Journal of Hydrology 605 (2022) 127286

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Modeling streamflow variability at the regional scale: (2) Development of a bespoke distributed conceptual model

Fabrizio Fenicia [a,*], Dennis Meißner [b], Jeffrey J. McDonnell [c,d]

[a] Swiss Federal Institute of Aquatic Science and Technology (Eawag), Dübendorf, Switzerland
[b] Federal Institute of Hydrology (BfG), Koblenz, Germany
[c] Global Institute for Water Security, University of Saskatchewan, Saskatoon, Canada
[d] School of Geography, Earth & Environmental Sciences, University of Birmingham, Birmingham, UK

ARTICLE INFO

This manuscript was handled by Emmanouil Anagnostou, Editor-in-Chief, with the assistance of Jesús Mateo-Lázaro, Associate Editor

Keywords:
Regional scale
Top-down
Distributed model
SUPERFLEX
Hypothesis testing

ABSTRACT

Regional scale distributed conceptual models are typically developed with a bottom-up approach, which is process-inclusive but prone to over-parameterization. Here we demonstrate a proof of concept top-down approach for distributed conceptual model development, intended to emphasize dominant streamflow generating processes and to fulfill the principle of model parsimony. A key challenge in applying the top-down approach to distributed model development is devising a model comparison experiment that is both informative and limited to a few model alternatives. Here, we show how such model comparisons can be informed by a perceptual model of key processes that control streamflow response variability at the regional scale. We demonstrate our approach for the 27,100 km^2 Moselle catchment, using the perceptual model developed in Part 1 of this two-part paper. We develop 5 distributed model structures for simulating daily streamflow at 26 subcatchments, and validate them on subcatchments that are not used during the calibration process. Our model comparisons illustrate how the spatial distribution of precipitation, lithology and topography affect the simulation of key signatures of streamflow response variability in the Moselle catchment, providing a basis to justify model decisions. Our analyses show how a minimally parameterized distributed model, with 12 calibration parameters, matches signatures of streamflow average ($r = 0.96$), baseflow index ($r = 0.86$), and hydrograph lag time (correct at 22 out of 26 subcatchments). Our proposed top-down approach contributes to improving distributed model development strategies, and can be used to develop parsimonious process based regional models elsewhere.

1. Introduction

Distributed hydrological models have been used by hydrologists since the first flood predictions on the Durnace River in France by Imbeaux (1892). Although developed mostly for small catchments (e.g. Loague, 2010), they are increasingly used at the regional or global scale (e.g. Paniconi and Putti, 2015; Adams and Pagano, 2016; Fatichi et al., 2016). While field work and process knowledge can help with the selection of an existing model or even structural development of a new distributed model at the small catchment scale, such selection or development options are few at regional scales and beyond (e.g. Loritz et al., 2018; Ehret et al., 2020). In these large catchments, even simple questions and decisions are not straightforward: Should a model have a coarser spatial resolution to limit its complexity, or a finer one, to enable a more detailed process representation? Should a model consider the variability in the properties of soil and lithology? And if so, how can these data be collected and incorporated parsimoniously into the model structure? Is topographical data sufficient to drive the model since it captures much of the co-evolved relations between geology, soils and upslope area and local slope angle? Should the unsaturated zone processes be described by the Richards equation, or by a simple bucket model approach? Should regional groundwater flow that extends beyond sub-catchment boundaries be considered, or it is irrelevant? Should a model require calibration, and if so, how can over-parameterization be avoided?

These are but a few of the questions that confront the distributed catchment modeler today. And the answers to these questions are rather ad hoc. This exposes the regional scale distributed model to the many

* Corresponding author.
E-mail address: fabrizio.fenicia@eawag.ch (F. Fenicia).

https://doi.org/10.1016/j.jhydrol.2021.127286
Received 13 July 2021; Received in revised form 12 November 2021; Accepted 27 November 2021
Available online 11 December 2021

criticisms that have been levied against them: the danger of over-parameterization when available data are not sufficient to constrain model parameters (e.g. Beven, 1989; Gupta et al., 1998); incorrect upscaling premises when the process based equations are derived at a scale different from that to which they are applied (e.g. Grayson et al., 1992; Kirchner, 2006), unfulfilled spatial extrapolation assumptions when the dominant processes observed at one catchment are assumed to be dominant elsewhere (e.g. McDonnell et al., 2007; Savenije, 2009). So, despite > 130 years of work since Imbeaux's first distributed catchment model on the Durnace, we are still wondering how to inform the many decisions that developing a distributed model requires.

Distributed models are typically developed through a "bottom-up" approach, which is rather inclusive in terms of processes that may be important in principle, but may lead to the incorporation of model elements that contribute little to the overall catchment response (Fatichi et al., 2016). In this study, we explore distributed model development through a "top-down" approach, hence focusing on a disaggregation of the system responses into their constitutive components (Sivapalan et al., 2003). It can be expected that the top-down approach will generally not lead to hyper-resolution models (Wood et al., 2011), but rather, to so-called semi-distributed models, which occupy an intermediate position in the continuum from lumped to fully distributed models (e.g. Boyle et al., 2001). Semi-distributed models seek an appropriate balance between process distribution and aggregation, in the attempt to maintain a connection with observable landscape properties, while still satisfying the principle of parsimony. These models are widely used in research and operation, particularly at the regional scale, as their ability to represent spatial patterns makes them amenable to process based interpretations, and their relatively limited computational requirements enable their use in operational setups. In terms of scientific applications, such models have proven useful to: (i) characterize the behavior of landscape sections with different macro-scale properties and behavior, depending for example on topography or geology (e.g. Savenije, 2010; Ehret et al., 2020), (ii) facilitate the application of regularization relationships or parameters and processes constraints based on expert knowledge, which reduce over-parameterization and increase the realism of spatially distributed simulations (e.g. Pokhrel et al., 2008; Gharari et al., 2014; Kelleher et al., 2017), (iii) enable multi-site validation, through the use of internal streamflow measurements (e.g. Dal Molin et al., 2020) or groundwater time series (e.g. Raneesh and Thampi, 2013), and (iv) enable the comparison with remotely sensed data, such as the spatial and temporal patterns of evaporation and total water storage anomalies, which can be used as additional criteria to verify process consistency (e.g. Mulder et al., 2015; Hulsman et al., 2021). Such applications illustrate that semi-distributed models are a versatile tool to explore regional scale catchment behavior. Moreover, semi-distributed models are a common choice in operational setups (Adams and Pagano, 2016). For example, the model LARSIM is used operationally in Germany (Demuth and Rademacher, 2016), PREVAH in Switzerland (Viviroli et al., 2009), and LISFLOOD in Europe (Smith et al., 2016).

The top-down approach typically relies on model comparisons and has been commonly applied using lumped models, providing useful insights on suitable ranges of model complexity needed to model streamflow (e.g. Jakeman and Hornberger, 1993; Jothityangkoon et al., 2001), or helping to interpret dominant processes in small catchments (e.g. McMillan et al., 2011; Fenicia et al., 2014; Yokoo et al., 2017). Correspondingly, it can be expected that its systematic application to distributed models may provide information about warranted complexity and provide insights on which dominant processes characterize streamflow variability at the regional scale. But if lumped models comparisons can easily involve hundreds of variants (e.g. Prieto et al.; Spieler et al., 2020), distributed models are generally more time consuming to devise and execute, and as a result, their comparison is often practically limited to a handful of alternatives (e.g. Gao et al., 2014; Gharari et al., 2014; Fenicia et al., 2016; Nijzink et al., 2016; Kelleher et al., 2017; Antonetti and Zappa, 2018; Dal Molin et al., 2020). In face of the larger space of decisions that distributed models entail (e.g. Fenicia et al., 2016), such model comparisons need to be strategically constructed in order to be informative.

In this study, we address the problem of building a model comparison experiment that informs key decisions in the development of a semi-distributed model for streamflow simulation. In order to develop a minimal yet informative model comparison experiment, we rely on a perceptual model of regional scale processes. In a companion paper (Fenicia and McDonnell, 2022, hereinafter referred to as FM2022), we outlined the approach for building such a perceptual model, leveraging data typically available at the regional scale. Here, we illustrate the process of translating that perceptual model into conceptual model decisions, and the use of controlled model comparisons to assist the ultimate model selection. Stringent model evaluation is essential in assessing the relative merits of the models that take part in such comparison experiments (e.g. Fenicia and Kavetski, 2021). In order to provide a comprehensive assessment of model performance, here we evaluate all models assessing the goodness of fit of both time series and signatures, hence, both in the "time" and in the "signature" domain (e.g. Hrachowitz et al., 2014; Kelleher et al., 2017; Kavetski et al., 2018). Moreover, we assess the models' ability to make predictions not only in time, but also in space, hence using subcatchments that are not used for calibration. Such space–time validation is a particularly stringent and revealing model evaluation instrument (Refsgaard and Knudsen, 1996).

The model development chain proposed here and in FM2022 enables the incorporation of expert knowledge into the model development process, which can be contributed both by the experimentalist and by the modeler. The "dialog between experimentalist and modeler" is considered an important instrument to increase model realism (Seibert and McDonnell, 2002). However, such dialog has mainly taken place at the hillslope- or headwater catchment scale. The development of distributed models, particularly at the regional scale, has been largely devoid of any dialog between experimentalist and modeler, mostly relegating the role of fieldwork to data collection for model parameterization (Burt and McDonnell, 2015). More generally, hydrological models have often been the result of very specific, often individual expertise. However, a participatory, rather than expert specific model development process can—as we will show—facilitate model understanding, criticism and revision. We call this a "bespoke model"—one that is tailor made for the regional watershed.

In FM2022 we developed a perceptual model of the 27,100 km^2 Moselle catchment, which explains the spatial variability of streamflow signatures observed at 26 gauged subcatchments. Briefly, this perceptual model established that precipitation, much more than evaporation or groundwater exchange, controls the spatial variability of average streamflow, bedrock permeability influences the partitioning between baseflow and quickflow, and topography and land use control hydrograph lag times. It also provided an assessment of which other landscape properties did not appear to have significant effect on streamflow variability. In this paper, we illustrate how this perceptual model can be used for informing a distributed model comparison experiment, aimed at determining and justifying a bespoke semi-distributed model for the catchment of interest. In order to alleviate the model building burden, the distributed models are developed within the flexible framework SUPERFLEX (Fenicia et al., 2011; Dal Molin et al., 2021).

Here we pursue the following objectives:

1. To illustrate the process of translating the Moselle perceptual model into a distributed conceptual model, and in particular how the perceptual model informs the multiple decisions that distributed models require.
2. To use model comparisons in space–time validation to test the major hypotheses about the process controls on the selected streamflow signatures posed by the perceptual model, and in particular:
   a. The effect of precipitation on average streamflow

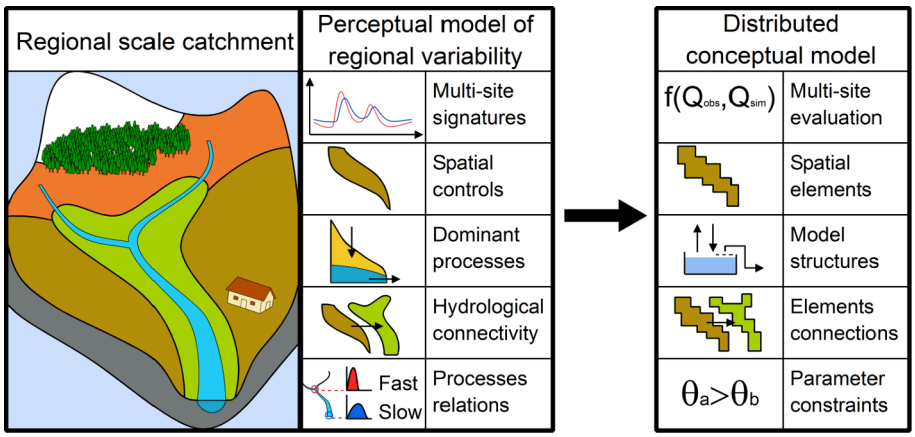

Fig. 1. Transition from a perceptual model of regional variability to a distributed conceptual model for a regional scale catchment.

Table 1
Summary of model variants.

Model variants	Motivation	# of HRUs	# of calibration parameters	# of landscape elements
M(HRU_1, Lag_0)	Benchmark model	1	7	26
M(HRU_1, Lag_1)	Include routing	1	9	26
M(HRU_2, Lag_1)	Add HRU	2	12	52
M(HRU_2, Lag_2)	Improve routing	2	13	52
M(HRU_3, Lag_2)	Add/tailor HRUs	3	12	75

b. The effect of lithology on the baseflow index
c. The effect of topography and land use on lag times

3. To discuss the added value of the distributed conceptual model compared to the underlying perceptual model of regional variability.

The paper is organized as follows. Section 2 describes the methods, and in particular the transition from the perceptual model development to the conceptual model, the model variants for hypothesis testing, and the model calibration and evaluation strategies. Section 3 describes the model results, both in the time- and in the key signature domain. Section 4 discusses the specific results of the model comparisons, and the general contribution of the proposed modeling approach. Section 5 summarizes the key conclusions.

2. Methods

2.1. From perceptual model of regional variability to distributed conceptual model

The main application of the distributed model developed in this study is to simulate streamflow at a number of points along the river network of a given catchment, hence at selected internal subcatchments. This application can be considered the most typical for a distributed regional catchment model, granted that such models can simulate variables other than streamflow (e.g. Samaniego et al., 2010; Hirpa et al., 2018; Athira and Sudheer, 2021).

In order to build a process-based distributed model with limited complexity for the selected application, we adhere to some common choices. In particular, we use semi-distributed models, hence designed to reproduce flow at a limited number of points along the river network (Boyle et al., 2001), and adopt the concept of on Hydrological Response Units (HRUs) (Leavesley et al., 1983) for distributing landscape properties.

Even restricting our scope to semi-distributed, HRU-based models, model development requires several additional decisions, which can be summarized as follows: (1) defining the HRUs, hence the spatial discretization approach; (2) specifying the model structures associated to each HRU; (3) establishing the connection between such model structures; (4) defining parameter constraints to improve model parsimony; (5) establishing model evaluation and diagnostic metrics.

In this study, this set of decisions is informed by a perceptual model. Fig. 1 indicates the correspondence between the information that a perceptual model might provide, and key decisions that distributed model building requires. FM2022 illustrated how a perceptual model of streamflow regional variability can be developed, in order to provide information on these specific points.

The process of translating a perceptual model into a conceptual model is not obvious or univocally defined. Attempting a direct mapping between perceptual and conceptual model would still involve many ad-hoc decisions, with the risk that the resulting conceptual model may be too simple or too complex based on the available data. Therefore, rather than directly translating the perceptual model into a conceptual model, here we take the approach of exploiting the perceptual model to inform a set of model variants, which are aimed to test a selection of the key hypotheses that the perceptual model embodies. The intention of this comparison is to gain insights into the effect of model decisions, and correspondingly, into the dominant processes that characterize regional scale catchment behavior in the study area.

2.2. Translation of the Moselle perceptual model into 5 distributed model variants

FM2022 developed a perceptual model for the Moselle catchment, which interprets the spatial variability in the streamflow response of its 26 subcatchments. In synthesis, this model established that (1) streamflow spatial variability can be expressed by a set of key signatures, characterizing streamflow average, baseflow index, and hydrograph lag times, (2) the spatial variability of these signatures is controlled by distinct climate or landscape properties, which need to be explicitly represented in a distributed model. In particular precipitation controls the spatial variability of streamflow average, lithology influences subsurface processes and eventually the baseflow vs. quickflow partitioning, and topography and land use control hydrograph lag times, and (3) conversely, the spatial variability of these signatures does not appear to be affected by other properties, which do not need to be spatially resolved in a model representation. In particular, vegetation and soil do not appear to play a major role in explaining streamflow signatures spatial variability. Moreover, the subcatchments appear to be "water tight", in the sense that regional groundwater flow extending beyond the

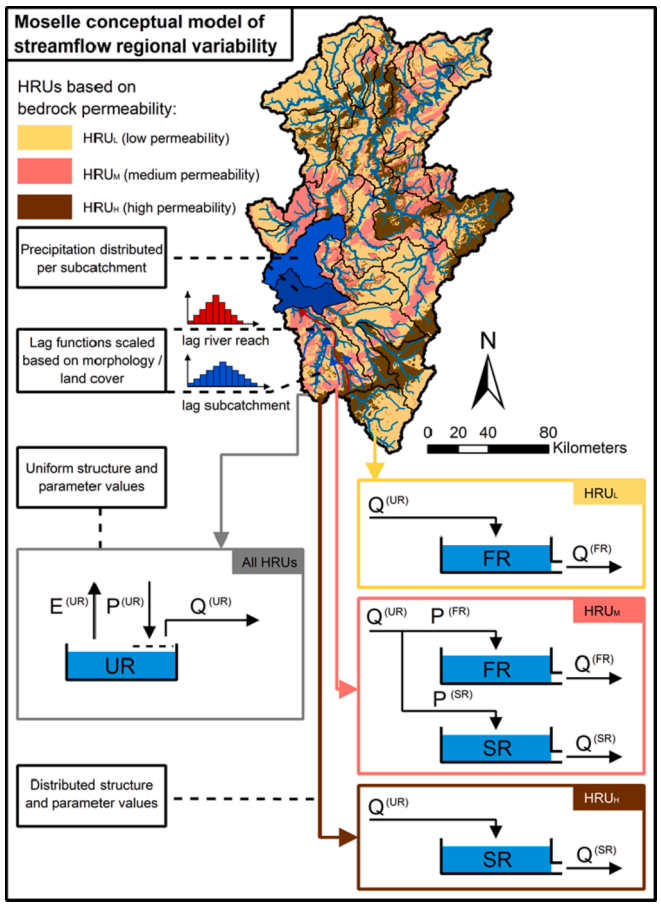

Fig. 2. Conceptual diagram of the model M(HRU_3,Lag_2), which is mapped from the perceptual model in FM2022.

subcatchment boundaries does not appear to be significant. These insights are here used to build a model comparison experiment that isolates the effect of these key hypotheses.

The model comparison experiment is in line with a top-down approach, starting from a relatively simple benchmark model, and introducing model modifications aimed at a more spatially explicit characterization of the catchment. The set of models includes 5 model variants, which are listed in Table 1. These models are labeled using the notation M(HRU_i,Lag_j), where M stands for model, HRU_i ($i = 1,2,3$) to a particular classification in hydrological response units, and Lag_j ($j = 0,1,2$) to a specific representation of the flow routing. In synthesis, the benchmark model is represented by M(HRU_1,Lag_0), which uses a single HRU, and no flow routing elements. This model has uniform parameters in space, therefore it is unable to characterize landscape heterogeneity, but has distributed states, and therefore it is able to characterize climate variability, and can produce distributed simulations. This model is progressively evolved, leading to M(HRU_1,Lag_1), which introduces a simple routing approach, then to M(HRU_2,Lag_1) which adds a landscape discretization into 2 HRUs, then to M(HRU_2,Lag_2), which introduces a more complex routing approach, and finally to the target model M (HRU_3,Lag_2), which considers 3 HRUs. In the following, we start by describing the most spatially resolved model M(HRU_3,Lag_2) (Section 2.2.1), and then introduce the simpler 4 model variants (Section 2.2.2). For ease of presentation, the model description focuses mainly on the concepts, relegating to the Appendix the mathematical formulations.

The study area and data have already been described in FM2022, and therefore only the relevant details for the present application are mentioned. All models use as forcing data the time series of precipitation, potential evaporation and temperature, and the streamflow is measured at 26 interior catchment points. Both the forcing data and the streamflow are considered at daily resolution. Maps of relevant catchment features as well their classification are the same as in FM2022. The calibration–validation approach is described in Section 2.3.2.

2.2.1. Target model

2.2.1.1. Model structure. This section describes the target model M (HRU_3,Lag_2), and its relationship to the perceptual model developed in FM2022, which was briefly summarized in Section 2.2 above. The model schematic is shown in Fig. 2, and model equations are detailed in the Appendix.

The minimum number of landscape elements that our distributed model requires is determined by the location of the points along the river network where streamflow simulations are needed. In the present case, this partitioning is based on the location of the 26 gauging stations where streamflow is observed (Fig. 2). For clarity, we define "total" subcatchment the total area drained by given point on the river network, and "incremental" subcatchment the incremental area from an upstream point on the river network. Hence, we divide the catchment into 26 incremental subcatchments.

The forcing data, hence precipitation, potential evaporation and temperature, are distributed per incremental subcatchment (Fig. 2). This decision is motivated by the perceptual model, which indicated that such discretization is sufficient to describe the long term water balance.

The landscape is further discretized into HRUs based on lithology. In particular, 3 HRUs are considered, representing the low (HRU_L), medium (HRU_M) and high (HRU_H) bedrock permeability classes. The basis for this decision is that the perceptual model indicated that this landscape discretization is necessary to account for the variability of the baseflow index signature.

Next, we assign different structures for each HRU, intended to reflect the associated dominant processes. We assume that lumped conceptual model structures considered adequate for precipitation-streamflow modeling would be rather inclusive for representing a given HRU. We therefore start with such a structure, and simplify it in order to account only for the processes that are perceived to be dominant in each HRU. As an inclusive model structure, we consider a 3 elements structure composed by three reservoirs: UR (unsaturated zone reservoir), intended to control the partitioning between precipitation and runoff, FR (fast reacting reservoir), which represent the quickflow generating processes, and SR (slow reacting reservoir), intended to represent the baseflow generating processes. An analog three elements structure was indicated by Jakeman and Hornberger (1993) as the "most commonly identified configuration" for streamflow simulation (see also Young, 2003). We then particularize this structure in order to describe HRU-specific dominant processes. In particular, the perceptual model suggests that as bedrock permeability goes from low to high, the subsurface processes progressively shift from shallow to deep, therefore triggering different hydrograph responses. We here assume that low bedrock permeability predominantly triggers a fast hydrograph response, and we therefore only consider the FR reservoir and we omit the SR reservoir. Medium bedrock permeability triggers both a fast and a slow response, and we therefore keep both the FR and SR reservoirs. High bedrock permeability predominantly triggers a slow hydrograph response, and we therefore keep the SR reservoir and omit the FR reservoir.

Although snow related processes do not have a major influence on streamflow generation, they are nonetheless occasionally present (see FM2022), and therefore we account for them. For this reason, each HRU model also includes a snow reservoir (WR) based on the degree day method, which accounts for snow accumulation and melting. This specific choice is also maintained for the other model variants.

Streamflow generated within each incremental subcatchment is obtained by first adding together the outflows from the individual HRU models, and then by propagating the resulting flow through a lag

function, intended to represent the incremental subcatchment routing (CL) (lag functions marked in blue color in Fig. 2). Streamflow at each total subcatchment needs to account also for the streamflow from upstream subcatchments. Incoming streamflow from each upstream subcatchment is first offset through a lag function, intended to represent the river routing (RL) through the incremental subcatchment (lag functions marked in red color in Fig. 2), and then added to the incremental subcatchment streamflow. Regional groundwater flow between subcatchments, as suggested by the perceptual model, is not considered, and the HRUs act in parallel.

The total number of landscape elements that need to be separately modeled is obtained by summing the number of HRUs present in each incremental subcatchment, as HRUs in different subcatchments need to be modeled separately, given that they receive different forcings. There are 26 subcatchments and 3 HRUs, however not all subcatchments contain all HRUs, resulting in a total of 75 landscape elements (Table 1).

2.2.1.2. Parameters constraints. The perceptual model can be further exploited to constrain the parameter space by providing parameters "regularization" relationships (e.g. Pokhrel et al., 2008). These constraints are described hereafter, and are also indicated in Fig. 2.

- HRU model structure parameters. We assume that some parameters associated to distinct HRU model structures are related to each other. In particular, the perceptual model suggests a similar behavior between HRUs in terms of partitioning precipitation between evaporation and runoff, which suggests that the parameters of the UR element can be kept common between HRU models (Fig. 2). Also the parameters of the snow reservoir (WR) are considered as spatially uniform.
- Subcatchment routing parameters. We assume that the parameters associated to the incremental subcatchment routing (CL) are a function of selected landscape properties and two global calibration parameters. Based on the analysis on time-to-peak variability in the perceptual model, we assume that the time parameter shaping the lag function in each incremental subcatchment increases with geometric distance, and decreases with the fraction of developed land use, and the 5% quantile of the slope (see Appendix).
- River routing parameters. We assume that the parameters associated to the river routing from upstream total subcatchments (RL) are a function of landscape properties and a global calibration parameter. Based on the analysis on the relative lag in the perceptual model, we assume that the time parameter shaping the lag function associated to river routing increases with the drainage distance, and decreases with the upstream contributing area (see Appendix).

Such parameter relationships strongly reduce the number of calibration parameters of the distributed model. Specifically, the model M(HRU_3,Lag_2) has 12 calibration parameters (Table 1 and Appendix), which are hereafter summarized, following the notation of indicating in the superscript the model element to which the parameter is associated: the degree-day snow parameter ($k^{(WR)}$), which is common to all spatial elements, 3 parameters characterizing the UR element ($c_E^{(UR)}$, $S_{Max}^{(UR)}$, $\beta^{(UR)}$), which are also common to all spatial elements, 1 parameter specific for the high permeability HRU_H ($k^{(SR,HRU_H)}$), characterizing the release rate of the SR element, 3 parameters for the medium permeability HRU_M ($D^{(HRU_M)}$, $k^{(FR,HRU_M)}$, $k^{(SR,HRU_M)}$), representing the split between FR and SR and the release rate of the 2 reservoirs, 1 parameter for the low permeability HRU_L ($k^{(FR,HRU_L)}$), characterizing the release rate of FR, 2 parameters ($a_1^{(CL)}$, $a_2^{(CL)}$) that scale the CL lag functions, and 1 parameter ($a^{(RL)}$) that scales the RL lag functions (see Appendix).

2.2.2. Model variants for hypothesis testing

The target model described above embeds several hypotheses about system characteristics that need to be represented in order to capture distributed streamflow responses. Although these hypotheses are motivated by the perceptual model, the translation from perceptual to conceptual model is not obvious. In order to provide a stronger basis to justify these hypotheses and ensure that they are supported by the data, 4 additional model structures are introduced. These model structures follow a path from simple to complex, in the spirit of a top-down approach, starting from a benchmark model intended to represent a null hypothesis about process controls on streamflow responses, and which is progressively evolved, in order to lead to the target model described above. The 4 model structures are described below, and summarized in Table 1.

- M(HRU_1,Lag_0) uses of a single HRU and no lag functions. Specifically, this model has subcatchment parameters that are uniform in space, hence it is unable to account for the heterogeneity in landscape attributes. This model is still distributed, as it has distributed states (per incremental subcatchment), hence, it is able to account for the spatial heterogeneity of climate. The lag functions characterizing the routing within the incremental subcatchments and from upstream subcatchments are also omitted. Hence, this model is unable to account for the associated lagging and dampening effects. As HRU structure, the full 3 reservoirs structures is used (previously used for HRU_M in M(HRU_3,Lag_2)). This model has 7 calibration parameters, associated to the individual HRU model structure: the degree-day snow parameter ($k^{(WR)}$), 3 parameters associated to UR ($c_E^{(UR)}$, $S_{Max}^{(UR)}$, $\beta^{(UR)}$), the split parameter (D), the parameters associated to FR ($k^{(FR)}$) and SR ($k^{(SR)}$). This model has 26 landscape elements, which are determined solely by the number of subcatchments (Table 1).
- M(HRU_1,Lag_1) introduces flow routing elements. Compared to the target model M(HRU_3,Lag_2), the routing is represented by the same lag functions, but the regularization relationships of these lag functions use less parameters and less landscape attributes. In particular, both the subcatchment routing elements (CL) and the river routing elements (RL) are scaled only according to drainage distance, each of them using a single calibration parameter (see Appendix). The use of drainage distance alone can be considered one of the simplest approaches to parameterize routing effects, as used for example by Lerat et al. (2012). This model has 9 calibration parameters, which are the same as for M(HRU_1,Lag_0) with the addition of a parameter for CL ($a^{(CL)}$) and a parameter for RL ($a^{(RL)}$), and 26 landscape elements (Table 1).
- M(HRU_2,Lag_1) introduces a partitioning of the landscape into 2 HRUs. Compared to M(HRU_3,Lag_2), this model maintains the high permeability HRU_H, and combines the other HRUs into a medium–low bedrock permeability HRU_{ML}. Both HRUs use the same 3 reservoirs structure used in M(HRU_1,Lag_0) and M(HRU_1,Lag_1). Hence this model does not tailor the model structures to the specific HRUs like M(HRU_3,Lag_2). Similarly to M(HRU_3,Lag_2), the parameters of the UR reservoir are kept uniform in space, whereas the split parameter and the parameters associated to FR and SR are distributed. This model has 12 parameters, which are the degree-day snow parameter ($k^{(WR)}$), 3 parameters associated to UR ($c_E^{(UR)}$, $S_{Max}^{(UR)}$, $\beta^{(UR)}$), which are common to both HRUs, 3 parameters for the high permeability HRU_H ($D^{(HRU_H)}$, $k^{(FR,HRU_H)}$, $k^{(SR,HRU_H)}$), characterizing the split, the FR and the SR reservoir respectively, the corresponding 3 parameters specific for the medium–low permeability HRU_{ML} ($D^{(HRU_{ML})}$, $k^{(FR,HRU_{ML})}$, $k^{(SR,HRU_{ML})}$), a parameter for CL ($a^{(CL)}$) and a parameter for RL ($a^{(RL)}$). As each of the 26 subcatchments contain both HRUs, this model uses 52 landscape elements (Table 1).
- M(HRU_2,Lag_2) introduces a more complex parameterization of the routing elements. In particular, this model uses the same regularization relationships for the lag functions as M(HRU_3,Lag_2), thus

involving more landscape attributes and one additional parameter associated to the RL lag functions ($a_1^{(CL)}$, $a_2^{(CL)}$). This model has 13 parameters, one parameter more than M(HRU_2,Lag_1) due to the additional parameters associated to the CL lag functions. Like M (HRU_2,Lag_1), this model has 52 landscape elements (Table 1).

The last step of this chain is represented by the target model M (HRU_3,Lag_2), already described in Section 2.2.1, which differs from M (HRU_2,Lag_2) for the addition of an HRU, and for the tailoring of the HRU model structures according to the perceived dominant process. We note that M(HRU_3,Lag_2) has 1 parameter less than M(HRU_2,Lag_2) in spite of the higher spatial resolution, due to the tailoring of the model structures to the HRU dominant processes.

Based on the premises of the perceptual model that informs the conceptual model decisions, the 5 model structures are associated to specific hypotheses about their relative merits. These hypotheses are as follows:

1. M(HRU_1,Lag_0), in spite of its simplicity, should be able to capture the long term water balance.
2. M(HRU_1,Lag_1) should enable the characterization of streamflow lag times.
3. M(HRU_2,Lag_1) should enable the characterization of baseflow variability.
4. M(HRU_2,Lag_2) should further improve the characterization of streamflow lag times.
5. M(HRU_3,Lag_2) should further improve the characterization of baseflow variability.

Hence the differences between the 5 model variants have an anticipated effect on an individual signature, and not on others. Correspondingly, the model comparison is expected to highlight influence factors on the selected streamflow signatures as well as their relative independence.

2.3. Model evaluation

This section describes the model evaluation approach. Section 2.3.1 specifies the objective function used for model calibration. Section 2.3.2 describes the calibration and space–time validation approach. Section 2.3.3 illustrates the assessment of model performance both in the time and in the signature domain.

Although a specific uncertainty analysis is not carried out, the model comparison provides a basis for a relative assessment of model uncertainties, which we consider sufficient for our purpose. In terms of parameters uncertainty, we rely on the assumption, typical of the top down approach, that additional complexity is justified by the data if it corresponds to an improvement in model simulations (Sivapalan et al., 2003). We also interpret significant differences between calibration and space–time validation performance as an indication of overfitting. In terms of uncertainties associated to model simulation, we rely on the consideration that an improvement in the selected likelihood function (Section 2.3.2) corresponds to tighter uncertainty bands.

The model comparison severs multiple related purposes, and in particular (1) to clarify the effect of model decisions, which is obtained by comparing the performance of pairs of models that differ in a controlled way, (2) to guide the ultimate model selection, which is approached by assessing whether the progression of models leads to an improvement in the model evaluation metrics, and (3) to evaluate the hypotheses of the perceptual model, as the model variants correspond to specific hypotheses of how dominant processes affect streamflow regional variability.

2.3.1. Likelihood function for model calibration

Semi-distributed models for streamflow simulation have been calibrated using different strategies (e.g. Wallner et al., 2012). The simplest strategy is the single-site approach, which consists in calibrating the models using data at an individual gauge, similarly to what is commonly done with lumped models (e.g. Gao et al., 2014; Gharari et al., 2014; Nijzink et al., 2016). This approach has the limitation that an individual time series may not contain sufficient information for the parameter identification of a distributed model, such as to disentangle the behavior of different HRUs. Hence, it may lead to poor parameter identifiability. An alternative calibration strategy is the sequential approach, which consists in calibrating the model from upstream to downstream, in a sequential manner (e.g. Ajami et al., 2004; Feyen et al., 2008; Lerat et al., 2012; de Lavenne et al., 2019). This approach, while easy to apply, has the disadvantage that it leads to many parameter sets (one set for each subcatchment), hence it makes the model on the one hand highly parameterized, and on the other hand highly dependent on calibration, and therefore difficult to regionalize. Finally, the multi-site approach consists in calibrating the model simultaneously at a number of stations (e.g. Zhang et al., 2010; Fenicia et al., 2016; Dal Molin et al., 2020). This approach overcomes the limitations of the previous approaches, and is adopted in the present study.

In order to enable model calibration, it is necessary to construct an objective function, which quantifies the goodness of fit of model predictions. This objective function is here derived from a Bayesian inference approach, where we assumed non-informative "flat" priors for model parameters, and we assumed that the residuals errors of transformed streamflow are statistically independent and can be approximated by a zero-mean Gaussian distribution:

$$\left(Q_{i,t}^{(\mathrm{Obs})}\right)^{\lambda} - \left(Q_{i,t}^{(\mathrm{Sim})}\right)^{\lambda} = N(0,\sigma) \tag{1}$$

where i indicates the subcatchment and varies from 1 to the number of selected subcatchments N_C, t indicates the time index, and varies from 1 to the number of data points $N_{T,i}$ in the subcatchment i, $Q^{(\mathrm{Sim})}$ and $Q^{(\mathrm{Obs})}$ are simulated and observed streamflow respectively, $N(0,\sigma)$ is the Gaussian distribution with zero mean and standard deviation σ, and the power λ is fixed at 0.5, implying a square root transformation, as recommended by McInerney et al. (2017).

Maximising the posterior parameter distributions under the assumption of Equation corresponds to maximising the Nash Sutcliffe efficiency of the square root of the streamflow:

$$F_{\mathrm{NS}} = 1 - \frac{\sum_{i=1}^{N_C}\sum_{t=1}^{N_{T,i}}\left(\sqrt{Q}_{i,t}^{(\mathrm{Sim})} - \sqrt{Q}_{i,t}^{(\mathrm{Obs})}\right)^2}{\sum_{i=1}^{N_C}\sum_{t=1}^{N_{T,i}}\left(\sqrt{Q}_{i,t}^{(\mathrm{Obs})} - ave\left(\sqrt{Q}_{1:N_C,1:N_T}^{(\mathrm{Obs})}\right)\right)^2} \tag{2}$$

where *ave* indicates the average, which is extended to the streamflow of all subcatchments.

2.3.2. Calibration and space–time validation strategy

The model evaluation strategy is motivated by testing the quality of model simulations in space–time validation, meaning using a time period and a set of subcatchments that the model does not see during calibration (e.g. Fenicia et al., 2016; Dal Molin et al., 2020). Using the words of Klemeš (1986), this validation may be called a proxy-basins split-sample test. For this purpose, the available data is organized as follows:

- The time period is partitioned into two 13-year periods: period 1, going from 1.09.1989 to 31.08.2002, and period 2 going from 1.09.2002 to 31.08.2015.
- The subcatchments are partitioned into two groups of 13 subcatchments each: group A and group B. The partition is made by sorting subcatchments by total area, and placing every second subcatchment in each group.

The organization of the data in two periods and two groups of

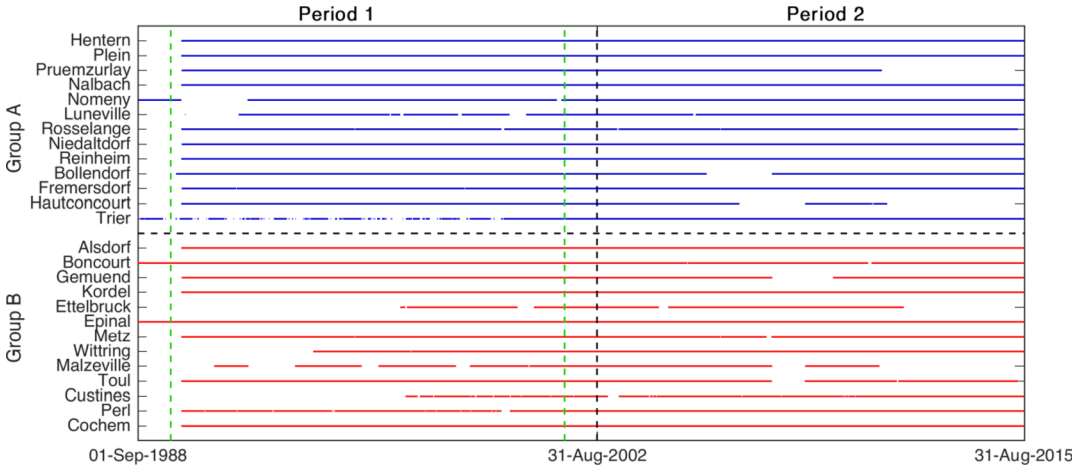

Fig. 3. Calibration-validation strategy. Each model is calibrated in each period and each group of subcatchments (i.e. in each of the 4 quadrants) and validated in the other period and group of subcatchments (i.e. in the opposite quadrant along the diagonal). Each calibration period is preceded by 1 year warm-up (delimitated by the green dashed line). The horizontal lines indicate data availability in each of the subcatchments (blue and red indicate the two groups), and gaps in the lines indicate missing values.

subcatchments is presented in Fig. 3, which also shows the data availability for each subcatchment (represented by the corresponding gauge name). Missing data are treated by considering in the objective function (Eq. (2)) only time indices where data is available. Fig. 1 and Table 1 in FM2022 shows the spatial location of the subcatchments and list their areal extension.

The model evaluation strategy proceeds as follows:

1. We calibrate the models on one group of subcatchments over a given time period, and validate it on the other group of subcatchments over the other time period, hence we perform space–time validation. With reference to Fig. 3, once a quadrant is chosen for calibration, moving horizontally on another quadrant means performing time validation, moving vertically means performing space validation, and moving diagonally means performing space–time validation. For example, if a model is calibrated on Period 1 – Group A, then Period 2 – Group A represents time validation, Period 1 – Group B represents space validation, and Period 2 – Group B represents space–time validation.
2. This process is repeated for all four combinations of calibration periods and subcatchment groups.
3. The streamflow time series from the four space–time validation scenarios are then concatenated, thus forming a single set of "predicted" streamflow time series spanning the entire observation period at each subcatchment. An analogous set of "concatenated" time series of "calibrated" streamflow is also constructed.

The results focus on comparing model performance in calibration (the least challenging scenario) and space–time validation (the most challenging scenario). As the two scenarios use the same underlying data, when moving from calibration to space–time validation, at least in terms of the calibration objective function, model performance can only degrade. The degree of this degradation may give an indication on whether any of the models is prone to overfitting the data.

2.3.3. *Performance metrics of streamflow simulations*

Model performance is analyzed both in the time domain and in the signature domain. For model evaluation in the time domain, we use the F_{NS} metric already defined in Equation (2), as it corresponds to the objective function used for model calibration.

For model evaluation in the signature domain, the following three streamflow signatures evaluation metrics are used, which are based on the analysis of streamflow spatial variability performed in FM2022:

Streamflow average correlation: The streamflow average at a given subcatchment is defined by:

$$\overline{Q}_k = \frac{1}{N_{\mathrm{T},k}} \sum_{t=1}^{N_{\mathrm{T},k}} Q_{k,t} \tag{3}$$

where Q is the streamflow, $k = 1 \ldots N_{\mathrm{C}}$ is the subcatchment index, with N_{C} representing the number of subcatchments, t is the time index, $N_{\mathrm{T},k}$ is the number of observations at the subcatchment k, and the overbar indicates the average over the observation period.

To assess model ability to simulate this signature, we consider the Pearson correlation r between observed and simulated streamflow average at the 26 subcatchments:

$$F_{\mathrm{Qavg}} = r\left(\overline{\boldsymbol{Q}}^{(\mathrm{Obs})}, \overline{\boldsymbol{Q}}^{(\mathrm{Sim})}\right) \tag{4}$$

We choose the correlation to quantify the alignment between the observed and predictive values, rather than the agreement of the absolute values, and we assess visually the agreement along the diagonal. We opt for using Pearson rather than Spearman correlation (as used in FM2022), as the agreement between observed and simulated quantities should ideally be linear.

Baseflow index correlation: The baseflow index is defined as:

$$Q_{\mathrm{BFI},k} = \sum_{t=1}^{N_{\mathrm{T},k}} Q_{k,t}^{(\mathrm{b})} \Big/ \sum_{t=1}^{N_{\mathrm{T},k}} Q_{k,t} \tag{5}$$

where $Q_{k,t}^{(\mathrm{b})}$ is calculated with the filter proposed by Lyne and Hollick (1979), using the same settings as in FM2022.

To assess model ability to simulate this signature, we consider the correlation r between observed and simulated baseflow index at the 26 subcatchments:

$$F_{\mathrm{Qbfi}} = r\left(\boldsymbol{Q}_{\mathrm{BFI}}^{(\mathrm{Obs})}, \boldsymbol{Q}_{\mathrm{BFI}}^{(\mathrm{Sim})}\right) \tag{6}$$

Also here, we chose the correlation to quantify the alignment between the observed and predictive values, rather than the agreement of the absolute values.

Hydrograph relative lag: In order to estimate the lag between observed and simulated hydrographs, we proceeded as follows:

- We calculate the Pearson cross-correlations $\boldsymbol{R}_{k,-N_{\mathrm{L}}:N_{\mathrm{L}}}$ between the observed and simulated hydrographs at each subcatchment k, shifting the observed hydrograph for lags $\pm N_{\mathrm{L}}$ of up to $\pm$ 20 d with respect to the simulated hydrograph.

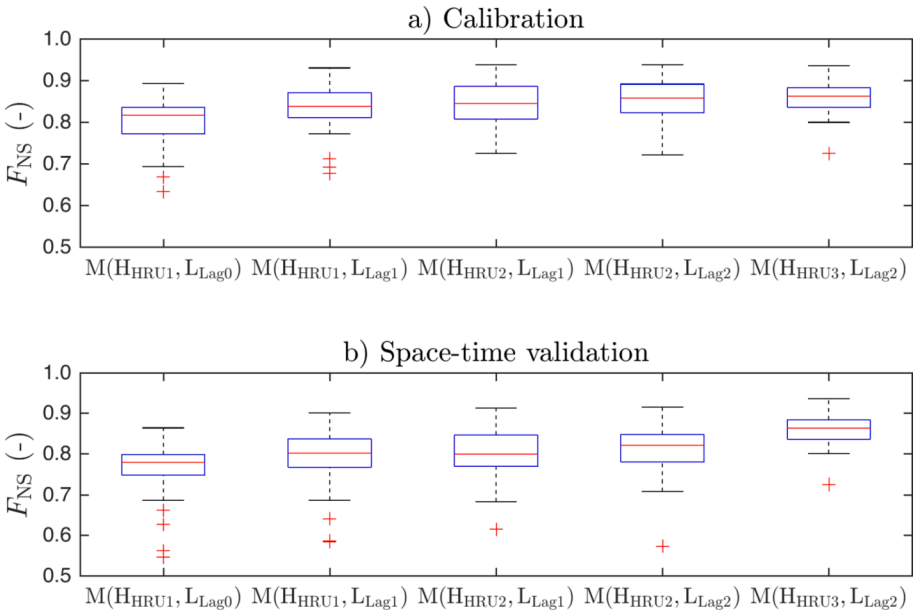

Fig. 4. Nash-Sutcliffe efficiency (F_{NS}) in calibration (panel a) and space–time validation (panel b). The x-axis indicates the 5 model variants being compared. F_{NS} is calculated for each individual subcatchment, and reported as box plots.

- We consider the lag $i_{L,k}$ that results in the maximum Pearson cross correlation:

$$Q_{Lag,k} = i_{L,k} \text{ such that } max\left(\boldsymbol{R}_{k,-N_L:N_L}\right) = R_{k,i_L} \tag{7}$$

The relative lag $Q_{Lag,k}$ should be zero if the lag is correctly simulated by the model (a positive relative lag would mean that the simulated hydrograph overestimates the lag).

We then simply sum the absolute values of the hydrograph relative lag for all the subcatchments:

$$F_{Qlag} = \sum_{k=1}^{N_C} \left| Q_{Lag,k} \right| \tag{8}$$

which is an integer equal or larger than zero, with lower values indicating better performance (i.e. a value of zero indicates that the lag is correctly simulated at all subcatchments).

3. Results

3.1. Model performance in the time domain

Fig. 4 shows box plots representing the variability of F_{NS} in the subcatchments under the calibration and space–time validation scenarios.

The following results can be observed:

1. In calibration, the averages for the 26 subcatchments of the F_{NS} for the models M(HRU$_1$,Lag$_0$), M(HRU$_1$,Lag$_1$), M(HRU$_2$,Lag$_1$), M(HRU$_2$, Lag$_2$) and M(HRU$_3$,Lag$_2$) were 0.80, 0.83, 0.85, 0.85, 0.86 respectively. Hence, there was a general increase in performance when

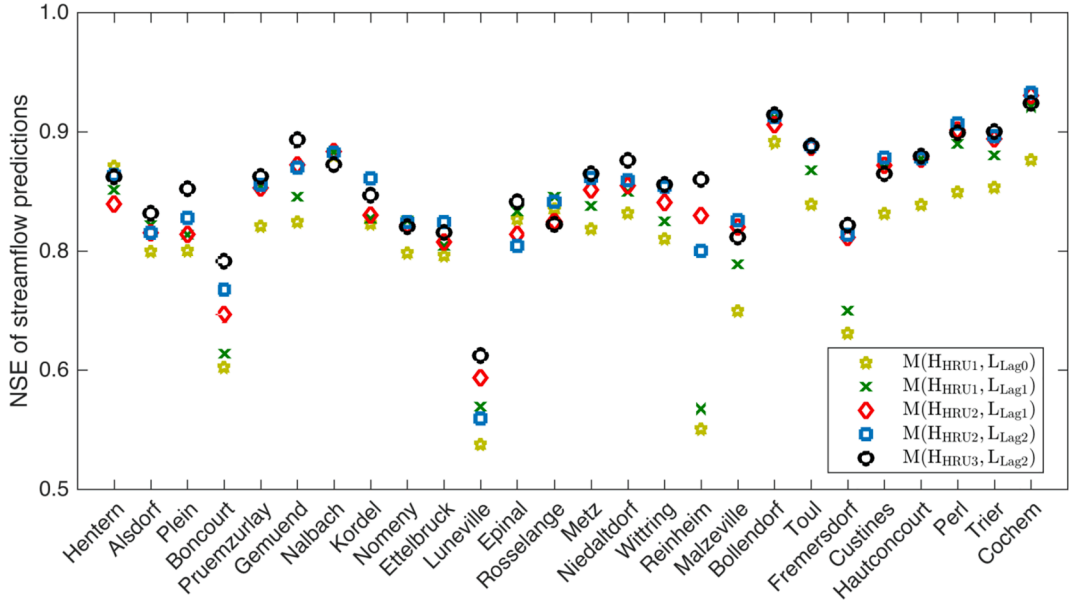

Fig. 5. Nash Sutcliffe efficiency (NSE) of streamflow predictions (in square root transformed space) for each individual subcatchment (sorted from smallest to largest) in space–time validation for the 5 model variants. The improvement in NSE is larger for the worse performing subcatchments.

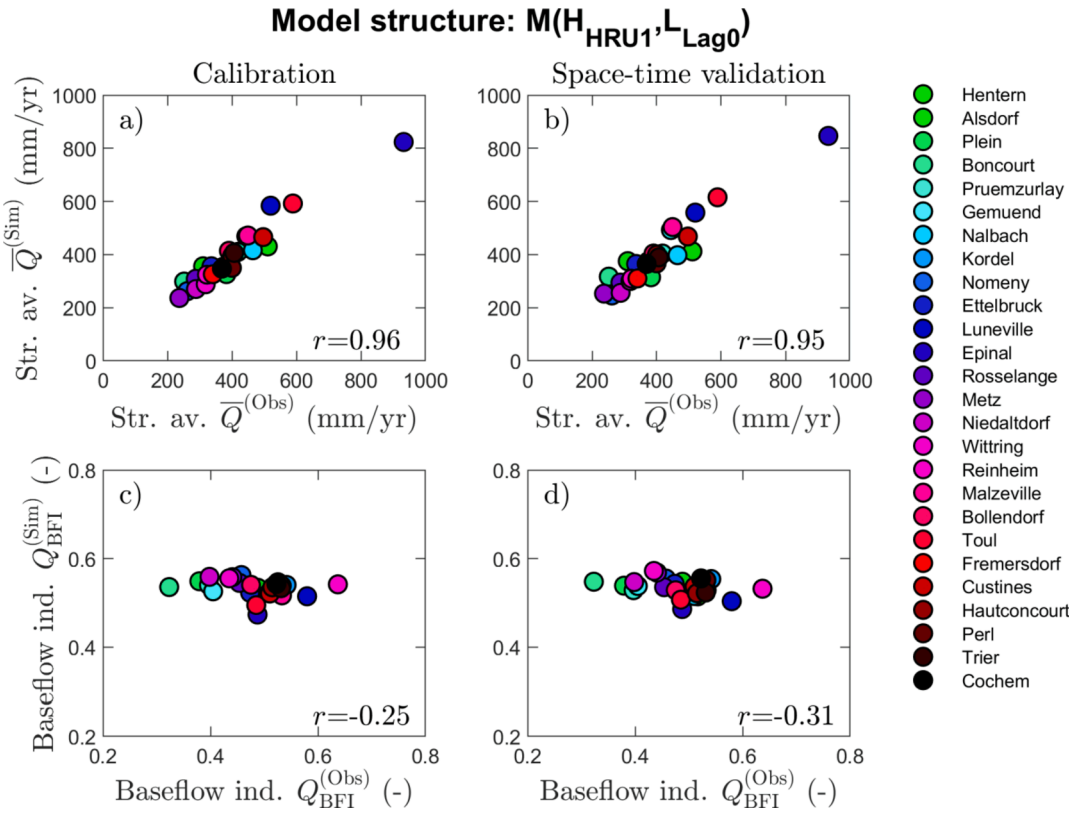

Fig. 6. Observed and modeled streamflow average (upper row) and baseflow index signatures (lower row) for M(HRU$_1$,Lag$_0$), in calibration (left column), and space–time validation (right column). The model captures the streamflow average signature, but does not capture the baseflow index signature.

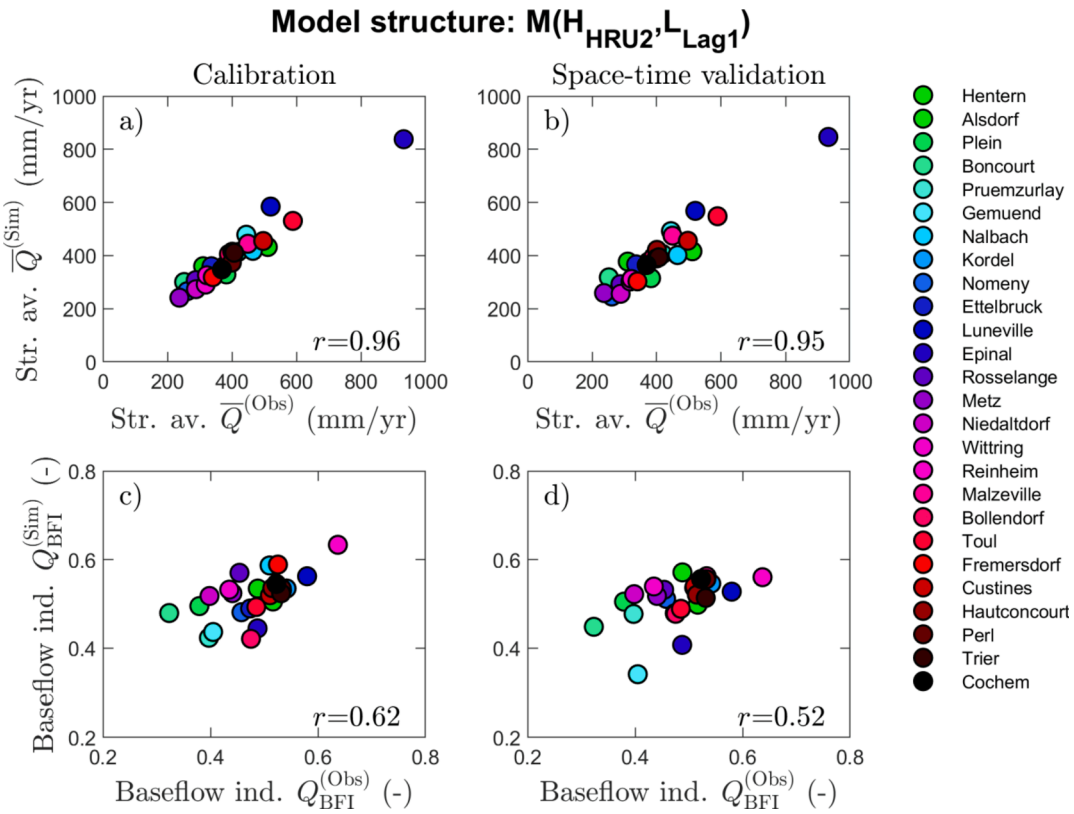

Fig. 7. Observed and modeled streamflow average (upper row) and baseflow index signatures (lower row) fir M(HRU$_2$,Lag$_1$), in calibration (left column), and space–time validation (right column). Compared to the single HRU models, there is a noticeable improvement in simulating the baseflow index.

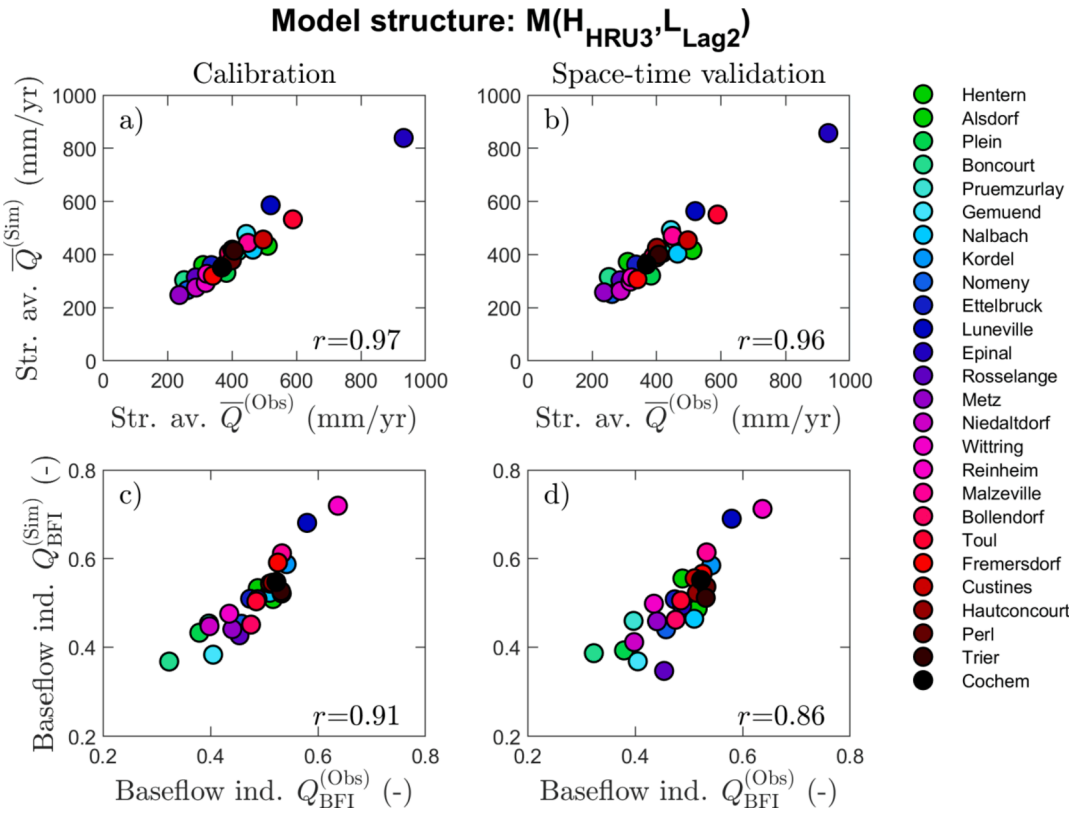

Fig. 8. Observed and modeled streamflow average (upper row) and baseflow index signatures (lower row) for M(HRU_3,Lag_2), in calibration (left column), and space–time validation (right column). Compared to the 2-HRUs models, there is a further significant improvement in simulating the baseflow index.

increasing the number of HRUs of the more complex models. The corresponding standard deviations of the F_{NS} values were $6.4{\cdot}10^{-2}$, $6.5{\cdot}10^{-2}$, $5.3{\cdot}10^{-2}$, $5.0{\cdot}10^{-2}$ and $4.5{\cdot}10^{-2}$, showing a decreasing tendency, which indicated a more stable model performance across subcatchments for the more complex models.

2. In space–time validation, the ranking was generally preserved, however with lower average and larger spread for all models. In particular, the F_{NS} average values were 0.76, 0.76, 0.79, 0.80, 0.81 and 0.82, while the corresponding standard deviations were $8.1{\cdot}10^{-2}$, $8.1{\cdot}10^{-2}$, $6.2{\cdot}10^{-2}$, $6.8{\cdot}10^{-2}$ and $5.5{\cdot}10^{-2}$.
3. The increments in performance between the models in terms of F_{NS} appeared minor. For example, the difference between the average F_{NS} values of the two HRU model M(HRU_2,Lag_2) and the 3 HRUs model M(HRU_3,Lag_2) was just 0.01.
4. The best performing model, hence M(HRU_3,Lag_2), in space–time validation, had F_{NS} values comprised between 0.64 (at Luneville, one of the smaller subcatchments) and 0.91 (at Cochem, the catchment outlet).

Hence, overall, the analysis in terms of F_{NS} indicated that the modeling progression led to an improvement of performance, as hypothesized in Section 2.2.2. However, based on this analysis alone, the differences in model performance did not appear very significant.

Fig. 5 shows the F_{NS} values for each individual subcatchment and model variant in space–time validation, thereby complementing the earlier assessment based on aggregated results. Differences in F_{NS} values between models were clearly catchment dependent. As a general tendency, subcatchments characterized by poorer performances when using the simpler model variants (e.g. Boncourt, Luneville, Reinheim and Fremersdorf) were the ones that experienced most of the improvement, while subcatchments that already presented a relatively high performance were relatively stable. Hence, the model improvements generally had an effect on fixing the outliers, rather than producing a uniform improvement for all subcatchments. This effect explains why on average, the improvement in F_{NS} values was rather minor. In order to better assess model differences and understand the associated causes, it is important to analyze the streamflow signatures, which provide the basis of a diagnostic approach to model evaluation, as shown in the subsequent section.

3.2. Model performance in the signature domain

In terms of streamflow signatures, we start by reporting the models' ability to simulate $\overline{\boldsymbol{Q}}^{(Obs)}$ and $\boldsymbol{Q}_{BFI}^{(Obs)}$. We explicitly report the simulations of models M(HRU_1,Lag_0), M(HRU_2,Lag_1) and M(HRU_3,Lag_2) in Fig. 6, Fig. 7, and Fig. 8 respectively, hence the initial model of the chain, and the subsequent models that showed the highest improvement in at least one of the signatures. This comparison is summarized in the following results:

1. All 5 models demonstrated a similar and very good ability to reproduce $\overline{\boldsymbol{Q}}^{(Obs)}$, with $F_{Qavg} > 0.95$ both in calibration and in space–time validation, and a good agreement along the diagonal line. This result implies that the differences in the structures of the 5 models did not have an effect on this signature, and that M(HRU_1, Lag_0), the simplest model, based on a single HRU, already contained the ability to simulate this signature.
2. The models differed significantly in their ability to reproduce $\boldsymbol{Q}_{BFI}^{(Obs)}$. In particular, M($HRU_1$,$Lag_0$) and M($HRU_1$,$Lag_1$) were unable to simulate differences between subcatchments ($F_{Qbfi} < 0$ in both cases), M(HRU_2,Lag_1) and M(HRU_2,Lag_2) provided a better match ($F_{Qbfi} = 0.52$ and 0.48 respectively in space–time validation), and M (HRU_3,Lag_2) significantly improved the fit ($F_{Qbfi} = 0.86$ in space–time validation, and a good agreement along the diagonal).
3. The performance of the models was lower in space–time validation than in calibration, however it decreased only slightly. In particular,

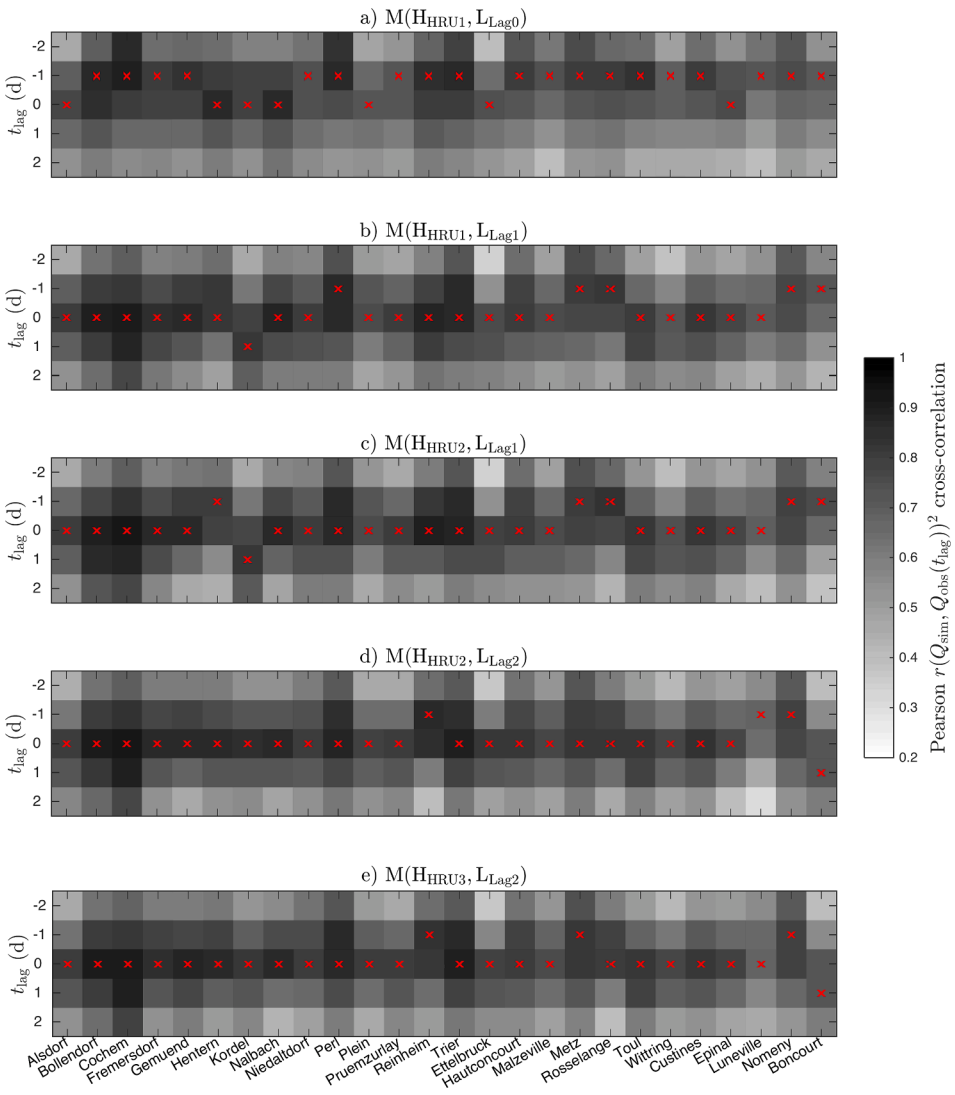

Fig. 9. Comparison of the model variants in their ability to simulate hydrograph lags in space–time validation. M(HRU_1,Lag_0) fails to correctly simulate the lag in 19 out of 26 subcatchments, as no routing elements are implemented. M (HRU_1,Lag_1) and M(HRU_2,Lag_1), which incorporate routing with a simple parameterization, significantly improve the performance, failing to reproduce the lag at 6 subcatchments. The more complex routing parameterization used in M (HRU_2,Lag_2) and M(HRU_3,Lag_2) further improves performance leading to 4 subcatchments where the lag is not correctly matched.

the performance of the models in terms of $\overline{\boldsymbol{Q}}^{(Obs)}$ remained more or less unchanged, whereas the models' ability to reproduce $\boldsymbol{Q}_{BFI}^{(Obs)}$ experienced a minor deterioration.

Fig. 9 illustrates the ability of the 5 models to correctly simulate hydrographs lags in space–time validation. Each column of a given panel shows the correlation between observed and simulated streamflow for different lags (t_{Lag}) of the simulated streamflow. The red cross indicates the lag for which the correlation was maximized. The cross should be at $t_{Lag} = 0$ if the lag is correctly simulated. If the red cross is at $t_{Lag} = -1$, it means that the simulated streamflow dynamics are anticipated of one time step (in this case one day) compared to the observed ones.

It can be observed that M(HRU_1,Lag_0), which has no lag functions, failed to correctly simulate the lag at 19 out of 26 subcatchments, where the lag was always anticipated of one day compared to the observations, leading to $F_{Qlag} = 19$ in space–time validation. M(HRU_1,Lag_1) and M (HRU_2,Lag_1), which use the simple parameterization of the lag function, led to a sensible improvement, failing to correctly simulate the lag at 6 of the subcatchments, with a lag of $\pm$ 1, and leading to $F_{Qlag} = 6$. M(HRU_2, Lag_2) and M(HRU_3,Lag_2), which adopt the more complex parameterization of the lag functions, further improved the representation of the lags, with $F_{Qlag} = 4$. It can be noted that models that share the same parameterization of the lag function were consistent both in the number of failures, which was exactly the same, and in the subcatchments where this failure appear, which was similar. The calibration performance was consistent with the performance in space–time validation with F_{Qlag} values of 16, 6, 6, 1 and 2, respectively for the 5 models.

Overall, the signatures analysis shows that model modifications had a targeted effect on an individual signature without affecting significantly the others. In particular, the distribution of the forcings data had an effect on capturing the long term streamflow average, the distribution of the landscape in HRUs had an effect on simulating the baseflow index, and the routing model and successive improvements predominantly affected the hydrograph lags.

By comparing the results in the signature domain from those in the time domain, it is apparent that the streamflow signatures provided a complementary perspective on the relative merits of the three models. In particular, they helped identify similarities and differences in model simulations, which were not immediately apparent with a metric of time series agreement such as the Nash Sutcliffe efficiency.

4. Discussion

4.1. Agreement of conceptual model results with perceptual model hypotheses

The distributed perceptual model developed in FM2022 has been employed to set-up a conceptual model comparison experiment, with

the purpose of clarifying the effect of model decisions, guiding model selection, and testing key hypotheses of how dominant processes affect streamflow regional variability.

Five model variants were compared based on the expectation that they would behave differently in their ability to simulate a selection of streamflow signatures, namely the streamflow average, the baseflow index, and the hydrograph lag. Our results indicated that (1) the benchmark model M(HRU_1,Lag_0), that accounted for the spatial variability of the forcings, but did not account for the spatial heterogeneity of the landscape, captured the streamflow average signatures as good as more complex models ($r = 0.95$ in space–time validation). This result indicates that climate rather than landscape affected streamflow average. (2) M(HRU_1,Lag_1) introduced routing elements, simply parameterized based on drainage distance. This modification greatly improved the simulation of hydrograph lags (leading from 19 to 6 subcatchments where the lag was wrongly simulated). This result suggests that such drainage distance acts as a primary control on hydrograph lags. (3) M(HRU_2,Lag_1), introduced a partitioning of the landscape into 2 HRUs, based on relative bedrock permeability, inferred from lithology. Such a partitioning enabled a characterization of the baseflow index signature (from $r < 0$ to $r = 0.52$ in space–time validation). This improvement suggests that lithology acts as a primary control on baseflow. (4) M(HRU_2,Lag_2) introduced a more complex parameterization of the routing elements, using additional topography and land use indicators, which further improved the characterization of lag times lags (reducing from 6 to 4 subcatchments where the lag was wrongly simulated). This outcome indicates the role of additional topographic and land use controls on lag times. (5) M(HRU_3,Lag_2) further discretized the landscape into 3 HRUs based on a finer resolution of relative bedrock permeability, and tailored the model structure to the perceived dominant processes in each HRU. This model improved the characterization of the baseflow index (from $r = 0.48$ to $r = 0.86$ in space–time validation), which confirms the control of lithology on baseflow.

In summary, the model comparison experiment confirms the key hypotheses of the perceptual model, indicating that (1) incorporating into the model structure information about spatial variability of precipitation, bedrock permeability, topography and land use was essential to capture the signatures of streamflow average baseflow index and hydrograph lag. Model structures that omitted this information had lower ability to reproduce these signatures; (2) the selected streamflow signatures responded to individual processes controls, as the model modifications that affected the processes associated to given signature, affected minimally the other signatures (notably the simulation of the streamflow average was essentially the same for all model variants).

It is interesting to note that the difference in simulation abilities of the 5 models was much more evident in the signatures domain than in the time domain. In terms of average Nash-Sutcliffe efficiency (F_{NS}), the ranking of performance in the time domain was consistent with the one in the signature domain, meaning that higher F_{NS} corresponded to better match of the streamflow signatures. However, the average F_{NS} of the 5 models varied within a very narrow range: 0.80 to 0.86 in calibration, and 0.76 to 0.82 in space–time validation, whereas the models' ability to match signatures went from very poor (near zero baseflow index correlations or lag wrong at the vast majority of the catchments) to relatively good (see above). Small differences in average F_{NS} are explained when considering that differences in F_{NS} values between models were clearly subcatchment dependent. Hence, model improvements generally had an effect on improving the performance of bad performing subcatchments, rather than producing a uniform improvement for all subcatchments. Improving the F_{NS} on a few subcatchments had a low impact on the overall average. Hence, one of the lessons learned from this model comparison was the importance of considering multiple model evaluation criteria and in particular hydrological signatures. Using signatures for model diagnostics has previously been encouraged in the hydrological literature (e.g. Sivapalan, 2006; Hrachowitz et al., 2014), but has rarely been taken up operationally. Our findings are a clear example of how a model evaluation metric such as the Nash Sutcliffe efficiency would be a poor indicator to detect differences in the models used, which are instead clear and significant in the signatures domain.

Although the model comparison experiment supported the hypotheses noted by the perceptual model, the "magnitude" of model improvements may not always appear consistent with the results of the analyses that underpinned the development of the perceptual model. In particular, in going from 2 to 3 HRUs the model improvement was relatively large (from Pearson $r = 0.48$ of M(HRU_2,Lag_2) to $r = 0.86$ of M(HRU_3,Lag_2) in space–time validation) compared to the improvement of using 2 to 3 HRUs when fitting the baseflow index in FM2022 (from Pearson $r = 0.87$ to $r = 0.90$). We explain this apparent inconsistency by noting that the models were calibrated to streamflow and not to the streamflow signatures, which were used for independent evaluation. It might therefore be that the simpler model was too simple in some other respects, which created a tradeoff in its ability to accommodate different objectives. The more complex model might reduce this tradeoff, and enable the model components to better comply with their intended function.

In terms of modeling hydrograph lags, we saw a big improvement when introducing such lag elements to a model that did not consider them (when going from M(HRU_1,Lag_0) to M(HRU_1,Lag_1), the subcatchments where the lag was wrongly simulated dropped from 19 to 6 in space–time validation). But we saw a small improvement when introducing a more complex lag parameterization (the drop was from 6 to 4). The models use two instruments to simulate such lags, being the routing that occurs within the incremental subcatchments, which is associated to the analysis on time-to-peak in FM2022, and the routing that occurs along the river stretches from upstream subcatchments, which is associated to the analysis on relative lags in FM2022. The analysis on relative lags of FM2022 did not show a large improvement in going from a simpler (used in M(HRU_i,Lag_1)) to a more complex formulation (used in M(HRU_i,Lag_2)), but the time-to-peak analysis of FM2022 showed a large improvement when going from a simpler to a more complex formulation (as shown in Table 4 in FM2022, the Spearman correlation went from 0.56 for the simpler formulation using only drainage distance, to 0.78 for the more complex formulation using 3 landscape attributes), which might have raised the expectation that the model improvement associated to a more complex parameterization of the lag functions was larger. This apparent disparity of results is explained considering that first, the current model application used daily data, whereas the analysis of the lags in FM2022 used hourly data, and for many subcatchments, both the time-to-peak and the relative lags assessed in FM2022 were sub-daily (see Figs. 5 and 9 in FM2022). Second, here the assessment of hydrograph lags was based on the overall timing agreement of observed and simulated time series, which combines all lags introduced by all model elements, hence all model reservoirs and lag functions collectively.

One might ask at this stage what the value of the conceptual model comparison experiment was beyond simply a confirmation of perceptual model hypothesis. First, a conceptual model can be predictive (e.g. it can predict streamflow time series in other time periods or different catchments), whereas a perceptual model is generally qualitative. Second, the conceptual model developed in this paper is an integrative model that has the ability to explain the combined relevant aspects of streamflow spatial variability. The individual regressions that were the basis of the perceptual models are a more "fragmented" representation of catchment behavior: each of them designed to mimic an individual aspect of streamflow spatial variability.

The model M(HRU_3,Lag_2) is the best performing model, and therefore the model of choice. However, it should not be interpreted as the end of the modeling process, but rather as a basis for further explorations. There are many possible refinements that can be tested and potentially incorporated, depending on the intended model objectives. These could include a more complex river routing module, or a more spatially refined representation of catchment properties. Our work has

followed the "top-down" philosophy of keeping the model as simple as possible for the task at hand. Pursuing the same philosophy, further model refinements should be tested based on what they add in terms of predictive abilities as well as requirements for the intended model use. In this spirit, a more complex processes representation should be warranted by an improved explanatory power, otherwise it would be unjustified by the data (e.g. Grayson et al., 1992).

4.2. Sources of uncertainty

Uncertainties in hydrological modeling arise from several sources, namely structure, parameters and observations, and propagate as uncertainties in model simulations (e.g. Liu and Gupta, 2007). There are several formal approaches that can be used to quantify such uncertainties (Beven and Binley, 1992; Kavetski et al., 2006; Montanari and Di Baldassarre, 2013), including recent approaches that enable the identification of individual model mechanisms in a Bayesian framework (Prieto et al., 2021). Such formal assessment was not conducted in this study as considered beyond scope. However, an argumentative assessment of these uncertainties can be provided on the basis of our analyses. In terms of model parameters, a general argument of the top-down approach is that additional complexity is justified by an improvement in model simulations (Sivapalan et al., 2003). As the progression through the 5 models resulted in an improvement in model simulations, the model modifications all appear to be supported by the available data and therefore are not expected to result into over-parameterization. Moreover, the performance of all models, both in the time and in the signature domain, degraded only little in space–time validation, which is also an indication that all models did not appear to over-fit the data. In terms of uncertainties in model simulations, given the correspondence between the objective function and a likelihood function associated to a streamflow error model (Section 2.3.1), an improvement in such objective function would also result in smaller uncertainties. Hence, the progression through the 5 models would gradually reduce the hydrograph uncertainty bands. In terms of model structure, the proposed controlled model comparison approach exposes model decisions to evaluation. Results confirm and reinforce the hypotheses of the perceptual model, which, as mentioned in FM2022, is also subjected to uncertainty.

Associated to the issue of uncertainty is issue of model transferability, hence, the extent to which the model can provide predictions outside calibration. An ideal objective of hydrological model building is to achieve model transferability in space, time and across scales. Our model development and evaluation approach suggests that the target model has such ability, albeit within a prescribed range. In particular, our space–time validation approach indicates that the model is able to extrapolate to different subcatchments within the same region, and different time periods than used in calibration, with minimal loss in performance. Moreover, the model bridges scales that vary within about two orders of magnitude in the 100 to the 10.000 km^2 range, hence from the smallest subcatchments to the total catchment area (see Table 1 in FM2022). However, extrapolations beyond the prescribed range, such as outside the study area or beyond the range of scales here considered, such as in headwater catchments, would not be granted and would require further analyses. Such analyses may be for example represented by repeating our model development approach to some of the smaller subcatchments with corresponding nested gauging stations, or to other catchments in a different location.

4.3. What do we learn moving to progressively larger scales?

Increasing scales reveal different influence factors on catchment response and associated dominant processes. This effect could be attributed the fact that, with increasing size, smaller scale variability tends to average out, and larger scale variability becomes progressively more visible and starts to play a role. For example, lithology or climatology is more uniform at the headwater scale, and becomes more heterogeneous at the regional scale, which makes their effects progressively more noticeable. Other properties may follow an opposite trend. For example, the fractions of different land uses may be variable at small scales, and tend to stabilize when upscaling to the regional scale. The effect of lithology on streamflow generation, in particular, is difficult to analyze when studying an individual catchment, where lithology is uniform. In a previous study on three headwater catchments in Luxembourg (Fenicia et al., 2014), the role of lithology on streamflow response became more visible though catchment comparisons, although difficult to isolate, as these catchments differed in many other aspects, including land use, area or topography. It became more obvious and discernible when considering a nested catchment setup, as in Fenicia et al. (2016) and in the current study. It is interesting to note that although the two latter applications differed in scale by an order of magnitude, the signature of lithology continued to be discernible and highly influential on streamflow generation. The current application also revealed the effects of variable climatology and the routing of the river network, which were considered progressively less important at smaller scales. In a different area (the Thur catchment in Switzerland) we confirmed the importance of lithology in affecting streamflow regional variability, but we also observed strong differences in streamflow seasonality (Dal Molin et al., 2020), which we did not observe in the current application. Other studies have shown the importance of regional groundwater flow, which here was considered negligible (e.g. Muñoz et al., 2016; Bouaziz et al., 2018). Hence catchment responses vary depending places and scales, which may affect individual modeling choices. Other studies on distributed model comparisons have stressed the importance of topography, and its effects on quickflow vs. baseflow partitioning, evaporation and lag times (e.g. Gao et al., 2014; Gharari et al., 2014; Nijzink et al., 2016). Our results confirm that topography affected lag times, which was incorporated in the parameterization of the lag functions. However, we could not confirm an effect on evaporation or baseflow. A direct comparison of our work with these studies is complicated by the fact that these studies have not considered lithology as a possible control on streamflow variability.

An important question when attempting to map process space to model space is at which scale such connection should take place. Many distributed models are based on the premise that "physical properties of the basic processes can only be retained at small spatial scales" (Martina et al., 2011). For example, MHM uses a parameter regionalization approach using data at the smallest possible scale (Samaniego et al., 2010). On the other hand, lumped model regionalization (e.g. Oudin et al., 2008) and catchment classification approaches (Addor et al., 2018) typically seek that connection using aggregated measures of landscape properties, hence directly at the scale of a catchment. Our results support the concept of regions of hydrological similarity or HRUs, and the idea that a relatively small number of HRUs enables a sufficient connection between processes and their model representation within a given region. Such premise is the basis of many semi-distributed models (e.g. Savenije, 2010; Dal Molin et al., 2020). Hence, it appears that the connection from process space to model space is in principle possible at multiple scales, depending on the scale of the processes that one is interested in representing, and provided that appropriate indicators of landscape properties are used, that are representative of the processes at that scale. As shown in FM2022 and elsewhere (e.g. Oudin et al., 2010; Gnann et al., 2021) devising such meaningful landscape indicators is nontrivial and requires careful analysis.

4.4. Calibration parameters in distributed regional scale models: Key takeaways?

In terms of the appropriate number of model calibration parameters for regional scale streamflow simulation, our results suggest that this number does not appear to scale up with the catchment size or with the number of subcatchments. The best performing model M(HRU_3,Lag_2)

had 12 calibration parameters. Earlier work in the Attert catchment, showed that a model with 11 calibration parameters was sufficient to reproduce relevant signatures of streamflow variability at 10 subcatchments (Fenicia et al., 2016). This number of parameters was higher than for a typical lumped streamflow simulation model (e.g. Jakeman and Hornberger (1993) concluded that the "permissible model complexity" contains around half a dozen parameters), but not orders of magnitude higher (e.g. Fenicia et al. (2014) found that simulating a complex double-peaked hydrograph response in a headwater catchments in Luxembourg required 11 parameters).

Such relatively limited complexity for a semi-distributed model may appear counterintuitive at first. Our explanation of this result is twofold. First, a semi-distributed model can be considered as many lumped models operating in parallel, but not all of the parameters that characterize these lumped models need to be spatially variable. Many of such parameters can share the same values, or their spatial variability can be prescribed by some parameter regularization relationship that depends on a few "global" parameters. In M(HRU_3,Lag_2), only the groundwater parameters were kept independent, but other parameters where either kept uniform (e.g. the snow or soil related parameters) or linked through some regularization relationship (e.g. the routing parameters). Second, HRU model structures can be tailored to the dominant processes of individual HRUs, resulting in being simpler than a typical lumped model. For example, 2 of the 3 HRUs of M(HRU_3,Lag_2) were characterized just by two buckets, as intended to model discharge with specific dynamics (fast or slow).

The number of calibration parameters that results from our model application appears low compared to other studies. For example, Pokhrel et al. (2008) needed to estimate 858 parameters of the grid based SACSMA model, which were reduced to 33 overall parameters using parameter regionalization relationships. Foglia et al. (2009) had 35 calibration parameters in their distributed version of TOPKAPI. Samaniego et al. (2010) related model parameters to observable characteristics and obtained 62 calibration parameters for the distributed MHM model. The disparity between our work and these other studies may be explained by the fact that we tailored our bespoke model to a specific catchment and scale, which enabled several simplifications. A general purpose model, as the ones in the studies mentioned above, need to consider whatever process may be important in principle, even if such processes may not be dominant in a given application.

Understanding what to account for and what to ignore when developing a distributed model is an important question. This paper and FM2022 proposed a new way to approach this question. Instead of starting by feeding a distributed model with data, our expert driven approach "looks at data first" and only then starts the modeling process. Such informed analysis of the data helps resolve many hydrological modeling questions before the model structure is determined—helping bring experimentalist insights into the initiation of the conceptual model construction process and execution of the model calibration approach. Although the ultimate model is specific to a given region and therefore not transferable to other places where other processes may dominate, the model development approach is indeed exportable to other areas. More generally, we suggest a shift from the quest for a general model, which may be an unattainable ideal given differences in model applications, to the quest for general model development approaches, which can systematically identify a suitable model for a given application. Our work shows that this can be possible at the regional scale, and that expert knowledge, supported by an experimentalist modeler dialog, can facilitate this process.

5. Conclusions

This study illustrated the development of a semi-distributed conceptual model for streamflow simulation at 26 subcatchments within the Moselle catchment. This bespoke model was informed by a perceptual model for the same catchment developed in FM2022. The perceptual model used 5 model variants of increasing complexity, designed to test specific model hypotheses. We evaluated the 5 models in a space–time validation, hence using a set of subcatchments and a time period not used for calibration.

From a methodological perspective, this study showed how the many decisions required for the development of a distributed model can be informed by a perceptual model. In particular, we showed how the understanding of landscape controls on streamflow spatial variability and how these vary spatially, can be translated into model decisions via HRUs, model structural components and model parameter constraints.

Our comparisons of the 5 model variants clarified the effect of individual model decisions on streamflow simulations, and informed the ultimate model selection. In particular, this model comparison was able to (i) identify a distributed model (in particular, M(HRU_3,Lag_2)) that matched key streamflow signatures in space–time validation, namely streamflow average ($r = 0.96$), baseflow index ($r = 0.86$), and hydrograph lag time (correct at 22 out of 26 subcatchments), and (ii) justify model decisions such as the distribution of precipitation per subcatchment, a landscape discretization based on bedrock permeability, the particularization of HRU model structures based on lithology-driven dominant processes, and a parameterization of lag functions based on topography and land use derived indices. With 12 calibration parameters, of which 5 are HRU-specific, and the others global or uniform in all HRUs, our final conceptual model structure represented a parsimonious representation of streamflow generating processes in the Moselle catchment.

We found that the added value of the conceptual model compared to the perceptual model alone, was that it helped provide a unique catchment description that explained several traits of the spatial variability of subcatchment response. While we acknowledge that our conceptual model is developed for an individual regional scale catchment, and its scope is therefore limited to the context for which it was developed. Nevertheless, the model development approach followed in this study is one that could be put into practice elsewhere. We believe this approach could contribute to better understanding regional scale catchment variability, and eventually, bespoke model development approaches that go beyond the current focus on generalizable models.

CRediT authorship contribution statement

Fabrizio Fenicia: Conceptualization, Methodology, Formal analysis, Writing - original draft. **Dennis Meißner:** Conceptualization, Methodology, Writing - review & editing. **Jeffrey McDonnell:** Conceptualization, Methodology, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Dmitri Kavetski for his input on the modeling work. Observed data can be accessed by the providers as indicated in the companion paper. The modeling code SuperflexPy is open source.

Appendix

The model forcing data are represented by precipitation P (mm/d), potential evaporation E_{Pot} (mm/d), and temperature T_C (°C). The equations of the 5 distributed models are summarized in the Tables A1–A5.

Table A1 summarizes the water balance equations. In particular M(HRU_1,Lag_0) and M(HRU_1,Lag_1) are both based on a single HRU which is represented by a 3 reservoirs structure, composed by the reservoirs UR (unsaturated), FR (fast) and SR (slow). UR splits the inflow $P^{(UR)}$ between a

Table A1
Water balance equations of the models used in the experiments (✓ and "-" indicate, respectively, presence or absence). S, P, Q, and E refer to storage, inflow, discharge and evaporation respectively. The superscripts (UR), (FR), (SR) and (HRU) refer to the UR, FR, and SR reservoir and the total HRU respectively.

Water balance equations	3-reservoir HRUs	2-reservoir HRU_L	2-reservoir HRU_H
$\frac{dS^{(UR)}}{dt} = P^{(UR)} - Q^{(UR)} - E^{(UR)}$	✓	✓	✓
$\frac{dS^{(FR)}}{dt} = P^{(FR)} - Q^{(FR)}$	✓	✓	–
$\frac{dS^{(SR)}}{dt} = P^{(SR)} - Q^{(SR)}$	✓	–	✓
$Q^{(UR)} = P^{(FR)}$	–	✓	–
$Q^{(UR)} = P^{(SR)}$	–	–	✓
$Q^{(UR)} = P^{(FR)} + P^{(SR)}$	✓	–	–
$Q^{(HRU)} = Q^{(FR)}$	–	✓	–
$Q^{(HRU)} = Q^{(SR)}$	–	–	✓
$Q^{(HRU)} = Q^{(FR)} + Q^{(SR)}$	✓	–	–

Table A2
Constitutive functions of the models used in the experiments (✓ and "-" indicate, respectively, presence or absence). The functions f are defined in Table A3. Parameters, fluxes, states and are defined in Tables A1 and A6. The parameter $m^{(UR)}$ is a threshold smoothing parameter (Kavetski and Kuczera, 2007) and is fixed to a value of 10^{-2} mm. The parameter $\alpha^{(FR)}$ is fixed at 2 (quadratic reservoir), whereas $\alpha^{(SR)}$ is fixed at 1 except in the 2-reservoir HRU_H, where it is fixed at 2.

Constitutive functions	3-reservoir HRUs	2-reservoir HRU_L	2-reservoir HRU_H
$\bar{S}^{(UR)} = S^{(UR)}/S^{(UR)}_{Max}$	✓	✓	✓
$Q^{(UR)} = P^{(UR)} f_p(\bar{S}^{(UR)} \mid \beta^{(UR)})$	✓	✓	✓
$E^{(UR)} = c^{(UR)}_E E_{Pot} f_m(\bar{S}^{(UR)} \mid m^{(UR)})$	✓	✓	✓
$P^{(SR)} = DQ^{(UR)}$	✓	–	–
$Q^{(FR)} = k^{(FR)} f_p(S^{(FR)} \mid \alpha^{(FR)})$	✓	✓	–
$Q^{(SR)} = k^{(SR)} f_p(S^{(SR)} \mid \alpha^{(SR)})$	✓	–	✓

Table A3
Constitutive functions.

Functions	Name
$f_p(x \mid \theta) = x^\theta$	Power function
$f_m(x \mid \theta) = \frac{x(1+\theta)}{x+\theta}$	Monod-type kinetics, adjusted so that $f_m(1 \mid \theta) = 1$

Table A4
Snow reservoir. S, P, and Q, refer to storage, inflow, and discharge. The superscript (WR) refers to WR (the snow reservoir), and the subscripts P and M refer to precipitation and melting (the two outflows from WR). The parameters $T^{(WR)}_{Cp}$ and $T^{(WR)}_{Cm}$ are fixed at 0 and 2 °C respectively.

Equation	Description
$\frac{dS^{(WR)}}{dt} = P - Q^{(WR)}_P - Q^{(WR)}_M$	Water balance
$P^{(UR)} = Q^{(WR)}_P + Q^{(WR)}_M$	Inflow to UR
$\frac{Q^{(WR)}_P}{P} = \begin{cases} 0, & T_C < T^{(WR)}_{Cp} \\ 1, & \text{otherwise} \end{cases}$	Constitutive equation[1]
$Q^{(WR)}_M = \begin{cases} 0, & T_C < T^{(WR)}_{Cm} \\ k^{(WR)} T_C, & \text{otherwise} \end{cases}$	Constitutive equation[1]

[1] Smoothed using the method in Kavetski and Kuczera (2007).

Table A5
Triangular lag function for flow routing. $P^{(L)}$ indicates the input, $Q^{(L)}$ indicates the output, $h^{(L)}$ is the lag function, the symbol * denotes the convolution operator, and $T^{(L)}$ is the time base of the lag function.

$Q^{(L)} = (P^{(L)} * h^{(L)})(t)$	Convolution
$h^{(L)}(t) = \begin{cases} 4t/(T^{(L)})^2, & 0 < t \leqslant T^{(L)}/2 \\ 4/T^{(L)}\left(1 - t/T^{(L)}\right), & T^{(L)}/2 < t \leqslant T^{(L)} \\ 0, & t > T^{(L)} \end{cases}$	Triangular lag function

Table A6
Overview of model parameters in the HRU model structures and corresponding calibration range.

Parameters	Units	Calibration range	Model
$k^{(WR)}$	mm/(d°C)	[0.01–10]	All
$c_E^{(UR)}$	–	[0.1–3.0]	All
$S_{Max}^{(UR)}$	mm	[0–600]	All
$\beta^{(UR)}$	–	[10^{-2}-10]	All
D	–	[0–1]	All
$k^{(FR)}$	$mm^{1-\alpha}\ d^{-1}$	[10^{-4}-1]	All
$k^{(SR)}$	$mm^{1-\alpha}\ d^{-1}$	[10^{-7}-10^{-1}]	All
$a_1^{(CL)}$	d	Calculated so that the maximum lag is shorter than 20 days	M(HRU_i,Lag_2)
$a_2^{(CL)}$	$d\ km^{-1}$	Calculated so that the maximum lag is shorter than 20 days	M(HRU_i,Lag_2)
$a^{(RL)}$	$d\ km^{-0.8}$	Calculated so that the maximum lag is shorter than 20 days	M(HRU_i,Lag_2)
$a^{(CL)}$	$d\ km^{-1}$	Calculated so that the maximum lag is shorter than 20 days	M(HRU_i,Lag_1)
$a^{(RL)}$	$d\ km^{-1}$	Calculated so that the maximum lag is shorter than 20 days	M(HRU_i,Lag_1)

portion that is stored, and eventually evaporates as $E^{(UR)}$, and a portion that eventually produces runoff, $Q^{(UR)}$. $Q^{(UR)}$ is then partitioned between $P^{(FR)}$ and $P^{(SR)}$, which enter FR and SR respectively. The outflows of FR and SR, $Q^{(FR)}$ and $Q^{(SR)}$, are summed together, and form the outflow of the individual HRU $Q^{(HRU)}$. M(HRU_2,Lag_1) and M(HRU_2,Lag_2) both have 2 HRUs, high permeability (HRU_H) and medium–low permeability (HRU_{ML}). Both HRUs are described by the same 3-reservoirs structure described above. M(HRU_3,Lag_2) has 3 HRUs, high, medium and low permeability (HRU_H, HRU_M, and HRU_L respectively). The HRU_H model structure excludes the FR reservoir, the HRU_M model structure has both FR and SR, while the HRU_L model structure excludes the SR reservoir.

Table A2 describes the models constitutive functions, with functions symbols defined in Table A3. UR partitions incoming precipitation between a portion that is stored and outflow using a power function, defined by a parameter $\beta^{(UR)}$. Evaporation from UR is proportional to the potential evaporation with a parameter $c_E^{(UR)}$, and approaches zero as the reservoir depletes depending on a smoothing parameter $m^{(UR)}$. The 3 reservoir structure partitions of the outflow from UR between the inflows to FR and SR depending on the parameter D. FR is parameterized as a nonlinear reservoir, which depends on a parameter $k^{(FR)}$ and a power $\alpha^{(FR)}$ fixed at 2 (hence a quadratic reservoir). SR is a linear reservoir, which depends on the parameter $k^{(SR)}$. The 2 reservoirs structures used for HRU_H and HRU_L in M(HRU_3,Lag_2) parameterize both SR and FR as quadratic reservoirs.

Table A4 describes the snow reservoir. Precipitation P is partitioned between a portion that is stored and rainfall $Q_P^{(WR)}$. Rainfall $Q_P^{(WR)}$ depends on a temperature threshold $T_{Cp}^{(WR)}$, which is fixed at 0° degrees Celsius. Snowmelt $Q_M^{(WR)}$ depends on a temperature threshold $T_{Cm}^{(WR)}$, which is fixed at 2° degrees Celsius, and on the degree-day parameter $k^{(WR)}$. The temperature thresholds as well as the effective melting rate are smoothed using the method in Kavetski and Kuczera (2007). The inflow to UR is given by the sum of rainfall and snowmelt.

For the routing elements, a triangular lag function, similarly to the HBV model (Lindström et al., 1997), is used. This function is described in Table A5, and depends on the time base parameter $T^{(L)}$. Routing elements are not present in M(HRU_1,Lag_0). The other models include the same number of routing elements, and differ for the regularization relationship used to scale the time base parameters $T^{(L)}$ for the various lag functions. M(HRU_1, Lag_1) and M(HRU_2,Lag_1) use a simpler parameterization, where the scaling of the routing depends on drainage distance alone e.g. Lerat et al., (2012):

- Incremental subcatchment routing elements (CL). We assume that the time parameter representing the routing of the incremental subcatchment i, is related to its landscape properties by the following expression:

$$T_i^{(CL)} = a^{(CL)} L_{DstD,i}^{(Top)} \tag{9}$$

where $a^{(CL)}$ is a calibration parameter, and $L_{DstD}^{(Top)}$, already defined in FM2022, indicates the maximum drainage distance of the incremental subcatchment.

- River routing elements (RL). We assume that the time parameter of the lag functions representing the routing of the river stretch i, which goes from the outlet of the subcatchment i (indicated with A) to the next downstream outlet (indicated with B), is related to landscape properties by the following expression:

$$T_i^{(RL)} = a^{(RL)} L_{AB,i} \tag{10}$$

where $a^{(RL)}$ is a calibration parameter, and the symbol L_{AB}, already defined in FM2022, represents the drainage distance between the two outlets A and B.

M(HRU_2,Lag_2) and M(HRU_3,Lag_2) use a more complex parameterization of the regularization relationships, motivated by the data analysis in the perceptual model, which is described as follows:

- Incremental subcatchment routing elements (CL). We assume that the time parameter representing the routing of the incremental subcatchment i, is related to its landscape properties by the following expression:

$$T_i^{(CL)} = a_1^{(CL)} + a_2^{(CL)} \frac{L_{DstG,i}^{(Top)}}{L_{Dev,i}^{(Lnd)} L_{Sl05,i}^{(Top)}} \tag{11}$$

where $a_1^{(CL)}$ and $a_2^{(CL)}$ are calibration parameters, and the symbols $L_{DstG}^{(Top)}$, $L_{Dev}^{(Lnd)}$ and $L_{Sl05}^{(Top)}$, already defined in FM2022, represent respectively the geometric distance, the fraction of developed land use, and the 5% quantile of the slope. This expression corresponds to Equation 10 in FM2022.

- River routing elements (RL). We assume that the time parameter of the lag functions representing the routing of the river stretch i, which goes from the outlet of the subcatchment i (indicated with A) to the next downstream outlet (indicated with B), is related to landscape properties by the following expression:

$$T_i^{(RL)} = a^{(RL)} \frac{L_{AB,i}}{A_{AB,i}^{\alpha}}$$

where $a^{(RL)}$ is a calibration parameter, and the symbols L_{AB}, A_{AB} and α, already defined in FM2022, represent respectively the distance along the river network between two points A and B, the average area between the total subcatchments at A and B, and the exponent constant (fixed at 0.1, see FM2022). This expression corresponds to Equation 12 in FM2022.

The absence of feedbacks between landscape elements greatly simplifies the solution of the system of mass balance differential equations, which can be solved step by step, using a fixed step implicit approximation Fenicia et al. (2011).

References

Adams, T.E., Pagano, T.C., 2016. Flood Forecasting a Global Perspective. Academic Press, 478 pp.

Addor, N., Nearing, G., Prieto, C., Newman, A.J., Le Vine, N., Clark, M.P., 2018. A ranking of hydrological signatures based on their predictability in space. Water Resour. Res. 54 (11), 8792–8812. https://doi.org/10.1029/2018WR022606.

Ajami, N.K., Gupta, H., Wagener, T., Sorooshian, S., 2004. Calibration of a semi-distributed hydrologic model for streamflow estimation along a river system. J. Hydrol. 298 (1), 112–135. https://doi.org/10.1016/j.jhydrol.2004.03.033.

Antonetti, M., Zappa, M., 2018. How can expert knowledge increase the realism of conceptual hydrological models? A case study based on the concept of dominant runoff process in the Swiss Pre-Alps. Hydrol. Earth Syst. Sci. 22 (8), 4425–4447. https://doi.org/10.5194/hess-22-4425-2018.

Athira, P., Sudheer, K.P., 2021. Calibration of distributed hydrological models considering the heterogeneity of the parameters across the basin: a case study of SWAT model. Environ Earth Sci 80 (4), 131. https://doi.org/10.1007/s12665-021-09434-8.

Beven, K., 1989. Changing ideas in hydrology — the case of physically-based models. J. Hydrol. 105 (1), 157–172. https://doi.org/10.1016/0022-1694(89)90101-7.

Beven, K., Binley, A., 1992. The future of distributed models: Model calibration and uncertainty prediction. Hydrol. Process. 6 (3), 279–298. https://doi.org/10.1002/hyp.3360060305.

Bouaziz, L., Weerts, A., Schellekens, J., Sprokkereef, E., Stam, J., Savenije, H., Hrachowitz, M., 2018. Redressing the balance: quantifying net intercatchment groundwater flows. Hydrol. Earth Syst. Sci. 22 (12), 6415–6434. https://doi.org/10.5194/hess-22-6415-2018.

Boyle, D.P., Gupta, H.V., Sorooshian, S., Koren, V., Zhang, Z., Smith, M., 2001. Toward improved streamflow forecasts: value of semidistributed modeling. Water Resour. Res. 37 (11), 2749–2759. https://doi.org/10.1029/2000WR000207.

Burt, T.P., McDonnell, J.J., 2015. Whither field hydrology? The need for discovery science and outrageous hydrological hypotheses. Water Resour. Res 51 (8), 5919–5928. https://doi.org/10.1002/2014wr016839.

Dal Molin, M., Kavetski, D., Fenicia, F., 2021. SuperflexPy 1.2.0: an open source Python framework for building, testing and improving conceptual hydrological models. Geosci. Model Dev. 2020, 1–39. https://doi.org/10.5194/gmd-2020-409.

Dal Molin, M., Schirmer, M., Zappa, M., Fenicia, F., 2020. Understanding dominant controls on streamflow spatial variability to set up a semi-distributed hydrological model: the case study of the Thur catchment. Hydrol. Earth Syst. Sci. 24 (3), 1319–1345. https://doi.org/10.5194/hess-24-1319-2020.

de Lavenne, A., Andréassian, V., Thirel, G., Ramos, M.-H., Perrin, C., 2019. A regularization approach to improve the sequential calibration of a semidistributed hydrological model. Water Resour. Res. 55 (11), 8821–8839. https://doi.org/10.1029/2018wr024266.

Demuth, N., Rademacher, S., 2016. Chapter 5 – Flood Forecasting in Germany — challenges of a federal structure and transboundary cooperation. In: Adams, T.E., Pagano, T.C. (Eds.), Flood Forecasting. Academic Press, Boston, pp. 125–151.

Ehret, U., van Pruijssen, R., Bortoli, M., Loritz, R., Azmi, E., Zehe, E., 2020. Adaptive clustering: reducing the computational costs of distributed (hydrological) modelling by exploiting time-variable similarity among model elements. Hydrol. Earth Syst. Sci. 24 (9), 4389–4411. https://doi.org/10.5194/hess-24-4389-2020.

Fatichi, S., et al., 2016. An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. J. Hydrol. 537, 45–60. https://doi.org/10.1016/j.jhydrol.2016.03.026.

Fenicia, F., McDonnell, J.J., 2022. Modeling streamflow variability at the regional scale: (1) Perceptual model development through signature analysis. J. Hydrol., 127287 https://doi.org/10.1016/j.jhydrol.2021.127287.

Fenicia, F., Kavetski, D., 2021. Behind every robust result is a robust method: perspectives from a case study and publication process in hydrological modelling. Hydrol. Process. 35 (8), e14266 https://doi.org/10.1002/hyp.14266.

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 (11), W11510. https://doi.org/10.1029/2010wr010174.

Fenicia, F., Kavetski, D., Savenije, H.H.G., Pfister, L., 2016. From spatially variable streamflow to distributed hydrological models: analysis of key modeling decisions. Water Resour. Res. 52 (2), 954–989. https://doi.org/10.1002/2015WR017398.

Fenicia, F., Kavetski, D., Savenije, H.H.G., Clark, M.P., Schoups, G., Pfister, L., Freer, J., 2014. Catchment properties, function, and conceptual model representation: is there a correspondence? Hydrological Processes 28 (4), 2451–2467. https://doi.org/10.1002/Hyp.9726.

Feyen, L.U.C., Kalas, M., Vrugt, J.A., 2008. 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. Hydrol. Sci. J. 53 (2), 293–308. https://doi.org/10.1623/hysj.53.2.293.

Foglia, L., Hill, M.C., Mehl, S.W., Burlando, P., 2009. Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function. Water Resour. Res. 45 (6) https://doi.org/10.1029/2008wr007255.

Gao, H., Hrachowitz, M., Fenicia, F., Gharari, S., Savenije, H.H.G., 2014. Testing the realism of a topography-driven model (FLEX-Topo) in the nested catchments of the Upper Heihe, China. Hydrol Earth Syst Sc 18 (5), 1895–1915. https://doi.org/10.5194/hess-18-1895-2014.

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., Savenije, H.H.G., 2014. Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration. Hydrol Earth Syst Sc 18 (12), 4839–4859. https://doi.org/10.5194/hess-18-4839-2014.

Gnann, S.J., McMillan, H.K., Woods, R.A., Howden, N.J.K., 2021. Including Regional Knowledge Improves Baseflow Signature Predictions in Large Sample Hydrology. Water Resour. Res. 57 (2) https://doi.org/10.1029/2020WR028354 e2020WR028354.

Grayson, R.B., Moore, I.D., McMahon, T.A., 1992. Physically based hydrologic modeling: 2. Is the concept realistic? Water Resour. Res. 28 (10), 2659–2666. https://doi.org/10.1029/92WR01259.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. Water Resour. Res. 34 (4), 751–763. https://doi.org/10.1029/97WR03495.

Hirpa, F.A., Salamon, P., Beck, H.E., Lorini, V., Alfieri, L., Zsoter, E., Dadson, S.J., 2018. Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data. J. Hydrol. 566, 595–606. https://doi.org/10.1016/j.jhydrol.2018.09.052.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H.H.G., Gascuel-Odoux, C., 2014. Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. Water Resour. Res. 50 (9), 7445–7469. https://doi.org/10.1002/2014wr015484.

Hulsman, P., Savenije, H.H.G., Hrachowitz, M., 2021. Learning from satellite observations: increased understanding of catchment processes through stepwise model improvement. Hydrol. Earth Syst. Sci. 25 (2), 957–982. https://doi.org/10.5194/hess-25-957-2021.

Imbeaux, E., 1892. La Durance: régime, crues et inondations, 200 pp., Dunod.

Jakeman, A.J., Hornberger, G.M., 1993. How much complexity is warranted in a rainfall-runoff model? Water Resour. Res. 29 (8), 2637–2649. https://doi.org/10.1029/93WR00877.

Jothityangkoon, C., Sivapalan, M., Farmer, D.L., 2001. Process controls of water balance variability in a large semi-arid catchment: downward approach to hydrological model development. J. Hydrol. 254 (1), 174–198. https://doi.org/10.1016/S0022-1694(01)00496-6.

Kavetski, D., Kuczera, G., 2007. Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration. Water Resour. Res. 43 (3) https://doi.org/10.1029/2006wr005195.

Kavetski, D., Kuczera, G., Franks, S.W., 2006. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resour. Res. 42 (3) https://doi.org/10.1029/2005WR004368.

Kavetski, D., Fenicia, F., Reichert, P., Albert, C., 2018. Signature-domain calibration of hydrological models using approximate bayesian computation: theory and comparison to existing applications. Water Resour. Res. 54 (6), 4059–4083. https://doi.org/10.1002/2017WR020528.

Kelleher, C., McGlynn, B., Wagener, T., 2017. Characterizing and reducing equifinality by constraining a distributed catchment model with regional signatures, local observations, and process understanding. Hydrol. Earth Syst. Sci. 21 (7), 3325–3352. https://doi.org/10.5194/hess-21-3325-2017.

Kirchner, J.W., 2006. Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. Water Resour. Res. 42 https://doi.org/10.1029/2005WR004362. W03S04.

Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrol. Sci. J. 31 (1), 13–24. https://doi.org/10.1080/02626668609491024.

Leavesley, G. H., R. W. Lichty, B. M. Troutman, and L. G. Saindon (1983), Precipitation-runoff modeling system: user's manual, USGS Water-Resources Investigations Report, 83-4238, 214 pp.

Lerat, J., Andréassian, V., Perrin, C., Vaze, J., Perraud, J.M., Ribstein, P., Loumagne, C., 2012. Do internal flow measurements improve the calibration of rainfall-runoff models? Water Resour. Res. 48 (2) https://doi.org/10.1029/2010wr010179.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., Bergström, S., 1997. Development and test of the distributed HBV-96 hydrological model. J. Hydrol. 201 (1), 272–288. https://doi.org/10.1016/S0022-1694(97)00041-3.

Liu, Y., Gupta, H.V., 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. Water Resour. Res. 43 (7) https://doi.org/10.1029/2006WR005756.

Loague, K. (Ed.) (2010), Rainfall-Runoff Modelling, Benchmark Papers in Hydrology ed., 506 pp., IAHS.

Loritz, R., Gupta, H., Jackisch, C., Westhoff, M., Kleidon, A., Ehret, U., Zehe, E., 2018. On the dynamic nature of hydrological similarity. Hydrol. Earth Syst. Sci. 22 (7), 3663–3684. https://doi.org/10.5194/hess-22-3663-2018.

Lyne, V., and M. Hollick (1979), Stochastic time-variable rainfall runoff modelling, Proceedings of the Hydrology and Water Resources Symposium, Perth, 10-12 September, 79(10), 89-92.

Martina, M.L.V., Todini, E., Liu, Z., 2011. Preserving the dominant physical processes in a lumped hydrological model. J. Hydrol. 399 (1), 121–131. https://doi.org/10.1016/j.jhydrol.2010.12.039.

McDonnell, J.J., et al., 2007. Moving beyond heterogeneity and process complexity: a new vision for watershed hydrology. Water Resour. Res. 43 (7), W07301. https://doi.org/10.1029/2006WR005467.

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., Kuczera, G., 2017. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. Water Resour. Res. 53 (3), 2199–2239. https://doi.org/10.1002/2016WR019168.

McMillan, H.K., Clark, M.P., Bowden, W.B., Duncan, M., Woods, R.A., 2011. Hydrological field data from a modeller's perspective: Part 1 Diagnostic tests for model structure. Hydrological Processes 25 (4), 511–522. https://doi.org/10.1002/Hyp.7841.

Montanari, A., Di Baldassarre, G., 2013. Data errors and hydrological modelling: the role of model structure to propagate observation uncertainty. Adv. Water Resour. 51, 498–504. https://doi.org/10.1016/j.advwatres.2012.09.007.

Mulder, G., Olsthoorn, T.N., Al-Manmi, D.A.M.A., Schrama, E.J.O., Smidt, E.H., 2015. Identifying water mass depletion in northern Iraq observed by GRACE. Hydrol. Earth Syst. Sci. 19 (3), 1487–1500. https://doi.org/10.5194/hess-19-1487-2015.

Muñoz, E., Arumí, J.L., Wagener, T., Oyarzún, R., Parra, V., 2016. Unraveling complex hydrogeological processes in Andean basins in south-central Chile: An integrated assessment to understand hydrological dissimilarity. Hydrol. Process. 30 (26), 4934–4943. https://doi.org/10.1002/hyp.11032.

Nijzink, R.C., Samaniego, L., Mai, J., Kumar, R., Thober, S., Zink, M., Schäfer, D., Savenije, H.H.G., Hrachowitz, M., 2016. The importance of topography-controlled sub-grid process heterogeneity and semi-quantitative prior constraints in distributed hydrological models. Hydrol. Earth Syst. Sci. 20 (3), 1151–1176. https://doi.org/10.5194/hess-20-1151-2016.

Oudin, L., Kay, A., Andréassian, V., Perrin, C., 2010. Are seemingly physically similar catchments truly hydrologically similar? Water Resour. Res. 46 (11) https://doi.org/10.1029/2009WR008887.

Oudin, L., Andréassian, V., Perrin, C., Michel, C., Le Moine, N., 2008. Spatial proximity, physical similarity, regression and ungaged catchments: a comparison of regionalization approaches based on 913 French catchments. Water Resour. Res. 44 (3) https://doi.org/10.1029/2007WR006240.

Paniconi, C., Putti, M., 2015. Physically based modeling in catchment hydrology at 50: Survey and outlook. Water Resour. Res. 51 (9), 7090–7129. https://doi.org/10.1002/2015wr017780.

Pokhrel, P., Gupta, H.V., Wagener, T., 2008. A spatial regularization approach to parameter estimation for a distributed watershed model. Water Resour Res 44 (12), 1–16. https://doi.org/10.1029/2007wr006615. W12419.

Prieto, C., D. Kavetski, N. Le Vine, C. Álvarez, and R. Medina Identification of dominant hydrological mechanisms using Bayesian inference, multiple statistical hypothesis testing and flexible models, Water Resour Res, n/a(n/a), e2020WR028338, doi: 10.1029/2020WR028338.

Prieto, C., Kavetski, D., Le Vine, N., Álvarez, C., Medina, R., 2021. Identification of dominant hydrological mechanisms using Bayesian inference, multiple statistical hypothesis testing, and flexible models. Water Resour. Res. 57 (8) https://doi.org/10.1029/2020WR028338 e2020WR028338.

Raneesh, K.Y., Thampi, S.G., 2013. A simple semi-distributed hydrologic model to estimate groundwater recharge in a humid tropical basin. Water Resour Manag 27 (5), 1517–1532. https://doi.org/10.1007/s11269-012-0252-5.

Refsgaard, J.C., Knudsen, J., 1996. Operational validation and intercomparison of different types of hydrological models. Water Resour. Res. 32 (7), 2189–2202. https://doi.org/10.1029/96wr00896.

Samaniego, L., Kumar, R., Attinger, S., 2010. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. Water Resour. Res. 46 (5) https://doi.org/10.1029/2008WR007327.

Savenije, H.H.G., 2009. HESS Opinions "The art of hydrology". Hydrol. Earth Syst. Sci. 13 (2), 157–161. https://doi.org/10.5194/hess-13-157-2009.

Savenije, H. H. G., 2010. HESS Opinions "Topography driven conceptual modelling (FLEX-Topo)", Hydrol Earth Syst Sc, 14(12), 2681-2692, doi: 10.5194/hess-14-2681-2010.

Seibert, J., McDonnell, J.J., 2002. On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. Water Resour. Res. 38 (11), 1241. https://doi.org/10.1029/2001WR000978.

Sivapalan, M., 2006. Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale, in Encyclopedia of Hydrological Sciences, edited, John Wiley & Sons, Ltd.

Sivapalan, M., Blöschl, G., Zhang, L., Vertessy, R., 2003. Downward approach to hydrological prediction. Hydrol. Process. 17 (11), 2101–2111. https://doi.org/10.1002/hyp.1425.

Smith, P.J., Pappenberger, F., Wetterhall, F., Thielen del Pozo, J., Krzeminski, B., Salamon, P., Muraro, D., Kalas, M., Baugh, C., 2016. Chapter 11 – On the operational implementation of the European Flood Awareness System (EFAS). In: Adams, T.E., Pagano, T.C. (Eds.), Flood Forecasting. Academic Press, Boston, pp. 313–348.

Spieler, D., Mai, J., Craig, J.R., Tolson, B.A., Schütze, N., 2020. Automatic model structure identification for conceptual hydrologic models. Water Resour. Res. 56 (9) https://doi.org/10.1029/2019WR027009 e2019WR027009.

Viviroli, D., Zappa, M., Gurtz, J., Weingartner, R., 2009. An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools. Environ. Modell. Software 24 (10), 1209–1222. https://doi.org/10.1016/j.envsoft.2009.04.001.

Wallner, M., Haberlandt, U., Dietrich, J., 2012. Evaluation of different calibration strategies for large scale continuous hydrological modelling. Adv. Geosci. 31, 67–74. https://doi.org/10.5194/adgeo-31-67-2012.

Wood, E.F., et al., 2011. Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. Water Resour. Res. 47 (5) https://doi.org/10.1029/2010WR010090.

Yokoo, Y., Chiba, T., Shikano, Y., Leong, C., 2017. Identifying dominant runoff mechanisms and their lumped modeling: a data-based modeling approach. Hydrol. Res. Lett. 11 (2), 128–133. https://doi.org/10.3178/hrl.11.128.

Young, P., 2003. Top-down and data-based mechanistic modelling of rainfall–flow dynamics at the catchment scale. Hydrol. Process. 17 (11), 2195–2217. https://doi.org/10.1002/hyp.1328.

Zhang, X., Srinivasan, R., Liew, M.V., 2010. On the use of multi-algorithm, genetically adaptive multi-objective method for multi-site calibration of the SWAT model. Hydrol. Process. 24 (8), 955–969. https://doi.org/10.1002/hyp.7528.