On the use of streamflow transformations for hydrological model calibration

Guillaume Thirel[1], Léonard Santos[1], Olivier Delaigue[1], and Charles Perrin[1]

[1]Université Paris-Saclay, INRAE, UR HYCAR, 92160 Antony, France

Correspondence: Guillaume Thirel (guillaume.thirel@inrae.fr)

Abstract. The calibration of hydrological models through the use of automatic algorithms aims at identifying parameter sets that minimize the deviation of simulations from observations (often streamflows). It is a widespread technique that has been the subject of much research in the past. Indeed, the choice of objective function (i.e. the criterion or combination of criteria to optimize) can significantly impact the parameter set values identified as optimal by the algorithm. Besides, the actual goal of the model application (flood or low-flow estimation, for instance) influences the way calibration is undertaken. This article discusses how mathematical transformations, which are sometimes applied to the target variable before calculating the objective function, impact model simulations. Such transformations, for example square root or logarithmic, aim at increasing the weight of errors made in specific ranges of the hydrograph. Typically, a logarithmic transformation tends to increase the fit of streamflows to lower values, compared to no transformation. We show in a catchment set that the impact of these transformations on the obtained time series can sometimes be different from ~~what could be expected~~ their expected behaviour. Extreme transformations, such as squared or inverse of squared transformations, lead to models that are specialized for extreme streamflows, but show poor performance outside the range of the targeted streamflows and are less robust. Other transformations, such as the power 0.2, the Box–Cox and the logarithmic transformations, can be qualified as more generalist, and show a good performance for the intermediate range of streamflows, along with an acceptable performance for extreme streamflows.

1   Introduction

Hydrological models are essential tools for hypothesis testing and process understanding (Rosbjerg and Madsen, 2006), but also for very practical applications such as flood or low-flow forecasting, water resources management or the assessment of climate change impact. Despite the long-lasting efforts of hydrologists, ~~so far no consensus has emerged for identifying a unique hydrological model fitting all purposes~~ there is consensus in the community that no universal hydrological model structure exists and it is doubtful whether it will ever be found. This has motivated a proliferation of flexible modeling platforms such as FUSE (Clark et al., 2008), SUPERFLEX (Fenicia et al., 2011), Noah-MP (Niu et al., 2011), SUMMA (Clark et al., 2015b, a, 2021b) , MARRMoT (Knoben et al., 2019), Raven (Craig et al., 2020) or airGR (Coron et al., 2017). Consequently, in order to fit spe-

cific applications such as those mentioned above, and due to the wide catchment diversity and the various targeted streamflow
25 ranges, performing a calibration of model parameters is generally necessary. The calibration process usually relies on the use of a criterion (sometimes a combination of criteria), i.e. a numerical metric of the model error, which is used as an objective function. The choice of this optimization criterion is subjective (Mendoza et al., 2016; Fowler et al., 2018; Melsen et al., 2019), since it depends on various aspects (application objective, model characteristics, etc.), and two different chosen criteria will impact differently on the calibration process and will lead to different optimal parameter sets and performances (Booij and
30 Krol, 2010). In addition, these criteria suffer from flaws leading to their incorrect use by modellers (Clark et al., 2021a), and each modeller has their own vision of what is a good model or hydrograph and how it translates into a numerical criterion (Crochemore et al., 2015).

While criteria are usually calculated for comparing raw simulated and observed streamflow time series, a wide panel of transformations ~~have~~ has been introduced in the literature (Bennett et al., 2013). These transformations consist in using a
35 mathematical function in order to transform both simulated and observed time series. These transformations rely on the fact that they distort the observed and simulated time series and their properties in such a way as to expect that the related errors are similarly distorted. This is illustrated in Fig. 1, where in panel a, the larger errors between the observed and simulated time series mostly occur for high-flow periods (pink shaded area), while in panel b, with log-transformed flows, these errors are much larger over low-flow periods (green shaded areas).
40 ~~More specifically, since~~ Since many metrics are squared metrics (e.g. root mean square error or Nash–Sutcliffe efficiency, Nash and Sutcliffe, 1970) and therefore are known to emphasize the most important errors (Sorooshian and Dracup, 1980), a large set of transformations were proposed for better representation of low flows. The most common transformations found in the literature are the square root transformation ~~(Oudin et al., 2006; Pushpalatha et al., 2012; Garcia et al., 2017; Song et al., 2019)~~, the logarithmic transformation ~~(Smakhtin et al., 1998; Houghton-Carr, 1999; Krause et al., 2005; Oudin et al., 2006; de Vos et al., 2010; S~~
45 , the reciprocal of squared root ~~(Chapman, 1964; Ding, 1966; Ishihara and Takagi, 1970)~~, the inverse ~~(Pushpalatha et al., 2012; Garcia et al.,~~ or other power–law transformations ~~(Dawdy and Lichty, 1968; Chiew et al., 1993). Some other works~~. Other studies used the Box–Cox transformation ~~(Box and Cox, 1964; Abdulla et al., 1999; Hogue et al., 2000; Duan et al., 2007; Vázquez et al., 2008b)~~. ~~Nicolle et al. (2014) focused on low-flow simulations and the different models they used were calibrated using~~ or combined several of the transformations listed above. A non-exhaustive list of references and the transformations they used are listed in
50 Table 1.

While the choice of transformations is wide and the theoretical basis is sound (as shown in Fig. 1), there is not an extensive literature discussing the merits of the transformation approach. ~~Nevertheless, some authors tried to investigate this issue. For instance,~~ Pushpalatha et al. (2012), in a review of the suitability of criteria for the evaluation of low flows, justify the use of transformations by several authors through the fact that "the sum of squared residuals calculated on the logarithms of flow
55 values" reduces "the biasing towards peak flows". They investigated which range of streamflows leads to the largest part of errors. Smith et al. (2014) showed that a transformation called "flow-corrected time", designed to provide greater weight to time periods with larger hydrologic flux, results in improved fits, compared to a baseline untransformed case and the logarithmic transformation, over the time periods that dominate hydrologic flux. Peña-Arancibia et al. (2015) showed that a squared root
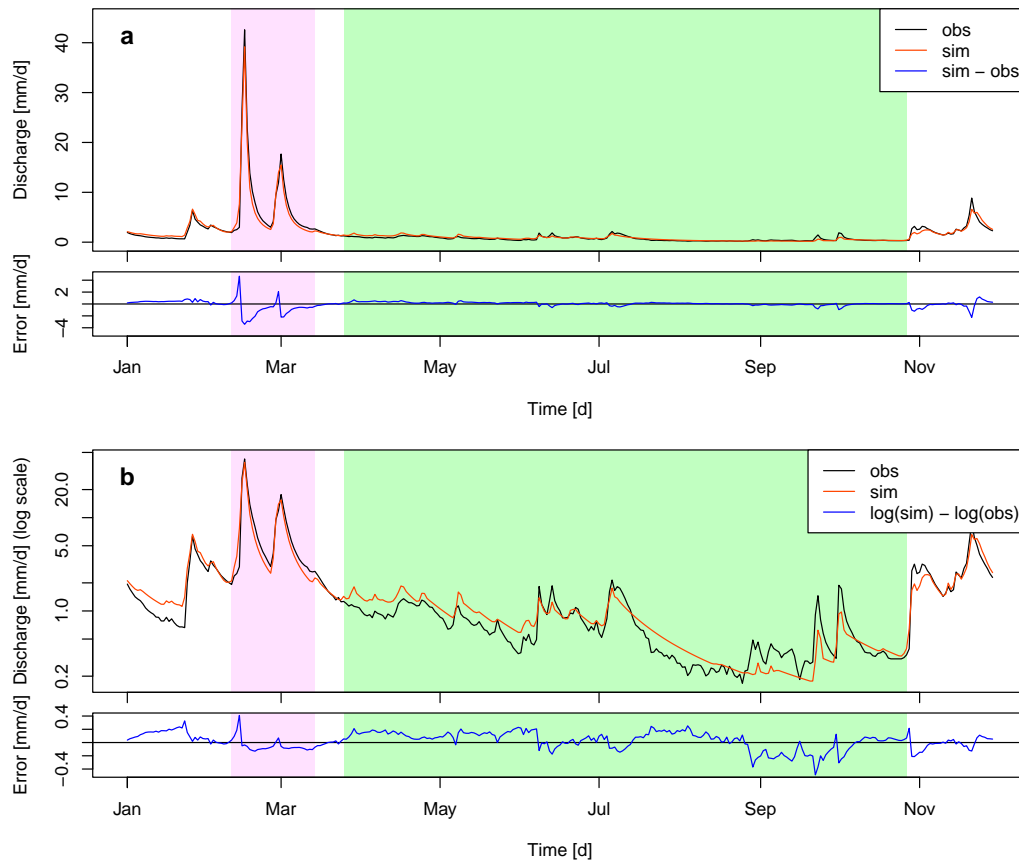
Figure 1. Panel a: Observed and simulated streamflow (with the GR4J model calibrated with NSE) time series (left axis) for the Fecht River at Wintzenheim and the related difference (i.e. error, right axis) ~~.~~ between 1 January 1990 and 29 November 1990. Panel b: The same observed and simulated streamflow time series plotted with a logarithmic scale and the difference in log-transformed observed and simulated streamflows. In the boxes, a low-flow period is highlighted in green, which shows that the error is minimal with no transformation (panel a) and much higher with a logarithmic transformation (panel b). The opposite is valid for high flows (in pink).

transformation with the Nash–Sutcliffe efficiency leads to ~~a better calibration and~~ calibration that provides higher performance

60   both on the calibration and on the evaluation period, and leads to a reduced parameter uncertainty than no transformation or a logarithmic transformation. Sadegh et al. (2018) investigated the role of several transformations in three catchments and two models and deduced that some data transformations might be more helpful for ~~evaluation and~~ the evaluation of model performance and the analysis of model behaviour than ~~model inference. Smith et al. (2014) showed that a transformation called "flow-corrected time", designed to provide greater weight to time periods with larger hydrologic flux, results in improved fits,~~

65   ~~compared to a baseline untransformed case and the logarithmic transformation, over the time periods that dominate hydrologic flux~~ for calibration.

Reference	$\sqrt{Q}$	$log(Q)$	$1/\sqrt{Q}$	Q^{-1}	Box-Cox	Other power-law transformations	Mix
Abdulla et al. (1999)					✓		
Beck et al. (2016)		✓					
Box and Cox (1964)					✓		
Chapman (1964)			✓				
Chiew et al. (1993)						✓	
Dawdy and Lichty (1968)						✓	
de Vos et al. (2010)		✓					
Ding (1966)			✓				
Duan et al. (2007)					✓		
Farmer and Vogel (2016)		✓					
Garcia et al. (2017)	✓	✓		✓			
Hogue et al. (2000)					✓		
Houghton-Carr (1999)		✓					
Ishihara and Takagi (1970)			✓				
Krause et al. (2005)		✓					
Nicolle et al. (2014)							✓
Oudin et al. (2006)	✓	✓					
Pechlivanidis et al. (2014)		✓					
Pushpalatha et al. (2012)	✓			✓			
Quesada-Montano et al. (2018)		✓					
Santos et al. (2018)		✓					
Seeger and Weiler (2014)		✓					
Smakhtin et al. (1998)		✓					
Song et al. (2019)	✓	✓					
Vázquez et al. (2008b)					✓		

Table 1. Non exhaustive list of references using streamflow transformations.

~~Still, most of the time~~To the best of our knowledge, the use and choice of transformation ~~are not~~ have not been thoroughly assessed. For example, Krause et al. (2005) ~~state~~ stated that they used the logarithmic transformation on the Nash–Sutcliffe efficiency "to reduce the problem of the squared differences [...]. Through the logarithmic transformation of the runoff values the peaks are flattened and the low flows are kept more or less at the same level. As a result the influence of the low flow values is increased in comparison to the flood peaks". Chiew et al. (1993) ~~use~~ used a power 0.2 transformation and justify it by the fact that "it generally leads to constant variances (values of $SIM^{0.2} - REC^{0.2}$ are similar for all flow volumes) in many of the

temperate catchments where models have been applied by the authors". Oudin et al. (2006) ~~report~~ reported that "it is common practice in hydrology to use a transformation on flows before optimization". Others only ~~state~~ stated that transformations are used "to remove the bias towards high flows" (Smakhtin et al., 1998), "to fit low flow periods" (Pechlivanidis et al., 2014) or to put "more weight on low flow" (Garcia et al., 2017).

While Fig. 1 ~~tends to illustrate these assertions , we feel that~~ illustrates these assertions to some degree, there is a lack of a general assessment of the impact of transformations on the calculation of criteria over diverse conditions. Consequently, ~~in this article~~ we aim to perform a systematic evaluation of the impact of transformations on a large set of catchments when calibrating hydrological models. We will not consider here metrics calculated with ~~streamflow alteration~~ specific streamflow selection procedures such as keeping only streamflow values under or over a threshold or the use of relative streamflow.

2 Material and method

2.1 Catchment ~~dataset~~ set and data

~~Data~~ We used data from 325 catchments around France (Chauveau et al., 2011) ~~were used~~ in order to i) generalize the conclusions drawn from this study (Gupta et al., 2014) and ii) possibly identify links between catchment characteristics and specific behaviours of transformations. These catchments were chosen for the low human impact on the precipitation–streamflow relationship and for the low rate of missing streamflow data (< 0.5 %) over the period of interest. Moreover, the catchments are spread throughout France (Figure 2), thus representing a wide variety of meteorological and hydrological conditions.

Precipitation and temperature data were retrieved from the Météo-France SAFRAN reanalysis (Vidal et al., 2010). Streamflow data were retrieved from the French HydroPortail database (https://hydro.eaufrance.fr/) (Leleu et al., 2014). Both meteorological and hydrological data were used at the daily time step. Records from 1985 to 2005 were used, with 1985–1995 as the calibration period and 1995–2005 as the independent evaluation period. The main characteristics of the 325 catchments are summarized in Table 2. It illustrates the high diversity of catchment characteristics encountered, with small to large catchments, various precipitation and temperature conditions, rainfed as well as snowfed catchments, and catchments facing low to high baseflow components. Moreover, the large range of land cover, slope and hydraulic length strengthens the diversity of possible catchment response. This table also shows that the climatic as well as hydrological conditions are similar between the two periods, with the evaluation period being only slightly warmer and wetter than the calibration period and other indicators showing only slight variations.
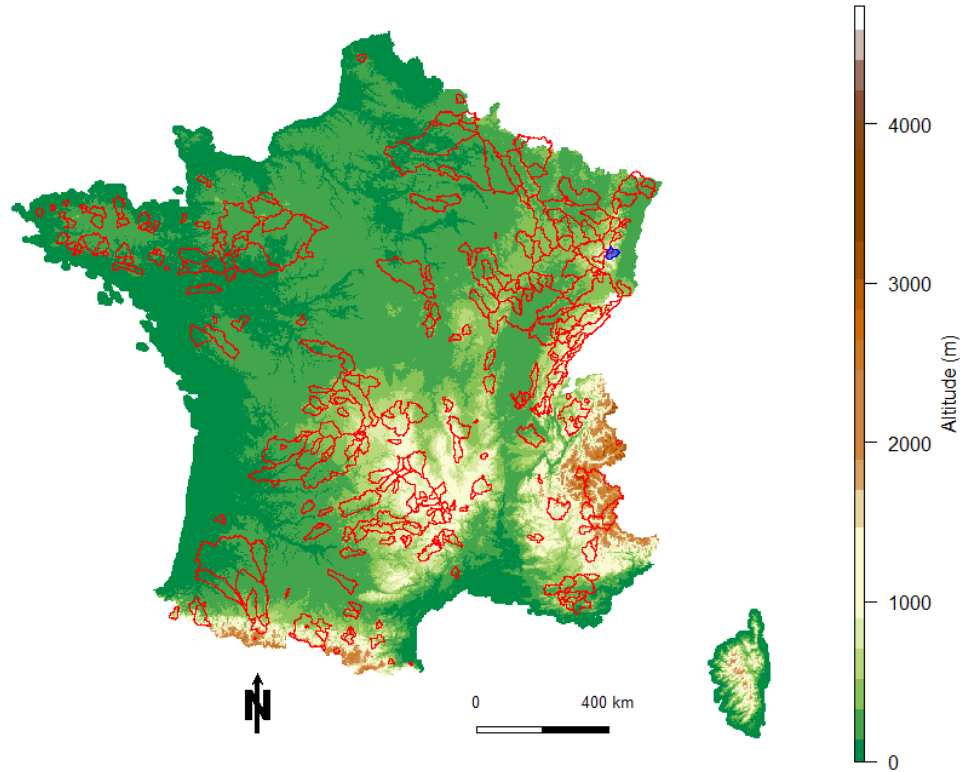
Figure 2. Map of the location of the 325 catchments used. The Fecht River at Wintzenheim, which is used as an example throughout this paper, is coloured in blue.

Characteristic	Period	Minimum	Median	Maximum
Surface area $[\mathrm{km}^2]$	-	5.3	225.5	13 483.5
Min. altitude $[\mathrm{m\,a.s.l.}]$	-	6.0	209.0	2 154.0
Median altitude $[\mathrm{m\,a.s.l.}]$	-	53.0	368.0	2 741.0
Max. altitude $[\mathrm{m\,a.s.l.}]$	-	93.0	784.0	3 997.0
Median slope $[\mathrm{deg}]$	-	1.1	7.4	35.8
Median hydraulic length $[\mathrm{km}]$	-	2.1	19.0	200.7
Artificial land cover $[\%]$	-	0.0	2.1	18.2
Agricultural land cover $[\%]$	-	0.0	54.2	97.7
Forest land cover $[\%]$	-	0.0	43.5	100.0
Mean annual precipitation $[\mathrm{mm\,y}^{-1}]$	Calibration	651	1 009	2 204
	Evaluation	691	1 025	2 077
Fraction of solid precipitation $[\%]$	Calibration	0.3	2.5	59.1
	Evaluation	0.0	2.2	50.3

2.2 Hydrological ~~modelling~~model

The GR4J model is a lumped conceptual daily rainfall–runoff model (Perrin et al., 2003). In this model, the effective precipitation is derived from the reduction in total precipitation by vegetation interception and by evapotranspiration from a soil moisture accounting production store. The effective precipitation is then routed through two unit hydrographs and one routing store. Groundwater exchange can occur from or to neighbouring catchments. A complete description of the model's equations is provided by Perrin et al. (2003).

This model contains four free parameters to calibrate against streamflow observations: the ~~maximal~~maximum capacity of the production store (X1, in mm), the groundwater potential exchange (X2, in $\mathrm{mm\,d^{-1}}$), the 1-day ahead routing store capacity (X3, in mm) and the time characteristics of the unit hydrographs (X4, in d).

For the catchments with a proportion of solid precipitation (considered here as precipitation occurring with negative air temperatures) greater than 10 % of the total precipitation, a snow model, CemaNeige, was used. This model is based on a degree–day approach and comprises two parameters to calibrate: the melt rate coefficient (K_f, in $\mathrm{mm\,^{\circ}C^{-1}\,d^{-1}}$) and a parameter regulating the energy of the snowpack (c_T, dimensionless). In order to ~~take into account the catchment heterogeneity~~consider intra-catchment variability, CemaNeige was applied to five elevation bands of equal area, which makes it possible to account for temperature and precipitation gradients (see Valéry et al., 2014, for more details).

~~While GR4J is the main model used in~~ In this work, ~~in order~~ we also use the GR6J model (Pushpalatha et al., 2011) to assess the transferability of the conclusions drawn~~, the GR6J (Pushpalatha et al., 2011) model is also used~~. GR6J adds two parameters to GR4J, X5 [-], which enables an inversion of the direction of the groundwater exchange throughout the year, and X6 [mm], which is the maximum capacity of an additional exponential store, whose purpose is to improve low-flow simulations.

All the calculations are made with the airGR R package (Coron et al., 2017, 2022). The built-in optimization algorithm, an initial parameter grid screening followed by a steepest gradient approach, is chosen due to its satisfactory performance with the GR models. All optimization criteria and streamflow transformations used in this work are embedded in airGR.

2.3 Optimization criteria

In order to assess the impact of transformations, the hydrological models are calibrated with several objective functions over the 1985–1995 period. However, in order to estimate how transformations impact the simulated time series, the 1995–2005 independent evaluation period is also used. In both cases, a one-year spin up period preceding the mentioned periods is used.

Three optimization criteria are chosen for their wide use in calibrating hydrological models: the well-known Nash–Sutcliffe efficiency (NSE, see Nash and Sutcliffe, 1970), the Kling–Gupta efficiency (KGE, see Gupta et al., 2009) and the modified Kling–Gupta efficiency (KGE', see Kling et al., 2012). The NSE concentrates most of the analyses of this work and the KGE and KGE' optimization criteria are used to assess the generality of the results. These three criteria are detailed in equations 1, 2 and 3.

$$E_{NSE} = 1 - \frac{\sum_{t=1}^{N}(Q_t^s - Q_t^o)^2}{\sum_{t=1}^{N}(\overline{Q^o} - Q_t^o)^2} \tag{1}$$

$$E_{KGE} = 1 - \sqrt{(r-1)^2 + (\frac{\overline{Q^s}}{\overline{Q^o}} - 1)^2 + (\frac{s_d(Q^s)}{s_d(Q^o)} - 1)^2} \tag{2}$$

$$E_{KGE'} = 1 - \sqrt{(r-1)^2 + (\frac{\overline{Q^s}}{\overline{Q^o}} - 1)^2 + (\frac{C_V(Q^s)}{C_V(Q^o)} - 1)^2} \tag{3}$$

~~with~~ being N the total number of days of the test period, Q_t^s and Q_t^o the simulated and observed streamflows, respectively, at time step t, $\overline{Q^o}$ (resp. $\overline{Q^s}$) the average observed (resp. simulated) streamflow over the period, r the correlation coefficient, s_d the standard deviation and C_V the coefficient of variation.

2.4 ~~Transformations~~ Streamflow transformations

The hydrological models are calibrated with the three above-mentioned criteria and different transformations of streamflows. Nine to 11 transformations are used (Table 3) for each criterion. In addition to the transformations mentioned in the Introduction, four additional transformations are used. The squared (Q^2) transformation is applied, as this can be used for focusing on floods (Tan et al., 2005), and its inverse (Q^{-2}) is applied as ~~this~~ it focuses on low flows. Furthermore, two composite criteria, $\frac{f(Q)+f(Q^{-1})}{2}$ and $\frac{f(Q)+f(log(Q))}{2}$ (with f standing for NSE, KGE or KGE'; Nicolle et al., 2014), are added since they can be used as a compromise between criteria focusing on too–specific ranges of streamflows. The two transformations containing the use of logarithm are not applied to KGE and KGE', as they cause numerical instabilities and unit-dependence, as shown by Santos et al. (2018). Regarding the Box–Cox transformation, equation 10 by Santos et al. (2018) is used to avoid the same issues as for the logarithmic transformation with a lambda value equal to 0.25 as suggested by Vázquez et al. (2008a).

3 ~~Methods~~

2.1 Evaluation methodology

In order to evaluate the impact of the transformations on model calibration, several analyses are made. They all rely on a common analysis framework, which aims at analysing the behaviour of transformations at every simulation time step. The general methodology, which is applied for each catchment and for each objective function, is detailed below and in Fig. 3 (for illustrative purposes for only two transformations):

1. the hydrological model is calibrated against observed streamflows for a catchment and with a given objective function, successively with different transformations (Fig. 3a),

Table 3. The 11 transformations used in this study and the criteria they are applied to. The abbreviations provided here are used in the figures and text.

Transformation	Abbreviation	NSE	KGE	KGE'
Q^2	2	✓	✓	✓
-	1	✓	✓	✓
$\sqrt{Q}$	0.5	✓	✓	✓
$Q^{0.2}$	0.2	✓	✓	✓
Box–Cox	$boxcox$	✓	✓	✓
$\frac{f(Q)+f(log(Q))}{2}$	$QlogQ$	✓		
$\frac{f(Q)+f(Q^{-1})}{2}$	$QinvQ$	✓	✓	✓
$log(Q)$	log	✓		
$1/\sqrt{Q}$	-0.5	✓	✓	✓
Q^{-1}	-1	✓	✓	✓
Q^{-2}	-2	✓	✓	✓

2. for each time step, the absolute error $|Q_t^s - Q_t^o|$ is calculated for the simulations obtained with the nine (or 11) transformations (Fig. 3b),

3. these daily absolute errors are ranked from the smallest to the largest among the nine (for KGE or KGE') or 11 (for NSE) simulations (Fig. 3c),

4. the time series of daily ranks are sorted according to the sorted observed streamflow time series (Fig. 3d),

5. the sorted ranks are aggregated over 200 sequential intervals of an equal number of time steps to smooth the results and facilitate the visual analysis. Three aggregations were made:

   – ~~calculation of the frequency of occurrence of all ranks (Fig. 3e),~~

   – extraction of the transformation with the most 'number 1' ranks (Fig. 3~~f~~e),

   – calculation of the average rank for each class (Fig. 3~~g~~f),

The use of ranks to classify the proximity between model simulations and streamflow observations could be criticized, since it gives the same importance to large and small errors. ~~However, ranks were found to be a good way of working on~~ This option was preferred to the use of direct (normalised) errors. Indeed, ranks allow to consider together various flow ranges, where the magnitude of errors can be very different. The impact of this methodological choice will be discussed later on.

The methodology above is applied catchment by catchment. Then, to aggregate results over the 325 catchments, we either identify the transformation with the most number 1 ranks or average the ranks over the 325 catchments.
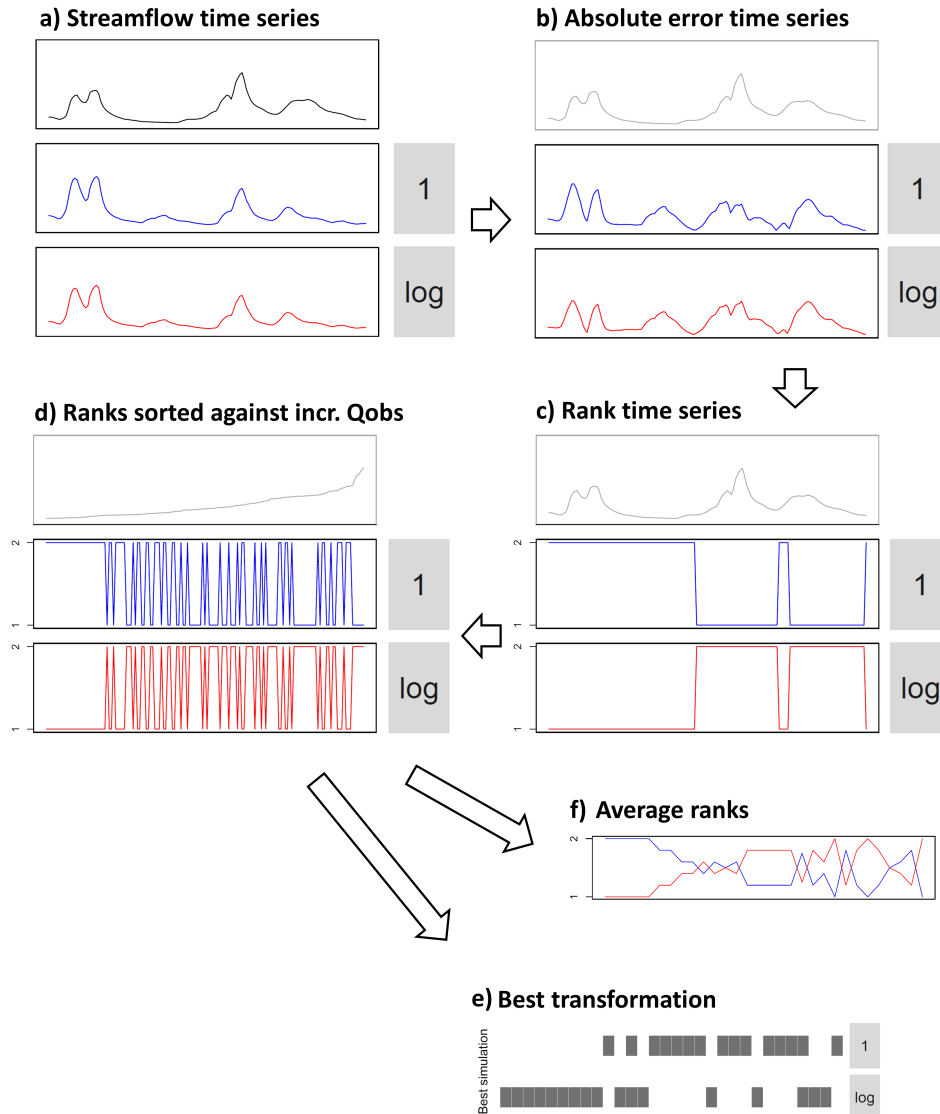
Figure 3. General methodology applied to assess the impact of transformations on the diverse ranges of streamflows. Here the example is shown over a short period and for only two transformations and a single catchment. In the study, the methodology is applied for nine to 11 transformations, 10-year periods and 200 intervals. a) Observed and simulated streamflow time series; b) Absolute values of errors in transformed streamflow time series; c) Ranking of error time series; d) Sorting of ranked time series according to increasing observed streamflows; For the next subplots, results are aggregated over intervals: e) ~~Calculation of the frequency of occurrence for each rank; f)~~ Identification of the transformation with most number 1 ranks for each interval; ~~e~~f) Calculation of average rank for each transformation and each interval. See the Methods section for more details.

3 ~~Results~~

When modellers choose an objective function (or, if relevant, a transformation), the main objective is to have a model fit for purpose, e.g. to be the best for low flows if the target is low flows. ~~Here~~ In the following, we evaluate the link between the objective function and transformation selected, and the accuracy of the model using the 200 flow intervals described above. We first perform this analysis on a single catchment, before applying it to 325 catchments.

3 Results

3.1 Analysis of the impact of transformations for a specific catchment

Figure 4 illustrates an example of the application of the methodology to a single catchment, the Fecht at Wintzenheim, for the GR4J model calibrated with the NSE objective function and for 11 different transformations. Here we show which transformation leads to the most number 1 ranks for each of the 200 intervals, for low flows (on the left) to high flows (on the right). It appears that some transformations are often ranked first (such as the -2, -1, 1 and 2 transformations). Conversely, some transformations rarely or never ranked first (such as $QlogQ$ or 0.2). In this figure, the transformations are represented in an order from a presupposed good representation of high flows (top row) to a presupposed representation of low flows (bottom row), because except for composite transformations, those are presented in an order of decreasing power. We can see that the logic is respected quite well, with transformations 2 and 1 being very well represented regarding intervals corresponding to high flows, and transformations -2 and -1 being very well represented regarding intervals corresponding to low flows. This does not preclude some transformations from being identified as the best one (or equally the best one, as ties are represented in Fig. 4) for unexpected intervals, such as transformation 2 that shows good results for some low-flow categories.
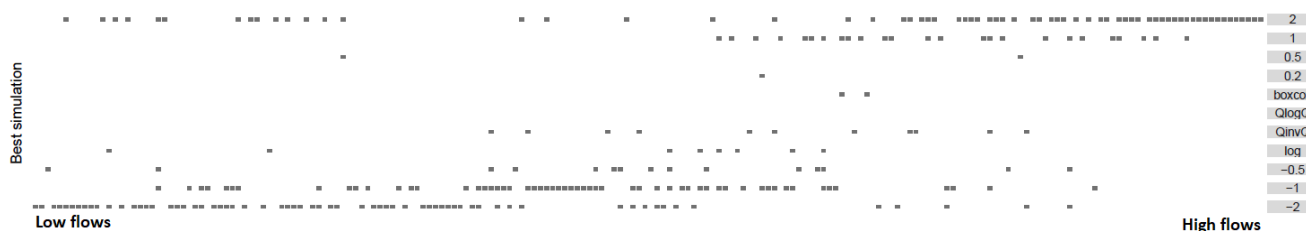
Figure 4. Identification of the simulation with the most number 1 ranks for 200 intervals ordered by increasing observed streamflows. Example for the Fecht River at Wintzenheim, over the calibration period (1985–1995), for the GR4J model calibrated with NSE. The CemaNeige model was not used. Each rectangle identifies for one interval which transformation(s) provides the most number 1 ranks, including ties.

~~Figure 4 only gives an incomplete overview of the behaviour of each transformation. Indeed, it is possible that some transformations behave reasonably well, while not being the closest to observations. Although having the best simulation for a given range of streamflows is interesting, having at one's disposal a simulation that behaves reasonably well for a rather~~

wide range of streamflows can also be interesting. Figure ?? provides elements for assessing the general performance of transformations, by showing the frequency of occurrence of the different transformations for all the possible ranks and for each

195 interval. In this figure we can identify two different types of transformations : those which are either among the best ones or among the worst ones, depending on the range of streamflows (e.g.2, 1, −1 or −2), and those which are generally neither the best ones nor the worst ones (e. g. 0.2, $boxcox$, $QlogQ$, etc.). However, it remains somewhat difficult to objectively qualify the performance of the 11 transformations A second representation is given in Fig. 5 with the average rank of transformations (i.e.

Frequency of appearance of each transformation for each rank (from 1 firstto 11 last) for 200 intervals ordered by increasing

200 observed streamflows. The frequency of appearance is calculated among all time steps for each interval (see Fig. 3e). Example for the Fecht River at Wintzenheim, over the calibration period (1985–1995), for the GR4J model calibrated with NSE.

A third representation is given in Fig. 5 with the average rank of transformations the average of the ranks of transformations over all the time steps of the concerned interval) for each of the 200 intervals. In this figure, we see that the transformations remain rather close together for low flows, with an average rank between 5 and 7. By contrast, the spread is larger for high

205 streamflows with average ranks between 4 and 9. Specifically, several transformations share the best average rank values for low flows, such as the −1, log and −0.5 transformations. Interestingly, the −2 transformation, which is supposedly the transformation giving the highest weight to low flows and was identified as the transformation with the most number 1 ranks for a high number of intervals in Fig. 4, only shows the best average rank for the very first interval, and then quickly shows a much worse average rank. This is in accordance with Fig. ??, where we clearly see that this transformation shows very bad

210 ranks quite often, indicating that it only fits a few time steps. This might indicate that the −2 transformation gives a high weight to errors over a limited number of time steps with the lowest streamflows (see Figure B1 in the Appendix for further analyses).

Regarding the middle range of streamflows, a couple of transformations show the best average rank, such as the −1, log and −0.5 transformations, but also progressively as streamflows get higher, the 0.2, $QinvQ$ and $boxcox$ transformations. Interestingly, this indicates that while being quite average most of the time , as highlighted in Fig. ??(analysis not shown here),

215 these transformations still have better average ranks than transformations with more occurrences of rank 1. Finally, regarding high flows, the 2, 1, $QlogQ$ and 0.5 transformations take the lead, quite logically.

3.2 Analysis of the impact of transformations for the 325 catchments

3.2.1 Analysis of the calibration period

While some trends could be identified in the analysis of a single catchment in the previous section, the results are impacted

220 by a rather high level of noise for successive intervals. To circumvent this issue, and to generalize the results, we perform a similar analysis over 325 catchments. This analysis is shown in Fig. 6 with the GR4J model calibrated with the NSE objective function. Results are presented for the calibration period. In this figure, the best simulation is identified for each catchment and for each interval according to the methodology presented in Fig. 3f. Then, for each interval, the simulation with the most number 1 ranks is labelled as the best. A clear pattern appears: The 2 transformation is the best for high flows, and the 1

225 transformation is the best for slightly lower flows. Regarding low flows, the best transformation for the most extreme flows is
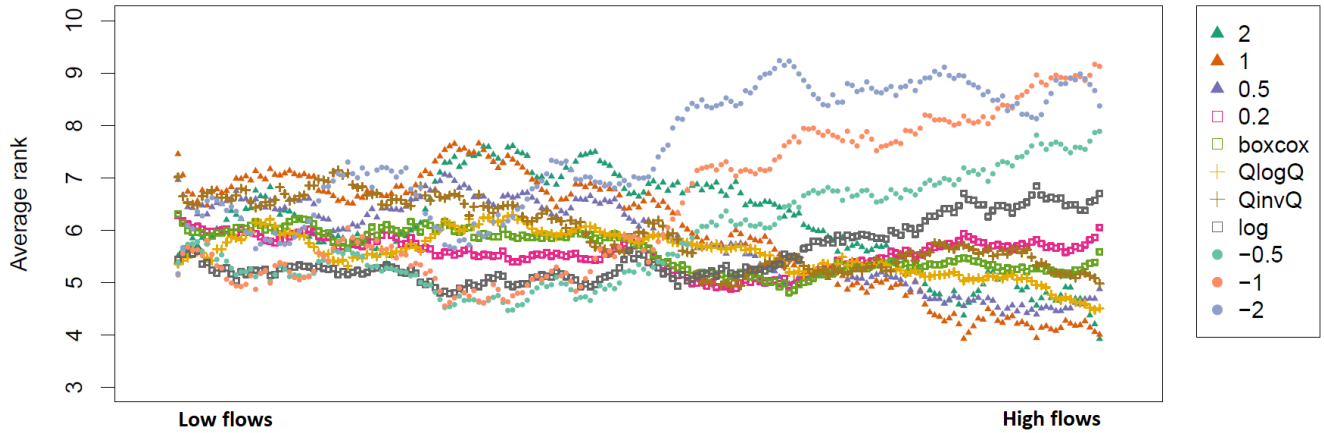
Figure 5. Average rank for each transformation. Example for the Fecht River at Wintzenheim, over the calibration period (1985–1995), for the GR4J model calibrated with NSE. A smoothing window (10-value moving average) is applied to improve legibility.

the -2 transformation, followed by the -1 and -0.5 transformations. This result confirms that the goal of transformations, which is to distort the streamflow time series, is easily reached when used for calibration. The only surprise is that just 5 out of the 11 transformations are identified as the best for at least one interval. However, the present analysis is binary and could lead to missing a more precise diagnosis.

Figure 6. Identification of the simulation with the most rank 1 for the 325 catchments and for 200 intervals ordered by increasing observed streamflows. To provide this analysis, the output of Fig. 3f is used, and for each interval, the number of number 1 ranks is cumulated over the 325 catchments to identify the simulation with the most number 1 ranks. The GR4J model is used and is calibrated with NSE. Results are shown for the calibration period.

    In order to better understand the behaviours of the different transformations, we show in Fig. 7 the interval-averaged rank of all transformations. The best average rank most of the time is between 4 and 5, except for high flows where it can reach 3.5
245   (the best being 1). We see that no transformation is always the best, even though some show a rather high interval-averaged rank throughout most of the intervals. Regarding the worst transformations, they show an interval-averaged rank around 8 to 9 out of 11; however, transformation 2 is clearly the worst transformation for low flows and −2 the worst one for high flows. Some trends can be observed. First, several transformations take the lead for low flows: the −2 transformation shows the best average rank for the very first intervals but quickly shows a worse average rank. The other leading transformations (i.e. with
250   best average rank), when going from the lowest flows to increasing flows, are successively transformations −1, −0.5, log, $boxcox$, 0.5 and finally 1. It is noteworthy that despite being identified as an excellent transformation for high flows in previous figures, transformation 2 is never the best transformation on average. This stems from the fact that even though it is the best high-flow transformation for many catchments, when it is not, its rank is rather low; we could qualify this transformation as an all-or-nothing transformation. Regarding the shape of the curves, we can distinguish four groups. First, transformations 2,
255   1 and 0.5 show a decreasing curve, i.e. have a best rank for high flows than for low flows. Conversely, transformations −2, −1 and −0.5 show an increasing curve, i.e. have a better rank for low flows than for high flows. Transformations $QlogQ$ and $QinvQ$ show the best ranks both for high and low flows, with the worst ranks for the mid-range of flows (arch-shaped curve). Finally, transformations 0.2, log and $boxcox$ show the best ranks for mid flows, and the lowest ranks for high and low flows (U-shaped curves).

260   Averaging the interval-averaged ranks over the 200 intervals provides an overview of the general ranking of transformations (Table 4). This analysis leads to the following ranking: transformations 0.2, log and $boxcox$ have the lowest (i.e. best) average rank, followed by 0.5, $QlogQ$, −0.5, $QinvQ$, 1, −1, −2 and 2. This result is interesting since most of the time, only one transformation is used by modellers for their application. If their application is very specific, this might make sense, as it is possible to identify transformations that outperform others. However, when their application is multi-purpose, it is very likely

265   that the transformation that is chosen only fits a limited range of streamflows. This is even more striking when we observe that the commonly used 1 transformation is very often applied for calibration despite being the best transformation for only a very limited portion of streamflow range, i.e. high flows. In addition, the transformations that show the best average rank are not the most widely used in the literature (0.2, log and $boxcox$).

Table 4. Interval-averaged ranks for a calibration of the GR4J model performed with NSE. Results are shown for the calibration period.

Transformation	Average rank
2	7.49
1	6.24
0.5	5.50
0.2	5.23
$boxcox$	5.25
$QlogQ$	5.59
$QinvQ$	6.12
log	5.30
-0.5	5.80
-1	6.39
-2	7.08

     The impact of using 200 instead of 100 or 500 intervals is illustrated in Appendix A, which shows that applying 200
270   intervals is a good compromise between too-coarse information and too-noisy information. It also shows that the results are only marginally impacted. We will therefore keep this number of intervals in the following. In addition, we will now only use figures similar to Fig. 7 (i.e. the evolution of average ranks through the range of streamflows) as it constitutes an efficient way of visualizing results and provides enough information to understand the behaviours of the transformations.

3.2.2   Are conclusions transferable to an independent period?

275   In the previous section, the results were presented for the calibration period, i.e. in optimal conditions to understand how transformations impact model simulations when used for model calibration. However, the purpose of using models is to apply them under conditions that are different from those they are calibrated on. Here we show the average range for the GR4J model calibrated with NSE over the evaluation period. The objective is to discuss whether the conclusions drawn for the calibration period are transferable to this independent period. Results are shown in Fig. 8.

280   Interestingly, the results are very similar to those of Fig. 7. The main discrepancy is that the average rank of the best transformation shows a higher value for the evaluation period than for the calibration period, and symmetrically the average rank of the worst transformation shows a lower value for the evaluation period. To phrase it differently, over the evaluation period, the transformations lead to simulations ~~coming from model/parameter combinations that lose some of their specificity~~that are less specific, i.e. closer to each other. For the lowest flows, the transformations log and 0.2 show the best average ranks, while for
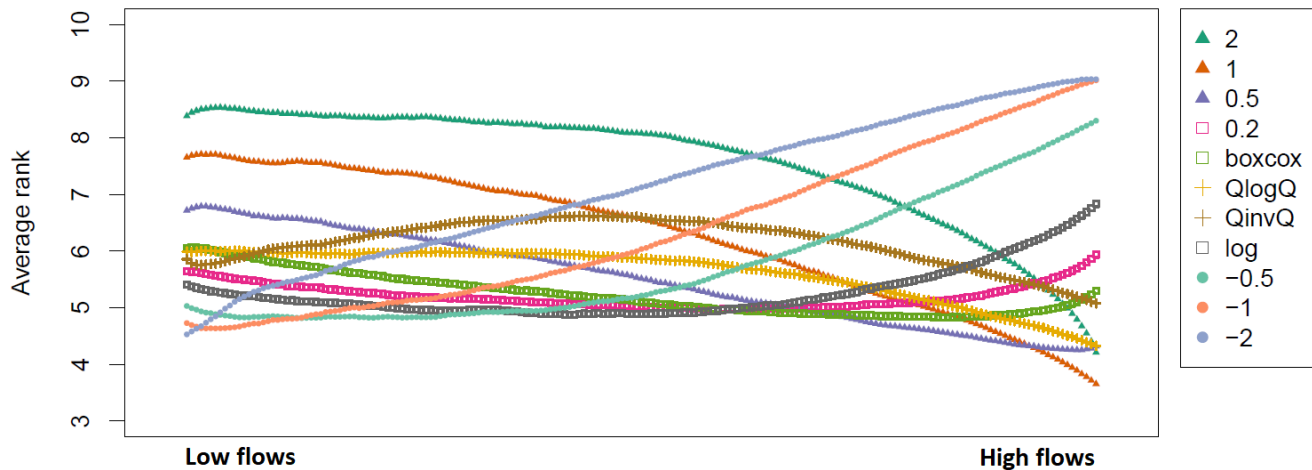
Figure 7. Interval-averaged rank over the 325 catchments and for 200 intervals ordered by increasing observed streamflows. To provide this analysis, the output of Fig. 3~~g~~ f is used, and for each interval, the mean rank is averaged over the 325 catchments. The GR4J model is used and is calibrated with NSE. Results are shown for the calibration period. A smoothing window (10-value moving average) is applied to improve legibility.

285 higher flows, transformations $boxcox$, 0.5 and 1 successively lead the pack. Transformations 2 and -2 are never at the top of the average ranks, indicating that in addition to being an all-or-nothing option for calibration, they are also poorly transferable to an independent period, even for their respective range of expertise. The averaging of the average ranks over the 200 intervals is shown in Table 5. This leads to the following ranking: transformations 0.2, $boxcox$ and log have the lowest average rank, followed by 0.5, $QlogQ$, -0.5, $QinvQ$, -1, 1, -2 and 2. Compared to the calibration period, this ordering is only marginally

290 modified and the average ranks over the 200 intervals are only slightly different. This indicates that, while not significantly modifying the conclusions of the previous analysis, over an independent period, the transformations lose specificity for their presupposed range of expertise, so as to gain performance for the rest of the streamflow range.

3.2.3 What is the impact of the choice of objective functions and the hydrological models?

All previous analyses were led with the GR4J model calibrated with NSE. In the following, we assess the impact of using

295 another hydrological model, GR6J, as well as two additional objective functions, KGE and KGE'. Although this model and these objective functions can be considered to be not drastically different from GR4J and NSE, we believe that they provide useful transferability information for the results. Indeed, these two new objective functions are increasingly reported in the literature, which justifies their use. The following analyses are made for the independent evaluation period only.

Figure 9 provides the interval-averaged ranks of the 11 transformations for the GR6J model calibrated with NSE, assessed

300 over the independent evaluation period. The general shape of this plot is rather similar to those of Fig. 8. The main differences concern transformation 2, which appears to ~~behave much worse~~ show much worse ranks than the other transformations for
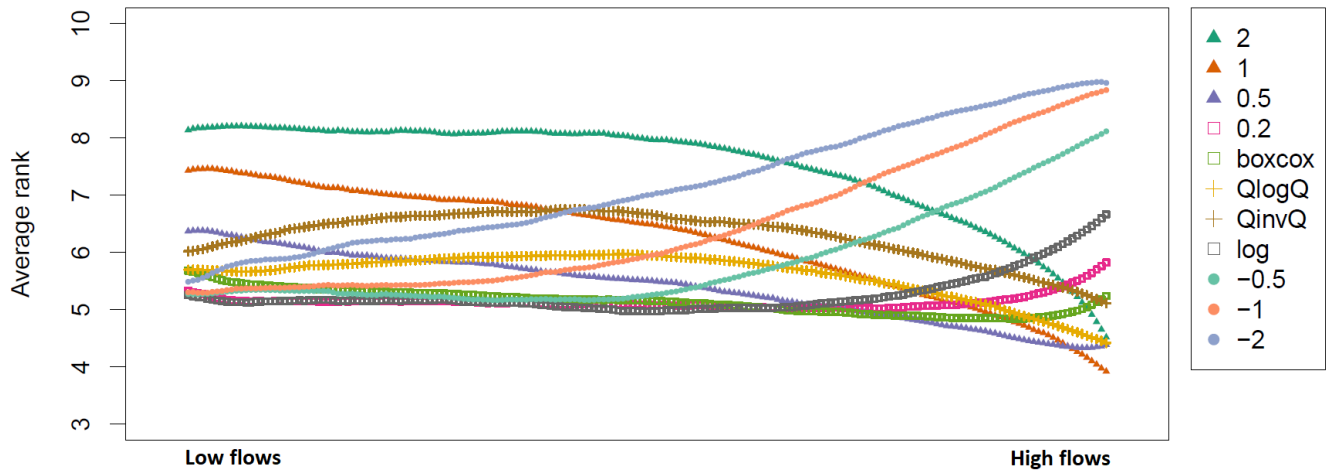
Figure 8. Average rank over the 325 catchments and for 200 intervals ordered by increasing observed streamflows. To provide this analysis, the output of Fig. 3g–f is used, and for each interval, the mean rank is averaged over the 325 catchments. The GR4J model is used and is calibrated with NSE. Results are shown for the independent evaluation period. A smoothing effect (10-value moving average) is applied to improve legibility.

Table 5. Interval-averaged ranks for a calibration of the GR4J model performed with NSE. Results are shown for the independent evaluation period.

Transformation	Average rank
2	7.44
1	6.19
0.5	5.40
0.2	5.16
boxcox	5.16
QlogQ	5.59
QinvQ	6.26
log	5.29
−0.5	5.88
−1	6.45
−2	7.16

GR6J than for GR4J, especially for low flows. This might stem from the fact that compared to GR4J, GR6J was mainly developed for improving low flows and therefore contains two more parameters to optimize, which focus on low-flow generating processes, and consequently could be less identifiable with transformation 2. We can also see that for most of the intervals (i.e. for the main part of the streamflow range), the best transformation is better identified for GR6J than for GR4J, as the latter

305

shows very close curves most of the time. Table 6 presents the values for the interval-averaged ranks. This table shows that the interval-averaged ranks are very similar for the two models except for transformation 2. We can therefore conclude that the analysis is only slightly dependent on the model used, although we must bear in mind that these models are partially similar.
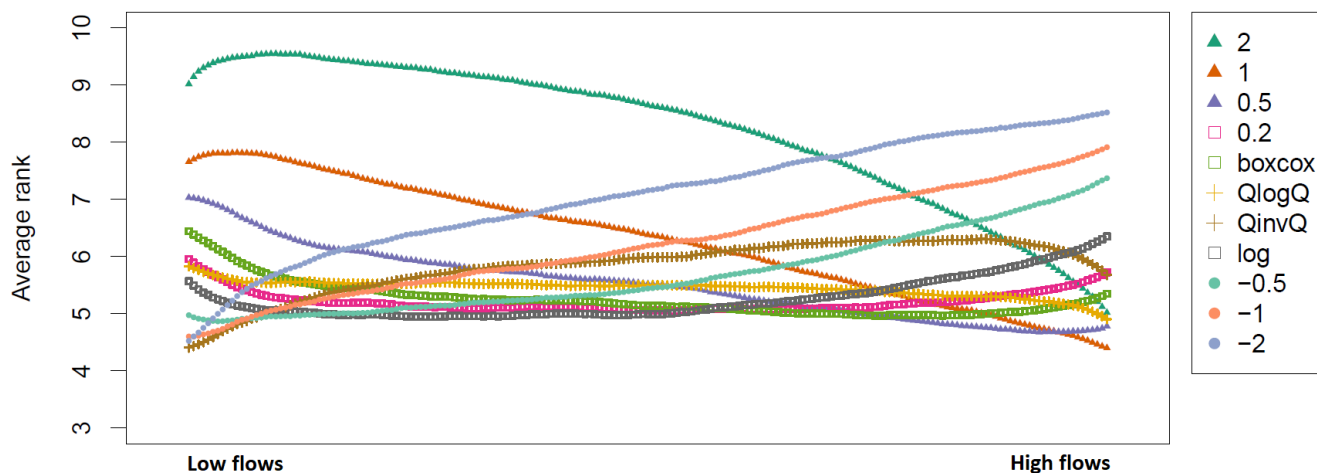
Figure 9. Same as Fig. 8 for the GR6J model calibrated with NSE.

Table 6. Interval-averaged ranks for a calibration of the two hydrological models performed with NSE. Results are shown for the independent evaluation period.

Transformation	GR4J	GR6J
2	7.44	8.15
1	6.19	6.28
0.5	5.40	5.54
0.2	5.16	5.25
boxcox	5.16	5.29
QlogQ	5.59	5.44
QinvQ	6.26	5.79
log	5.29	5.29
−0.5	5.88	5.71
−1	6.45	6.23
−2	7.16	7.04

Figure 10 shows how the interval-averaged ranks of transformations evolve when we use different objective functions. Interestingly, these two panels are similar to each other and to Fig. 8 (although the reader should note the absence of the two log-dependent transformations for KGE and KGE'). This means that the use of transformations seems to lead to similar ranges

of streamflows that are targeted by calibration whatever the objective function used, in particular for the very common NSE, KGE and KGE'. Here we do not show the interval-averaged ranks over the 200 intervals, as the number of transformations differs between NSE on the one hand, and KGE and KGE' on the other hand.
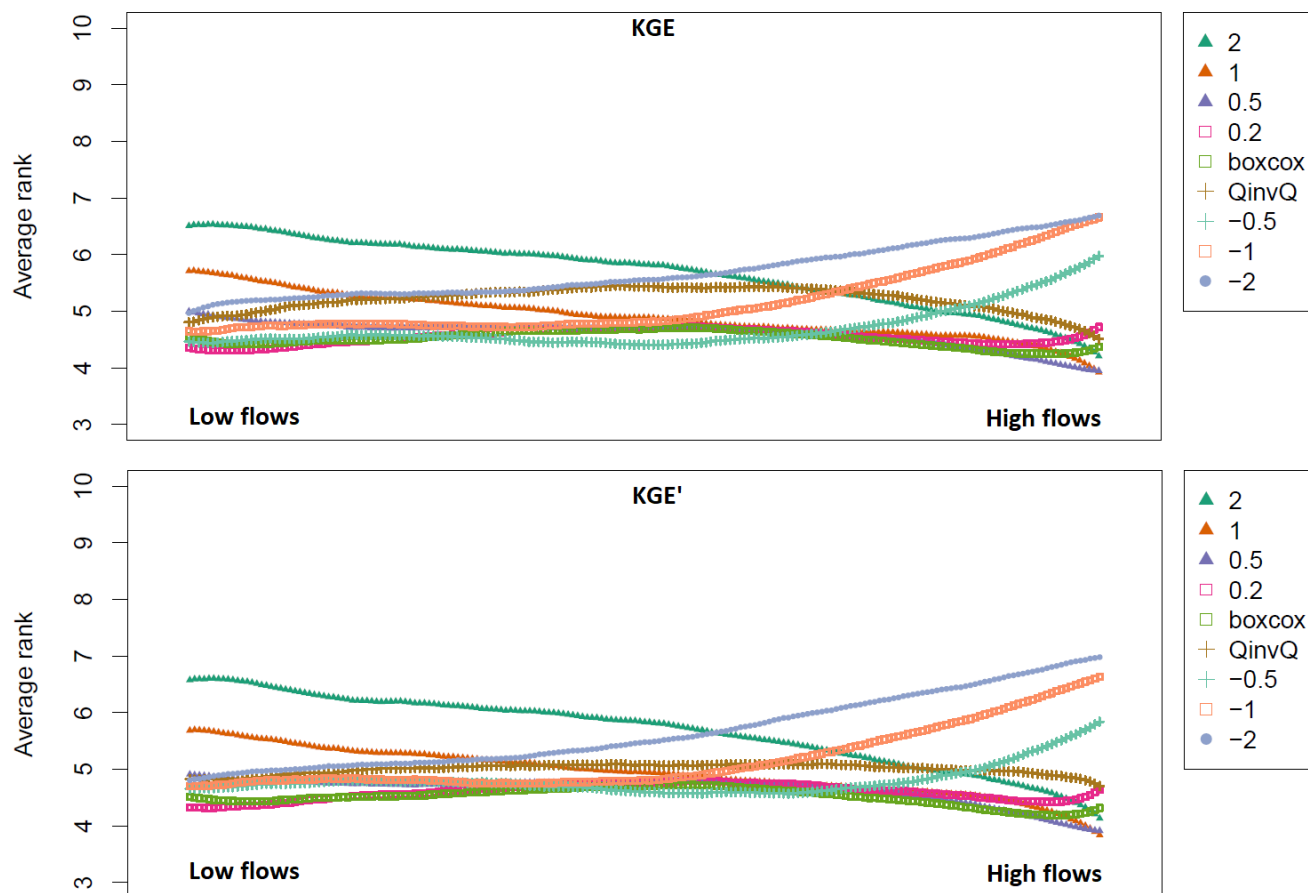
Figure 10. Same as Fig. 8 for the GR4J model and the independent evaluation period calibrated with KGE and KGE'. Note that the two transformations using a *log* transformation were not used with the KGE and KGE' objective functions.

3.3 Links between catchment characteristics and transformations

We tried to identify links between catchment characteristics and the performance of transformations to better understand the behaviour of the different transformations but also to potentially help prescribe transformations knowing the catchment characteristics. To do so, we used the Spearman correlation to analyse how the interval-averaged ranks of transformations for each catchment could relate to the catchment characteristics listed in Table 2. The GR4J model calibrated with NSE was used

here. In order to maximize the possibility of identifying strong links, only the catchment characteristics and interval-averaged ranks over the calibration period were used.

~~Unfortunately, only a few correlations could be identified.~~ The most important correlations were found between the BFI values and the $QlogQ$ (correlation equal to 0.56), $boxcox$ (0.51), 0.2 (0.50) and log (0.38) transformations. ~~Anti-correlations~~ Negative correlations were found between BFI and the -1 (-0.32) and -0.5 (-0.30) transformations. We observed that the central slope of the flow duration curve shows correlations of the same order of magnitude, but opposite to the one shown by the BFI, with exactly the same transformations. This indicates that these indicators are somehow well related. All the other catchment characteristics ~~showed~~ show no correlation over 0.30 or under -0.30, signifying a weak explanation of the performance of transformations with these characteristics. ~~Regarding BFI~~ Using the stats::cor.test() function in R, we found that the correlations lower than -0.11 and higher than 0.11 were all significant (i.e. p-values < 0.05), whereas none of the other ones were significant. Regarding BFI and the central slope of the flow duration curve, although the above-mentioned correlation values are interesting, they are not sufficient to permit any prescription of transformations for a specific kind of catchment. To deepen this analysis, a co-inertia analysis (Dolédec and Chessel, 1994; Dray et al., 2003) was undertaken on two principal component analyses made, respectively, for the table of catchment characteristics and the table of transformations, but could not show any further informative interdependence between catchment characteristics and transformations.

3.4   Impact of the method used to compare transformations

In this work, the transformations have been compared using ranks. The idea behind choosing to work with ranks, instead of direct (normalised) errors was to i) be less impacted by different orders of errors magnitudes between catchments or ranges of streamflows, and to ii) answer the question of which transformations are the best ones, rather than how good transformations are. Assigning ranks can however have the effect that a rank difference of 1 can both signify a small error or a larger one. Nevertheless, ranks are accounted for time step by time step, meaning that if differences between two simulations are very low, there can be changes in the order quite easily, which then results in similar average ranks over the intervals.

The impact of using absolute differences rather than ranks to compare transformations has been assessed in Thirel et al. (2023). The general shape of curves obtained with this alternative method were found to be similar, although some discrepancies were observed. The large errors of some outlier transformations made it difficult to identify differences between the other transformations. While this could lead to the conclusion that these transformations can be used interchangeably because they seem to lead to very similar errors, it is important to notice that the very high normalised MSE of one or two transformations lead to smaller differences for the other transformations. In other words, the results using MSE are impacted by the large error of some transformations and leads to less informative results. We therefore believe that the methodology based on ranks used in this study provides more readable results.

4 Conclusions

This study explored the impact of mathematical transformations applied to streamflow time series prior to the computation of objective functions for the calibration of hydrological models. Such transformations are often used to focus on different ranges of streamflows, but their actual impact has rarely been assessed. On the basis of the GR4J rainfall–runoff model and 11 transformations for the Nash–Sutcliffe efficiency criterion, we analysed the impact of transformations on streamflow simulations regarding the difference against observations. We ranked the 11 transformations for each time step and then aggregated the results at different scales. This first analysis on the Fecht River at Wintzenheim showed that, in general, the transformations indeed have the best ranks for the range of streamflows they are presupposed to focus on (e.g. the squared transformation focuses on high flows and shows good ranks for high flows). However, it was shown that some extreme transformations (squared and its inverse) were rather binary, i.e. were either very highly ranked or very poorly ranked, resulting in average ranks not being among the best for most of the streamflow range. In addition, the results also showed that some transformations can have a satisfactory performance for a range of streamflows they are not aimed at and with no clear reason, justifying further analyses on a larger set of 325 catchments.

This larger set of catchments allowed us to generalize the results and to smooth the transformation-ranking relationship. The analysis showed that only a few transformations were identified as being most frequently the best over the 325 catchments, with transformation -2 being the best for low flows, followed by transformations -1 and -0.5, while transformations 1 and 0.5 were most often the best ones for high flows. Complementary to this binary analysis, the calculation of the averaged ranks over the 325 catchments showed that the -2 transformation is only the best for very low streamflows, meaning that for many catchments, it is often a poorly performing transformation even for rather low flows. Symmetrically, the 2 transformation was found to be only efficient for very high flows. Some more intermediate transformations, such as the 0.2, log and $boxcox$ transformations, seem to be less specific but well-performing transformations, quite often being among the best transformations for high, mid and low flows.

Although first tested for the Nash–Sutcliffe criterion objective function, with the GR4J model and over the calibration period, this analysis was performed for two additional objective functions, one additional hydrological model and for the independent evaluation period. The results were only slightly modified, showing the robustness of the analysis.

The results of this study may have important implications for hydrological modellers. They show that no a priori assumption on streamflow transformations can be taken as warranted. Indeed, some transformations that are focused on extreme streamflows are shown to lead to calibrated models that are indeed better for these ranges over the calibration period, but that are poorly robust, i.e. that no longer necessarily perform well for this range of streamflows for an independent evaluation period. This might stem from the fact that these transformations rely on a limited number of time steps (see Figure B2 in the Appendix). In addition, these transformations are shown to lead to models that fit only a limited small range of streamflows. Conversely, some other transformations show a high performance for a large range of streamflows, and still lead to a reasonable performance for extreme streamflows. Namely, transformations 0.2, $boxcox$ and log show the best average rank both for the calibration and the evaluation period and may represent adequate transformations to use for many applications. These results

should encourage modellers to evaluate the streamflow transformations they use when calibrating hydrological models. The reader may, however, note that complementary aspects may be investigated, such as model robustness when applying hydrological models to climate change applications, flood peak or timing or other models time steps while working on flash flood modelling, for example.

Appendix A: On the impact of the number of intervals used

The whole range of streamflows was split into 200 intervals of equal length. The number of intervals could have an impact on the results. Consequently, we analyse in Fig. A1 how the use of 100 or 500 intervals may impact the conclusions. It appears in this figure that the general shape of all curves remains similar when modifying the number of intervals. However, when 500 intervals are used, the curves are less smoothed. This is understandable, since for smaller intervals the results can be more noisy. The main difference stems from the extremes (high and low flows). Indeed, the -2 transformation does not reach the top position for low flows and 100 intervals, while it does so for 200 intervals. Symmetrically, the 2 transformation is very close to the top position for 500 intervals and high flows, while it is not so for 100 intervals. We believe that although they are real, these differences do not modify the conclusions that were drawn in this manuscript.

Appendix B: On the contribution of largest error days of the total error

To assess the contribution of largest error days of the total error, the concept of fractional contribution to the squared error was used (Newman et al., 2015). The fractional contribution of squared error for the 1, 10, 100, 1000 days with the most error, was calculated for the 11 transformations for GR4J calibrated with NSE over the 325 catchments (Figure B2). It is very clear on this figure that extreme transformations rely on a more limited number of time steps than other transformations (see B for further analyses). A more complete description of the methodology used to compute the fractional error as well as an example on a single river are shown in Thirel et al. (2023).

Appendix C: Maps of catchment characteristics

In this Appendix, the statistics presented in Table 2 are provided as maps.

Author contributions. GT and LS conceptualized the study. GT developed the methodology. GT and OD implemented the analyses and visualizations. GT prepared the manuscript with contributions from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.
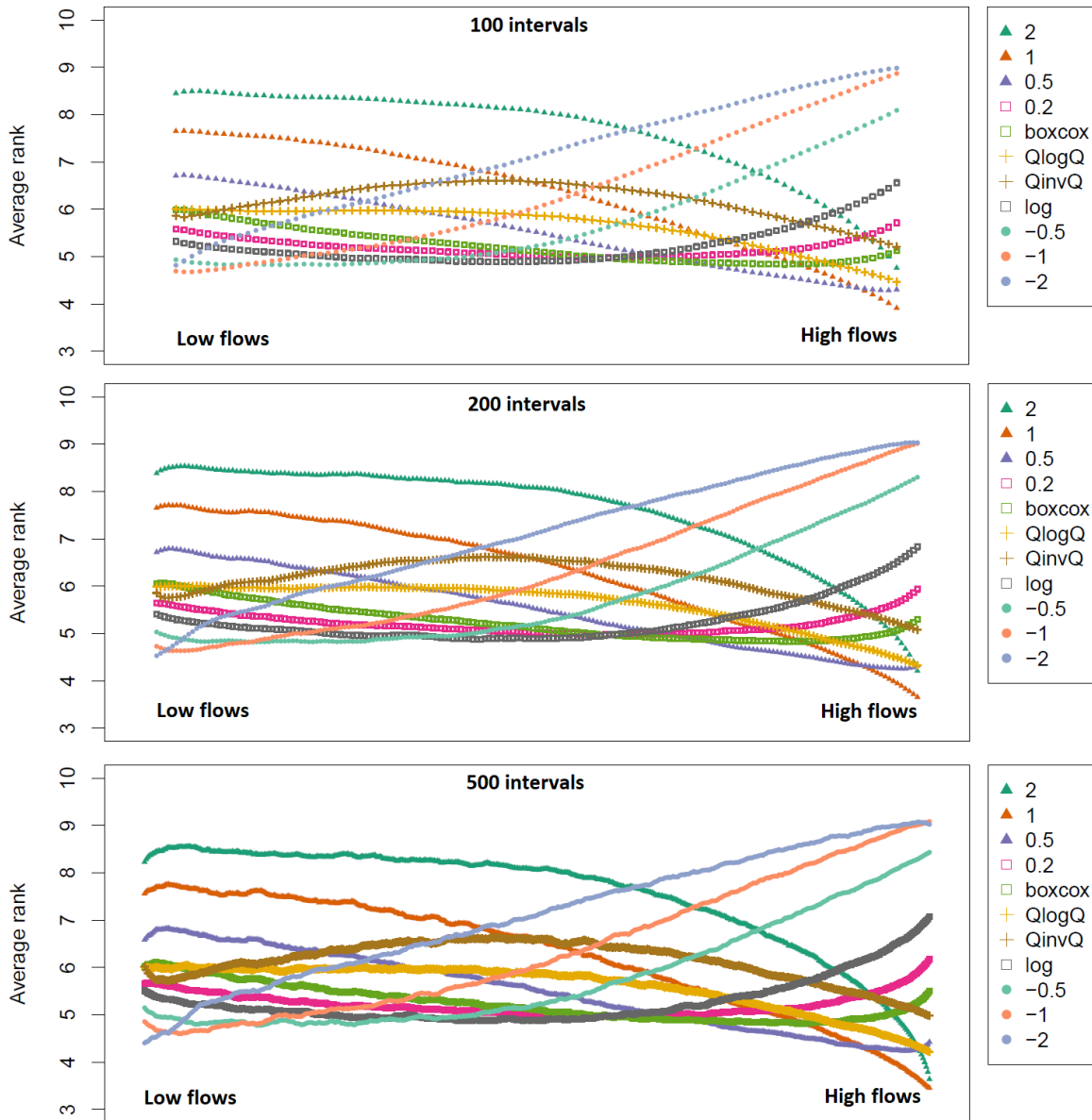
Figure A1. Same as Fig. 7 for the GR4J model and the calibration period, for three different numbers of intervals. The model was calibrated with NSE.

Acknowledgements. The SCHAPI and Météo-France are thanked for providing the hydro-meteorological dataset. Laurent Strohmenger and 410 Vazken Andréassian are thanked for their feedback on this study.
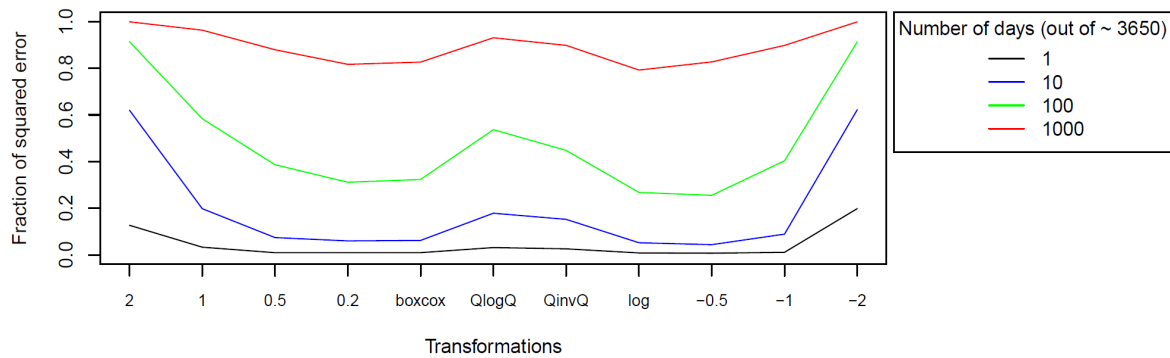
Figure B1. Fractional contribution for the GR4J model calibrated with NSE for the Fecht River at Wintzenheim for different numbers of time steps having the highest fractional contribution. Results shown on the calibration period.
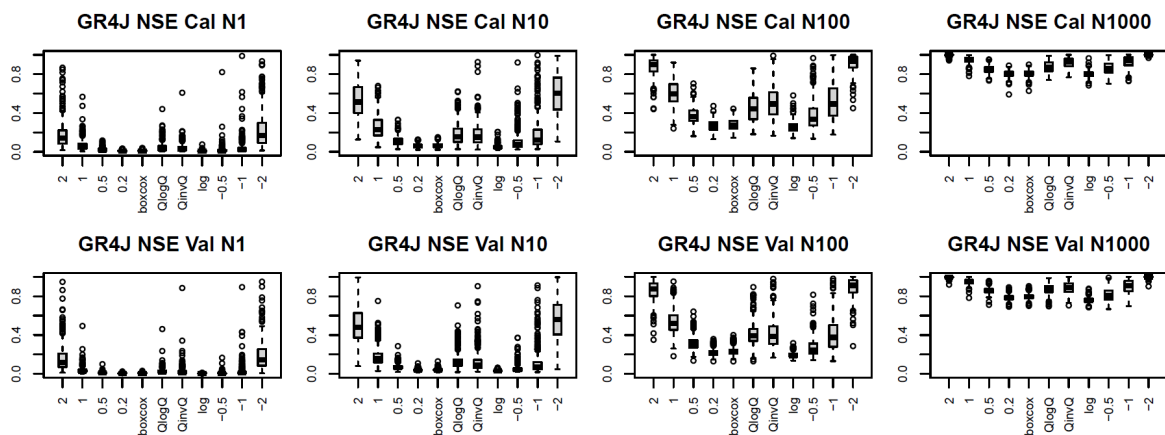
Figure B2. Fractional contribution for the GR4J model calibrated with NSE over the 325 stations. N1 to N1000 represent the number of time steps having the highest fractional contribution. Cal means calibration period, and Val means evaluation period.

References

Abdulla, F., Lettenmaier, D., and Liang, X.: Estimation of the ARNO model baseflow parameters using daily streamflow data, Journal of Hydrology, 222, 37–54, https://doi.org/10.1016/S0022-1694(99)00096-7, 1999.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, L. A.: Global-scale regional-

415   ization of hydrologic model parameters, Water Resources Research, 52, 3599–3622, https://doi.org/10.1002/2015WR018247, 2016.

Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., and Andréassian, V.: Characterising performance of environmental models, Environmental Modelling and Software, 40, 1–20, https://doi.org/10.1016/j.envsoft.2012.09.011, 2013.

Booij, M. J. and Krol, M. S.: Balance between calibration objectives in a conceptual hydrological model, Hydrological Sciences Journal, 55, 1017–1032, https://doi.org/10.1080/02626667.2010.505892, 2010.

Box, G. E. P. and Cox, D. R.: An analysis of transformations, Journal of the Royal Statistical Society. Series B (Methodological), pp. 211–252, 1964.

Chapman, T. G.: Effects of ground-water storage and flow on the water balance, in: Proceedings of "Water resources, use and management", Symposium held at Canberra by Australian Academy of Science, pp. 291 – 301, Melbourne Univ.Press, Victoria, 1964.

Chauveau, M., Bourgin, P., Peschard, J., and Coron, L.: Base de données d'observations hydroclimatiques à l'échelle de bassins versants français, https://webgr.inrae.fr/base-de-donnees, Université Paris-Saclay, INRAE, UR HYCAR, Équipe Hydrologie des bassins versants, Antony, 2011.

Chiew, F., Stewardson, M., and McMahon, T.: Comparison of six rainfall-runoff modelling approaches, Journal of Hydrology, 147, 1 – 36, https://doi.org/10.1016/0022-1694(93)90073-I, 1993.

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

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

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

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R., and Papalexiou, S. M.: The Abuse of Popular Performance Metrics in Hydrologic Modeling, Water Resources Research, 57, e2020WR029 001, https://doi.org/https://doi.org/10.1029/2020WR029001, 2021a.

Clark, M. P., Zolfaghari, R., Green, K. R., Trim, S., Knoben, W. J. M., Bennett, A., Nijssen, B., Ireson, A., and Spiteri, R. J.: The Numerical Implementation of Land Models: Problem Formulation and Laugh Tests, Journal of Hydrometeorology, 22, 1627 – 1648, https://doi.org/https://doi.org/10.1175/JHM-D-20-0175.1, 2021b.

Copernicus: Corine Land Cover (CLC), Version 2020_20u1, 2012.

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, https://CRAN.R-project.org/package=airGR, R package version 1.7.0, 2022.

Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., Mai, J., Serrer, M., Sgro, N., Shafii, M., Snowdon, A. P., and Tolson, B. A.: Flexible watershed simulation with the Raven hydrological modelling framework, Environmental Modelling & Software, 129, 104 728, https://doi.org/https://doi.org/10.1016/j.envsoft.2020.104728, 2020.

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

Dawdy, D. R. and Lichty, R. W.: Methodology of Hydrologic Model Building, Intl. Assoc. Hydr. Sci. Publ., 81, 1968.

de Vos, N. J., Rientjes, T. H. M., and Gupta, H. V.: Diagnostic evaluation of conceptual rainfall–runoff models using temporal clustering, Hydrological Processes, 24, 2840–2850, https://doi.org/10.1002/hyp.7698, 2010.

460 Ding, J.: Discussion of "Inflow hydrographs from large unconfined aquifers", by Ibrahim, H. A. and Brutsaert, W., J. Irrig. Drain. Div. Am. Soc. Civ. Eng., 92, 104 – 107, 1966.

Dolédec, S. and Chessel, D.: Co-inertia analysis: an alternative method for studying species–environment relationships, Freshwater Biology, 31, 277–294, https://doi.org/10.1111/j.1365-2427.1994.tb01741.x, 1994.

Dray, S., Chessel, D., and Thioulouse, J.: Co-inertia analysis and the linking of ecological data tables, Ecology, 84, 3078–3089, 465 https://doi.org/10.1890/03-0178, 2003.

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

Farmer, W. H. and Vogel, R. M.: On the deterministic and stochastic use of hydrologic models, Water Resources Research, 52, 5619–5633, https://doi.org/10.1002/2016WR019129, 2016.

470 Farr, T., Rosen, P., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D.: The Shuttle Radar Topography Mission, Reviews of Geophysics, 45, https://doi.org/10.1029/2005RG000183, 2007.

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

475 Fowler, K., Peel, M., Western, A., and Zhang, L.: Improved Rainfall-Runoff Calibration for Drying Climate: Choice of Objective Function, Water Resources Research, 54, 3392–3408, https://doi.org/https://doi.org/10.1029/2017WR022466, 2018.

Garcia, F., Folton, N., and Oudin, L.: Which objective function to calibrate rainfall–runoff models for low-flow index simulations?, Hydrological Sciences Journal, 62, 1149–1166, https://doi.org/10.1080/02626667.2017.1308511, 2017.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: 480 Implications for improving hydrological modelling, Journal of Hydrology, 377, 80 – 91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

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

Hogue, T. S., Sorooshian, S., Gupta, H., Holz, A., and Braatz, D.: A Multistep Automatic Calibration Scheme for River Forecasting Models, 485 Journal of Hydrometeorology, 1, 524 – 542, https://doi.org/10.1175/1525-7541(2000)001<0524:AMACSF>2.0.CO;2, 2000.

Houghton-Carr, H. A.: Assessment criteria for simple conceptual daily rainfall-runoff models, Hydrological Sciences Journal, 44, 237–261, https://doi.org/10.1080/02626669909492220, 1999.

Ishihara, T. and Takagi, F.: A study on the variation of low flow, Bulletin, Disaster Prevention Research Institute, 15, 75 – 98, http://hdl. handle.net/2433/124698, 1970.

490 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.

Knoben, W. J. M., Woods, R. A., and Freer, J. E.: A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data, Water Resources Research, 54, 5088–5109, https://doi.org/https://doi.org/10.1029/2018WR022913, 2018.

Knoben, W. J. M., Freer, J. E., Fowler, K. J. A., Peel, M. C., and Woods, R. A.: Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations, Geoscientific Model Development, 12, 2463–2480, https://doi.org/10.5194/gmd-12-2463-2019, 2019.

Krause, P., Boyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Advances in Geosciences, 5, 89–97, https://doi.org/10.5194/adgeo-5-89-2005, 2005.

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.

McMillan, H., Westerberg, I., and Branger, F.: Five guidelines for selecting hydrological signatures, Hydrological Processes, 31, 4757–4761, https://doi.org/https://doi.org/10.1002/hyp.11300, 2017.

Melsen, L. A., Teuling, A. J., Torfs, P. J., Zappa, M., Mizukami, N., Mendoza, P. A., Clark, M. P., and Uijlenhoet, R.: Subjective modeling decisions can significantly impact the simulation of flood and drought events, Journal of Hydrology, 568, 1093–1104, https://doi.org/https://doi.org/10.1016/j.jhydrol.2018.11.046, 2019.

Mendoza, P. A., Clark, M. P., Mizukami, N., Gutmann, E. D., Arnold, J. R., Brekke, L. D., and Rajagopalan, B.: How do hydrologic modeling decisions affect the portrayal of climate change impacts?, Hydrological Processes, 30, 1071–1095, https://doi.org/https://doi.org/10.1002/hyp.10684, 2016.

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.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrology and Earth System Sciences, 19, 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Le Lay, M., 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.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, Journal of Geophysical Research: Atmospheres, 116, https://doi.org/https://doi.org/10.1029/2010JD015139, 2011.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., and Michel, C.: Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations, Water Resources Research, 42, https://doi.org/10.1029/2005WR004636, 2006.

Pechlivanidis, I. G., Jackson, B., McMillan, H., and Gupta, H.: Use of an entropy-based metric in multiobjective calibration to improve model performance, Water Resources Research, 50, 8066–8083, https://doi.org/10.1002/2013WR014537, 2014.

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.

Pelletier, A., Andréassian, V., and Delaigue, O.: baseflow: Computes Hydrograph Separation, https://doi.org/10.15454/Z9IK5N, https://cran.r-project.org/package=baseflow, R package version 0.13.2, 2021.

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.

Peña-Arancibia, J. L., Zhang, Y., Pagendam, D. E., Viney, N. R., Lerat, J., van Dijk, A. I., Vaze, J., and Frost, A. J.: Streamflow rating uncertainty: Characterisation and impacts on model calibration and performance, Environmental Modelling and Software, 63, 32–44, https://doi.org/doi.org/10.1016/j.envsoft.2014.09.011, 2015.

Pushpalatha, R., Perrin, C., Moine, N. L., 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.

Quesada-Montano, B., Westerberg, I. K., Fuentes-Andino, D., Hidalgo, H. G., and Halldin, S.: Can climate variability information constrain a hydrological model for an ungauged Costa Rican catchment?, Hydrological Processes, 32, 830–846, https://doi.org/10.1002/hyp.11460, 2018.

Rosbjerg, D. and Madsen, H.: Concepts of Hydrologic Modeling, chap. 10, American Cancer Society, https://doi.org/10.1002/0470848944.hsa009, 2006.

Sadegh, M., Shakeri Majd, M., Hernandez, J., and et al: Effects of Data Transformation on Bayesian Inference of Watershed Models, Water Resour Manage, 32, 1867–1881, https://doi.org/doi.org/10.1007/s11269-018-1908-6, 2018.

Santos, L., Thirel, G., and Perrin, C.: Technical note: Pitfalls in using log-transformed flows within the KGE criterion, Hydrology and Earth System Sciences, 22, 4583–4591, https://doi.org/10.5194/hess-22-4583-2018, 2018.

Seeger, S. and Weiler, M.: Reevaluation of transit time distributions, mean transit times and their relation to catchment topography, Hydrology and Earth System Sciences, 18, 4751–4771, https://doi.org/10.5194/hess-18-4751-2014, 2014.

Smakhtin, V., Sami, K., and Hughes, D.: Evaluating the performance of a deterministic daily rainfall-runoff model in a low-flow context, Hydrological Processes, 12, 797–811, https://doi.org/10.1002/(SICI)1099-1085(19980430)12:5<797::AID-HYP632>3.0.CO;2-S, 1998.

Smith, T., Marshall, L., and McGlynn, B.: Calibrating hydrologic models in flow-corrected time, Water Resources Research, 50, 748–753, https://doi.org/doi.org/10.1002/2013WR014635, 2014.

Song, J.-H., Her, Y., Park, J., and Kang, M.-S.: Exploring parsimonious daily rainfall-runoff model structure using the hyperbolic tangent function and Tank model, Journal of Hydrology, 574, 574 – 587, https://doi.org/10.1016/j.jhydrol.2019.04.054, 2019.

Sorooshian, S. and Dracup, J. A.: Stochastic parameter estimation procedures for hydrologie rainfall-runoff models: Correlated and heteroscedastic error cases, Water Resources Research, 16, 430–442, https://doi.org/10.1029/WR016i002p00430, 1980.

Stewart, I. T., Cayan, D. R., and Dettinger, M. D.: Changes toward Earlier Streamflow Timing across Western North America, Journal of Climate, 18, 1136 – 1155, https://doi.org/https://doi.org/10.1175/JCLI3321.1, 2005.

Tan, K., Chiew, F. H. S., Grayson, R., Scanlon, P., and Siriwardena, L.: Calibration of a Daily Rainfall-Runoff Model to Estimate High Daily Flows, in: In Zerger, A. and Argent, R.M. (eds) MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, pp. 2960–2966, 2005.

Thirel, G., Santos, L., Delaigue, O., and Perrin, C.: Reply on RC2, Hydrology and Earth System Sciences, 2023.

Valéry, A., Andréassian, V., and Perrin, C.: 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 – Sensitivity analysis of the CemaNeige snow accounting routine on 380 catchments, Journal of Hydrology, 517, 1176 – 1187, https://doi.org/10.1016/j.jhydrol.2014.04.058, 2014.

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, 2010.

570 Vázquez, R. F., Willems, P., and Feyen, J.: Improving the predictions of a MIKE SHE catchment-scale application by using a multi-criteria approach, Hydrological Processes, 22, 2159–2179, https://doi.org/https://doi.org/10.1002/hyp.6815, 2008a.

Vázquez, R. F., Willems, P., and Feyen, J.: Improving the predictions of a MIKE SHE catchment-scale application by using a multi-criteria approach, Hydrological Processes, 22, 2159–2179, https://doi.org/10.1002/hyp.6815, 2008b.
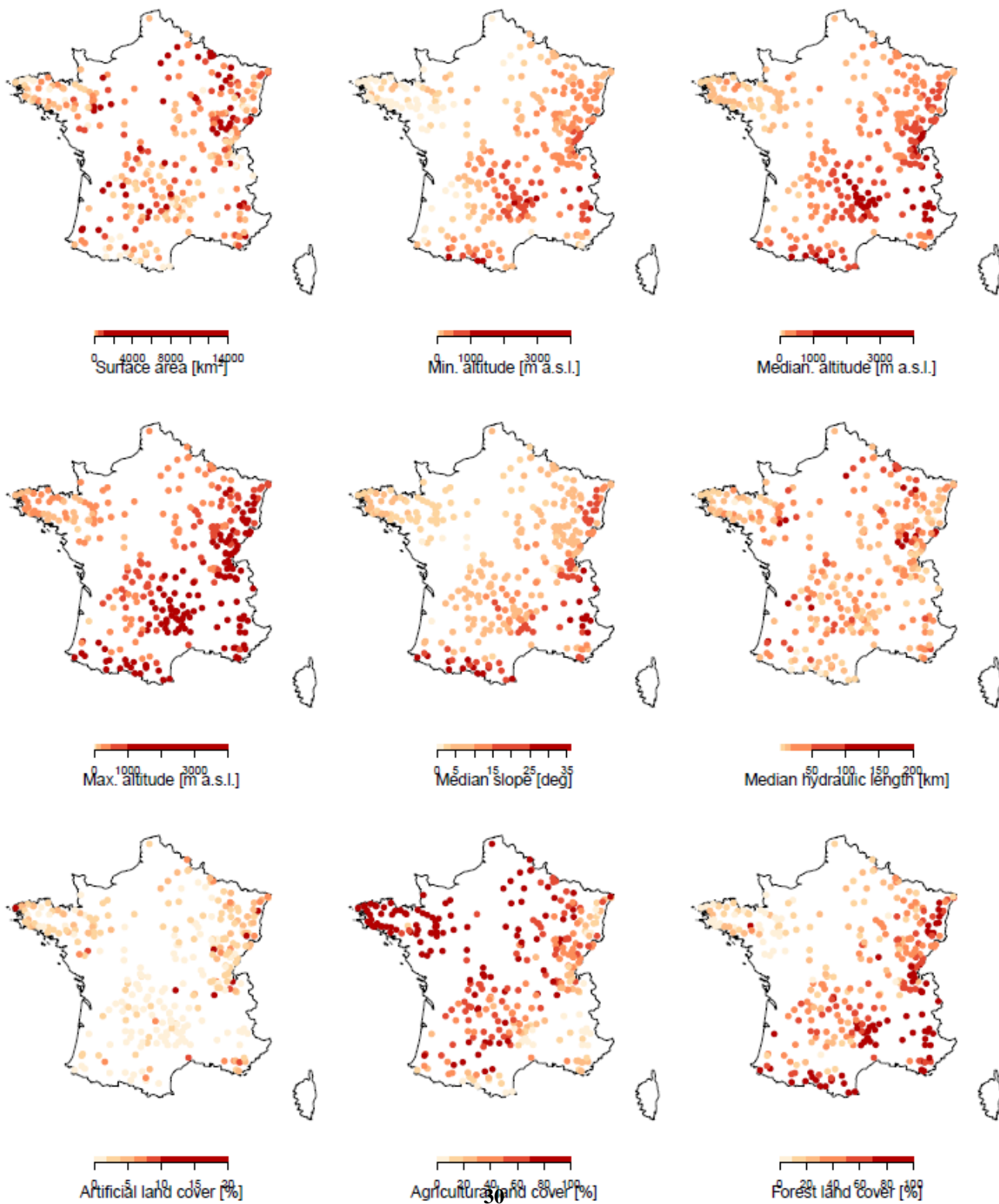
Figure C1. Maps of physical characteristics.
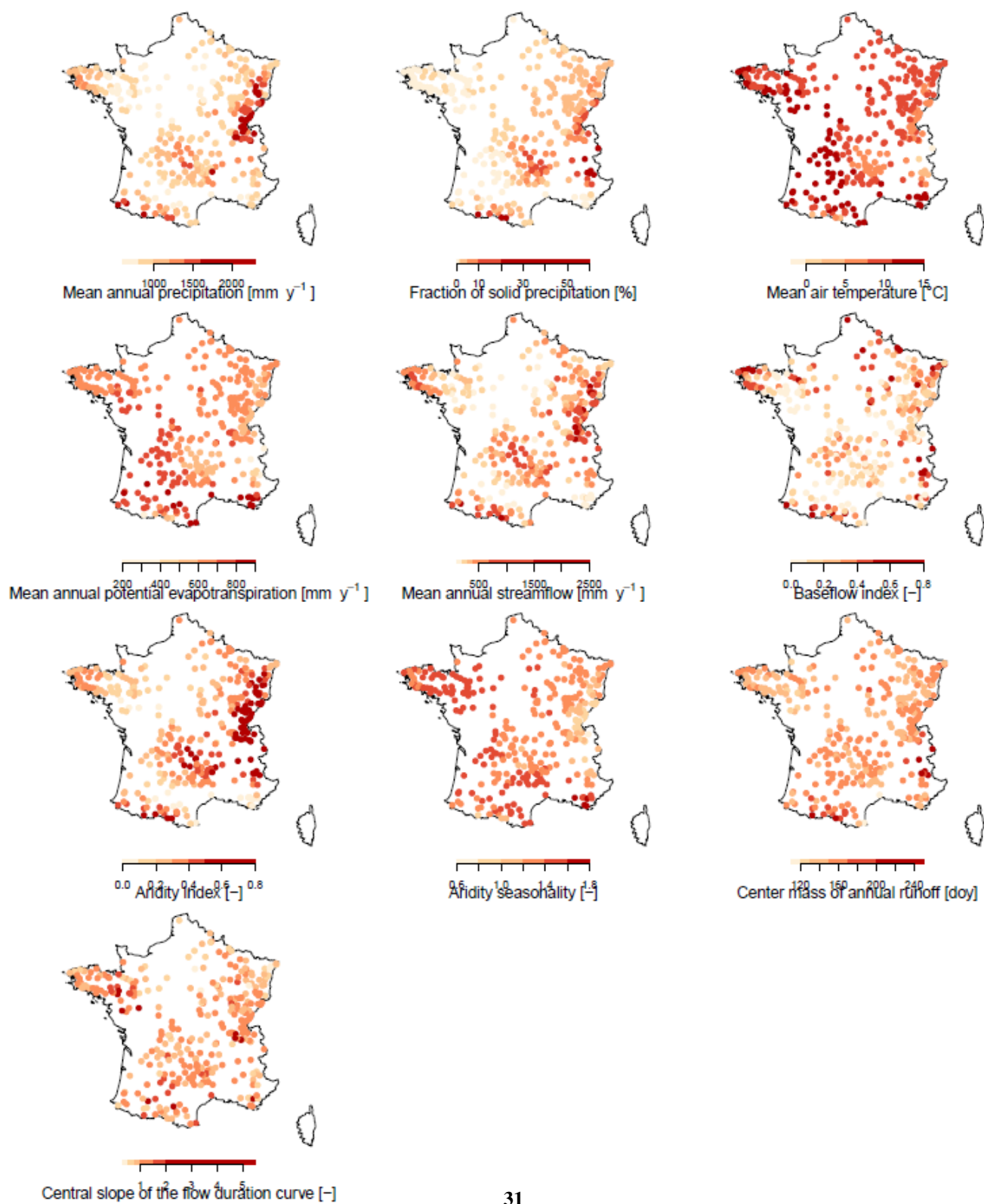
Figure C2. Maps of indicators over the calibration period.
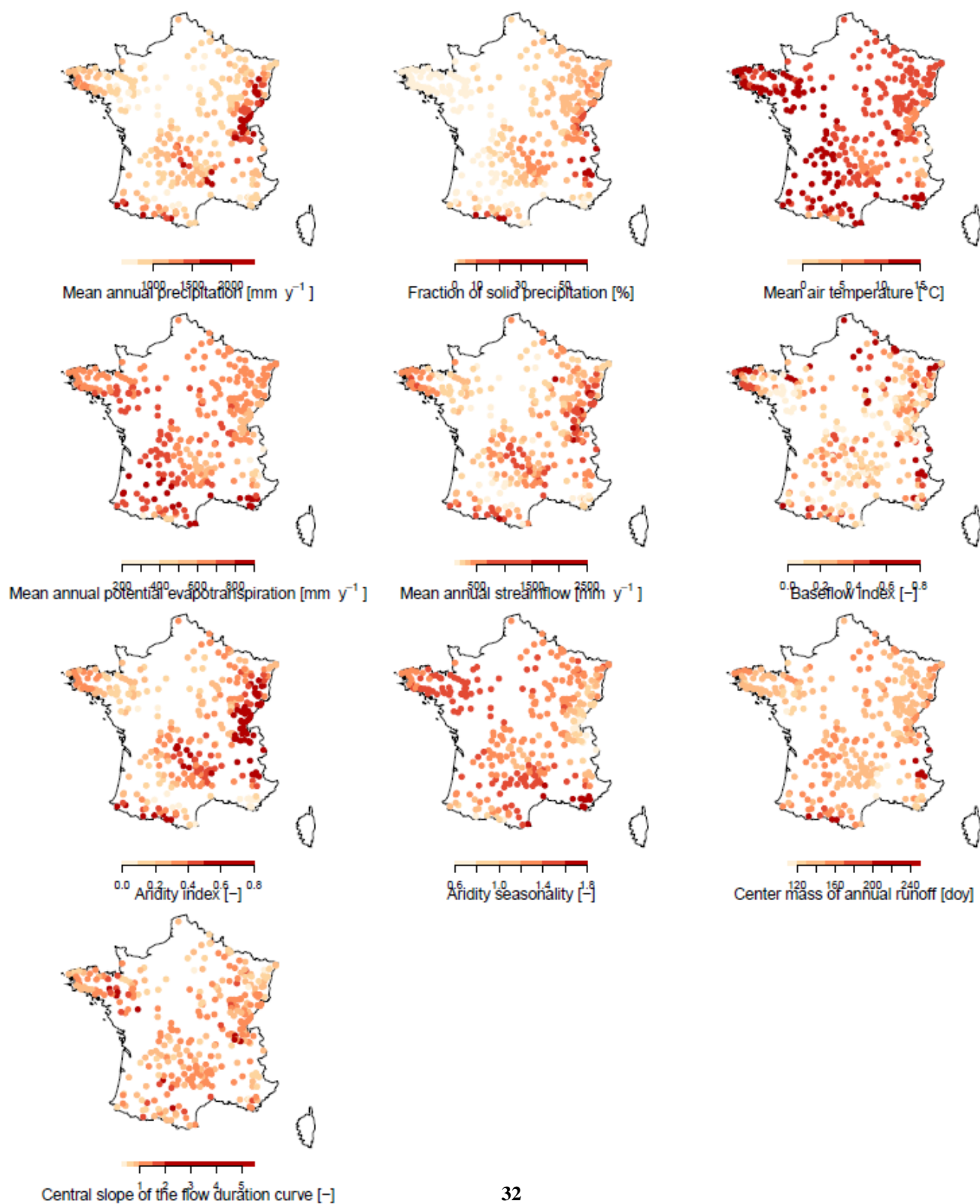
Figure C3. Maps of indicators over the evaluation period.