always the best models on their own. However, the converse does not seem to be true since a model with good individual performance always seems to be a key element of the multi-model.
4. The largest improvement of a multi-model approach over the single models comes from the simple average combination of two models. The performance gain when increasing the number of models in the multi-model becomes limited when more than four different models are combined.
5. The simplified mixed multi-model (based on a sub-sample of only four models applied to a lumped and a semi-distributed framework) outperform the benchmark (based on the 39 available models applied to a lumped framework) but does not reach the performances obtained with a full mixed multi-model (based on the 39 available models applied to a lumped and a semi-distributed framework).

These conclusions are valid in the modelling framework used. As a reminder, in this work we aimed to obtain simulations that represent a wide range of streamflow and that can be applied to a large number of French catchments with limited human influences.

It would be interesting to test other deterministic methods to combine models such as random forest, artificial neural network or long short-term memory networks, which are increasingly applied in hydrology (Li *et al.*, 2022). Another perspective of this work would be to test the semi-distributed multi-model approach in a probabilistic framework by considering the different models as a hydrological ensemble in order to quantify the uncertainties related to the models. It would also be relevant to conduct this study in a forecasting framework by combining a hydrological ensemble with a meteorological ensemble.

Although we have worked in the context of a large hydrological sample, the catchments are exclusively located in continental France. Testing the semi-distributed multi-model approach on catchments under other hydroclimatic conditions may therefore

be useful. For example, applying the multi-model approach to different snow modules for taking into account snow melt is food for thought regarding high mountain catchments. It should be noted that the Matryoshka dolls approach developed in this study allows for only a simple division of the catchments. A more complex semi-distribution may be more relevant, especially

555   in places where the spatial variability of rainfall is high. The catchments with human influences were removed from our sample because they do not have a natural hydrological behaviour. However, semi-distribution often enables a better representation of streamflow in these areas which are difficult to model.

Appendix A. Event selection methodology: flood events

The selection procedure of flood events was based on the methodology developed/used by Astagneau *et al.* (2022). It is an

560   automated procedure selecting peak flows exceeding the 95% quantile and setting the beginning and the end of the flood event to 20% and 25%, respectively, of the flood peak. The starting window has been slightly extended by a few hours in the case of flash floods, characterized by a rapid rise in water levels (Figure A1).

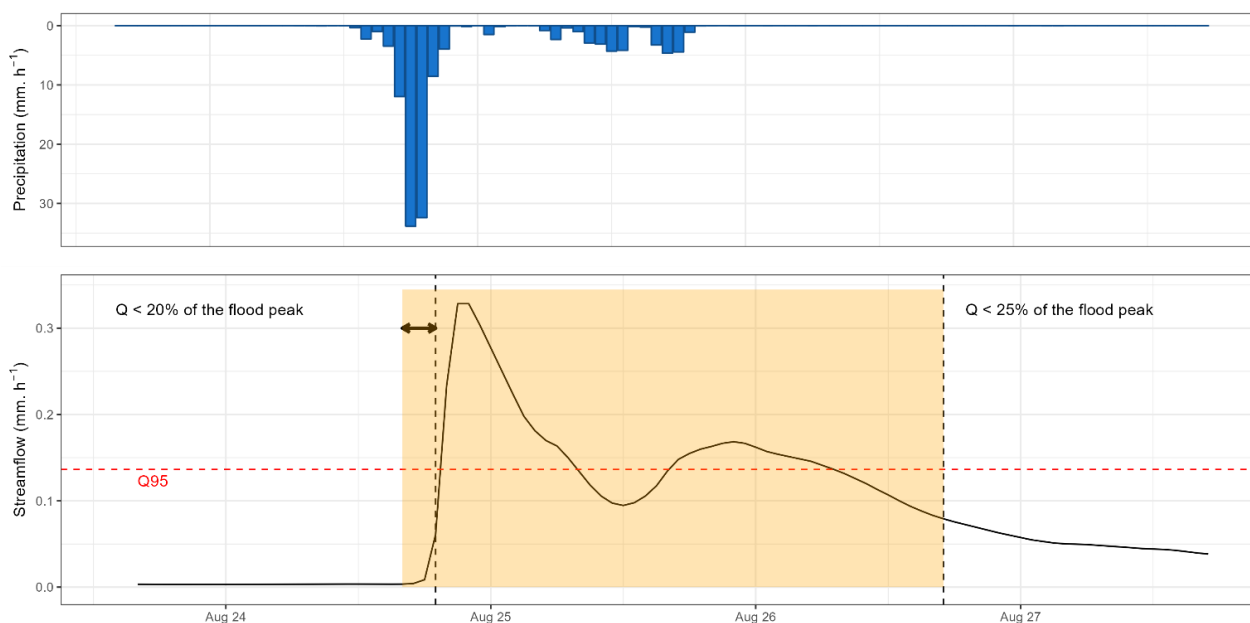



Figure A1: Flood event from 24 to 26 August 2009 on the Dore River at Saint-Gervais-sous-Meymont (K287191001). The yellow
565   **time window corresponds to the automatically selected event. The dashed red line represents the detection threshold set to streamflow quantile 95 %. The dashed black lines refer to the initial entry and end dates of the events (respectively below 20 % and 25 % of the flood peak). The arrow here shows an extension of the event start due to the rapid rise of water level.**

Each selected event was visually inspected to mitigate the errors associated with automatic selection. This step is particularly important for large catchments with inter-annual processes. In order to obtain consistent statistics between the catchments, a

570 maximum of 50 events (25 in calibration and 25 in evaluation) was set. In the end, 5,447 events were selected. Figure A2 shows the distribution of the number of flood events on each catchment.

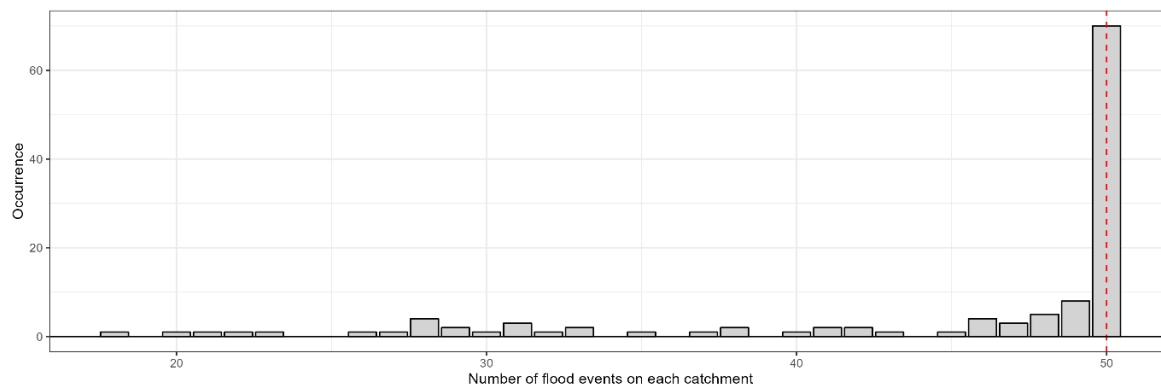



Figure A2: Distribution of the number of flood events on each catchment. The red line shows the maximum number allowed.

Appendix B. Event selection methodology: low-flow events

575 The selection procedure of low-flow events was based on the methodology developed/used by Caillouet *et al.* (2017). It is an automated procedure selecting periods under a threshold (fixed here to the 10% quantile) and aggregating the intervals corresponding to the same event thanks to the severity index (Figure B1).

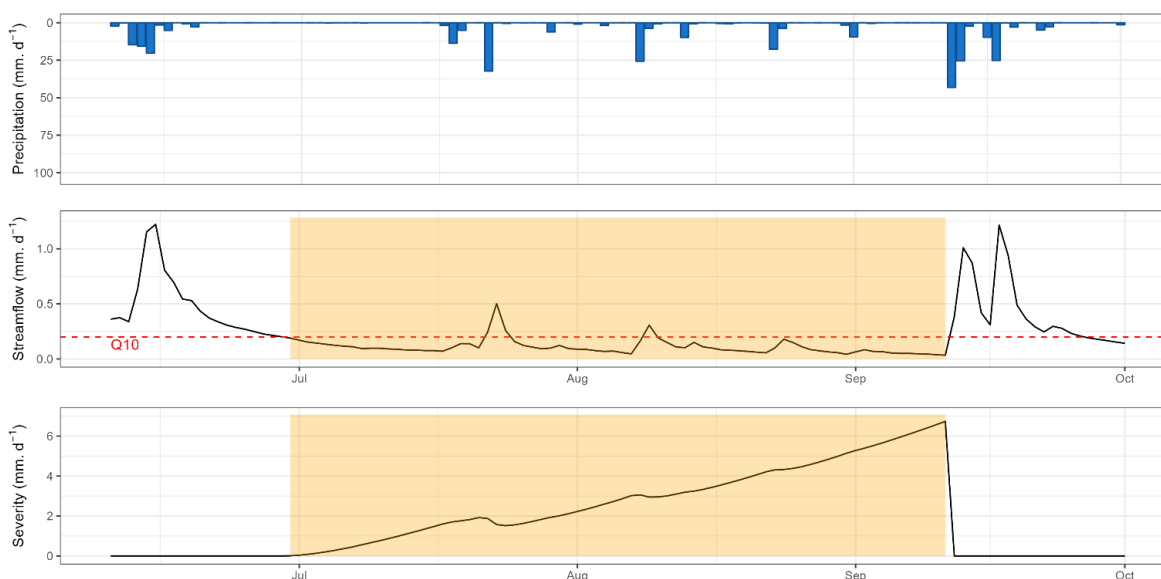



Figure B1: Low-flow event from summer 2015 on the Dore River at Saint-Gervais-sous-Meymont (K287191001). The yellow time
580 **window corresponds to the automatically selected event. The dashed red line represents the detection threshold set to the 10%**
quantile for streamflow.

Each selected event was visually inspected to mitigate the errors associated with automatic selection. This step is rather difficult because the quality of the low-flow data is quite heterogeneous (e.g. influenced by noise) from one catchment to another. In the end, 1,332 events were selected. Figure B2 shows the distribution of the number of low-flow events on each catchment.

585

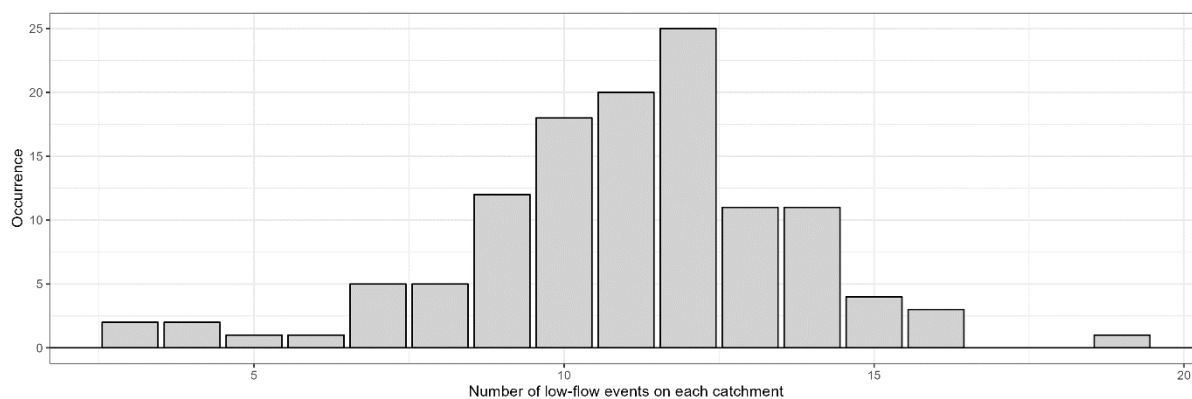



Figure B2: Distribution of the number of low-flow events on each catchment.

Author contributions. CP, CT, GT, SL and VA conceptualized the study. CP and CT developed the methodology. CT and
590   OD developed the model code. CT performed the simulations and analyses. CT prepared the manuscript with contributions from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

595   **Acknowledgements.** The authors wish to thank Météo-France and the SCHAPI, for making available the climate and hydrological data used in this study. CNR and INRAE are thanked for co-funding the PhD grant of the first author. Paul Astagneau and Laurent Strohmenger are also thanked for their advice on the manuscript. The authors thank Isabella Athanassiou for copyediting the manuscript.

600 **References**

Ajami, N.K., Duan, Q., Gao, X., Sorooshian, S., 2006. Multimodel Combination Techniques for Analysis of Hydrological Simulations: Application to Distributed Model Intercomparison Project Results. J. Hydrometeorol. 7, 755–768. https://doi.org/10.1175/JHM519.1

Ajami, N.K., Duan, Q., Sorooshian, S., 2007. An integrated hydrologic Bayesian multimodel combination framework: 605 Confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resour. Res. 43. https://doi.org/10.1029/2005WR004745

Andréassian, V., Hall, A., Chahinian, N., Schaake, J., 2006. Introduction and synthesis: Why should hydrologists work on a large number of basin data sets?, in: Large Sample Basin Experiments for Hydrological Parametrization : Results of the Models Parameter Experiment- MOPEX. IAHS Red Books Series N° 307. AISH, pp. 1–5.

610 Arsenault, R., Gatien, P., Renaud, B., Brissette, F., Martel, J.-L., 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. J. Hydrol. 529, 754–767. https://doi.org/10.1016/j.jhydrol.2015.09.001

Artigue, G., Johannet, A., Borrell, V., Pistre, S., 2012. Flash flood forecasting in poorly gauged basins using neural networks: case study of the Gardon de Mialet basin (southern France). Nat. Hazards Earth Syst. Sci. 12, 3307–3324. 615 https://doi.org/10.5194/nhess-12-3307-2012

Astagneau, P.C., Bourgin, F., Andréassian, V., Perrin, C., 2022. Catchment response to intense rainfall: Evaluating modelling hypotheses. Hydrol. Process. 36, e14676. https://doi.org/10.1002/hyp.14676

Atkinson, S.E., Woods, R.A., Sivapalan, M., 2002. Climate and landscape controls on water balance model complexity over changing timescales. Water Resour. Res. 38, 50-1-50–17. https://doi.org/10.1029/2002WR001487

620 Bergström, S., Forsman, A., 1973. Development of a conceptual deterministic rainfall-runoff model. Nord. Hydrol 4, 147–170. https://doi.org/10.2166/nh.1973.0012

Beven, K., 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Adv. Water Resour. 16, 41–51. https://doi.org/10.1016/0309-1708(93)90028-E

Beven, K., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology / Un modèle à base 625 physique de zone d'appel variable de l'hydrologie du bassin versant. Hydrol. Sci. Bull. 24, 43–69. https://doi.org/10.1080/02626667909491834

Block, P.J., Souza Filho, F.A., Sun, L., Kwon, H.-H., 2009. A Streamflow Forecasting Framework using Multiple Climate and Hydrological Models1. JAWRA J. Am. Water Resour. Assoc. 45, 828–843. https://doi.org/10.1111/j.1752-1688.2009.00327.x

630 Bourgin, F., Ramos, M.H., Thirel, G., Andréassian, V., 2014. Investigating the interactions between data assimilation and post-processing in hydrological ensemble forecasting. J. Hydrol. 519, 2775–2784. https://doi.org/10.1016/j.jhydrol.2014.07.054

Caillouet, L., Vidal, J.-P., Sauquet, E., Devers, A., Graff, B., 2017. Ensemble reconstruction of spatio-temporal extreme low-flow events in France since 1871. Hydrol. Earth Syst. Sci. 21, 2923–2951. https://doi.org/10.5194/hess-21-2923-2017

635 Clark, M.P., Nijssen, B., Lundquist, J.D., Kavetski, D., Rupp, D.E., Woods, R.A., Freer, J.E., Gutmann, E.D., Wood, A.W., Brekke, L.D., Arnold, J.R., Gochis, D.J., Rasmussen, R.M., 2015a. A unified approach for process-based hydrologic modeling: 1. Modeling concept. Water Resour. Res. 51, 2498–2514. https://doi.org/10.1002/2015WR017198

Clark, M.P., Nijssen, B., Lundquist, J.D., Kavetski, D., Rupp, D.E., Woods, R.A., Freer, J.E., Gutmann, E.D., Wood, A.W., Gochis, D.J., Rasmussen, R.M., Tarboton, D.G., Mahat, V., Flerchinger, G.N., Marks, D.G., 2015b. A unified 640 approach for process-based hydrologic modeling: 2. Model implementation and case studies. Water Resour. Res. 51, 2515–2542. https://doi.org/10.1002/2015WR017200

Clark, M.P., Slater, A.G., Rupp, D.E., Woods, R.A., Vrugt, J.A., Gupta, H.V., Wagener, T., Hay, L.E., 2008. Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. Water Resour. Res. 44. https://doi.org/10.1029/2007WR006735

645 Cormary, Y., Guilbot, A., 1974. Etude des rélations pluie-débit sur trois bassins versants d'investigation.

Coron, L., Delaigue, O., Thirel, G., Dorchies, D., Perrin, C., Michel, C., 2021. airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling. R package version 1.6.12. https://doi.org/10.15454/EX11NA

Coron, L., Perrin, C., Delaigue, O., Thirel, G., 2022. airGRplus: Additional Hydrological Models to the "airGR" Package. R package version 0.9.14.7.9001.

650 Coron, L., Thirel, G., Delaigue, O., Perrin, C., Andréassian, V., 2017. The suite of lumped GR hydrological models in an R package. Environ. Model. Softw. 94, 166–171. https://doi.org/10.1016/j.envsoft.2017.05.002

Crochemore, L., Perrin, C., Andréassian, V., Ehret, U., Seibert, S.P., Grimaldi, S., Gupta, H., Paturel, J.-E., 2015. Comparing expert judgement and numerical criteria for hydrograph evaluation. Hydrol. Sci. J. 60, 402–423. https://doi.org/10.1080/02626667.2014.903331

655 de Lavenne, A., Thirel, G., Andréassian, V., Perrin, C., Ramos, M.H., 2016. Spatial variability of the parameters of a semi-distributed hydrological model, in: 7th International Water Resources Management Conference of ICWRS, Proceedings of the International Association of Hydrological Sciences. Bochum, Germany, pp. 87–94. https://doi.org/10.5194/piahs-373-87-2016

Delaigue, O., Génot, B., Lebecherel, L., Brigode, P., Bourgin, P.-Y., 2020. Database of watershed-scale hydroclimatic
660 observations in France. INRAE, HYCAR Research Unit, Hydrology group, Antony.

Duan, Q., Ajami, N.K., Gao, X., Sorooshian, S., 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. Adv. Water Resour. 30, 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Dufeu, E., Mougin, F., Foray, A., Baillon, M., Lamblin, R., Hebrard, F., Chaleon, C., Romon, S., Cobos, L., Gouin, P., Audouy, J.-N., Martin, R., Poligot-Pitsch, S., 2022. Finalisation of the French national hydrometric data information system
665 modernisation operation (Hydro3). LHB 0, 2099317. https://doi.org/10.1080/27678490.2022.2099317

Fenicia, F., Kavetski, D., Savenije, H.H.G., 2011. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. Water Resour. Res. 47. https://doi.org/10.1029/2010WR010174

Ficchì, A., Perrin, C., Andréassian, V., 2019. Hydrological modelling at multiple sub-daily time steps: Model improvement via flux-matching. J. Hydrol. 575, 1308–1327. https://doi.org/10.1016/j.jhydrol.2019.05.084

670 Garçon, R., 1999. Modèle global pluie-débit pour la prévision et la prédétermination des crues. Houille Blanche 88–95. https://doi.org/10.1051/lhb/1999088

Georgakakos, K.P., Seo, D.J., Gupta, H., Schaake, J., Butts, M.B., 2004. Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. J. Hydrol. 298, 222–241. https://doi.org/10.1016/j.jhydrol.2004.03.037

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance
675 criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gupta, H.V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., Andréassian, V., 2014. Large-sample hydrology: a need to balance depth with breadth. Hydrol. Earth Syst. Sci. 18, 463–477. https://doi.org/10.5194/hess-18-463-2014

Her, Y., Chaubey, I., 2015. Impact of the numbers of observations and calibration parameters on equifinality, model
680 performance, and output and parameter uncertainty. Hydrol. Process. 29, 4220–4237. https://doi.org/10.1002/hyp.10487

Jakeman, A.J., Littlewood, I.G., Whitehead, P.G., 1990. Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. J. Hydrol. 117, 275–300. https://doi.org/10.1016/0022-1694(90)90097-H

685 Kavetski, D., Fenicia, F., Clark, M.P., 2011. Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment. Water Resour. Res. 47. https://doi.org/10.1029/2010WR009525

Khakbaz, B., Imam, B., Hsu, K., Sorooshian, S., 2012. From lumped to distributed via semi-distributed: Calibration strategies for semi-distributed hydrologic models. J. Hydrol., The Distributed Model Intercomparison Project (DMIP) - Phase
690 2 Experiments in the Oklahoma Region, USA 418–419, 61–77. https://doi.org/10.1016/j.jhydrol.2009.02.021

Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrol. Sci. J. 31, 13–24. https://doi.org/10.1080/02626668609491024

Kumar, A., Singh, R., Jena, P.P., Chatterjee, C., Mishra, A., 2015. Identification of the best multi-model combination for simulating river discharge. J. Hydrol. 525, 313–325. https://doi.org/10.1016/j.jhydrol.2015.03.060

695 Li, D., Marshall, L., Liang, Z., Sharma, A., 2022. Hydrologic multi-model ensemble predictions using variational Bayesian deep learning. J. Hydrol. 604, 127221. https://doi.org/10.1016/j.jhydrol.2021.127221

Liu, Y., Gupta, H.V., 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. Water Resour. Res. 43. https://doi.org/10.1029/2006WR005756

Lobligeois, F., 2014. Mieux connaître la distribution spatiale des pluies améliore-t-il la modélisation des crues ? Diagnostic
700 sur 181 bassins versants français (These de doctorat). AgroParisTech.

Lobligeois, F., Andréassian, V., Perrin, C., Tabary, P., Loumagne, C., 2014. When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation using 3620 flood events. Hydrol. Earth Syst. Sci. 18, 575–594. https://doi.org/10.5194/hess-18-575-2014

Loumagne, C., Vidal, J., Feliu, C., Torterotot, J., Roche, P., 1995. Procédures de décision multimodèle pour une prévision des
705 crues en temps réel: Application au bassin supérieur de la Garonne. Rev. Sci. Eau J. Water Sci. 8, 539–561. https://doi.org/10.7202/705237ar

Mathevet, T., 2005. Quels modèles pluie-débit globaux au pas de temps horaire ? Développements empiriques et comparaison de modèles sur un large échantillon de bassins versants (phdthesis). Doctorat spécialité Sciences de l'eau, ENGREF Paris.

710 McMillan, H., Krueger, T., Freer, J., 2012. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. Hydrol. Process. 26, 4078–4111. https://doi.org/10.1002/hyp.9384

Moore, R.J., Clarke, R.T., 1981. A distribution function approach to rainfall runoff modeling. Water Resour. Res. 17, 1367–1382. https://doi.org/10.1029/WR017i005p01367

Moradkhani, H., Sorooshian, S., 2008. General Review of Rainfall-Runoff Modeling: Model Calibration, Data Assimilation,
715 and Uncertainty Analysis, in: Sorooshian, S., Hsu, K.-L., Coppola, E., Tomassetti, B., Verdecchia, M., Visconti, G. (Eds.), Hydrological Modelling and the Water Cycle: Coupling the Atmospheric and Hydrological Models, Water Science and Technology Library. Springer, Berlin, Heidelberg, pp. 1–24. https://doi.org/10.1007/978-3-540-77843-1_1

Nielsen, S.A., Hansen, E., 1973. Numerical simulation of the rainfall-runoff process on a daily basis.
720 https://doi.org/10.2166/NH.1973.0013

O'Connell, P.E., Nash, J.E., Farrell, J.P., 1970. River flow forecasting through conceptual models part II - The Brosna catchment at Ferbane. J. Hydrol. 10, 317–329. https://doi.org/10.1016/0022-1694(70)90221-0

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., Michel, C., 2006. Dynamic Averaging of Rainfall-Runoff Model Simulations from Complementary Model Parameterizations. Water Resour. Res. 42. https://doi.org/10.1029/2005WR004636

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., Loumagne, C., 2005. Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. J. Hydrol. 303, 290–306. https://doi.org/10.1016/j.jhydrol.2004.08.026

Pechlivanidis, I., Jackson, B., Mcintyre, N., Wheater, H., 2011. Catchment scale hydrological modelling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications. Glob. Int. J. 13, 193–214.

Perrin, C., 2000. Vers une amélioration d'un modèle global pluie-débit (phdthesis). Institut National Polytechnique de Grenoble - INPG.

R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Saadi, M., Oudin, L., Ribstein, P., 2021. Physically consistent conceptual rainfall–runoff model for urbanized catchments. J. Hydrol. 599, 126394. https://doi.org/10.1016/j.jhydrol.2021.126394

Santos, L., 2018. Que peut-on attendre des Super Modèles en hydrologie ? Évaluation d'une approche de combinaison dynamique de modèles pluie-débit (phdthesis). Doctorat en Hydrologie, AgroParisTech.

Schaake, J.C., Hamill, T.M., Buizza, R., Clark, M., 2007. HEPEX: The Hydrological Ensemble Prediction Experiment. Bull. Am. Meteorol. Soc. 88, 1541–1548. https://doi.org/10.1175/BAMS-88-10-1541

Shamseldin, A.Y., O'Connor, K.M., Liang, G.C., 1997. Methods for combining the outputs of different rainfall runoff models. J. Hydrol. 197, 203–229. https://doi.org/10.1016/S0022-1694(96)03259-3

Smith, M.B., Seo, D.-J., Koren, V.I., Reed, S.M., Zhang, Z., Duan, Q., Moreda, F., Cong, S., 2004. The distributed model intercomparison project (DMIP): motivation and experiment design. J. Hydrol., The Distributed Model Intercomparison Project (DMIP) 298, 4–26. https://doi.org/10.1016/j.jhydrol.2004.03.040

Sugawara, M., 1979. Automatic calibration of the tank model / L'étalonnage automatique d'un modèle à cisterne. Hydrol. Sci. Bull. 24, 375–388. https://doi.org/10.1080/02626667909491876

Tabary, P., Dupuy, P., L'Henaff, G., Gueguen, C., Moulin, L., Laurantin, O., Merlier, C., Soubeyroux, J.-M., 2012. A 10-year (1997-2006) reanalysis of quantitative precipitation estimation over France: Methodology and first results. IAHS-AISH Publ. 351, 255–260.

Thiéry, D., 1982. Utilisation d'un modèle global pour identifier sur un niveau piézométrique des influences multiples dues à diverses activités humaines. https://doi.org/10.1080/02626668209491102

Turcotte, R., Fortier Filion, T.-C., Lacombe, P., Fortin, V., Roy, A., Royer, A., 2010. Simulation hydrologique des derniers jours de la crue de printemps: le problème de la neige manquante. Hydrol. Sci. J. 55, 872–882. https://doi.org/10.1080/02626667.2010.503933

van den Berge, L.A., Selten, F.M., Wiegerinck, W., Duane, G.S., 2011. A multi-model ensemble method that combines imperfect models through learning. Earth Syst. Dyn. 2, 161–177. https://doi.org/10.5194/esd-2-161-2011

van Esse, W.R., Perrin, C., Booij, M.J., Augustijn, D.C.M., Fenicia, F., Kavetski, D., Lobligeois, F., 2013. The influence of conceptual model structure on model performance: a comparative study for 237 French catchments. Hydrol. Earth Syst. Sci. 17, 4227–4239. https://doi.org/10.5194/hess-17-4227-2013

Vaze, J., Chiew, F.H.S., Perraud, J.M., Viney, N., Post, D., Teng, J., Wang, B., Lerat, J., Goswami, M., 2011. Rainfall-Runoff Modelling Across Southeast Australia: Datasets, Models and Results. Australas. J. Water Resour. 14, 101–116. https://doi.org/10.1080/13241583.2011.11465379

765

Velázquez, J.A., Anctil, F., Ramos, M.H., Perrin, C., 2011. Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. Adv. Geosci. 29, 33–42. https://doi.org/10.5194/adgeo-29-33-2011

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., Soubeyroux, J.-M., 2010. A 50-year high-resolution atmospheric reanalysis over France with the Safran system. Int. J. Climatol. 30, 1627–1644. https://doi.org/10.1002/joc.2003

770

Wan, Y., Chen, J., Xu, C.-Y., Xie, P., Qi, W., Li, D., Zhang, S., 2021. Performance dependence of multi-model combination methods on hydrological model calibration strategy and ensemble size. J. Hydrol. 603, 127065. https://doi.org/10.1016/j.jhydrol.2021.127065

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., Hinkelmann, R., 2021. Ensemble machine learning paradigms in hydrology: A review. J. Hydrol. 598, 126266. https://doi.org/10.1016/j.jhydrol.2021.126266

775