On constraining a lumped hydrological model with both piezometry and streamflow: results of a large sample evaluation

Antoine Pelletier[1, 2] and Vazken Andréassian[2]

[1]École des Ponts, Marne-la-Vallée, France
[2]Université Paris-Saclay, INRAE, UR HYCAR, Antony, France

Correspondence: Antoine Pelletier (antoine.pelletier@inrae.fr)

Abstract. The role of aquifers in the seasonal and multiyear dynamics of streamflow is undisputed: in many temperate catchments, aquifers store water during the wet periods and release it all year long, making a major contribution to low flows. The complexity of groundwater modelling has long prevented surface hydrological modellers from including groundwater level data, especially in lumped conceptual rainfall–runoff models. In this article, we investigate whether using groundwater level data in the daily GR6J model, through a composite calibration framework, can improve the performance of streamflow simulation. We tested the new calibration process on 107 French catchments. Our results show that these additional data are superfluous if we look only at model performance for streamflow simulation. However, parameter stability is improved and the model shows a surprising ability to simulate groundwater levels with a satisfying performance, in a wide variety of hydrogeological and hydroclimatic contexts. Finally, we make several recommendations regarding the model calibration process to be used, according to the hydrogeological context of the modelled catchment.

1 Introduction

1.1 Why use piezometry in low-flow modelling?

"Geology is the fundamental base of hydrology" (Castany, 1963): what happens ~~in the sub-soil~~ under the surface is an essential part of the behaviour of many hydrological systems. At the catchment scale, aquifers have the ability to store water in the long run and to release it afterwards, thereby contributing to streamflow. The hydrological processes taking place underground, ~~whose complexity is not straightforward to describe~~which are complex and therefore difficult to faithfully and simply picture, are often aggregated in surface hydrology models and represented by a simple reservoir, which fills during each rainfall event and slowly empties during rainless periods. This conceptualisation is called into question by the ability of underground water to contribute heavily to flood events – see e.g. Habets et al. (2010), Roche et al. (2012) or Guérin et al. (2019) – but it remains an acceptable representation of aquifer-river exchanges during droughts. Indeed, the fundamental role of aquifers in supporting river flows during the dry season is well-known: the trailblazer hydrologist Maillet (1905) observed it on several springs in the Paris basin. More recently, Carlier et al. (2018) and Wirth et al. (2020) linked low-flow statistics to hydrogeological descriptors in Swiss catchments and reported that their low-flow behaviour was heavily dependent on the hydrogeological context, with a particular role of sandstone and quaternary aquifers in inter-seasonal water storage; Tague and Grant (2009) and Hayashi

(2020) showed the buffering role of small aquifers in mountainous catchments and underlined their ability to support low flows and to supplement the snow reservoir that is dried up by climate change. Tobin and Schwartz (2020) and Käser and Hunkeler (2016) highlighted that even aquifers with a small spatial extent at the catchment scale ~~can~~ could support low flows significantly, even during long dry periods. Tracer studies (see e.g. Soulsby et al., 2006; Tobin and Schwartz, 2020) confirmed that groundwater contributes significantly to streamflow during the dry season and that the extent of this contribution depends on the hydrogeological configuration, i.e. the geological nature of the catchment's subsoil. The *buffering* or *storage* role of aquifers contributes to the phenomenon known as *catchment memory*, i.e. the smoothing of the input climatic signal by the catchment response (Tomasella et al., 2008; Lo and Famiglietti, 2010; Creutzfeldt et al., 2012). Using other words, Roche et al. (2012) highlight that, at least in temperate regions, severe droughts are often the result of several drier-than-normal years that lead to aquifers reaching exceptionally low levels.

Despite the level of evidence of the role of aquifers in low-flow dynamics, many hydrological modelling tools that are commonly – and quite successfully – used to simulate and forecast droughts have no explicit representation of groundwater dynamics. The cultural differences between hydrogeologists and surface hydrologists, highlighted e.g. by Barthel (2014), contribute to this situation: different systems with different characteristics and different problems to be solved lead to different models whose coupling is not straightforward. In particular, the main goal of surface hydrology modelling – streamflow – is almost directly and dynamically accessible, which makes elementary calibration of all kinds of models possible, while measuring the state of an aquifer is only possible using a limited number of piezometers that measure the hydraulic head at a point. Satellite remote sensing is now able to monitor groundwater changes (Swenson et al., 2006; Syed et al., 2008) but the temporal availability and the spatial resolution of such products limit their use in hydrological modelling at local to regional scales.

The difficulty in using piezometric data is one of the reasons why hydrologists often prefer to retrieve the river-groundwater flux by solving the inverse problem, i.e. using the surface data to infer the state of the aquifer. The most common approach is hydrograph separation, which consists in splitting streamflow into two components: a slow one, named *baseflow* and a quick one, named *quickflow*. Baseflow is then regarded as the result of the slowest hydrological processes operating in the catchment, generally underground processes. This approach can be useful for analysing the hydrological behaviour of large sets of catchments and a high proportion of baseflow in total streamflow is often correlated with a geological context favourable to a high contribution of aquifers (Pelletier and Andréassian, 2020). However, assimilating conceptual baseflow into aquifer contribution is generally unsuitable (Beven, 1991), since it results from a confusion between catchment time response and water molecule transit time (McDonnell and Beven, 2014). To provide a hydrological model with new information about the catchment state, here its underground state, it is necessary to provide new data, such as, when it is available, piezometry.

1.2 What are the existing modelling approaches?

Hydrological models are often classified depending on their level of spatial discretisation – *lumped* versus *distributed* models – and their ambition to represent more or less explicitly the physical processes taking place in the catchment – *empirical* and *conceptual* versus *physically based* models – (Roche et al., 2012). Lumped models have no spatial discretisation at the catchment

scale – i.e. the catchment is treated as a single unit with spatially averaged descriptors – whereas distributed models discretise the catchment into grid units, each of them described by several variables (Beven, 2012). Semi-distributed models constitute an intermediate option, in which the catchment under study is divided into sub-catchments, each of them becoming the object of lumped computations (see e.g. de Lavenne et al., 2016). Physically based or process-based models strive to reproduce the physical processes taking place in the catchment, by solving a version of fluid mechanics equations, while conceptual models ~~develop~~ are based on their own empirical equations to reproduce the total water balance without any reductionist ambition. Because every grid element of a distributed model needs to be parametrised, it usually carries a large number of parameters that cannot all be calibrated on observations and need to be set a priori; lumped models, on the other hand, often have a smaller number of parameters that are easier to calibrate automatically from observations. Since physical laws need to be solved at local scale and lumped models are generally designed for their simplicity in operational purposes, there is a general correspondence between distributed and physically based models on the one hand and lumped and conceptual models on the other (Beven, 2012).

Hydrogeological models dedicated to groundwater simulation are generally classified as physically based – see Mackay et al. (2014) for a rare example of a conceptual model. Therefore, the surface/groundwater interaction is more naturally represented in physically based distributed hydrological models (Dassargues et al., 1999). At local scale, this can be achieved by fluid mechanics equations (Bartlett and Porporato, 2018) but at catchment scale, distributed models generally use simplified versions of these equations. Barthel and Banzhaf (2015) performed an extensive review of models taking into account the surface/groundwater interaction at regional scale. We will not summarise the review here, but a salient point is the distinction between fully coupled schemes, where equations are solved simultaneously for surface and groundwater flows (see e.g. HydroGeoSphere by Brunner and Simmons, 2011), and loosely coupled schemes, where several models are coupled only via the exchange of results (see e.g. Isba-Modcou in Habets et al., 2010). All these approaches are difficult to implement on large sets of catchments, because of parametrisation requirements.

Using conceptual lumped rainfall–runoff models to simulate the surface/groundwater interaction is less straightforward, since fluid mechanics equations cannot be used; a conceptual representation of the aquifer, often using a reservoir, is therefore necessary. Water exchange with an aquifer can be computed solving the inverse problem, i.e. inferring the fluxes from the amount of water needed by the model to close the water budget – see e.g. Perrin et al. (2003), Le Moine (2008), Le Moine et al. (2008) and Herron and Croke (2009) – but it is far from sufficient for simulating the actual level of an aquifer. Bergström and Sandberg (1983) added a groundwater simulation module to the HBV model (Bergström and Forsman, 1973) and implemented it on three aquifers; they obtained a satisfactory performance in reproducing past piezometric time series, despite parametrisation issues caused by computation cost, which are no longer mentioned in recent studies (Széles et al., 2020), considering advances in computer science. Thiéry (1988) used the *ground* reservoir of the Gardenia model (Thiéry, 2014) to simulate and forecast the piezometry of the Paris basin chalk aquifer, using a linear regression between the reservoir levels and the aquifer levels. Borzì et al. (2019) designed a modified version of the IHACRES model (Jakeman and Hornberger, 1993) with an explicit representation of a volcanic deep aquifer in Sicily, through an additional conceptual reservoir. In order to represent the specific role of groundwater in intermittent streams, Moore and Bell (2002) added a piezometry simulation module to the

PDM rainfall–runoff model (Moore, 1999), which was able to represent pumped abstractions. The path followed by Hughes (2004) and Efstratiadis et al. (2008) is intermediate, with a semi-distributed conceptual hydrological model connected to a semi-distributed – with a different spatial discretisation – conceptual aquifer representation; this model is easier to implement and needs fewer data than a fully distributed one, allowing for many experiments simulating anthropogenic influence, but it is far from straightforward to implement on any catchment with few data.

These modelling schemes have shown noteworthy simulation abilities for both aquifers and streamflow. However, they have not been tested on large sets of catchments in various contexts to value their robustness and generalisation capacity. Moreover, groundwater simulation is, in most hydrological modelling studies, a side product of rainfall–runoff modelling. There is little evidence on how the addition of groundwater data can actually help obtain a better streamflow simulation.

1.3 How are measured data used in hydrological modelling?

Most hydrological models are parametric and their parameters are calibrated using measured streamflow data (Roche et al., 2012). To find the best set of parameters with which to reproduce the streamflow time series, a calibration criterion, which is a function of measured and simulated – or forecasted – streamflow, is optimised, the most common one being the Nash–Sutcliffe efficiency or NSE (Nash and Sutcliffe, 1970). Gupta et al. (2009) and Kling et al. (2012), investigating the drawbacks of NSE, proposed another criterion, henceforth known as Kling–Gupta efficiency (KGE), which is a Euclidean combination of three criteria that all compare measured and simulated streamflow. Computed with untransformed time series, these criteria are focused on the peaks of the hydrograph; to get a better calibration on the lower part of the latter, i.e. low flows, streamflow time series can be transformed using concave functions (Pushpalatha et al., 2012), such as square root or logarithm.

Traditional calibration approaches are generally *single-objective*, i.e. only one objective function is used. However, all criteria can be regarded as flawed, since they focus on only one aspect of the hydrograph representation. Linear or Euclidean combinations of criteria can be used (Nicolle et al., 2014), for instance the mean between NSE and KGE, which is called *composite calibration*. *Multi-objective calibration* (Madsen, 2003) tries to optimise several criteria at the same time. It is generally impossible to get a unique optimal set of parameters as the result of a multi-objective calibration problem; a Pareto front, i.e. an ensemble of parameter sets, is formed, each one representing a different compromise between objective functions. For operational purposes, it is necessary to choose a parameter set in this Pareto front, generally using a determined weighting between objective functions – either a linear combination or a Euclidean distance to a reference point – which is similar to composite calibration.

Complex, distributed hydrological models, especially when they claim to be physically based, often explicitly simulate physical variables. Therefore, measured data can be directly associated to these variables without having to build an observation ~~function~~ operator between model variables and measured data and designing a calibration process for such models is more straightforward. Even if they are rarely available on large sets of catchments, in-field measurements are often used in models for specific instrumented catchments. For instance, the isoWATFLOOD model (Stadnyk et al., 2013; Stadnyk and Holmes, 2020) is calibrated using both streamflow and isotopic – $\delta^{18}O$ - data, but a visual – and thus, rather subjective – evaluation of calibration by the modeller is necessary; Jian et al. (2017) used, in a catchment where only few streamflow measurements

were available, river level data and added three new parameters to a hydrological model to simulate the rating curve. Whereas in-field measurements are not always common, satellite data are broadly available around the world and numerous studies have used them in hydrological models: Immerzeel and Droogers (2008) used satellite evaporation to calibrate the SWAT distributed model through a composite criterion and got a closer representation of actual evaporation and less equifinality in parameter determination; Mostafaie et al. (2018) performed a multi-objective calibration using NSE for streamflow and total water storage from GRACE satellite data; Milzow et al. (2011) combined several satellite datasets – surface soil moisture, radar altimetry and total water storage – to calibrate a semi-distributed model in a catchment with few streamflow measurements through a composition of nine criteria; Demirel et al. (2019) explored different combinations of objective functions, computed on several satellite products measuring soil moisture and water storage, to calibrate a conceptual model, with little gain on the streamflow simulation performance; Dembélé et al. (2020) performed a composite calibration of a distributed model with four datasets – measured streamflow and satellite evaporation, soil moisture and water storage – and improved the model representation of processes at the expense of a small degradation of the streamflow simulation performance.

Using other data than streamflow is less straightforward in empirical or conceptual models that do not explicitly simulate physical fluxes or states. A particular state of the model is generally linked to the available physical variable. In catchments affected by snow and/or glaciers, related data – i.e. snow depth or glacier ~~state~~ thickness – can be used in model calibration (Riboust et al., 2018; Tiel et al., 2020). Beyond calibration, extra data can be assimilated into the model to correct its trajectory during runtime; several studies showed an improved performance of hydrological models with assimilation of soil moisture (Aubert et al., 2003a, b; Oudin et al., 2003) or snowpack (Thirel et al., 2013).

As far as piezometry is concerned, distributed hydrological models are rarely calibrated using piezometry time series. Most gridded models have a physical parametrisation: parameter values are, directly or indirectly, inferred from local properties measured in situ (Moreda et al., 2006) – for instance, topography, soil types, vegetation or geological properties. At a pinch, the parameter set can be adjusted, with a limited variation margin adapted to the physicalness of parameters, to better represent streamflow; but given the often large number of parameters to be adjusted, distributed models cannot be fully calibrated without suffering from equifinality (Beven, 1993). In these conditions, several studies underlined the possibility of calibrating a distributed hydrological model using both piezometry and streamflow, with semi-automatic (Feyen et al., 2000; El-Nasr et al., 2005; Li et al., 2017) or automatic multi-objective calibration procedures (Khu et al., 2008). Lumped conceptual models, with a reduced number of parameters, are easier to calibrate directly without prior determination of the parameters. When a particular state of the model, in general a groundwater reservoir, can be coerced to a measured piezometry time series, calibration using both piezometry and streamflow is possible, generally through a linear composite objective function combining criteria on streamflow and piezometry (Thiéry, 1988; Moore and Bell, 2002; Széles et al., 2020). Despite significant improvements in piezometry simulation, these studies found that adding piezometric information to the calibration process did not significantly impact streamflow simulation.

1.4 Scope of the paper

In view of the undisputed role of aquifers in low-flow dynamics in many catchments, it seems reasonable to try to improve the performance of a hydrological model by adding piezometric data to the calibration process. However, most of the approaches reviewed in the previous section are difficult to implement due to a relatively large number of parameters and because the performance gain offered by the new data has not been assessed on a large set of catchments, which is necessary for model evaluation (Barthel and Banzhaf, 2015).

In this study, we aim to develop a new modelling approach based on a simple structure with an easy parametrisation, assessed on a large sample of catchments to ensure the generality of conclusions. We propose an adaptation of the structure of the conceptual daily rainfall–runoff model GR6J (Pushpalatha et al., 2011) to make it simulate groundwater table levels. Since no element of the existing model structure was designed to explicitly simulate groundwater level ~~, an~~ and because of the huge scale gap between point piezometric measurements and aquifer-scale storage volumes, we could not propose a physically explicit solution; thus, we investigated an empirical adaptation of the model structure~~is necessary~~. Section 2 recounts the process that led to designing this adaptation and the calibration and evaluation schemes of the new model. Section 3 presents the hydroclimatic dataset of 107 catchments over mainland France that was used to evaluate the new calibration with respect to the original one, performed only on streamflow. Section 4 summarises the results and proposes recommendations for model calibration in various contexts.

2 Hydroclimatic dataset

2.1 Context

The French mainland territory hosts a large diversity of climatic, topographic and geological contexts, with catchments representing various hydrological and hydrogeological configurations. Several major aquifers are known to have a significant influence on surface waters, especially on low flows. The Paris basin, with its pile of secondary and tertiary sedimentary formations, hosts several major aquifers for surface hydrology: the Late Cretaceous chalk aquifer is known to govern the multiyear dynamics of the Somme and part of the Seine and Loire basins, with a noteworthy long flood event after the exceptionally wet years of 1999 and 2000 (Pinault et al., 2005; Habets et al., 2010); the Beauce tertiary limestone aquifer controls the hydrology of a key agricultural region astride the Loire and the Seine basin, with a major groundwater contribution to low flows (Lalot et al., 2015); the Cenomanian sand aquifer in the Perche region, which is directly connected to the Eure and Huisne basins, is regarded as an essential groundwater reserve for the region and its declining trend is a major threat for Perche rivers (Lenhardt et al., 2009). The second largest French sedimentary basin, the Aquitaine Basin, has a more complex configuration with thick multi-layer aquifers covered by poorly permeable formations, such as Pyrenean molasses. The outcropping areas of these formations are visible in figure 1.

The large sedimentary basins are not the only geological areas in France which host aquifers that are of interest for surface hydrology modelling. Aquifers located in alluvial plains, such as the international Rhineland aquifer – and its French part in

the Alsace plain quaternary alluvium – or the Bresse graben gravels, play a major role in the streamflow dynamics of the Saone and the Rhine basins. As highlighted in the Introduction, even small alluvial aquifers outside plains can have an influence on rivers: for example, several small left-bank tributaries of the Rhone are mostly ruled by the Bièvre moraine aquifer, with visible consequences on water quality (Bel et al., 1999). Regions in which geological formations are composed of metamorphic or igneous rocks, such as Brittany or the Ardennes, can host fractured bedrock aquifers, linked to surface rivers. The wide monitoring network of rivers and groundwater in France, described below, allowed us to select a test dataset of catchments which is representative of this diversity.

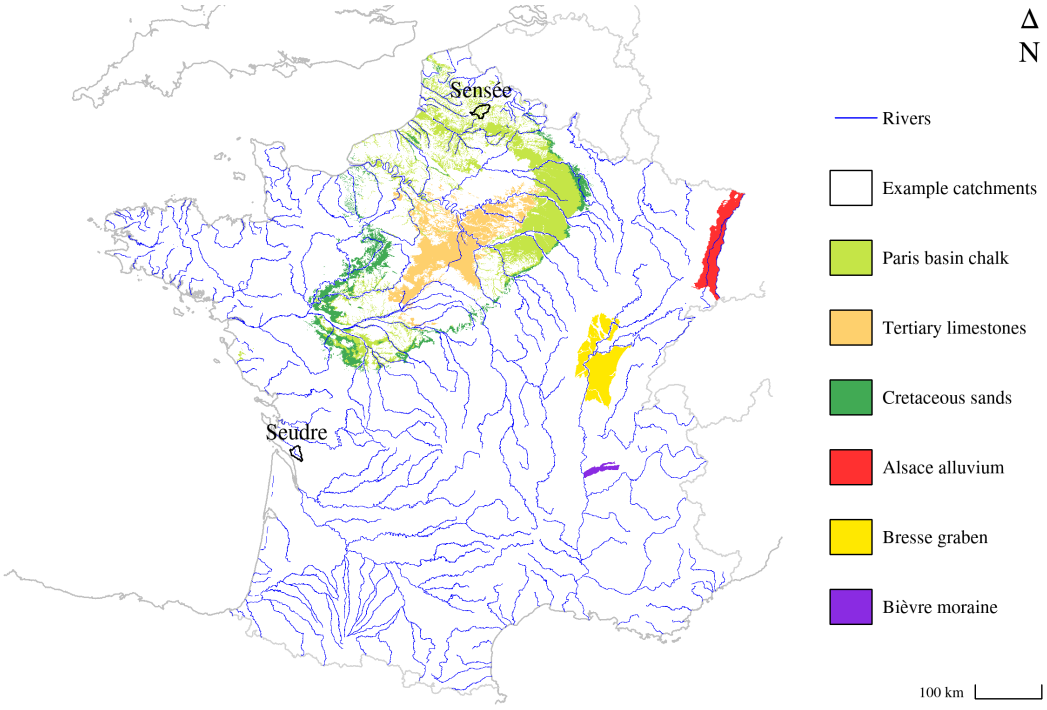

Figure 1. Outcropping areas of several major aquifers in mainland France and locations of example catchments shown in figure 2.

2.2 Data sources

Climatic data – daily cumulative precipitation, average temperature and fraction of solid precipitation – were taken from the SAFRAN (*Système d'Analyse Fournissant des Renseignements Adaptés à la Nivologie*) re-analysis (Vidal et al., 2009) by Météo France; they are available at a daily time step for the 1958–2018 period. Daily potential evaporation was computed using the formula by Oudin et al. (2005). Streamflow data were retrieved from the French national database Banque Hydro (Leleu et al., 2014; SCHAPI, 2021). These hydroclimatic data are aggregated at the catchment scale and at a daily time step for mainland France in the HydroSafran database (Delaigue et al., 2021), maintained by INRAE (*Institut national de recherche pour l'agriculture, l'alimentation et l'environnement*).

Groundwater level data are from the French national database ADES (BRGM, 2021) (*Accès aux données sur les eaux souterraines*), which gathers piezometric data from many providers in the French territory. Selected piezometers were taken from two reference networks, to ensure the quality of data: RNESOUPMOBRGM (national quantitative monitoring network managed by BRGM, the French national geological survey) and RNESP (heritage national network for groundwater monitoring).

2.3 Dataset selection

Catchments were selected on the basis of data availability criteria, ~~exposed~~ shown below, and an analysis of the hydrogeological context through he French national reference cartography of hydrogeological formations BDLISA (Brugeron et al., 2018) (*Base de données des limites des systèmes aquifères*). For each catchment, one or several piezometers were chosen, assessing the connection of the monitored aquifers with surface water bodies. First, using the provided metadata, each piezometer extracted from the ADES database was associated with a hydrogeological entity in BDLISA, representing an aquifer. Catchments in which anthropogenic activities – dams, major direct withdrawals or inflows – are known to have a significant influence on streamflow and catchments in which more than 10% of precipitation falls as snow were discarded. Then, for each catchment, piezometers associated with aquifers emerging inside the catchment boundaries were listed and maps – see an example in figure 2 – were produced to assess the importance of each hydrogeological formation for the catchment. Piezometers associated with formations with outcropping or sub-outcropping areas representing less than 5 % of the catchment area were discarded, along with those located on the wrong side of underground watersheds ~~,~~ – i.e. where groundwater does not flow to the catchment outlet but to another catchment – as identified by BDLISA.

After this spatial selection, the available groundwater level and streamflow data were examined. An initial visual inspection of the time series was performed to eliminate data of too low quality, relying on the expertise of database maintainers. Then, catchments and piezometers were selected according to the following criteria :

- At least 20 years of continuously available streamflow data with less than 10 % of missing data;
- At least 20 years of continuously available groundwater level data with less than 10 % of missing data;
- At least 10 years of continuous contemporaneity between streamflow and groundwater level.

Figure 2 shows two situations encountered at this stage: on the left, the Sensée river is connected to one monitored aquifer but three piezometers are available. In this case, the piezometer with the longest time series, with respect to contemporaneity with streamflow data, was selected to represent the aquifer. On the right, the Seudre river is connected to two monitored aquifers with one piezometer for each one; in that case, the two piezometers are kept. The choice of keeping only one piezometer per aquifer in the catchment was made for the sake of simplicity; in most catchments, when visually comparing the dynamics of the groundwater level time series, no major difference was encountered between piezometers monitoring the same aquifer within the same catchment. A correlation study between groundwater level time series led to the same conclusions.

Finally, this selection process yielded to a set of 107 catchments and 160 piezometer/catchment pairs. The majority of catchments – 73 – are associated with only one piezometer; 22 of them with two; eight of them with three; one of them with

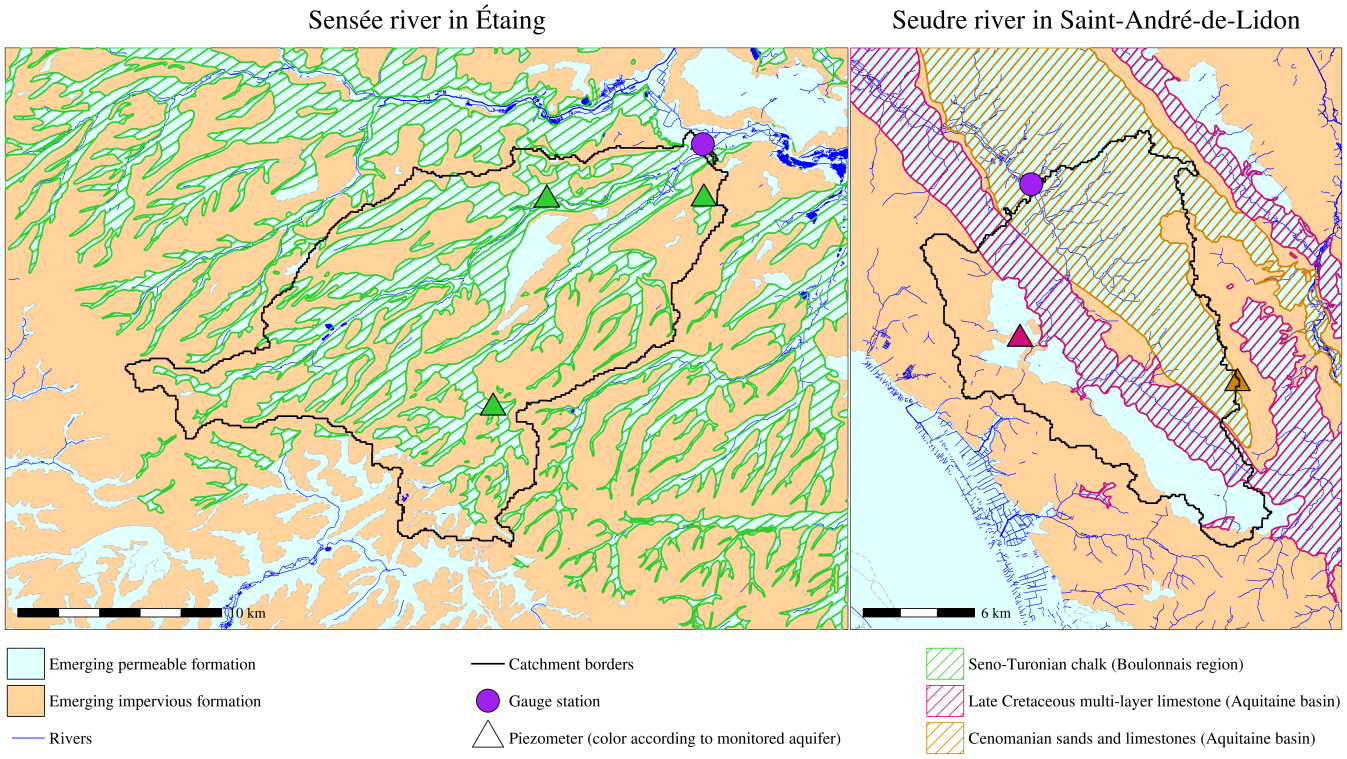

Figure 2. Two examples of hydrogeological maps of catchments used for dataset selection. Their location in the French mainland territory is shown in figure 1.

Table 1. Geographical characteristics of the 107 catchments dataset

	Catchment area (km^2)	**Outlet altitude (m)**	**Mean altitude (m)**	**Maximum altitude (m)**
Minimum	27.0	0	39	65
1st quarter	168.2	27	113	169
Median	326.0	62	136	236
Mean	617.0	81	175	236
3rd quarter	685.7	114	202	330
Maximum	7,907	367	667	1421

four and three of them with five piezometers. Tables 1 and 2 show geographical and hydrogeological characteristics of the set. The necessity to choose catchments which are not anthropogenically regulated led to a selection mainly composed of small headwater catchments, representative of the climatic diversity of the French territory. Dismissing the mountainous catchments

Table 2. Hydrological characteristics of the catchment dataset. The aridity index is defined as the quotient of annual rainfall and annual potential evaporation (PET). Catchment yield is the quotient of annual streamflow and annual rainfall.

	Mean annual streamflow (mm)	Mean annual rainfall (mm)	Mean annual potential evaporation (mm)	Catchment yield (%)	Aridity index
Minimum	29	626	600	4.6	0.90
1st quarter	146	723	638	21	1.09
Median	210	808	658	27	1.18
Mean	238	828	667	28	1.25
3rd quarter	316	921	693	34	1.40
Maximum	795	1,413	792	56	2.35

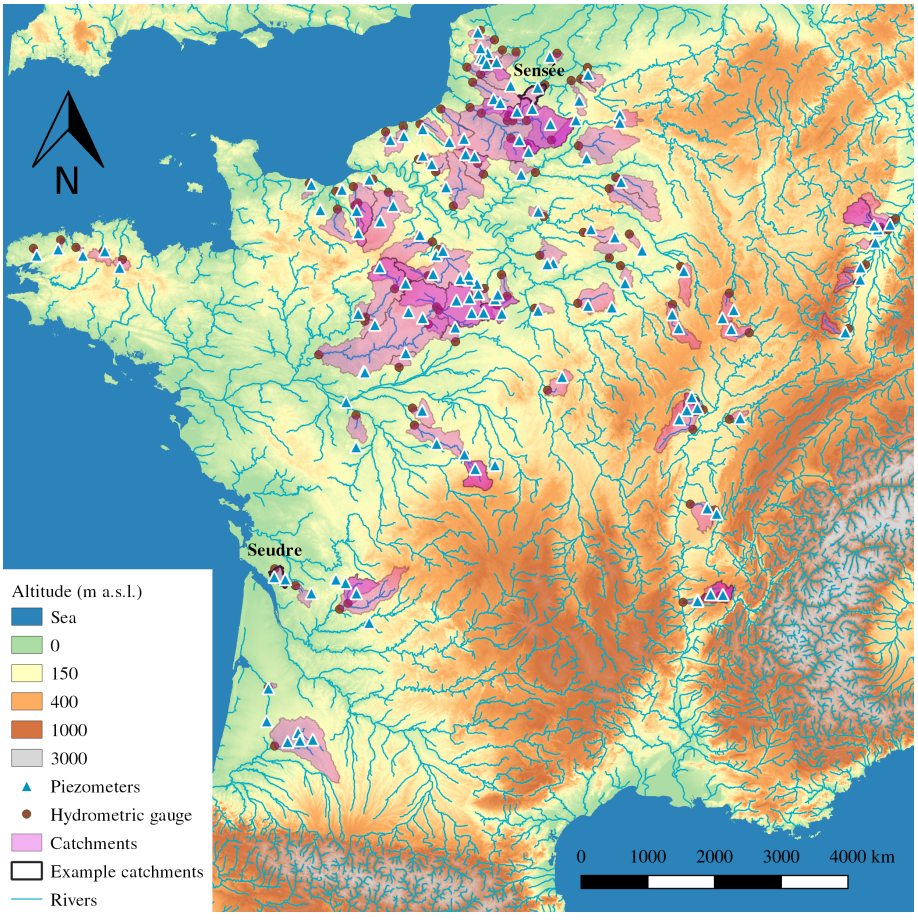

Figure 3. Map of the catchment and piezometer dataset. Example catchments of figure 2 are shown.

to avoid the influence of snow favoured the selection of lowland catchments, although several Vosges catchments, whose downstream part is linked to the Alsace plain aquifer, reach maximum altitudes above 1,000 m. However, the average altitude remains low enough for the solid precipitation fraction not to overtake 10 ~~%of solid precipitation~~ % of total precipitation at the catchment scale. The variability of the mean annual potential evaporation is low, since the dataset does not contain ~~neither~~ high-altitude catchments – in which low yearly PET values are observed – ~~nor~~ or catchments located in the South-East of France – where the highest PET values are reached (Brigode et al., 2021).

Figure 3 shows a map of the selected catchments and piezometers. The northern part of mainland France, especially the Paris basin, is over-represented because of data availability; in particular, the chalk and tertiary limestone aquifers in this basin are the hydrogeological formations that have been monitored for the longest time in the territory. However, attention was paid to represent the diversity of hydrogeological contexts, with smaller local aquifers or fractured bedrock aquifers, in order to assess the proposed modelling approach in the widest possible range of configurations.

3 Methodology

3.1 Presentation of the original GR6J model

3.1.1 General presentation

GR6J – for *modèle du Génie Rural à 6 paramètres Journalier* – is a daily six-parameter rainfall–runoff model. It was developed by Pushpalatha et al. (2011), as an evolution of previous GR4J (Perrin et al., 2003) and GR5J (Le Moine, 2008) versions, using a conceptual description of the hydrological processes taking place in the catchment: the model structure, visible in black in figure 4, is composed of stores, unit hydrographs and empirical equations that link them. The model is lumped and operates at a daily time step, taking as inputs precipitation P and potential evaporation E, averaged on the time step and the spatial extent of the catchment. ~~Potential evaporation is computed using the formula by Oudin et al. (2005).~~ GR6J is also parametric, i.e. for each catchment, 6 independent parameters have to be identified. All variables and parameters are expressed either as water depth, in millimetres, or are unitless.

This section does not intend to report the modelling tests that have led to the development of the GR6J structure ~~,~~ since the original paper by Michel (1983) – such discussions can be found in Perrin et al. (2003), Le Moine (2008) or Pushpalatha et al. (2011). A summary description of the model computations is available in appendix A; a table of variables is available in appendix C. Computing codes can be found in the open-source `airGR` package (Coron et al., 2017, 2021) available in R (Slater et al., 2019; R Core Team, 2021).

3.1.2 Parametrisation strategy

For each catchment, the model is calibrated to fit measured streamflow: the six parameters are determined through an optimisation process, by minimising an error criterion between measured and simulated streamflow, in a reference period. ~~Commonly used criteria are the Nash–Sutcliffe efficiency or NSE (Nash and Sutcliffe, 1970), the Kling–Gupta efficiency or~~

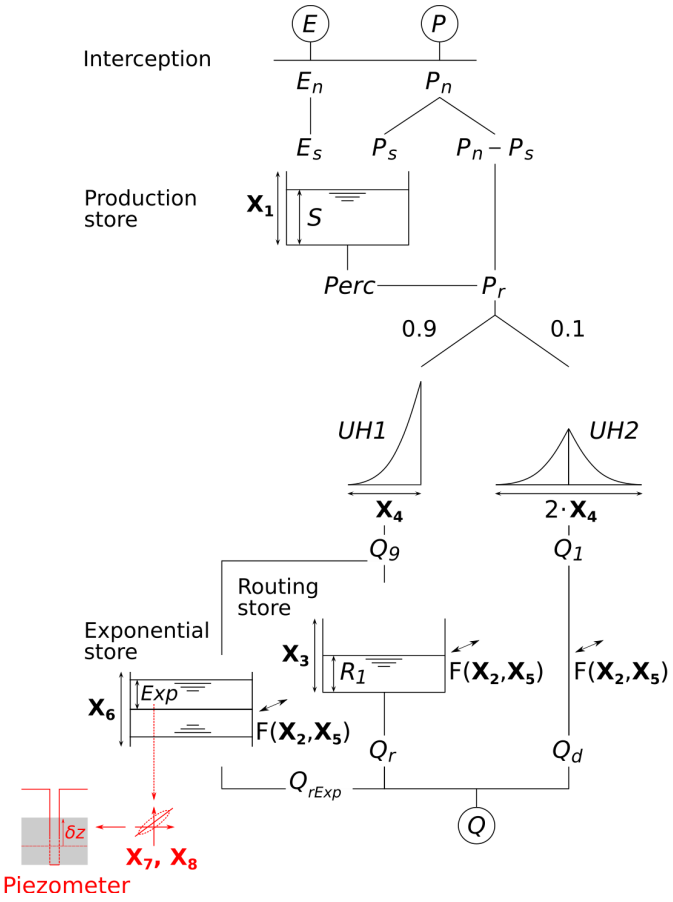

Figure 4. Structure of the original GR6J model – in black – and built-in piezometry simulation module – in red. See table C for the meaning of parameters and variables.

~~KGE (Gupta et al., 2009; Kling et al., 2012) and the root mean square error (RMSE).~~ In this study, the Nash–Sutcliffe efficiency or NSE (Nash and Sutcliffe, 1970) was used for streamflow.

Since the six parameters have very different dimensions and variation ranges, each of them is transformed with a bijective function to fit into the $[-9.99; 9.99]$ interval. Thereby, the optimisation space for optimal parameter research becomes $[-9.99; 9.99]^6$, which helps most optimisation algorithms find the global optimum. Detailed transformations and ranges are available in appendix B. Several optimisation algorithms are used to calibrate the GR6J model, examples can be found in Coron et al. (2021).

3.2 Study of the model correlation with piezometry

To adapt the existing model structure for groundwater level simulation, we followed an approach similar to other lumped conceptual models, i.e. using a store as a representation of the aquifer, regarding the water content in the store as a proxy

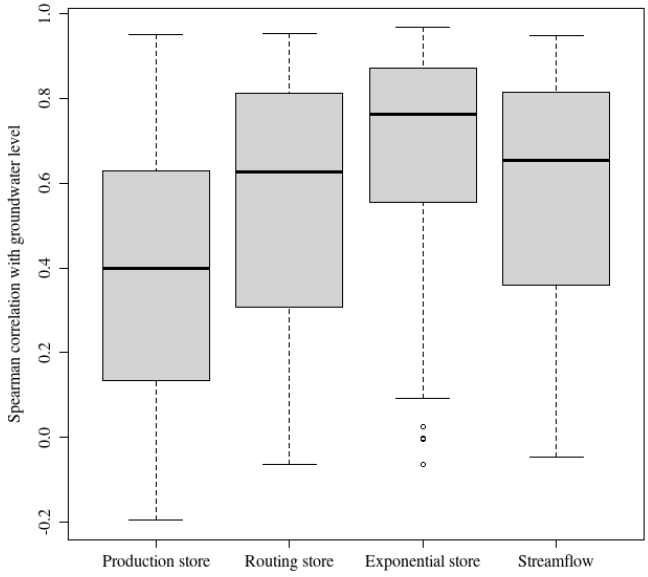

Figure 5. Distributions of correlations between piezometry and several states of the model

for groundwater level. With this aim in mind, a correlation study was performed on the dataset presented in section 2, in order to identify which of the three conceptual reservoirs of the model structure was the most correlated with piezometry. The production store is part of the production function, which computes the balance between rainfall and evaporation to determine the amount of water available for streamflow~~; the~~ . The routing and the exponential stores are part of the routing function, which models the time repartition of this available water to simulate streamflow.

GR6J was calibrated for the 107 catchments of the dataset using the Nash–Sutcliffe efficiency criterion computed on the square root of streamflow. The algorithm by Michel (1991), as implemented in `airGR` R package (Coron et al., 2021; R Core Team, 2021), was used on the whole period of available climatic data (1958–2018). Then, the time series of model states obtained – the levels of the three conceptual stores and simulated streamflow as control data – and the groundwater level time series were aggregated at a monthly time step, to avoid problems caused by missing piezometry measurements. Afterwards, for each of the 160 catchment/piezometer pairs, Spearman's correlation (Spearman, 1907) between piezometry and each state series was computed; the results are summarised as boxplots in figure 5. The exponential store (Michel et al., 2003) – see figure 4 for a description of the model – is the most correlated with piezometry and, moreover, it is the only store to be more correlated with groundwater level than with streamflow. The median correlation obtained is 0.762 and 80% of pairs reach a value higher than 0.5.

A high Spearman correlation may highlight a non-linear relationship, since it is a rank correlation. However, it does not seem to be the case here: other investigations not detailed here show that the relationship between the exponential store content and

the groundwater level can be regarded as linear, all the more so as the correlation is high. Therefore, it was decided to use the exponential store to simulate piezometry, with an adapted scheme presented in the following section.

3.3 Adaptation of the model scheme

A built-in module is added to the existing model structure to simulate groundwater level. The streamflow simulation chain is not modified, but a new output is added to the model, through a linear transformation of the exponential store level.

Groundwater level absolute values strongly depend on the piezometer location – its altitude, but also its position with respect to the catchment topography. Indeed, two piezometers monitoring the same aquifer and therefore representing the same dynamics can have different mean levels and their fluctuations can have different ranges – for instance, if the first one is located on a plateau while the second one is on a slope. To avoid having to take into account these problems, it was decided to work with normalised groundwater level δz, where z is absolute groundwater level, $\bar{z}$ is its mean and σ_z its standard deviation:

$$\delta z = \frac{z - \bar{z}}{\sigma_z} \tag{1}$$

To represent the relationship between the exponential store level Exp – in mm – and simulated normalised groundwater level $\delta_{z,sim}$, several polynomial relationships were investigated. It appeared that using a function of degree 2 or more was not useful to improve performance. Therefore, an affine relationship is added to the model, with two additional parameters X_7 and X_8, using the following equation:

$$\delta_{z,sim} = \frac{1}{X_7}\left(\frac{Exp}{X_6} + X_8\right) \tag{2}$$

Simulated piezometry z_{sim} can be computed by reversing equation 1:

$$z_{sim} = \sigma_z \delta_{z,sim} + \bar{z} \tag{3}$$

X_7 is the *groundwater linear coefficient*; trials have shown that it generally takes values between 0 and 1 but can reach 4. X_8 is called the *groundwater linear offset* and takes non-negative values, with an upper bound at 20. The new built-in module is shown in red in figure 4.

3.4 Composite calibration strategy

Now that two additional parameters have been added to the model structure to simulate piezometry, it is necessary to determine their value through an adapted parametrisation strategy. A composite objective function is chosen for calibration, using a linear combination of a criterion on streamflow – the Nash–Sutcliffe efficiency computed on the square root of streamflow – and a criterion on piezometry, called $ZError$ and defined as:

$$ZError = 1 - \sum_t \left(\delta_{z,sim}(t) - \delta_{z,obs}(t)\right)^2 \quad (4)$$

Computations detailed in appendix D show that this criterion is in fact Nash-Sutcliffe efficiency, expressed for groundwater level instead of streamflow. Since the two criteria on streamflow and piezometry have the same variation ranges – $]-\infty;1]$ – and the same properties, the objective function C for composite calibration can be taken as a linear combination of the two criteria, with a weight α:

$$C(\alpha) = \alpha\ ZError + (1-\alpha)\ NSE \quad (5)$$

α can take any value between 0 and 1: $\alpha = 0$ means that the calibration is performed only on streamflow and $\alpha = 1$ only on piezometry. In order to find a compromise between these two objectives, 51 values are explored from 0 to 1 by a step of 0.02. For each value of α, $C(\alpha)$ is maximised as a function of eight parameters. The parameter space transformations described in appendix B are used to convert the optimisation space into the hypercube $[-9.99;9.99]^8$. The differential evolution global optimisation algorithm – implemented in the `RcppDE` R package (Price et al., 2006; Mullen et al., 2011; Ardia et al., 2011a, b; Eddelbuettel, 2018; Slater et al., 2019; Ardia et al., 2020; R Core Team, 2021) – is then executed to find the global optimal point for the eight parameters.

3.5 Split-sample test evaluation scheme

To assess the effect on streamflow simulation performance of the new calibration scheme described above, a split-sample test (Klemeš, 1986) is conducted for each catchment/piezometer pair of the assessment dataset described in section 2. For each pair, the available data are divided into two time periods P_1 and P_2 of equal length, defined so as to encompass the same number of data points for which both groundwater level and streamflow are available. Thereby, both periods contain the same amount of information and can be equally used for calibration and validation. The exact duration of periods depends on the pair, since the data availability time periods are diverse: durations spread from 5.6 to 28.5 years by period. Before each period, a warm-up timespan of 5 years is set: the model is run on this period but the resulting simulated values are not used to compute criteria.

After determining these periods, the adapted model structure is calibrated on P_1 using $C(\alpha)$, for each value of α; the parameter set obtained is then used to run the model on P_2 and compute several validation criteria. Then, the periods are switched and the same procedure is executed. The following validation criteria are used:

- $NSE(\sqrt{Q})$ to evaluate the model performance on the whole streamflow spectrum;
- $NSE(\sqrt[3]{Q})$ to evaluate the model performance on low-flows. It was preferred to zero-diverging transformations such as $\frac{1}{Q}$ or $\log(Q)$ to avoid numerical problems with very low streamflow values;
- $ZError$ to assess the model performance in groundwater level simulation.

Since the evaluation is performed for validation, the results presented in section 4 are, unless otherwise specified, validation results.

To assess the benefit of using groundwater level data in the calibration process, the distributions of evaluation criteria values need to be compared to reference ones. For streamflow, the value $\alpha = 0$ corresponds to the original calibration framework, only performed on observed streamflow data. Parameters X_7 and X_8 are only used to simulate normalised groundwater level and therefore, when $\alpha = 0$, the ~~sensibility~~ sensitivity of the calibration criterion to their values is zero. Thus, they are randomly determined by the stochastic optimisation algorithm and no relevant normalised groundwater level is simulated, except a random affine transformation of the exponential store level which cannot be compared to observed data. Therefore, another reference distribution than the one obtained for $\alpha = 0$ is needed to evaluate groundwater level simulation. The value $\alpha = 1$ is used, since it is the case in which the model is calibrated only with observed groundwater level data and no streamflow; we thus expect the best theoretically possible groundwater level simulation performance for this value of α.

The differences between evaluation criteria distributions are evaluated visually and then, in order to objectify them, a Wilcoxon–Mann–Whitney test (Wilcoxon, 1945; Mann and Whitney, 1947; Bauer, 1972) is conducted. The distributions obtained for the values of α are compared with the reference ones: $\alpha = 0$ for NSE; $\alpha = 1$ for ZError. Therefore, for each value of α, two tests are conducted: one to assess whether the streamflow simulation performance has significantly deteriorated and one to evaluate whether the performance of groundwater level simulation is significantly lower than the one obtained for $\alpha = 1$.

To assess the influence of the geological context, the test dataset of 160 catchment/piezometer pairs was divided into six groups, detailed in table 3. The groups were established in accordance with the hydrogeological formation attributed to each piezometer, in the BDLISA reference inventory by Brugeron et al. (2018). This classification may look arbitrary or inaccurate, since each piezometer corresponds to an idiosyncratic local situation; however, such a subgroup analysis of the test dataset highlights the influence of geology on the model performance~~, as seen in figure 17~~.

Table 3. Groups of catchment/piezometer pairs, gathered by geological context

Number	Number of pairs	Description
1	26	Quaternary alluvia
2	11	Bedrock and Triassic sandstones
3	72	Chalk and Cretaceous limestones
4	20	Paleogene and Neogene limestones
5	19	Jurassic limestones
6	12	Cretaceous sands

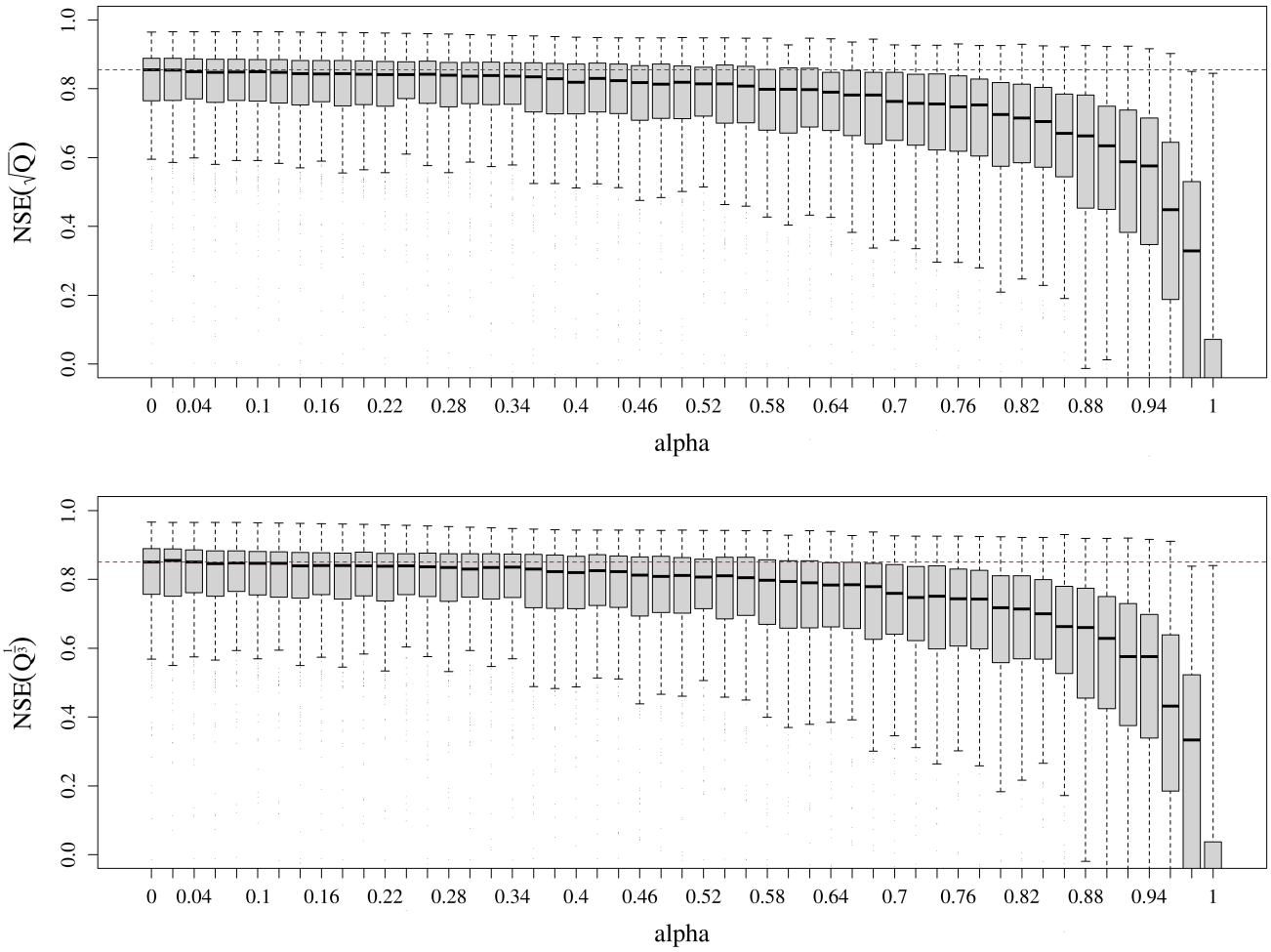

Figure 6. Distributions of the NSE criterion values, obtained in validation by the modified model on the 107 catchments, for the 51 values of α, the criterion weight. The red dashed line indicates the median NSE value for the original calibration strategy $\alpha = 0$. Values below 0 were cut off for readability.

4 Results and discussion

4.1 Is low-flow simulation improved?

Model performance for streamflow simulation in validation is not improved by the proposed calibration scheme. Figure 6 shows that the distribution of Nash–Sutcliffe efficiency computed on the square root of streamflow does not appear to change significantly for values of α under 0.34; for higher values, the performances deteriorate, but it is surprising to note that they slowly decrease while increasing α and they remain acceptable until $\alpha = 0.84$ – even though the loss of about 0.2 is signifi-

cant. Beyond these values, the performances have considerably deteriorated: the calibration cannot be suitably performed on groundwater level time series only; this is an expected result, since the dynamics of groundwater level and streamflow signal are different. The same trend is observed with performances in low flows, assessed through Nash–Sutcliffe efficiency computed on the cubic root of streamflow.

4.2 Is the model able to simulate groundwater levels?

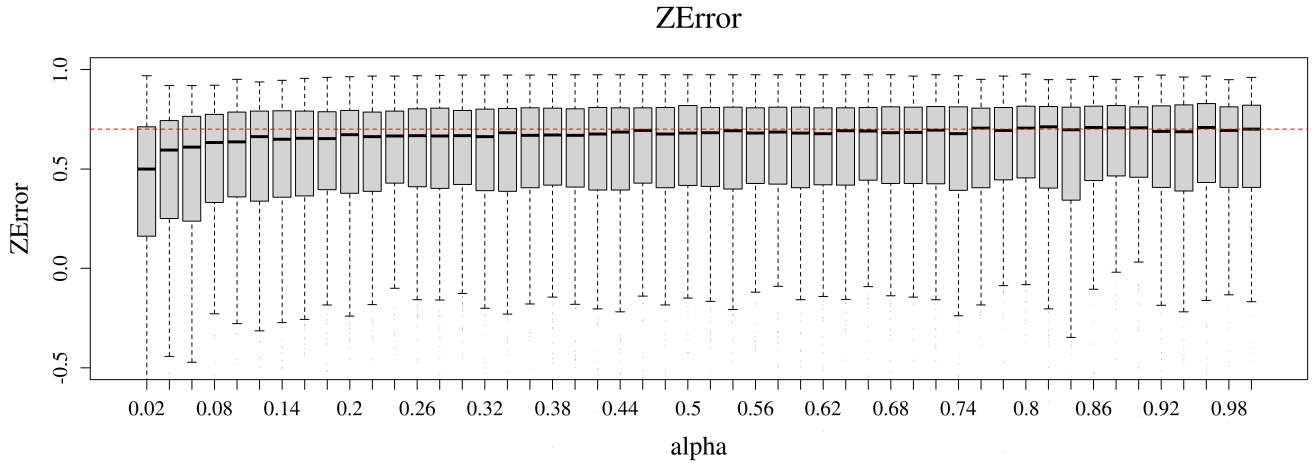

Figure 7. Distributions of the ZError criterion values obtained, for 50 values of α, the criterion weight. $\alpha = 0$ was discarded since no groundwater level simulation is performed in that case. The red dashed line indicates the median ZError value obtained with $\alpha = 1$. Values below -0.5 were cut off for readability.

The model appears to be able to simulate groundwater levels with a satisfactory performance. Figure 7 shows the distributions of the ZError criterion for the 51 values of α, compared to the theoretically maximum possible performance which is obtained with $\alpha = 1$ – i.e. a calibration performed only on groundwater level with no streamflow information. The distribution of ZError values appears to be similar for all α values above 0.34, with a median ZError around 0.70. For α between 0.12 and 0.34, the performance decreases slightly but remains close to the best possible performance, with median ZError around 0.66. Finally, for α under 0.1, with very little groundwater level information added to the calibration process, the performance is much lower, but even for $\alpha = 0.02$, it is acceptable, with a median ZError around 0.5.

4.3 Recommended calibration framework and examples

Results of the statistical evaluation of differences between performance criteria distributions are presented as p-values in figure 8, with a significance threshold of 5%.

It appears that for values of α greater than 0.22, streamflow simulation performance has significantly deteriorated; for α lower than 0.12, groundwater level simulation performance is significantly below that obtained for higher values of α.

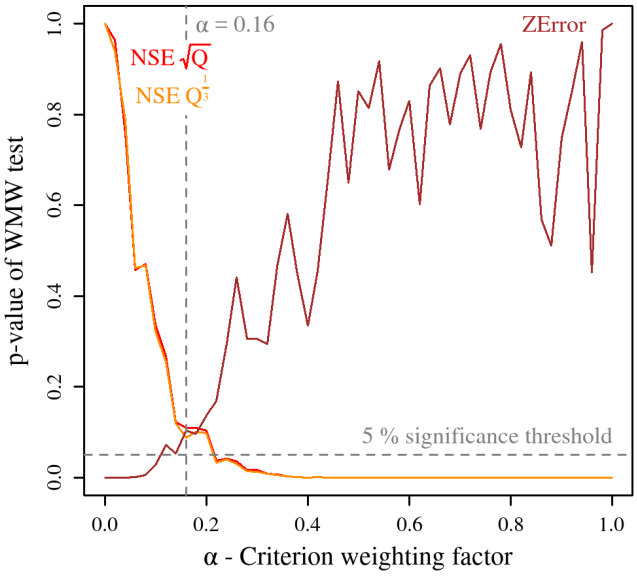

Figure 8. p-values of the Wilcoxon–Mann–Whitney (WMW) tests comparing the criteria value distributions obtained for the 51 values of α with reference ones.

A narrow interval – α between 0.14 and 0.2 – corresponds to values for which the model performance for both outputs is comparable to reference distributions: the model is as good at streamflow simulation as the original model and it cannot be better at groundwater level simulation. Therefore, it was decided to choose the value $\alpha = 0.16$ as the recommended calibration framework, even though any value in the described interval could be chosen without significantly changing the results.

Figures 9 and 10 present an example of the new calibration framework applied to the Sensée catchment in Étaing, in the north of mainland France. The Sensée River is a tributary of the Scheldt (*Escaut*), which is influenced by the Seno-Turonian chalk aquifer – see figure 2. This catchment is an outlier with respect to the model performance distribution, since Nash–Sutcliffe efficiency is improved by 0.14 for the period shown in the figures. However, this large difference is difficult to visualise, since the two simulated hydrographs are close. The ZError obtained is average – 0.630 – and multiyear groundwater dynamics are reproduced, but the model struggles to simulate the peaks of the observed piezometry time series.

Another example catchment is presented in figures 11, 12 and 13: the Seudre River in Saint-André-de-Lidon. It is a small coastal river located in Saintonge, linked to two regional aquifers of the Aquitaine basin – see figure 2: the Cenomanian sands and limestones and the Late Cretaceous multi-layer limestones, each one being monitored by one selected piezometer. Figure 11 shows the results on streamflow: adding piezometry to the calibration process did not significantly improve the performance and the simulated hydrographs are not distinguishable. However, the results for groundwater simulation shown in figures 12 and 13 are satisfactory, with respective ZError values of 0.734 for Cenomanian sands and limestones and 0.787 for Late

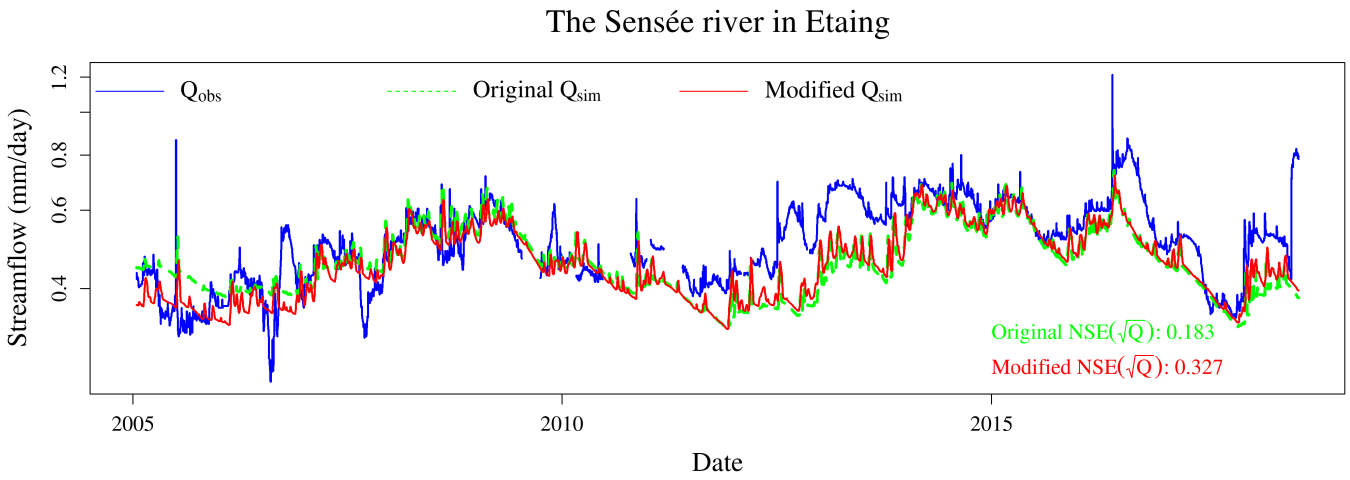

Figure 9. Observed and simulated streamflow of the Sensée River in Étaing, obtained with the original and the new calibration frameworks. Log-scale is used to focus on low flows.

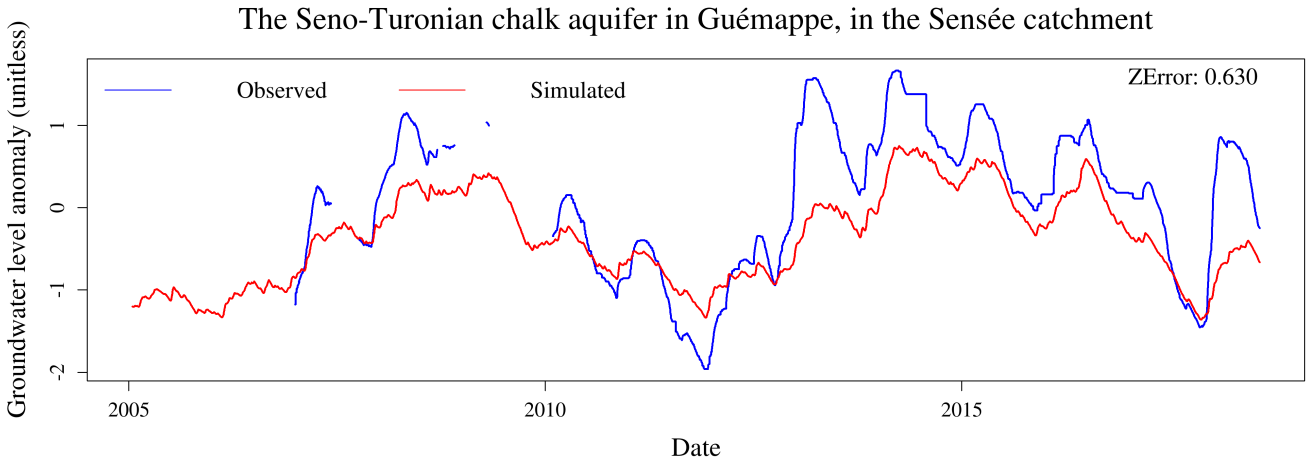

Figure 10. Observed and simulated groundwater level in the Sensée catchment, obtained with the new calibration framework.

Cretaceous limestone. In addition, the main failure period for groundwater level simulation – between 2010 and 2012 – in which piezometry in underestimated, is also unsatisfactory for streamflow simulation, since the model is unable to reproduce the whole variability of the hydrograph during this period.

4.4 Is the new parametrisation stable?

The parametrisation stability between periods is another measure of the robustness of the model: if the parameter values depend on the calibration period, it will cast doubt on the model capacity to extrapolate streamflow values outside this period and thus, to be used, for instance, as a forecasting tool. The split-sample test allows us to assess this stability by comparing the parameter

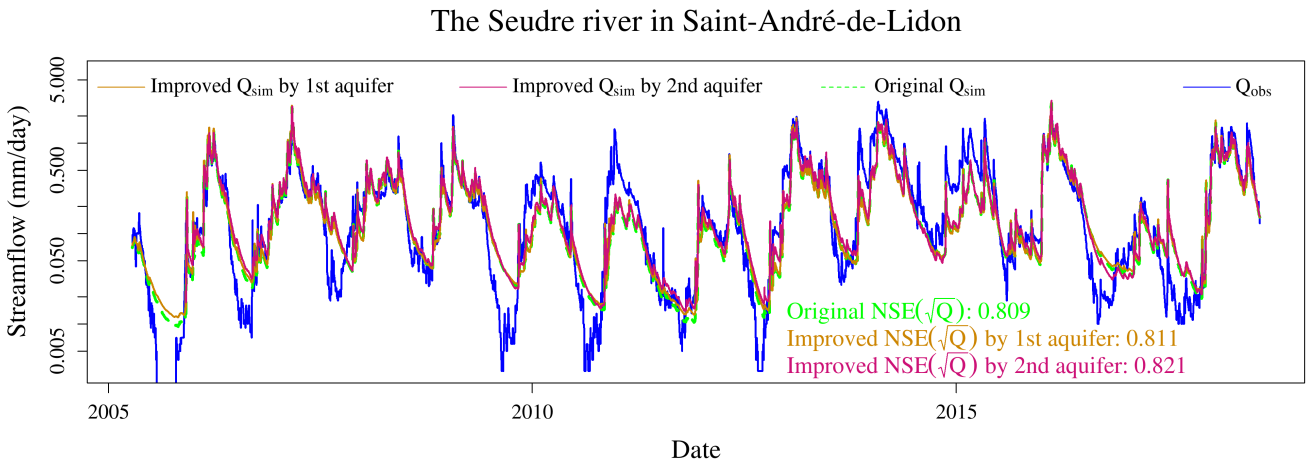

Figure 11. Observed and simulated streamflow of the Seudre River in Saint-André-de-Lidon, obtained with the original and the new calibration frameworks, for the two selected piezometers in the catchment. Log-scale is used to focus on low flows.

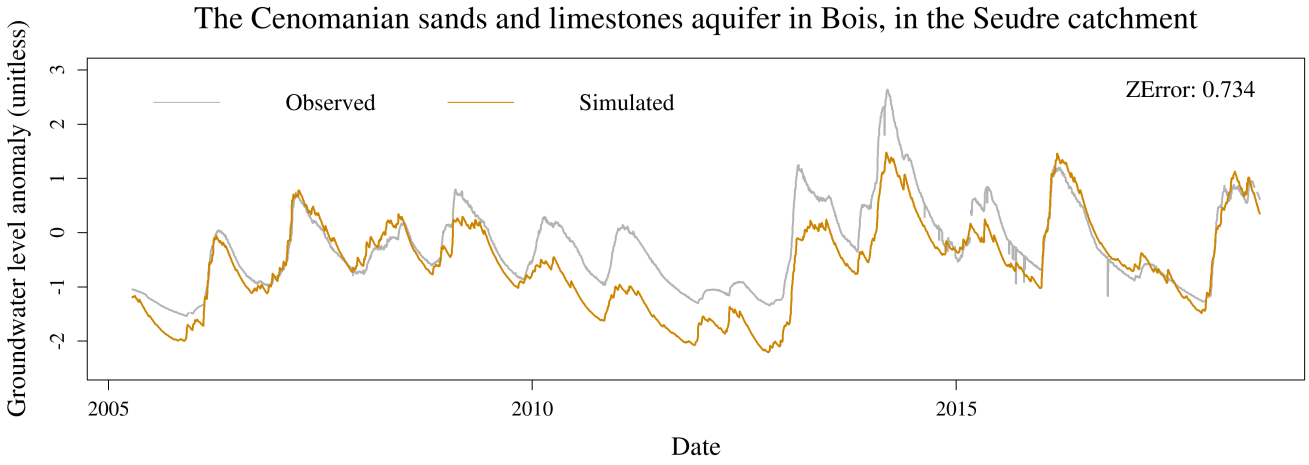

Figure 12. Observed and simulated groundwater level in the Seudre catchment, obtained with the new calibration framework: first piezometer.

values between the two calibration periods P_1 and P_2. Figure 14 shows the results of this comparison for the six parameters of GR6J and the original calibration framework – obtained with $\alpha = 0$; figure 15 does so for the modified calibration, with the two added parameters. For each parameter, the Pearson correlation between the values obtained for the two periods was computed.

425 The original calibration framework leads to a rather stable parametrisation, except for the exchange threshold X_5 with a non-significant correlation between the two periods. The modified calibration, using groundwater level data, yields more stable parameter values, with increased correlations between the two periods, except for the two parameters ruling the inter-catchment exchange function, X_2 and X_5, for which the correlations have slightly deteriorated. The two added parameters, X_7 and X_8,

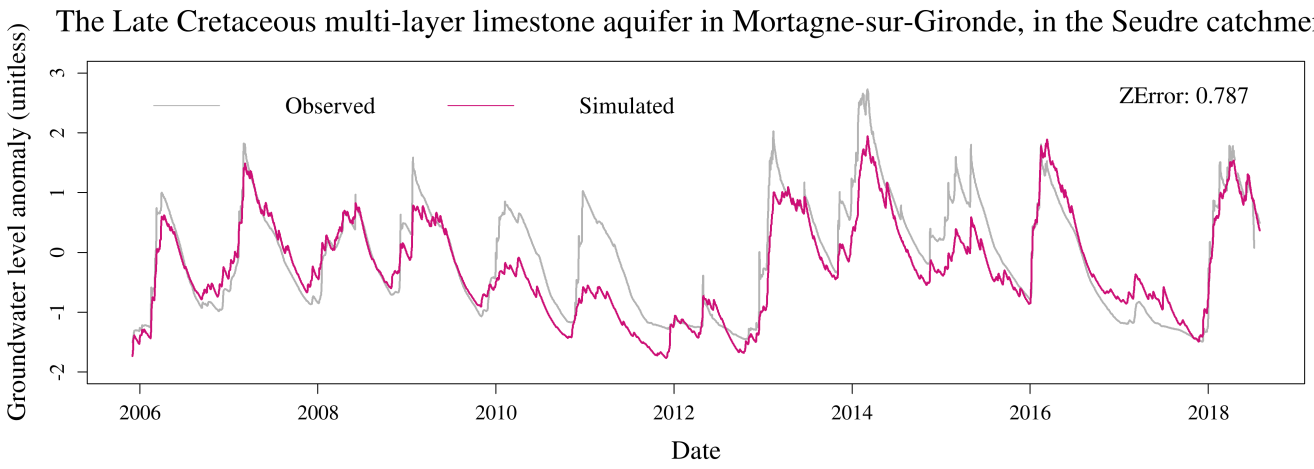

Figure 13. Observed and simulated groundwater level in the Seudre catchment, obtained with the new calibration framework: second piezometer.

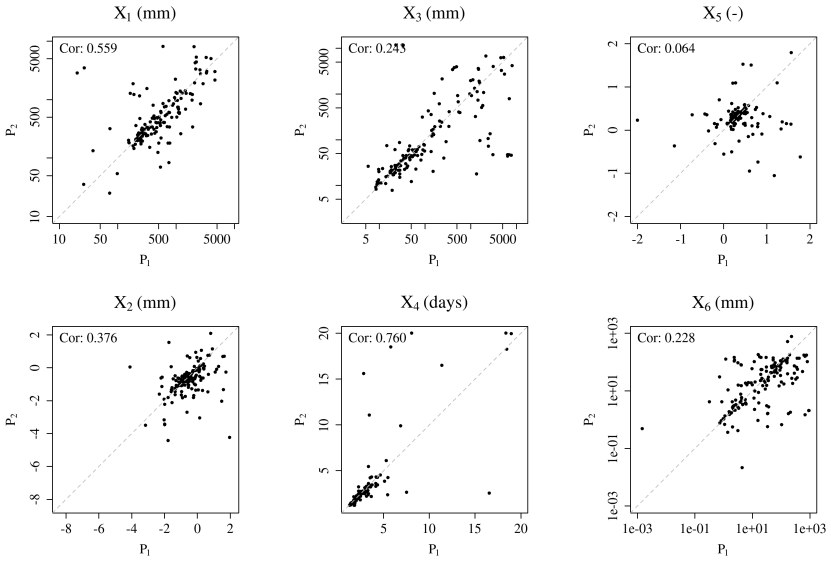

Figure 14. Comparison of the values obtained for the six model parameters through calibration on the two periods of the split-sample test, with the original calibration framework – $\alpha = 0$. Log-scale is used for visual readability of some plots. Pearson correlation between periods is indicated.

are also very stable between periods, with a correlation of 0.73. Since the modified calibration framework is a new constraint on the routing function, it is not surprising to note that the three routing parameters – X_3, X_4 and X_6 – become significantly more stable between periods, which is a sign of an improved model robustness.

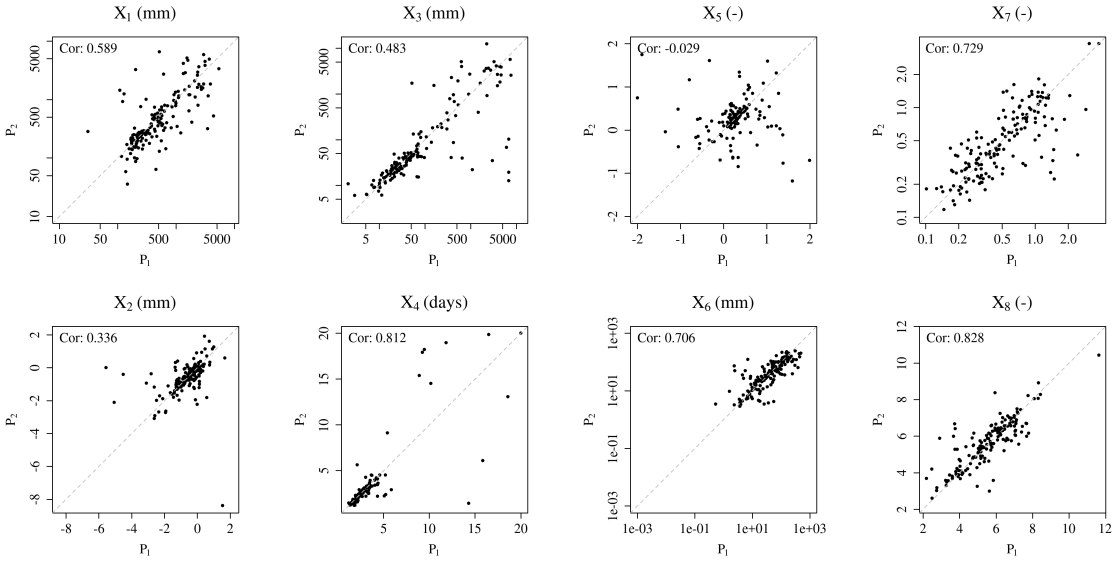

Figure 15. Comparison of the values obtained for the eight model parameters through calibration on the two periods of the split-sample test, with the composite calibration framework – $\alpha = 0.16$. Log-scale is used for visual readability of some plots. Pearson correlation between periods is indicated.

The difficult transferability of the exchange function parameter values, i.e. the relevance of using them while they were calibrated on another period of time, was highlighted by de Lavenne et al. (2016). This function is sometimes regarded as the flux between catchments as they are defined by surface topography, which may not correspond to underground watersheds – for instance, in karstic contexts, see e.g. Le Moine et al. (2008) – or as a representation between the catchment and an externalised aquifer. But it is merely used by the model as a way to correct the global water budget. Poncelet (2016) underlined the relatively marginal role of the exchange threshold X_5, introduced by Le Moine (2008), in the general performance of the model. The stability issues exposed by the present study highlight the need for further development of this exchange function to take into account the henceforth explicit representation of groundwater level through the exponential reservoir.

4.5 Is performance dependent on the regional and (hydro)geological context?

There is no clear spatial pattern in the results shown in figure 16. Since the streamflow simulation performance differences between the original and the composite calibration frameworks are small – and non-significant – the geographical distributions of their performance are similar. High values of performance criteria are noted in the Aquitaine basin, in Brittany, in Upper Champagne and for the downstream tributaries of the Loire River – Maine and Indre basins. Lower values of NSE are found in the Beauce plain, in the Somme basin, in the inland part of the North region – mostly in the Scheldt (*Escaut*) basin – and in the Saone and Rhone basin, with the particular case of the Bièvre morainic plain in which the minimum performance is reached.

Other parts of the Paris basin, the North Sea coastal rivers and the Alsace plain have a mixed situation but generally do not reach the extreme points of the NSE distributions.

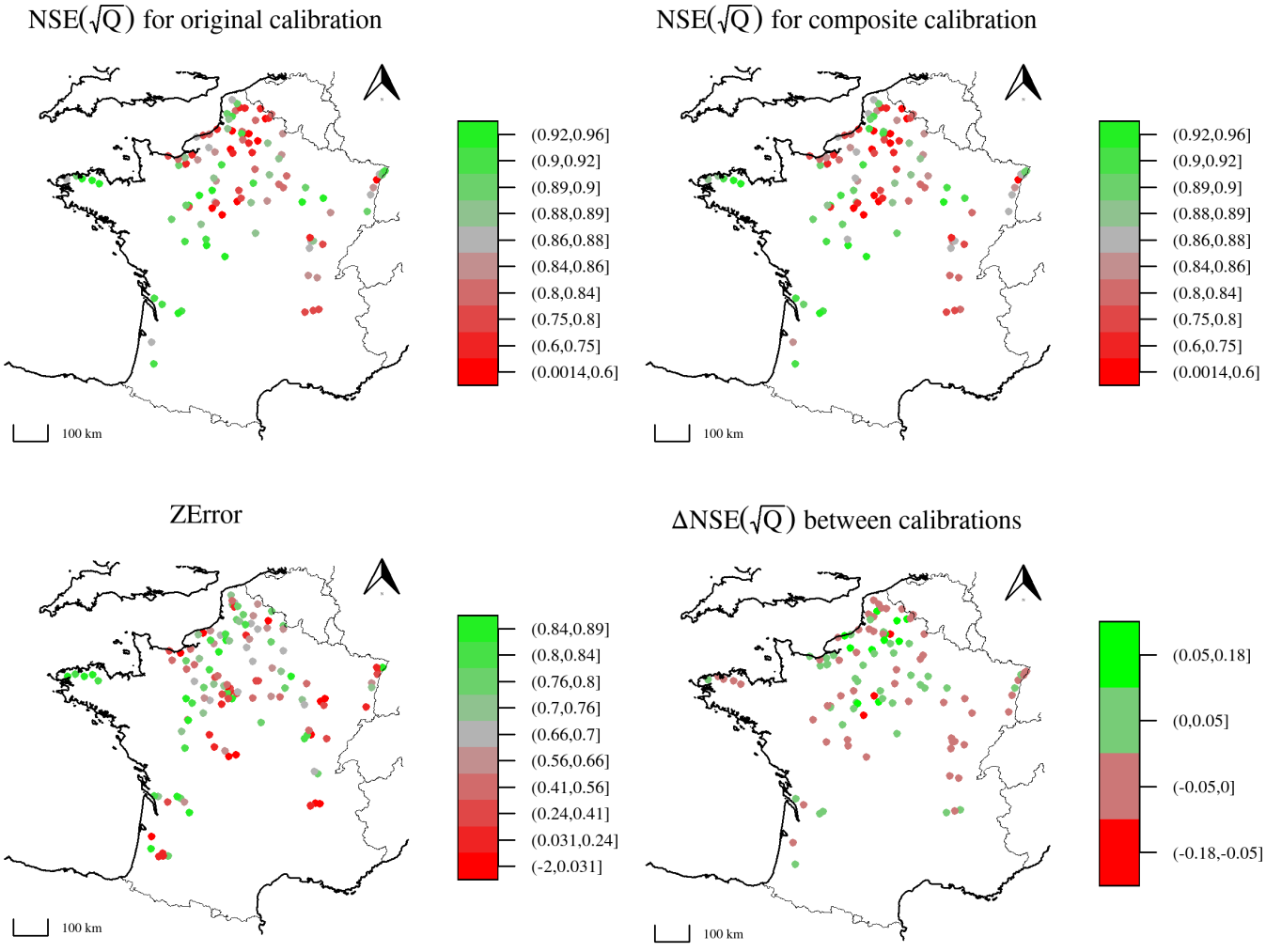

Figure 16. Map of the results. From top-left to bottom-right: value of the Nash–Sutcliffe efficiency at each gauge station, computed on the square root of streamflow, for the original calibration, i.e. $\alpha = 0$; value of the Nash–Sutcliffe efficiency at each gauge station, computed on the square root of streamflow, for the composite calibration, with $\alpha = 0.16$; value of the ZError criterion at each piezometer, for $\alpha = 0.16$; difference between NSE obtained with the composite and the original calibration frameworks. For each point, the maximum value among catchment/piezometer pairs was chosen.

Catchments in which the performance gain between the two calibration frameworks is significant, i.e. beyond 0.05, are all located either on the Picardy and Normandy chalk or in the Beauce plain. It is interesting to note this significant improvement

is observed in catchments in which the initial model performances was low. However, these areas also host catchments for which the composite calibration framework produces a significant deterioration of streamflow simulation performance.

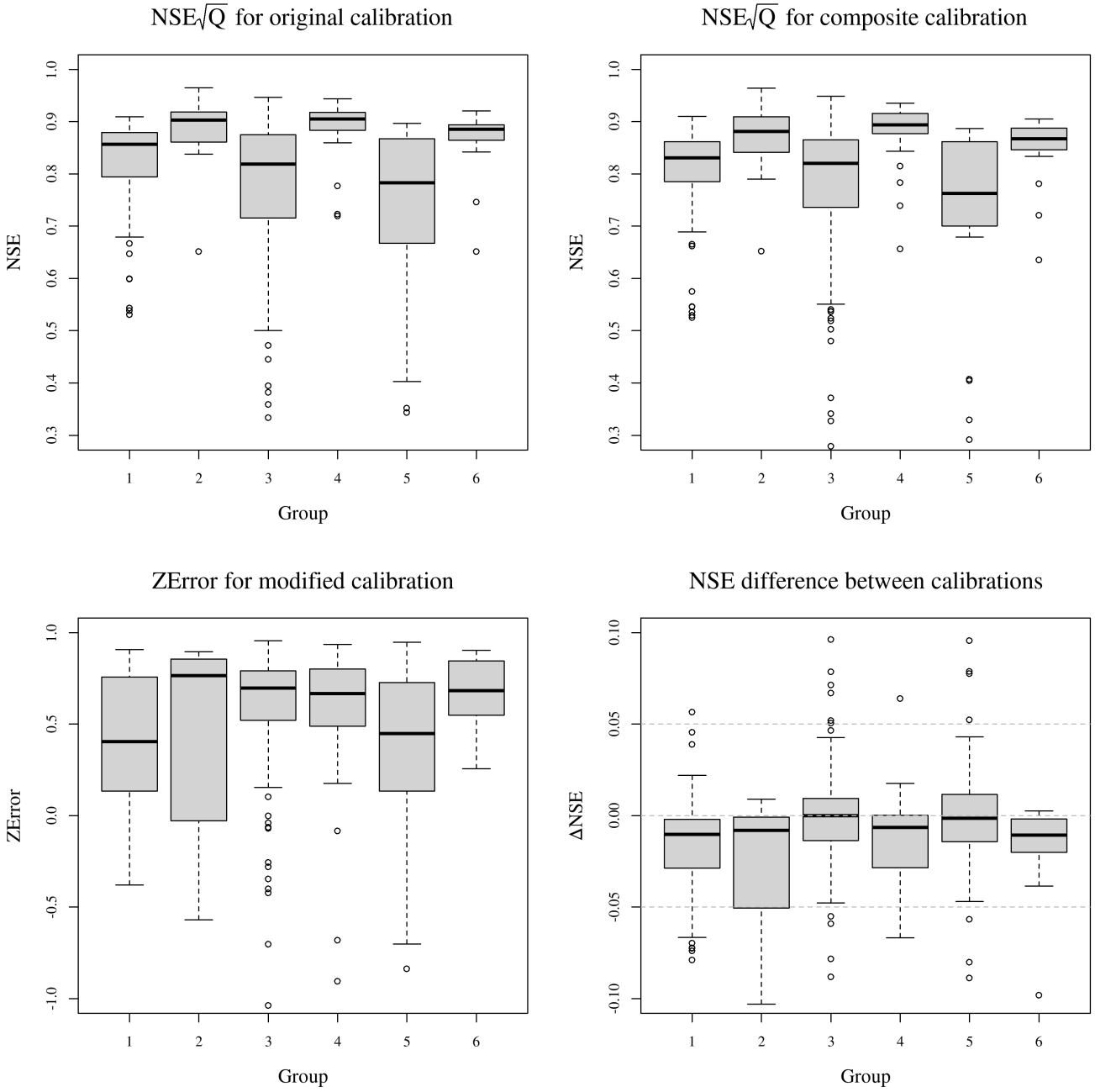

Figure 17. Distributions of results for groups detailed in table 3. From top-left to bottom-right: value of the Nash–Sutcliffe efficiency for each pair, computed on the square root of streamflow, for the original calibration, i.e. $\alpha = 0$; value of the Nash–Sutcliffe efficiency for each pair, computed on the square root of streamflow, for the composite calibration, with $\alpha = 0.16$; value of the ZError criterion for each pair, for $\alpha = 0.16$; difference between NSE obtained with the composite and the original calibration frameworks.

As for the groundwater level simulation, the performance does not follow the same spatial distribution as streamflow. High ZError values are observed in Brittany and in the western part of the Paris basin, along a crescent running from Artois to Touraine. Lower scores are reached in the Bièvre plain, in Upper Champagne and in the extreme south of Paris basin on the Massif Central piedmont. Other regions have mixed results with no clear spatial pattern.

The sub-group analysis yields clearer results, visible on figure 17. For the absolute streamflow simulation performance, the original calibration framework yields high values of NSE for groups 2, 4 and 6, medium ones for group 1 and lower scores for groups 3 and 5. These patterns are found again for the composite calibration framework, even though the distribution of performances for group 5 is narrower. Regarding the difference between the two calibration frameworks, a significant improvement for a small part of the dataset is observed in groups 3 and 5 too, with no deterioration of the median performance in the group, while in groups 1, 2, 4 and 6, the median performance is reduced. A significant decrease in performance is observed for a quarter of group 2 and more than a decile of group 4. As for groundwater level simulation, groups 3, 4 and 6 have narrow distributions centred around a high median score – around 0.7 – while other groups have much wider distributions including simultaneously high, medium and low scores.

The analysis of results between groups of catchments with similar hydrogeological contexts allows to formulate general recommendations to model users, ~~exposed~~ shown in section 5.2. However, the differences between sub-groups have not been successfully linked to hydrogeological characteristics of aquifers, such as permeability or transmissivity. In fact, these data are ~~difficulty~~ difficultly available over the French territory, except in experimental, ~~heavily-instrumented~~ extensively instrumented catchments.

4.6 Are these results model-specific? Or dataset-specific?

The results presented in the previous sections can be seen as disappointing with reference to the objective of the study. Bringing ~~a~~ new information, observed groundwater level data, to the GR6J model yielded no improvement in streamflow simulation performance. Of course, a question arises that we do not wish to avoid in this paper: is this disappointing conclusion model-specific, i.e. is it due to the conceptual nature of the GR6J model? Would a less conceptual and more descriptive model have yielded more satisfactory results? Would a more heavily parameterised model have yielded more satisfactory results? Let us first answer this second question: equifinality is a plague in all modelling efforts, and we would not claim as a success an operation that would consist in improving marginally the situation of a model that was previously impossible to calibrate. Thus, we reject the critique on model complexity as unworthy for a modeller. Concerning the physical realism of the model, no a priori conclusion can be drawn on more physically-based models performance without any empirical evaluation on a large set of catchments. However, the fact that the exponential reservoir – introduced in GR6J structure to represent the slow aquifer transfers – represents either well or very well the dynamics of piezometers on a large catchment set cannot be the sheer consequence of luck. If the piezometric measurements are well represented, both on the calibration and the validation period, this means that our mathematical representation is adequate to describe the underlying physical processes, even without having been designed to do so.

Similar studies performed on other conceptual models did not result in different conclusions: Thiéry (1988) does not mention an improvement in streamflow simulation when calibrating GARDÉNIA with groundwater level data; the study by Moore and Bell (2002) on the PDM model is not conclusive since the new model structure is not compared to a reference one; finally, the calibration framework proposed by Széles et al. (2020) for HBV model gave a similar conclusion to the one of this study: using groundwater level data for calibration helps representing aquifer storage in the model, but did not result in improving streamflow simulation.

Regarding the catchment dataset, we tried to constitute a set of catchments that is the widest possible with respect to our data sources and the selection criteria exposed in section 2.3. A selection bias in the present study is possible, mainly because aquifers regarded as important for surface water resources are the ones that have been monitored for the longest time, in the largest number of measurement points: the chalk aquifer in Picardy, the Alsace plain alluvium of the Beauce tertiary limestones. However, the catchment dataset used in this study is diverse enough to draw general conclusions, at least in climatic and hydrogeological conditions similar to the ones observed in mainland France. Further evaluation on different contexts, in other countries, would help putting our results into perspective.

5 Conclusions

5.1 Synthesis

The study presented here concerns the implementation of a new calibration procedure for an existing streamflow simulation model, GR6J; it is not about the development of a completely new model. For each catchment, among all parameter sets that yield equivalent streamflow simulations, we identified a particular parameter set which is able to simulate, additionally, groundwater level. This new modelling capacity does not induce a significant deterioration in the streamflow simulation performance, neither does it improve it, except in a few particular cases. However, an advantage of the composite calibration framework was highlighted: since we identified a particular parameter set among equivalent sets for streamflow, we probably reduced equifinality in the model calibration, which is suggested by the parameter stability improvement. We may thus expect a more robust model, even if a specific equifinality study would help enforce this conclusion.

The results presented in this paper can be seen as *truly encouraging* – realistic representation of the piezometric variability as one of the states of the model – but scientific honesty requires us to mention that to us they were – at least initially – *truly disappointing*, because we aimed at improving the overall representation of streamflow through inclusion of piezometric information and not the other way around.

5.2 Recommendations to users: which calibration should be used in which context?

The analyses performed in this study lead to the following recommendations for the GR6J model calibration:

- in most catchments, no improvement in streamflow simulation is expected using a composite calibration framework with groundwater level data;

- in catchments in which the original model already performs well, adding groundwater level data to the calibration is probably useless to improve streamflow simulation performance;
- in catchments in which the model reaches lower validation scores, a performance improvement is possible but not probable and it is most likely to happen in a chalk or tertiary limestone context;
- the model, with composite calibration, is able to simulate groundwater level with satisfactory performance for chalk, tertiary limestones and Cretaceous sand aquifers;
- groundwater level simulation is more uncertain for other geological contexts (quaternary formations, bedrock, Triassic sandstones or Jurassic limestones). Good results have been observed in the bedrock context of Brittany.

5.3 Perspectives

Beyond streamflow simulation, being able to simulate groundwater level using such a lumped conceptual model – much simpler and lighter to implement than usual groundwater models – is likely to lead to new uses of GR6J. Thereby, since GR6J is part of the operational low-flow forecasting platform Premhyce (Nicolle et al., 2020; Tilmant et al., 2020), it is conceivable to use it as a groundwater level sub-seasonal forecasting tool in some chosen points in France, which is crucial for an anticipative management of groundwater resources. Further studies are needed to evaluate the framework in forecasting mode; a data assimilation process may be necessary to improve the forecast liability and smoothness. Although this study does not include any modification of the streamflow simulation scheme, it offers an overview of possible modifications: the division coefficient between the routing and the exponential stores remained fixed in the present study and may become a new model parameter to rule the size of the aquifer–river flux; the role of the exchange function needs to be clarified and its formulation has to become more stable and readable.

Code and data availability. Streamflow data are available on the Banque HYDRO website (SCHAPI, 2021), their use is limited to particular conditions exposed on the website. Climatic data are avaiblable upon request to Météo France for research use. Groundwater level data are available on the ADES website (BRGM, 2021); their use is conditioned to the Etalab open licence. The original version of GR6J is available in the open-source R package airGR (Coron et al., 2021). The national hydrogeological reference map is available on the BD LISA website https://bdlisa.eaufrance.fr (Brugeron et al., 2018).

Appendix A: Detailed operation of the GR6J model

A1 Production function

The production function is mainly composed of a production store, whose capacity X_1 is the first parameter of the model. Inputs are P the daily rainfall depth and E the daily potential evaporation. Rainfall is *neutralised* by evaporation to compute net rainfall P_n and net evaporation E_n, through a case disjunction:

- If $P > E$, then $P_n = P - E$ and $E_n = 0$;
- Otherwise, $P_n = 0$ and $E_n = E - P$.

If P_n is positive, a part of it, P_s, feeds the production store, which has a level S and a parameter X_1:

$$P_s = \frac{X_1 \left(1 - \left(\frac{S}{X_1}\right)^2\right) \tanh\left(\frac{P_n}{X_1}\right)}{1 + \frac{S}{X_1} \tanh\left(\frac{P_n}{X_1}\right)} \; ; \; E_s = 0 \tag{A1}$$

Otherwise, a part E_s of E_n is taken from the production store:

$$E_s = \frac{S \left(2 - \frac{S}{X_1}\right) \tanh\left(\frac{E_n}{X_1}\right)}{1 + \left(1 - \frac{S}{X_1}\right) \tanh\left(\frac{E_n}{X_1}\right)} \; ; \; P_s = 0 \tag{A2}$$

The content of the production store is then updated by $S = S - E_s + P_s$. Part of the water content of the production store $Perc$ percolates to the routing function:

$$Perc = S \left(1 - \left(1 + \left(\frac{4S}{9X_1}\right)^4\right)^{-\frac{1}{4}}\right) \tag{A3}$$

The content of the production store is updated again by $S = S - Perc$. The quantity of water P_r that reaches the routing part of the model is finally $P_r = Perc + P_n - P_s$.

A2 Unit hydrographs

P_r is divided into two components: 90% are routed through the one-sided unit hydrograph UH_1 and the remaining 10%, through a two-sided unit hydrograph UH_2. The cumulated ordinates of the unit hydrographs $SH_1(t)$ and $SH_2(t)$ are determined by the basetime X_4, for $t \in \mathbb{N}$:

$$SH_1(t) = \begin{cases} 0 & \text{if } t = 0 \\ \left(\frac{t}{X_4}\right)^{\frac{5}{2}} & \text{if } 0 < t < X_4 \\ 1 & \text{if } t \geq X_4 \end{cases} \tag{A4}$$

$$SH_2(t) = \begin{cases} 0 & \text{if } t = 0 \\ \frac{1}{2}\left(\frac{t}{X_4}\right)^{\frac{5}{2}} & \text{if } 0 < t < X_4 \\ 1 - \frac{1}{2}\left(2 - \frac{t}{X_4}\right)^{\frac{5}{2}} & \text{if } X_4 \leq t < 2X_4 \\ 1 & \text{if } t \geq 2X_4 \end{cases} \tag{A5}$$

Ordinates $UH_1(t)$ and $UH_2(t)$ are then computed differentiating the cumulated ordinates:

$$UH_1(t) = SH_1(t) - SH_1(t-1) \; ; \; UH_2(t) = SH_2(t) - SH_2(t-1) \tag{A6}$$

Finally, the respective outputs of the first unit hydrograph Q_9 and the second one Q_1 are computed through a convolution of P_r:

$$Q_9(t) = 0.9 \sum_{k=1}^{\lfloor X_4 \rfloor + 1} UH_1(k) P_r(t-k+1) \tag{A7}$$

$$Q_1(t) = 0.1 \sum_{k=1}^{\lfloor 2X_4 \rfloor + 1} UH_2(k) P_r(t-k+1) \tag{A8}$$

A3 Routing stores

This part of the model structure is composed of two branches, that of the stores – fed by Q_9 from the first unit hydrograph – and the direct branch – fed by Q_1 from the second unit hydrograph. In the stores' branch, Q_9 is partitioned between the two stores, with 60% for the routing store and 40% for the exponential store. A potential exchange $Exch$ is computed from the water content of the routing stores $Rout$, its capacity X_3 and the exchange parameters X_2 and X_5:

$$Exch = X_2 \left(\frac{Rout}{X_3} - X_5 \right) \tag{A9}$$

This flux can be negative, zero or positive. Since the routing store cannot have a water content $Rout$ under zero, the actual exchange flux from the routing store $AExch_1$ is limited by the content of the latter, which gives the following equation:

$$AExch_1 = \begin{cases} Exch & \text{if } Rout + 0.6\, Q_9 + Exch \geq 0 \\ -Rout - 0.6\, Q_9 & \text{otherwise} \end{cases} \tag{A10}$$

The routing reservoir is then filled with:

$$Rout = Rout + Q_9 + AExch_1 \tag{A11}$$

And the output Q_R of the routing reservoir is computed as:

$$Q_R = Rout \left(1 - \left(1 + \left(\frac{Rout}{X_3} \right)^4 \right)^{-\frac{1}{4}} \right) \tag{A12}$$

The water content of the reservoir is finally updated as $Rout = Rout - Q_R$.

As for the exponential store, it is a bottomless reservoir whose water content Exp can be negative. Therefore, no case disjunction is necessary and the store can be filled with:

$$Exp = Exp + 0.4\, Q_9 + Exch \tag{A13}$$

Its output is computed, using its capacity X_6, as:

$$Q_{Rexp} = X_6 \log\left(1 + \exp\left(\frac{Exp}{X_6}\right)\right) \tag{A14}$$

The exponential store can now be updated using $Exp = Exp - Q_{Rexp}$.

The second branch, fed by Q_1, is also subject to exchange $AExch_2$ with a case disjunction:

$$AExch_2 = \begin{cases} Exch & \text{if } Q_1 + Exch \geq 0 \\ -Q_1 & \text{otherwise} \end{cases} \tag{A15}$$

The output of the second branch Q_d can now be computed using $Q_d = Q_1 - AExch2$. The simulated streamflow Q_{sim} is finally computed by adding the components from the three branches:

$$Q_{sim} = Q_R + Q_{Rexp} + Q_d \tag{A16}$$

Appendix B: Parameter ranges and transformations used for original and modified GR6J calibration

Table B1. Parameter ranges and transformation functions for GR6J model calibration. To make calibration easier, the parameter original search ranges, exposed below, are transformed to $[-9.99, 9.99]$ by each transformation function. Found values are then re-transformed into parameter values using reciprocal transformation. Details can be found in section 3.1.2.

Parameter	Unit	Description	Search range	Transformation function	Reciprocal transformation
X_1	mm	Production store capacity	$\mathbb{R}_+^*$	$x \mapsto \log(x)$	$x \mapsto \exp(x)$
X_2	mm/day	Inter-catchment exchange coefficient	$[-9.99; 9.99]$	Id	Id
X_3	mm	Routing store capacity	$\mathbb{R}_+^*$	$x \mapsto \log(x)$	$x \mapsto \exp(x)$
X_4	days	Unit hydrographs time base	$[0.5; 20]$	$x \mapsto 9.99 + 19.98\left(\frac{x-20}{19.5}\right)$	$x \mapsto 20 + 19.5\left(\frac{x-9.99}{19.98}\right)$
X_5	unitless	Inter-catchment exchange threshold	$[-2; 2]$	$x \mapsto 5.0x$	$x \mapsto x/5.0$
X_6	mm	Exponential store capacity	$\mathbb{R}_+^*$	$x \mapsto \log(x)$	$x \mapsto \exp(x)$
X_7	unitless	Groundwater linear coefficient	$]0; 4]$	$x \mapsto 20\sqrt{\tanh x} - 10$	$x \mapsto \operatorname{argtanh}\left(\left(\frac{x+10.0}{20.0}\right)^2\right)$
X_8	unitless	Groundwater linear offset	$]0; 20[$	$x \mapsto x - 10$	$x \mapsto x + 10$

Appendix C: Table of variables

Table C1. Table of variables used in the document

Variable	Unit	Description
α	unitless	Composite calibration weight
$AExch_1$	mm/day	Actual exchange of the routing store
$AExch_2$	mm/day	Actual exchange of the direct branch
$C(\alpha)$	unitless	Composite calibration objective function
δ_z	unitless	Normalised groundwater level
$\delta_{z,obs}$	unitless	Observed normalised groundwater level
$\delta_{z,sim}$	unitless	Simulated normalised groundwater level
E	mm/day	Daily potential evaporation used as model input
E_n	mm/day	Net evaporation
E_s	mm/day	Part of evaporation withdrawn from the production store
$Exch$	mm/day	Potential exchange flux
Exp	mm	Exponential store level
NSE	unitless	Nash-Sutcliffe efficiency
P	mm/day	Daily rainfall used as model input
P_n	mm/day	Net rainfall
P_r	mm/day	Flux reaching the routing part of the model
P_s	mm/day	Part of rainfall filling the production store
$Perc$	mm/day	Percolation flux
Q_1	mm/day	Output of the two-sided unit hydrograph
Q_9	mm/day	Output of the one-sided unit hydrograph
Q_d	mm/day	Output of the direct branch
Q_R	mm/day	Output of the routing store
Q_{Rexp}	mm/day	Output of the exponential store
Q_sim	mm/day	Daily simulated streamflow
$Rout$	mm	Routing store level
S	mm	Production store level
$SH_1(t)$	unitless	Cumulative ordinates of the one-sided unit hydrograph
$SH_2(t)$	unitless	Cumulative ordinates of the two-sided unit hydrograph
$UH_1(t)$	unitless	Ordinates of the one-sided unit hydrograph
$UH_2(t)$	unitless	Ordinates of the two-sided unit hydrograph
X_1	mm	Production store capacity
X_2	mm/day	Inter-catchment exchange coefficient
X_3	mm	Routing store capacity
X_4	days	Unit hydrographs time base
X_5	unitless	Inter-catchment exchange threshold
X_6	mm	Exponential store capacity
X_7	unitless	Groundwater linear coefficient
X_8	unitless	Groundwater linear offset
z	m NGF	Absolute groundwater level
$\bar{z}$	m NGF	Mean absolute groundwater level
$ZError$	unitless	Error criterion on groundwater level
z_{obs}	m NGF	Absolute observed groundwater level
z_{sim}	m NGF	Absolute simulated groundwater level

Appendix D: Equivalence between ZError and Nash-Sutcliffe efficiency

The model structure proposed here does not simulate absolute groundwater level, but only its normalised version. Then, by reversing the normalisation equation 1, the equation 3 allows to get the absolute groundwater level. By combining equations 1, 3 and 4, the following expression of $ZError$ is found:

$$ZError = 1 - \sum_t \left(\frac{z_{sim} - \overline{z}}{\sigma_z} - \frac{z_{obs} - \overline{z}}{\sigma_z} \right)^2 \tag{D1}$$

Which gives:

$$ZError = 1 - \frac{\sum_t \left(z_{sim} - z_{obs} \right)^2}{\sigma_z^2} \tag{D2}$$

By definition of standard deviation, we have:

$$\sigma_z^2 = \sum_t \left(z_{obs} - \overline{z} \right)^2 \tag{D3}$$

By combining the two previous equations, we get:

$$ZError = 1 - \frac{\sum_t \left(z_{sim} - z_{obs} \right)^2}{\sum_t \left(z_{obs} - \overline{z} \right)^2} \tag{D4}$$

Which is exactly the definition of Nash-Sutcliffe efficiency or NSE (Nash and Sutcliffe, 1970), expressed for groundwater level instead of streamflow. This shows the correspondence between ZError and NSE.

Author contributions. Both authors conceptualised the method. AP performed the tests on the dataset and developed the computing code. Both authors wrote the paper.

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

Acknowledgements. This work is part of the CIPRHES project, funded by the French national research agency ANR (grant #ANR-20-CE04-0009), and a PhD funded by the French ministry of Environmental Transition (MTE). The stream network shown in maps is from the BD CARTO database (IGN, 2021). We would like to thank Delphine ALLIER (BRGM) for the help in selecting the test dataset and Jean-Baptise BOISSONNAT (INRAE Antony) and Benoît GÉNOT (INRAE Antony, now at U.R.B.S.) for the database maintenance. We extend

our warmest thanks to Charles PERRIN (INRAE Antony), Paul ASTAGNEAU (INRAE Antony) and François BOURGIN (INRAE Antony) for their thorough proofreading which considerably improved the manuscript. Last but not least, computing codes could not have been developed without the precious expertise and availability of Olivier DELAIGUE (INRAE Antony).

References

Ardia, D., Arango, J. O., and Gomez, N. G.: Jump-Diffusion Calibration using Differential Evolution, Wilmott Magazine, 55, 76–79, https://mpra.ub.uni-muenchen.de/id/eprint/27852, 2011a.

Ardia, D., Boudt, K., Carl, P., Mullen, K. M., and Peterson, B. G.: Differential Evolution with DEoptim: An Application to Non-Convex Portfolio Optimization, The R Journal, 3, 27–34, https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Ardia~et~al.pdf, 2011b.

Ardia, D., Mullen, K. M., Peterson, B. G., and Ulrich, J.: DEoptim: Differential Evolution in R, https://CRAN.R-project.org/package=DEoptim, version 2.2-5, 2020.

Aubert, D., Loumagne, C., and Oudin, L.: Sequential assimilation of soil moisture and streamflow data in a conceptual rainfall–runoff model, Journal of Hydrology, 280, 145–161, https://doi.org/10.1016/s0022-1694(03)00229-4, 2003a.

Aubert, D., Loumagne, C., Oudin, L., and Hégarat-Mascle, S. L.: Assimilation of soil moisture into hydrological models: the sequential method, Canadian Journal of Remote Sensing, 29, 711–717, https://doi.org/10.5589/m03-042, 2003b.

Barthel, R.: HESS Opinions "Integration of groundwater and surface water research: an interdisciplinary problem?", Hydrology and Earth System Sciences, 18, 2615–2628, https://doi.org/10.5194/hess-18-2615-2014, 2014.

Barthel, R. and Banzhaf, S.: Groundwater and Surface Water Interaction at the Regional-scale – A Review with Focus on Regional Integrated Models, Water Resources Management, 30, 1–32, https://doi.org/10.1007/s11269-015-1163-z, 2015.

Bartlett, M. S. and Porporato, A.: A Class of Exact Solutions of the Boussinesq Equation for Horizontal and Sloping Aquifers, Water Resources Research, 54, 767–778, https://doi.org/10.1002/2017WR022056, 2018.

Bauer, D. F.: Constructing Confidence Sets Using Rank Statistics, Journal of the American Statistical Association, 67, 687–690, https://doi.org/10.1080/01621459.1972.10481279, 1972.

Bel, F., Lacroix, A., Mollard, A., David, C., Beaudoin, N., Mary, B., Vachaud, G., Vauclin, M., and Garino, B.: Une approche interdisciplinaire, pluri-échelle, multipartenaire des pollutions diffuses de l'eau : l'expérience de La Côte Saint-André (Isère), La Houille Blanche, pp. 72–79, https://doi.org/10.1051/lhb/1999074, 1999.

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

Bergström, S. and Sandberg, G.: Simulation of Groundwater Response by Conceptual Models, Hydrology Research, 14, 71–84, https://doi.org/10.2166/nh.1983.0007, 1983.

Beven, K.: Hydrograph separation?, in: Proc.BHS Third National Hydrology Symposium, pp. 3.2–3.8, Institute of hydrology, 1991.

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

Beven, K.: Rainfall-Runoff Modelling, John Wiley & Sons, Ltd, https://doi.org/10.1002/9781119951001, 2012.

Borzì, I., Bonaccorso, B., and Fiori, A.: A Modified IHACRES Rainfall-Runoff Model for Predicting the Hydrologic Response of a River Basin Connected with a Deep Groundwater Aquifer, Water, 11, 2031, https://doi.org/10.3390/w11102031, 2019.

BRGM: ADES: portail national d'accès aux données sur les eaux souterraines, https://ades.eaufrance.fr/, 2021.

Brigode, P., Génot, B., Lobligeois, F., and Delaigue, O.: Summary sheets of watershed-scale hydroclimatic observed data for France, https://doi.org/10.15454/UV01P1, 2021.

Brugeron, A., Paroissien, J., and Tillier, L.: Référentiel hydrogéologique BDLISA version 2 : Principes de construction et évolutions, Rapport final RP-67489-FR, BRGM, http://infoterre.brgm.fr/rapports/RP-67489-FR.pdf, 2018.

Brunner, P. and Simmons, C. T.: HydroGeoSphere: A Fully Integrated, Physically Based Hydrological Model, Ground Water, 50, 170–176, https://doi.org/10.1111/j.1745-6584.2011.00882.x, 2011.

Carlier, C., Wirth, S. B., Cochand, F., Hunkeler, D., and Brunner, P.: Geology controls streamflow dynamics, Journal of Hydrology, 566, 756–769, https://doi.org/10.1016/j.jhydrol.2018.08.069, 2018.

Castany, G.: Traité pratique des eaux souterraines, Dunod, 1963.

Coron, L., Thirel, G., Delaigue, O., Perrin, C., and Andréassian, V.: The Suite of Lumped GR Hydrological Models in an R package, Environmental Modelling and Software, 94, 166–171, https://doi.org/10.1016/j.envsoft.2017.05.002, 2017.

Coron, L., Delaigue, O., Thirel, G., Dorchies, D., Perrin, C., and Michel, C.: airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling, https://doi.org/10.15454/EX11NA, r package version 1.6.10.4, 2021.

Creutzfeldt, B., Ferré, T., Troch, P., Merz, B., Wziontek, H., and Güntner, A.: Total water storage dynamics in response to climate variability and extremes: Inference from long-term terrestrial gravity measurement, Journal of Geophysical Research: Atmospheres, 117, https://doi.org/10.1029/2011JD016472, 2012.

Dassargues, A., Maréchal, J. C., Carabin, G., and Sels, O.: On the necessity to use three-dimensional groundwater models for describing impact of drought conditions on streamflow regimes, in: Hydrological Extremes : Understanding, Predicting, Mitigating, edited by Press, I., pp. 165–170, 1999.

de Lavenne, A., Thirel, G., Andréassian, V., Perrin, C., and Ramos, M.-H.: Spatial variability of the parameters of a semi-distributed hydrological model, Proceedings of the International Association of Hydrological Sciences, 373, 87–94, https://doi.org/10.5194/piahs-373-87-2016, 2016.

Delaigue, O., Génot, B., Lebecherel, L., Brigode, P., and Bourgin, P.-Y.: Base de données hydroclimatiques observées à l'échelle de la France, INRAE, UR HYCAR, https://webgr.inrae.fr/base-de-donnees, 2021.

Dembélé, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., and Schaefli, B.: Improving the Predictive Skill of a Distributed Hydrological Model by Calibration on Spatial Patterns With Multiple Satellite Data Sets, Water Resources Research, 56, https://doi.org/10.1029/2019wr026085, 2020.

Demirel, Özen, Orta, Toker, Demir, Ekmekcioğlu, Tayşi, Eruçar, Sağ, Sarı, Tuncer, Hancı, Özcan, Erdem, Koşucu, Başakın, Ahmed, Anwar, Avcuoğlu, Vanlı, Stisen, and Booij: Additional Value of Using Satellite-Based Soil Moisture and Two Sources of Groundwater Data for Hydrological Model Calibration, Water, 11, 2083, https://doi.org/10.3390/w11102083, 2019.

Eddelbuettel, D.: RcppDE: Global Optimization by Differential Evolution in C++, https://CRAN.R-project.org/package=RcppDE, r package version 0.1.6, 2018.

Efstratiadis, A., Nalbantis, I., Koukouvinos, A., Rozos, E., and Koutsoyiannis, D.: HYDROGEIOS: a semi-distributed GIS-based hydrological model for modified river basins, Hydrology and Earth System Sciences, 12, 989–1006, https://doi.org/10.5194/hess-12-989-2008, 2008.

El-Nasr, A. A., Arnold, J. G., Feyen, J., and Berlamont, J.: Modelling the hydrology of a catchment using a distributed and a semi-distributed model, Hydrological Processes, 19, 573–587, https://doi.org/10.1002/hyp.5610, 2005.

Feyen, L., Vázquez, R., Christiaens, K., Sels, O., and Feyen, J.: Application of a distributed physically-based hydrological model to a medium size catchment, Hydrology and Earth System Sciences, 4, 47–63, https://doi.org/10.5194/hess-4-47-2000, 2000.

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

Guérin, A., Devauchelle, O., Robert, V., Kitou, T., Dessert, C., Quiquerez, A., Allemand, P., and Lajeunesse, E.: Stream-Discharge Surges Generated by Groundwater Flow, Geophysical Research Letters, 46, 7447–7455, https://doi.org/10.1029/2019GL082291, 2019.

Habets, F., Gascoin, S., Korkmaz, S., Thiéry, D., Zribi, M., Amraoui, N., Carli, M., Ducharne, A., Leblois, E., Ledoux, E., Martin, E., Noilhan, J., Ottlé, C., and Viennot, P.: Multi-model comparison of a major flood in the groundwater-fed basin of the Somme River (France), Hydrology and Earth System Sciences, 14, 99–117, https://doi.org/10.5194/hess-14-99-2010, 2010.

Hayashi, M.: Alpine Hydrogeology: The Critical Role of Groundwater in Sourcing the Headwaters of the World, Groundwater, 58, 498–510, https://doi.org/doi.org/10.1111/gwat.12965, 2020.

Herron, N. and Croke, B.: Including the influence of groundwater exchanges in a lumped rainfall-runoff model, Mathematics and Computers in Simulation, 79, 2689–2700, https://doi.org/10.1016/j.matcom.2008.08.007, 2009.

Hughes, D. A.: Incorporating groundwater recharge and discharge functions into an existing monthly rainfall–runoff model/Incorporation de fonctions de recharge et de vidange superficielle de nappes au sein d'un modèle pluie-débit mensuel existant, Hydrological Sciences Journal, 49, https://doi.org/10.1623/hysj.49.2.297.34834, 2004.

IGN: BD CARTO, Institut national de l'information géographique et forestière, https://geoservices.ign.fr/documentation/donnees/vecteur/bdcarto, 2021.

Immerzeel, W. and Droogers, P.: Calibration of a distributed hydrological model based on satellite evapotranspiration, Journal of Hydrology, 349, 411–424, https://doi.org/10.1016/j.jhydrol.2007.11.017, 2008.

Jakeman, A. J. and Hornberger, G. M.: How much complexity is warranted in a rainfall-runoff model?, Water Resources Research, 29, 2637–2649, https://doi.org/10.1029/93wr00877, 1993.

Jian, J., Ryu, D., Costelloe, J. F., and Su, C.-H.: Towards hydrological model calibration using river level measurements, Journal of Hydrology: Regional Studies, 10, 95–109, https://doi.org/10.1016/j.ejrh.2016.12.085, 2017.

Khu, S.-T., Madsen, H., and di Pierro, F.: Incorporating multiple observations for distributed hydrologic model calibration: An approach using a multi-objective evolutionary algorithm and clustering, Advances in Water Resources, 31, 1387–1398, https://doi.org/10.1016/j.advwatres.2008.07.011, 2008.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrological Sciences Journal, 31, 13–24, https://doi.org/10.1080/02626668609491024, 1986.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424-425, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Käser, D. and Hunkeler, D.: Contribution of alluvial groundwater to the outflow of mountainous catchments, Water Resources Research, 52, 680–697, https://doi.org/10.1002/2014WR016730, 2016.

Lalot, E., Curie, F., Wawrzyniak, V., Baratelli, F., Schomburgk, S., Flipo, N., Piegay, H., and Moatar, F.: Quantification of the contribution of the Beauce groundwater aquifer to the discharge of the Loire River using thermal infrared satellite imaging, Hydrology and Earth System Sciences, 19, 4479–4492, https://doi.org/10.5194/hess-19-4479-2015, 2015.

Le Moine, N.: Le bassin versant de surface vu par le souterrain : une voie d'amélioration des performances et du réalisme des modèles pluie-débit ?, Ph.D. thesis, Université Pierre et Marie Curie, Paris, France, cemagref Antony, 2008.

Le Moine, N., Andréassian, V., and Mathevet, T.: Confronting surface- and groundwater balances on the La Rochefoucauld-Touvre karstic system (Charente, France), Water Resources Research, 44, https://doi.org/10.1029/2007wr005984, 2008.

Leleu, I., Tonnelier, I., Puechberty, R., Gouin, P., Viquendi, I., Cobos, L., Foray, A., Baillon, M., and Ndima, P.-O.: La refonte du système d'information national pour la gestion et la mise à disposition des données hydrométriques, La Houille Blanche, pp. 25–32, https://doi.org/10.1051/lhb/2014004, 2014.

Lenhardt, F., Doucet, N., Boisson, M., and Billault, P.: The Cenomanian Sands aquifer model: an effective groundwater management tool, Tech. rep., SOGREAH, http://feflow.info/fileadmin/FEFLOW/content_tagung/TagungsCD/papers/5.pdf, 2009.

Li, S., Gitau, M., Engel, B. A., Zhang, L., Du, Y., Wallace, C., and Flanagan, D. C.: Development of a distributed hydrological model to facilitate watershed management, Hydrological Sciences Journal, 62, 1755–1771, https://doi.org/10.1080/02626667.2017.1351029, 2017.

Lo, M.-H. and Famiglietti, J. S.: Effect of water table dynamics on land surface hydrologic memory, Journal of Geophysical Research, 115, https://doi.org/10.1029/2010JD014191, 2010.

Mackay, J., Jackson, C., and Wang, L.: A lumped conceptual model to simulate groundwater level time-series, Environmental Modelling & Software, 61, 229–245, https://doi.org/10.1016/j.envsoft.2014.06.003, 2014.

Madsen, H.: Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives, Advances in Water Resources, 26, 205–216, https://doi.org/10.1016/s0309-1708(02)00092-1, 2003.

Maillet, E.: Essais d'hydraulique souterraine & fluviale, A. Hermann, http://archive.org/details/essaisdhydrauli00mailgoog, 1905.

Mann, H. B. and Whitney, D. R.: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, The Annals of Mathematical Statistics, 18, 50–60, https://doi.org/10.1214/aoms/1177730491, 1947.

McDonnell, J. J. and Beven, K.: Debates—The future of hydrological sciences: A (common) path forward? A call to action aimed at understanding velocities, celerities and residence time distributions of the headwater hydrograph, Water Resources Research, 50, 5342–5350, https://doi.org/10.1002/2013WR015141, 2014.

Michel, C.: Que peut-on faire en hydrologie avec modèle conceptuel à un seul paramètre ?, La Houille Blanche, pp. 39–44, https://doi.org/10.1051/lhb/1983004, 1983.

Michel, C.: Hydrologie appliquée aux petits bassins versants ruraux, Cemagref, 1991.

Michel, C., Perrin, C., and Andréassian, V.: The exponential store: a correct formulation for rainfall—runoff modelling, Hydrological Sciences Journal, 48, 109–124, https://doi.org/10.1623/hysj.48.1.109.43484, 2003.

Milzow, C., Krogh, P. E., and Bauer-Gottwein, P.: Combining satellite radar altimetry, SAR surface soil moisture and GRACE total storage changes for hydrological model calibration in a large poorly gauged catchment, Hydrology and Earth System Sciences, 15, 1729–1743, https://doi.org/10.5194/hess-15-1729-2011, 2011.

Moore, R. J.: Real-Time Flood Forecasting Systems: Perspectives and Prospects, pp. 147–189, Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-58609-5_11, 1999.

Moore, R. J. and Bell, V. A.: Incorporation of groundwater losses and well level data in rainfall-runoff models illustrated using the PDM, Hydrology and Earth System Sciences, 6, 25–38, https://doi.org/10.5194/hess-6-25-2002, 2002.

Moreda, F., Koren, V., Zhang, Z., Reed, S., and Smith, M.: Parameterization of distributed hydrological models: learning from the experiences of lumped modeling, Journal of Hydrology, 320, 218–237, https://doi.org/10.1016/j.jhydrol.2005.07.014, 2006.

Mostafaie, A., Forootan, E., Safari, A., and Schumacher, M.: Comparing multi-objective optimization techniques to calibrate a conceptual hydrological model using in situ runoff and daily GRACE data, Computational Geosciences, 22, 789–814, https://doi.org/10.1007/s10596-018-9726-8, 2018.

Mullen, K., Ardia, D., Gil, D., Windover, D., and Cline, J.: DEoptim: An R Package for Global Optimization by Differential Evolution, Journal of Statistical Software, 40, 1–26, https://doi.org/10.18637/jss.v040.i06, 2011.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Lay, M. L., Besson, F., Soubeyroux, J.-M., Viel, C., Regimbeau, F., Andréassian, V., Maugis, P., Augeard, B., and Morice, E.: Benchmarking hydrological models for low-flow simulation and forecasting on French catchments, Hydrology and Earth System Sciences, 18, 2829–2857, https://doi.org/10.5194/hess-18-2829-2014, 2014.

Nicolle, P., Besson, F., Delaigue, O., Etchevers, P., François, D., Lay, M. L., Perrin, C., Rousset, F., Thiéry, D., Tilmant, F., Magand, C., Leurent, T., and Jacob, É.: PREMHYCE: An operational tool for low-flow forecasting, Proceedings of the International Association of Hydrological Sciences, 383, 381–389, https://doi.org/10.5194/piahs-383-381-2020, 2020.

Oudin, L., Weisse, A., Loumagne, C., and Hégarat-Mascle, S. L.: Assimilation of soil moisture into hydrological models for flood forecasting: a variational approach, Canadian Journal of Remote Sensing, 29, 679–686, https://doi.org/10.5589/m03-038, 2003.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall–runoff model?, Journal of Hydrology, 303, 290–306, https://doi.org/10.1016/j.jhydrol.2004.08.026, 2005.

Pelletier, A. and Andréassian, V.: Hydrograph separation: an impartial parametrisation for an imperfect method, Hydrology and Earth System Sciences, 24, 1171–1187, https://doi.org/10.5194/hess-24-1171-2020, 2020.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, https://doi.org/10.1016/s0022-1694(03)00225-7, 2003.

Pinault, J.-L., Amraoui, N., and Golaz, C.: Groundwater-induced flooding in macropore-dominated hydrological system in the context of climate changes, Water Resources Research, 41, https://doi.org/10.1029/2004WR003169, 2005.

Poncelet, C.: Modélisation hydrologique en contexte non jaugé: développements méthodologiques et conceptualisation, Ph.D. thesis, Université Pierre et Marie Curie, Paris, 2016.

Price, K. V., Storn, R. M., and Lampinen, J. A.: Differential Evolution - A Practical Approach to Global Optimization, Natural Computing, Springer-Verlag, https://doi.org/10.1007/3-540-31306-0, iSBN 540209506, 2006.

Pushpalatha, R., Perrin, C., Le Moine, N., Mathevet, T., and Andréassian, V.: A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, Journal of Hydrology, 411, 66–76, https://doi.org/10.1016/j.jhydrol.2011.09.034, 2011.

Pushpalatha, R., Perrin, C., Moine, N. L., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, Journal of Hydrology, 420-421, 171–182, https://doi.org/10.1016/j.jhydrol.2011.11.055, 2012.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2021.

Riboust, P., Thirel, G., Moine, N. L., and Ribstein, P.: Revisiting a Simple Degree-Day Model for Integrating Satellite Data: Implementation of Swe-Sca Hystereses, Journal of Hydrology and Hydromechanics, 67, 70–81, https://doi.org/10.2478/johh-2018-0004, 2018.

Roche, P.-A., Miquel, J., and Gaume, E.: Hydrologie quantitative, Springer Paris, https://doi.org/10.1007/978-2-8178-0106-3, 2012.

SCHAPI: Banque HYDRO, http://www.hydro.eaufrance.fr/, ministère de la Transition Écologique, 2021.

Slater, L. J., Thirel, G., Harrigan, S., Delaigue, O., Hurley, A., Khouakhi, A., Prosdocimi, I., Vitolo, C., and Smith, K.: Using R in hydrology: a review of recent developments and future directions, Hydrology and Earth System Sciences, 23, 2939–2963, https://doi.org/10.5194/hess-23-2939-2019, 2019.

Soulsby, C., Tetzlaff, D., Rodgers, P., Dunn, S., and Waldron, S.: Runoff processes, stream water residence times and controlling landscape characteristics in a mesoscale catchment: An initial evaluation, Journal of Hydrology, 325, 197–221, https://doi.org/10.1016/j.jhydrol.2005.10.024, 2006.

Spearman, C.: Demonstration of Formulae for True Measurement of Correlation, The American Journal of Psychology, 18, 161, https://doi.org/10.2307/1412408, 1907.

Stadnyk, T. A. and Holmes, T. L.: On the value of isotope-enabled hydrological model calibration, Hydrological Sciences Journal, 65, 1525–1538, https://doi.org/10.1080/02626667.2020.1751847, 2020.

Stadnyk, T. A., Delavau, C., Kouwen, N., and Edwards, T. W. D.: Towards hydrological model calibration and validation: simulation of stable water isotopes using the isoWATFLOOD model, Hydrological Processes, 27, 3791–3810, https://doi.org/10.1002/hyp.9695, 2013.

Swenson, S., Yeh, P. J.-F., Wahr, J., and Famiglietti, J.: A comparison of terrestrial water storage variations from GRACE with in situ measurements from Illinois, Geophysical Research Letters, 33, https://doi.org/10.1029/2006GL026962, 2006.

Syed, T. H., Famiglietti, J. S., Rodell, M., Chen, J., and Wilson, C. R.: Analysis of terrestrial water storage changes from GRACE and GLDAS, Water Resources Research, 44, https://doi.org/10.1029/2006WR005779, 2008.

Széles, B., Parajka, J., Hogan, P., Silasari, R., Pavlin, L., Strauss, P., and Blöschl, G.: The Added Value of Different Data Types for Calibrating and Testing a Hydrologic Model in a Small Catchment, Water Resources Research, 56, https://doi.org/10.1029/2019wr026153, 2020.

Tague, C. and Grant, G. E.: Groundwater dynamics mediate low-flow response to global warming in snow-dominated alpine regions, Water Resources Research, 45, https://doi.org/10.1029/2008WR007179, 2009.

Thirel, G., Salamon, P., Burek, P., and Kalas, M.: Assimilation of MODIS Snow Cover Area Data in a Distributed Hydrological Model Using the Particle Filter, Remote Sensing, 5, 5825–5850, https://doi.org/10.3390/rs5115825, 2013.

Thiéry, D.: Forecast of changes in piezometric levels by a lumped hydrological model, Journal of Hydrology, 97, 129–148, https://doi.org/10.1016/0022-1694(88)90070-4, 1988.

Thiéry, D.: Logiciel GARDÉNIA, version v8.2. Guide d'utilisation, BRGM, Orléans, France, https://www.brgm.fr/sites/default/files/documents/2020-11/logiciel-gardenia-v8-2-rp-62797-fr-notice.pdf, bRGM report RP-62797-FR, 2014.

Tiel, M., Stahl, K., Freudiger, D., and Seibert, J.: Glacio-hydrological model calibration and evaluation, WIREs Water, 7, https://doi.org/10.1002/wat2.1483, 2020.

Tilmant, F., Nicolle, P., Bourgin, F., Besson, F., Delaigue, O., Etchevers, P., François, D., Lay, M. L., Perrin, C., Rousset, F., Thiéry, D., Magand, C., Leurent, T., and Jacob, É.: PREMHYCE : un outil opérationnel pour la prévision des étiages, La Houille Blanche, 106, 37–44, https://doi.org/10.1051/lhb/2020043, 2020.

Tobin, B. W. and Schwartz, B. F.: Quantifying the role of karstic groundwater in a snowmelt-dominated hydrologic system, Hydrological Processes, 34, 3439–3447, https://doi.org/10.1002/hyp.13833, 2020.

Tomasella, J., Hodnett, M. G., Cuartas, L. A., Nobre, A. D., Waterloo, M. J., and Oliveira, S. M.: The water balance of an Amazonian micro-catchment: the effect of interannual variability of rainfall on hydrological behaviour, Hydrological Processes, 22, 2133–2147, https://doi.org/10.1002/hyp.6813, 2008.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, International Journal of Climatology, 30, 1627–1644, https://doi.org/10.1002/joc.2003, 2009.

Wilcoxon, F.: Individual Comparisons by Ranking Methods, Biometrics Bulletin, 1, 80, https://doi.org/10.2307/3001968, 1945.

Wirth, S. B., Carlier, C., Cochand, F., Hunkeler, D., and Brunner, P.: Lithological and Tectonic Control on Groundwater Contribution to Stream Discharge During Low-Flow Conditions, Water, 12, https://doi.org/10.3390/w12030821, 2020.