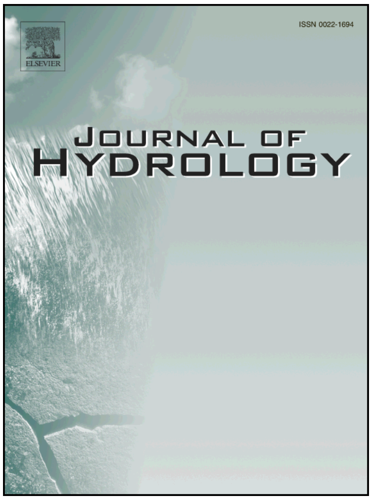

Research papers

The hazards of split-sample validation in hydrological model calibration

Richard Arsenault, François Brissette, Jean-Luc Martel

PII: S0022-1694(18)30714-5
DOI: https://doi.org/10.1016/j.jhydrol.2018.09.027
Reference: HYDROL 23124

To appear in: *Journal of Hydrology*

Received Date: 10 May 2018
Revised Date: 7 September 2018
Accepted Date: 10 September 2018

Please cite this article as: Arsenault, R., Brissette, F., Martel, J-L., The hazards of split-sample validation in hydrological model calibration, *Journal of Hydrology* (2018), doi: https://doi.org/10.1016/j.jhydrol.2018.09.027

The hazards of split-sample validation in hydrological model calibration

Richard Arsenault[1]*, François Brissette[1] and Jean-Luc Martel[1]

[1]École de technologie supérieure, Department of Construction Engineering, 1100 Notre-Dame Ouest, Montréal (Québec) Canada H3C 1K3

* Corresponding author: richard.arsenault@etsmtl.ca

Highlights

- Calibration on full time series is shown to be more robust than split-sample methods
- 30 Bootstrapping tests on 6 cases provide evidence towards this method being optimal
- Verification on 10 independent catchment-model pairs support the conclusions
- Caveats of split-sampling on model performance are demonstrated
- Length of the calibration period is proportional to the parameter set robustness

Keywords

Model validation, model calibration, model performance, hydrological modeling, split-sample testing

Abstract

This paper investigates the issues related to the use of validation in hydrological model calibration. Traditionally, models are calibrated and then assessed on an independent period (split-sample) to determine their adequacy in simulating streamflow as compared to observations. In this study, two hydrological models and three North American catchments are used to evaluate the effects of using validation to assess the model parameters' robustness on the model's actual simulation capabilities and accuracy in simulating streamflow. The length of the calibration period is increased from 1 to 16 years, and for each case a large number of randomly selected combinations of years are used for calibration and for validation using the Nash-Sutcliffe Efficiency metric. The calibrated model is then run on an independent 8-year test-period to assess the model's actual performance in simulation mode in unknown conditions. The process is bootstrapped 30 times to ensure the robustness of the results. The tests pit the calibration/validation methods on increasing calibration period lengths against a full calibration on the entire available dataset. Results show that the calibration on the full dataset is the optimal strategy as it generates the most robust parameter sets, provides the best model accuracy on an independent testing period and does not require assumption-making on the modeler's part. The calibrated parameter sets for each test-case were evaluated using the relative bias and correlation metrics, which revealed that the method transfers well to these two other metrics. Results also demonstrate the pitfalls of the commonly used split-sampling strategy, where good parameter sets may be discarded due to model performance discrepancies between calibration and validation periods. The conclusions point to the need to use as many years as possible in the calibration step and to entirely disregard the validation aspect under certain conditions.

1. Introduction

Hydrological models are amongst the most commonly used environmental models. They are commonly used for streamflow forecasting (e.g. Arsenault et al. 2016a; Day 1985; Nash and Sutcliffe 1970; Wood et al. 2002), climate change impact studies (Bergström et al. 2001; Chen et al. 2011a; Chen et al. 2011b; Jiang et al. 2007; Middelkoop et al. 2001; Minville et al. 2008) and to better understand processes at various scales on catchments covering a wide array of sizes, from small urban parcels all the way up to the continental scale (Band et al. 1993; Fatichi et al. 2016; Gupta et al. 2014; Singh and Woolhiser 2002). Many of the physical processes governing water and energy balance at the parcel or catchment scale are not well understood or in need of unavailable data and, as a result, all hydrological models require some level of parameterization (Andreassian et al. 2006; Jakeman and Hornberger 1993). Consequently, a calibration and validation process is generally needed to assess the hydrological model performance on any given catchment.

There is confusion in the literature about the precise meaning of model validation. Model calibration, validation, evaluation and even verification have been used interchangeably in some papers to describe if a model is adequate at representing observations. In particular, the term "validation" has been used to mean different things by different authors. In this paper the definitions of Refsgaard et al. (2005) are used since they are closest to the general agreement in the hydrological community.

1.1. Definitions

In the modeling community, model validation, model verification or model evaluation is the process by which a comparison is made between model outputs and observations to evaluate the adequacy of the model (Legates and McCabe 1999). Some authors differentiate verification (computer code accuracy in solving equations) from validation (ability of the model to represent the underlying physical process). Model verification and validation should be seen as the process for which one verifies if a model is adequate and with the assumption of similar performance if the model is used in similar conditions (KlemeŠ 1986). Furthermore, a distinction can be made between various types of validation (e.g. scientific vs performance validation as discussed in Biondi et al. (2012)). Performance validation (ability to represent observations) is the one that is typically made for hydrological models. A review of environmental performance evaluation is presented in Bennett et al. (2013). In the hydrological modeling community, validation has a much stricter sense ascribed to the process of demonstrating the model ability to perform outside of its training period. This is the definition of Refsgaard et al. (2005) which is adopted in this paper for the sake of consistency with the work of many other authors in hydrology.

Model calibration consists in adjusting model parameters over a training period, either manually or automatically (Arsenault et al. 2014; Boyle et al. 2000; Duan et al. 1994; Gupta et al. 2014), so that model outputs match observations as closely as possible. The adequacy of parameters adjustment is typically based on a single objective function representing the similarity between model outputs and observations. An objective function is a mathematical equation to be minimized in order to insure maximal similarity between model outputs and observations. In hydrological modeling, the most commonly used objective function is the Nash-Sutcliffe Efficiency metric (*NSE*; Nash and Sutcliffe 1970) which minimizes the root mean square error between modeled and observed streamflows. Many other objective functions have also been proposed (e.g. Garcia et al. 2017; Gupta et al. 2009; Moriasi et al. 2007). It should be noted that model calibration implies some level of model verification (or validation in its wide sense) since a bad performance would normally stop the modeling process and the need for validation (in its stricter sense as used in this paper).

Following calibration, a validation is typically performed over a different period to ensure parameter transferability and model robustness. This practice of using two different periods for calibration and validation is referred to as the split-sample approach.

1.2. Split-sample approach

The split-sample approach is a classic method which is central to KlemeŠ (1986) hierarchical scheme for validating hydrological models. It has been implemented in many ways over the years, with each variant aiming to target a specific calibration goal. The most common split-sample approach in the literature is the two-period method, by which the calibration and validation periods are split in two sections of approximately equal length. This method has the advantage of being easy to implement and minimizes the model runtime which is very helpful when the hydrological model is computationally intensive. For similar reasons, sometimes the calibration period is longer than the validation period or vice-versa. Tolson and Shoemaker (2007) implemented a 6-year calibration and two independent validation series of 3 years and 1 year respectively on the Cannonsville Reservoir catchment using the SWAT2000 model. They used these two different validation periods to verify the model's adequacy on contrasting hydrological conditions and found that the validation *NSE* was higher than the calibration *NSE*, leading to the conclusion that the parameter optimization was robust. Arnold et al. (2012) recommends using a long-enough period to encompass varying conditions, e.g. dry/wet years. Larabi et al. (2018) used eight calibration years followed by four validation years in an implementation of a Functional Data Analysis (FDA) based calibration scheme. Liu et al. (2018) specifically tackle the challenges related to sample-splitting and provide an in-depth analysis of the shortcomings of using this approach.

The second option is a mixed-bag approach by which calibration and validation years are sampled randomly throughout the dataset, sometimes using bootstrapping methods to resample and draw robust parameter sets, other times by identifying unusual events to use during calibration (Razavi and Tolson 2013; Singh and Bárdossy 2012). This method has the advantage of generating multiple parameter sets that can be analyzed against the calibration and validation performance, but the tradeoff is that the computing time for the calibration aspect grows linearly with the number of calibration years and the length of time over which they are spread out. In the case where some calibration years are discontinuous, the model must still be run on the entire time-series between the first and last calibration year to ensure that the model states are consistent in time. Wallner et al. (2012) used this method to calibrate the HEC-HMS hydrological model on five catchments in Germany for evaluating regionalization approaches. They calibrated on the years 2004, 2007 and 2008, and the validation was performed on the 2005-2006 period. However, no reason was given as to why these years were selected instead of using a half-and-half method, for example. Singh and Bárdossy (2012) compare the calibration performance using the entire dataset to a subset of unusual events, and find that the results are only marginally worse than when using the full dataset. Razavi and Tolson (2013) implemented a method that selects surrogate years representative of the entire time series and terminates the model simulation if the objective function does not seem promising after the first few evaluations, immediately switching to the new parameter set to be evaluated. Gharari et al. (2013) divided the time series in multiple small blocks (e.g. by week, by month, by wetness, etc.) and performed calibration of the HyMod model on each block. The best parameter sets are discovered by analyzing the parameter distributions for each subset and are then validated on independent periods.

Finally, the last split-sample method presented here is the odd-even approach. This method was developed as a response to calibration under non-stationary conditions. In essence, the method is based on calibrating the hydrological model on the odd years of the dataset and the validation is performed on the even years (or vice-versa). In this setup, the hydrological model is exposed to the non-stationary trend in both the calibration and validation steps, which in theory should allow the model to integrate the trend information in the parameter set and then use this information on the validation period. Essou et al. (2016) and Arsenault et al. (2017) used the odd/even split-sample testing to take into consideration the trends arising in long climate time-series during calibration. Gowda et al. (2012) used the same technique when calibrating the ADAPT model because the first and last halves of the time series were respectively wetter and drier than the average. Using the odd-even method ensured sampling the entire spectrum of available values.

It is also important to note that in this paper, the split-sample approach is used solely for calibration and validation of a hydrological model. In some papers (i.e. Butts et al. 2004), split-sampling is used to compare different hydrological model structures. This application of split-sampling is out of the scope of this paper and is thus not investigated.

1.3. Length of the calibration period

The question of the length of the calibration period has been the subject of many studies (Razavi and Tolson 2013). Vrugt et al. (2006) found that increasing the length of the calibration dataset improved the performance of the SAC-SMA hydrological model. Juston et al. (2009) and Perrin et al. (2007) investigated various sampling strategies to select a calibration dataset and found that sampling could be optimized so that calibration performance may reach that of calibrating over the full period. Razavi and Tolson (2013) came to similar conclusions using a surrogate shorter calibration period. In both papers, calibration over the entire dataset was considered the benchmark for comparison. van der Spek and Bakker (2017) looked at the length of the calibration period on the performance of a groundwater hydraulics model. They found the length of the calibration period to be more important than the frequency of observations. For an equal number of observations, a long record was clearly superior to a shorter one with a higher observation frequency.

Although a few authors imply that calibration over the full length of the available dataset may be preferable for parameter identifiability (e.g. Singh and Bárdossy 2012) the overwhelming majority of published work still advocates the split sample strategy (e.g. Daggupati et al. 2015; Gaborit et al. 2015; Garavaglia et al. 2017; Gaur et al. 2017; Liu et al. 2018; Moriasi et al. 2015; Newman et al. 2015).

1.4. Research Objectives

In this paper it is argued that in gray-box hydrological model (the type most commonly used in hydrology) where calibration is needed (whether manual or automatic) and which uses a standard “performance” model validation only, in those cases, validation is not needed and in fact detrimental to model performance. Refsgaard et al. (2005) called upon the community to investigate and develop better model validation methods, but in parallel, other authors hint to the idea that model calibration on the entire period is the best option to obtain a robust, time-transferable parameter set (Singh and Bárdossy 2012). In this paper, the issue is investigated in detail and the effect of using validation on model performance on an independent testing period is evaluated. More specifically, the question as to which option is the optimal choice between using a validation period and calibrating on the entire time-series (with no validation) is answered.

2. Study area and data

Three catchments' metadata (drainage area and outline) and streamflow time series from the MOPEX (Duan et al. 2006) and CANOPEX (Arsenault et al. 2016b) databases covering

North America were considered in this study. The meteorological data needed as input for the hydrological models (i.e. precipitation and temperature) were obtained from different observed datasets. For the two catchments located in the United States, daily precipitation and temperature were extracted from the gridded dataset (interpolated on a 0.125° x 0.125° grid) of the University of Santa Clara (Maurer et al. 2002). Regarding the third catchment located within Canada, precipitation and temperature were obtained from Natural Resources Canada (NRCAN) gridded datasets (interpolated on a 0.083° x 0.083° grid; Hutchinson et al. 2009). Figure 1 shows the location of the study sites as well as their elevation profiles. These catchments were selected based on four criteria:

1) At least 25 years of continuous data must be available with little to no missing values, i.e. less than 5%, in both the hydrometric and meteorological observation records;
2) There must be at least some snowfall on the catchment on a regular basis;
3) The hydrological models should perform to different levels based on the Nash-Sutcliffe Efficiency (*NSE*; Nash and Sutcliffe 1970) to explore cases where the models would be more or less challenged to produce good hydrographs.
4) The observation period must not contain a trend in observed streamflow as detected by a Mann-Kendall test (Kendall 1975) at a 95% confidence level.

The latter condition was imposed in order to allow random sampling of years whilst considering them as independent.

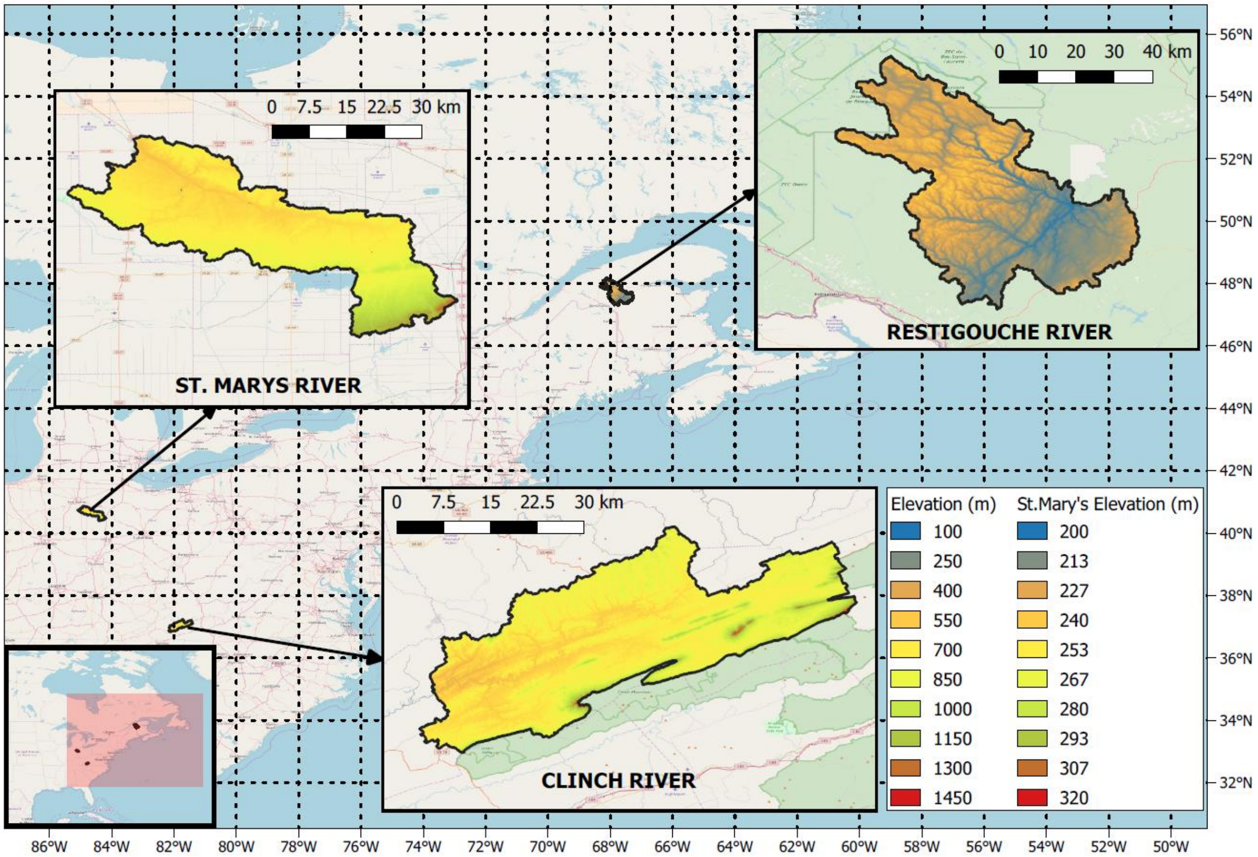

Figure 1. The three selected catchments' locations and elevation profiles (m). The St. Mary's River catchment is much flatter than the other two catchments, which explains why it has its own elevation legend.

Time and computing resources constraints limited the number of catchments that could be explored in this study; therefore, the catchments were selected in such a way as to maximize the variability in the test cases. Table 1 shows the selected catchment physical and hydrometeorological characteristics.

Table 1. The three selected catchments' characteristics and mean annual hydrometeorological variables.

	Restigouche River	Clinch River	St. Mary's River
Catchment characteristics:			
Drainage area (km^2)	3160	1380	1608
Average altitude (m)	343	760	258
Latitude centroid (degrees)	47.7	37.0	30.6
Mean hydrometeorological variables:			
Annual total precipitation (mm)	1070	1146	975
Rainfall	693	1046	888
Snowfall	377	100	87
Annual daily temperature (°C)			
Minimum	-2.6	3.7	5.6
Maximum	8.2	17.5	16.1
Mean	2.8	11.1	10.8
Annual daily streamflow (m^{3}s^{-1})	70	19	17

For each catchment, a 25-year period (1986-2010) of data was selected to standardize the tests as detailed in the methodology. This was to allow a 1-year model warm-up period, an 8-year test period and 16 years for conducting the traditional model calibration and validation steps.

Finally, although the most intensive analysis in this paper was performed using these three catchments, the dataset was extended to eight catchments in the discussion section of this paper.

3. Methods

In this section, the hydrological models are described, as are the main methodological steps and underlying hypotheses.

3.1. Hydrological models

Two lumped hydrological models of varying complexity, namely the 6-parameter GR4J-CN model and the 21-parameter HMETS model were used in order to provide a stronger basis to generalize the results and evaluate the proposed methodology robustness. The GR4J-CN and HMETS models' structures are described in the following sections.

3.1.1. GR4J-CN

The GR4J (standing for modèle du Génie Rural à 4 paramètres Journaliers – daily rural engineering model with four parameters; Perrin et al. 2003) model is a simple parsimonious 4-parameter lumped model operating at a daily scale. This rainfall-runoff model structure is composed of a production and a routing store.

Since the simulation of the snow processes in a requirement in this work, the GR4J model has been coupled with the CemaNeige degree-day snow model (Valéry 2010). CemaNeige is a 2-parameter module that simulates the evolution of the snow cover and the different processes leading to snowmelt.

This results in a 6-parameter hydrological model referred to as GR4J-CN in this work. The required inputs for the GR4J-CN model are continuous time series of daily precipitation, mean temperature and potential evapotranspiration. Potential evapotranspiration has been estimated based on the work of Oudin et al. (2005) and the resulting formula is based on extraterrestrial radiation.

This model structure (GR4J-CN with potential evapotranspiration estimated with the Oudin formula) has been used in numerous hydrological studies and has shown strong performance in the simulation of daily streamflow over North American catchments (Poissant et al. 2017; Troin et al. 2015; Troin et al. 2018; Velázquez et al. 2015).

3.1.2. HMETS

The HMETS (Hydrological Model – École de technologie supérieure; Martel et al. 2017) is a simple and efficient model that has been developed for educational applications. This lumped-conceptual model uses two connected reservoirs to simulate the vadose and phreatic zones. Streamflow is computed as the sum of surface, delayed, hypodermic and groundwater flows.

Here the snowmelt processes are simulated by the 10-parameter degree-day model developed by Vehviläinen (1992). This model simulates the evolution of the snowpack notably through the melting and the refreezing processes. Similarly to the GR4J-CN model, the Oudin formula (Oudin et al. 2005) has been used to estimate the potential evapotranspiration required by HMETS.

HMETS has up to a total of 21 free parameters that can be optimized during the calibration. While several of these parameters can be initially fixed for a more parsimonious model, all 21 parameters were kept for the calibration in this work, to offer an opposite counterpart to the parsimonious GR4J-CN model. The needed inputs consist of daily precipitation, minimum and maximum temperature, as well as the potential evapotranspiration.

The HMETS model with the Oudin formula has also been applied over multiple hydrological studies using North American catchments and has shown to provide a good performance (Troin et al. 2015; Troin et al. 2018).

3.2. Model calibration algorithm and objective function

In this project, the Covariance-Matrix Adaptation Evolution Strategy (CMA-ES; Hansen and Ostermeier 1996, 2001) was implemented to automate the hydrological model parameter calibration step. CMA-ES was shown to be a robust algorithm that requires little to no hyperparameter tuning, making it ideal for this project (Arsenault et al. 2014). CMA-ES's strength comes from the internal learning of a second-order model that represents the objective functions' response surface. It does not require any assumptions on response surface convexity or smoothness and performs well even on ill-conditioned functions. Its robustness was therefore a strong asset for this project due to the large impact that the optimization algorithm performance could have on the end-result. A budget of 10 000 model evaluations was selected based on the work of Arsenault et al. (2014) to ensure convergence towards an optimal parameter set.

In all cases, the *NSE* metric was used to calibrate, evaluate and test the hydrological model performance. While it does have some drawbacks, such as weighting the peak flows more heavily, it is still considered as a good overall objective function in hydrological model simulation (McCuen et al. 2006). The objective function used in CMA-ES, which attempts to minimize an objective function, was unity minus the *NSE* value (1-*NSE*). In practice, the optimization algorithm attempts to attain the minimum possible value of zero, returning the perfect *NSE* value of 1. The *NSE* itself is calculated as follows:

$$NSE = 1 - \frac{\sum_{i-1}^{n}\left(Q_{o,i} - Q_{s,i}\right)^2}{\sum_{i=1}^{n}\left(Q_{o,i} - \overline{Q_{o,\imath}}\right)^2} \qquad (1)$$

where *NSE* is the Nash-Sutcliffe Efficiency metric, Q_o is the observed streamflow, Q_s is the simulated streamflow and the *i* index represents the simulation day. The optimization algorithm then used the [1-*NSE*] value to perform the optimization itself. The *NSE* was preferred over other metrics such as Kling-Gupta Efficiency (*KGE*; Gupta et al. 2009) simply due to the prevalence of *NSE* in the literature (Jain and Sudheer 2008), which allows for a better visualization of the scores from the hydrological modeling community. The results are not expected to be influenced by this choice.

3.3. Strategy to evaluate the impact of calibration and validation

Typically, hydrologists will divide their observation database in two parts: a calibration period for parameter tuning, and a second period for the hydrological model validation. This separation follows many forms, including the split-sample or half-half method (KlemeŠ 1986) and the odd/even year method (Arsenault et al. 2017; Gowda et al. 2012). Usually, the parameter set that is found during calibration is first evaluated on the calibration period, and on the validation period in a second step. The objective functions between both periods are then compared. An important drop in performance between the calibration and validation periods will raise flags and over-parameterization or data quality problems might be suspected. In any case, another parameter set might be evaluated using another sample until a robust parameter set is found for both the calibration and validation periods. It is also customary to switch the calibration and validation periods to verify if the underlying observed data are at cause. Once the hydrologist is satisfied with the calibrated parameter set, the model can then be used for simulation and forecasting applications.

The main drawback of using such sample-based calibration and validation methods is that the parameter set is conditioned on a subset of the available data, possibly depriving itself from important information that is lost in the validation phase. This is the research question this work aims to answer.

To see if that is truly the case, a strategy commonly used in neural-network calibration and testing was implemented for hydrological modeling (Guo et al. 2017), in which the data are separated into three parts rather than two. The third part is reserved for testing the data on a period independent from the calibration and validation periods. In the case of this study, the 25 years of data were divided according to the method detailed in the pseudocode of Figure 2.

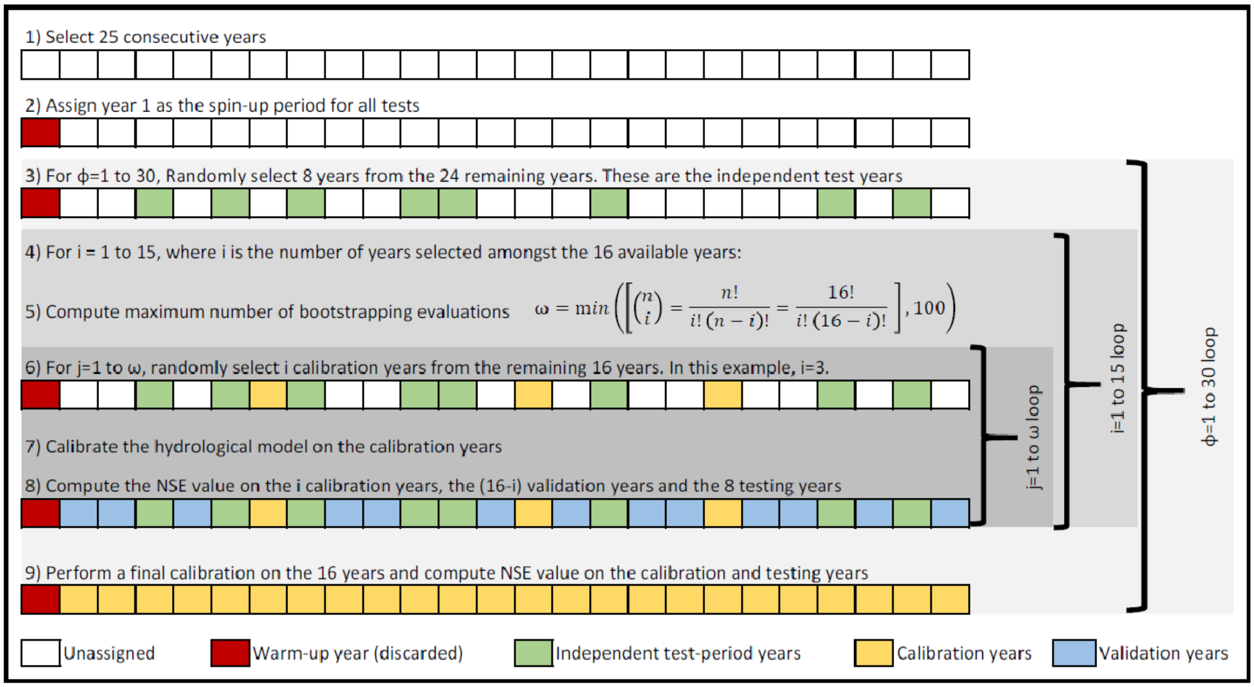
1) Select 25 consecutive years
2) Assign year 1 as the spin-up period for all tests
3) For ϕ=1 to 30, Randomly select 8 years from the 24 remaining years. These are the independent test years
4) For i = 1 to 15, where i is the number of years selected amongst the 16 available years:
5) Compute maximum number of bootstrapping evaluations
$\omega = \min\left(\left[\binom{n}{i} = \frac{n!}{i!\,(n-i)!} = \frac{16!}{i!\,(16-i)!}\right], 100\right)$
6) For j=1 to ω, randomly select i calibration years from the remaining 16 years. In this example, i=3.
7) Calibrate the hydrological model on the calibration years
8) Compute the NSE value on the i calibration years, the (16-i) validation years and the 8 testing years
9) Perform a final calibration on the 16 years and compute NSE value on the calibration and testing years
j=1 to ω loop
i=1 to 15 loop
ϕ=1 to 30 loop
Unassigned
Warm-up year (discarded)
Independent test-period years
Calibration years
Validation years

Figure 2. Pseudocode of the methodology implemented in this study.

The nine steps are carried out as follows:

STEP 1: For each catchment and hydrological model used in this study, 25 years of continuous streamflow and their corresponding input data are prepared.

STEP 2: The hydrological models are always run starting at the beginning of the first year and the *NSE* values never include the first year as it is considered a warm-up period to bring the model states to realistic internal variable states. Therefore, this year is always simulated by the hydrological model, but is never computed in the *NSE* scores. The warm up period of one year was considered sufficient so that the uncertainty of initial conditions becomes negligible compared to the uncertainty of parameters (Huard and Mailhot 2008).

STEP 3: To ensure robustness of the results, the entire process (Steps 3-9) is repeated 30 times. For each φ ϵ [1,30], a new random set of 8 independent test years is selected.

STEP 4: Given that there are 16 remaining years to perform calibration and validation, $i \in [1,15]$ years are selected from the 16 years to evaluate the impact of the number of calibration years on the performance over the independent test period. Here *i* is limited to 15 as the calibration on the entire 16-year set is performed later, at step 9.

STEP 5: To make sure that the process is not affected by random “luck of the draw”, multiple trials are performed for each *i*. When $i = 1$, there are a total of 16 possible selections of 1 year from a set of 16, but when $i = 8$ years from 16 must be selected, then there are 12870 possible combinations. It would be extremely computationally intensive to perform the entire set of combinations in calibration, keeping in mind that the process must be repeated for each *i* and each φ for three catchments and two hydrological models. Therefore, to keep computing time reasonable and still explore the parameter space, the maximum number of bootstrap evaluations ω is fixed at 100.

STEP 6: At this step, *i* years are selected at random in the available 16 years. These will be used for calibrating the hydrological model. This step, along with steps 7-8, are repeated ω times.

STEP 7: With the *i* years selected, the model calibration is performed following the method described in Section 3.2.

STEP 8: With the calibrated model parameters, compute the *NSE* metric values on the *i* calibration years, the 8 testing years and the (16-*i*) validation years.

STEP 9: Once all *i* and *j* are exhausted, the hydrological model is calibrated on the 16 years and compute the *NSE* scores on the calibration and testing periods. Obviously since all years are used in calibration, there are no remaining years to perform the validation step.

The results were then analyzed by comparing the *NSE* values on the independent testing periods as a function of the number of calibration years. In total, 239940 calibration experiments were performed in this work following the strategy described above.

4. Results

The first results aim to display the evolution of the *NSE* over the independent testing period as a function of the number of calibration years. Figure 3 shows the distributions of *NSE* values obtained on the test period for one of the 30 independent testing periods φ. Each of the boxplots represents all of the *NSE* values ω for each combination of years *i* used during calibration. For example, the first boxplots hold 16 points (16 combinations of 1 year from a set of 16), the last boxplot also holds 16 points (16 combinations of 15 years from a set of 16) and all boxplots in between are limited to 100 combinations. The horizontal line that cuts through the boxplots presents the *NSE* value on the independent test period when the 16 years are used for calibration.

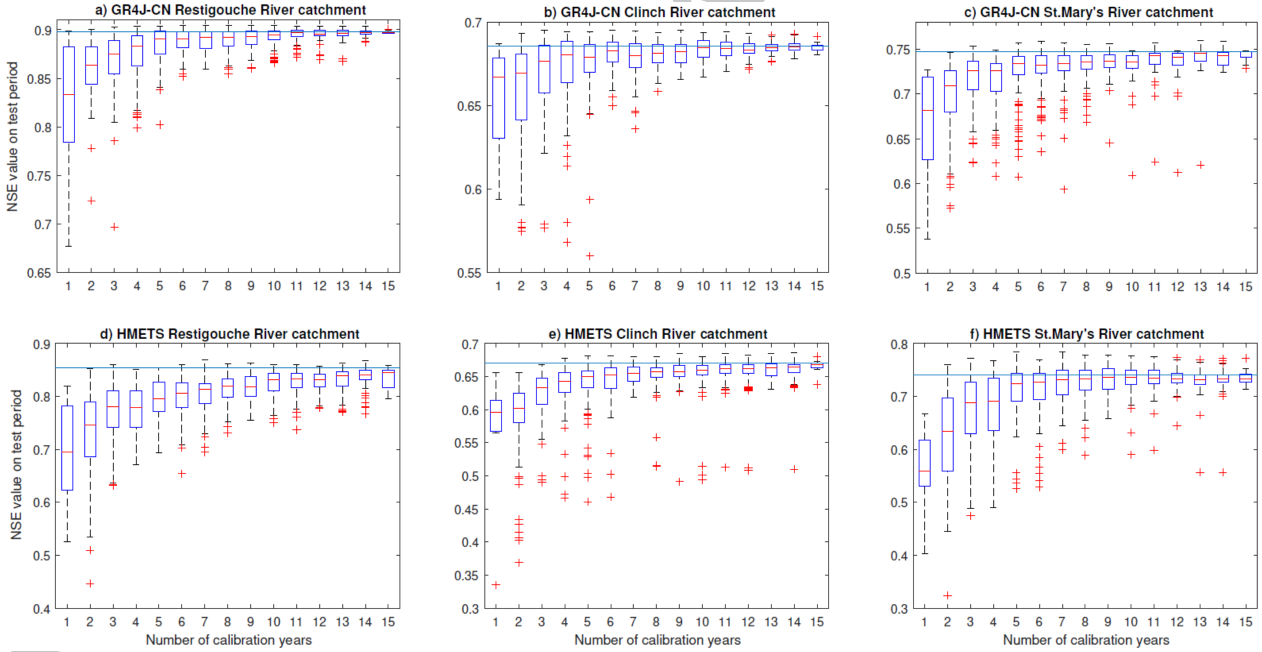

Figure 3. Test-period *NSE* values with increasing number of contributing calibration years for the GR4J-CN model (top row) and HMETS model (bottom row) for the three catchments. The horizontal line represents the test-period *NSE* when all 16 remaining years are used during

calibration. Boxplot box edges represent the 25th and 75th quantiles, and the whiskers represent +/-2.7 standard deviations from the mean.

There is a clear trend which shows that using more years in calibration improves the overall skill on the independent testing period, which should be intuitive. Note that the results for the other 29 testing periods φ are similar (not shown) in that they all converge towards higher values as more years are selected.

It can also be seen here that a small number of calibrations using fewer than the 16 years of available data (located in the upper whisker of most box plots) perform better on the test-period compared the calibration using all 16 years (the continuous line in Figure 3). This is to be expected, as some of these calibrations end up using all the years that are more representative of the test-period, and leaving out the information that is not improving the simulation over this same test-period. Even though this increase in the *NSE* value seems like the optimal solution, it is impossible to know *a priori* which years will be more representative of the future conditions, therefore the optimal solution is the one that has the highest median value and the smallest variance.

Figure 4 presents the calibration and validation *NSE* values for an increasing number of contributing calibration years. Once again, results for only one of the 30 testing periods φ are shown here but the same results are systematically found for all models and catchments.

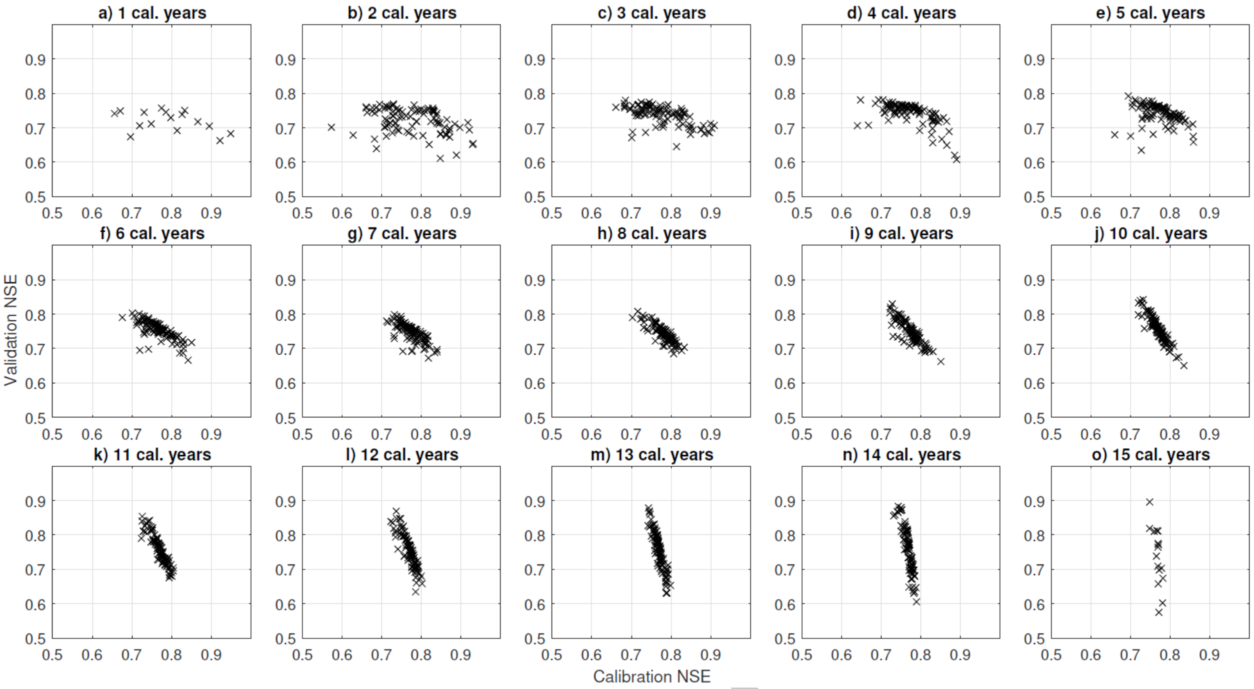

Figure 4. Calibration (x-axis) and validation (y-axis) *NSE* values for independent testing period 1, on the Clinch River catchment and for the HMETS model. Each panel presents the results for a specific number of calibration years, increasing monotonically from 1 to 15. The case of 16 calibration years is not shown because there are no validation years left when calibrating on all 16 available years.

It can be seen that there is a systematic evolution of the calibration and validation relationship, where at first the calibration skill is defined by the quality of the randomly selected year and the validation is relatively constant due to the much larger size, which gives it inertia and resistance to large changes in performance. As more years are used during the calibration, the trend reverses progressively to the point where all but one year is used in calibration and the validation skill becomes more volatile depending on that particular validation year's characteristics. This shows that selecting a good parameter set in calibration and discarding it if it does not perform as well in validation is a risky proposition that supposes that all years are hydrologically similar. Depending on the random draw of the calibration and validation periods, this could lead to false methodological hypotheses. For example, it could be decided to discard a parameter set that would have been "good" if the problematic year had been drawn in the calibration period rather than in the validation period.

The results were then compared on a more macro scale, looking at the 30 testing periods φ at the same time. Figure 5 shows the results of the test-period *NSE* value when calibrating the model on 15 of the 16 available calibration years on the Clinch River catchment. Results are similar on the other two catchments and are not shown. Therefore, each boxplot has 16 points. The results can then be compared to the test-period *NSE* value obtained when all 16 years are used in calibration, identified by a cross. It can be seen that in most cases, calibrating on the full set of data is more robust than calibrating on 15 of the 16 years. Furthermore, there is no case in which calibrating on 16 years performs worse than the worst case amongst the 16 points in any boxplot.

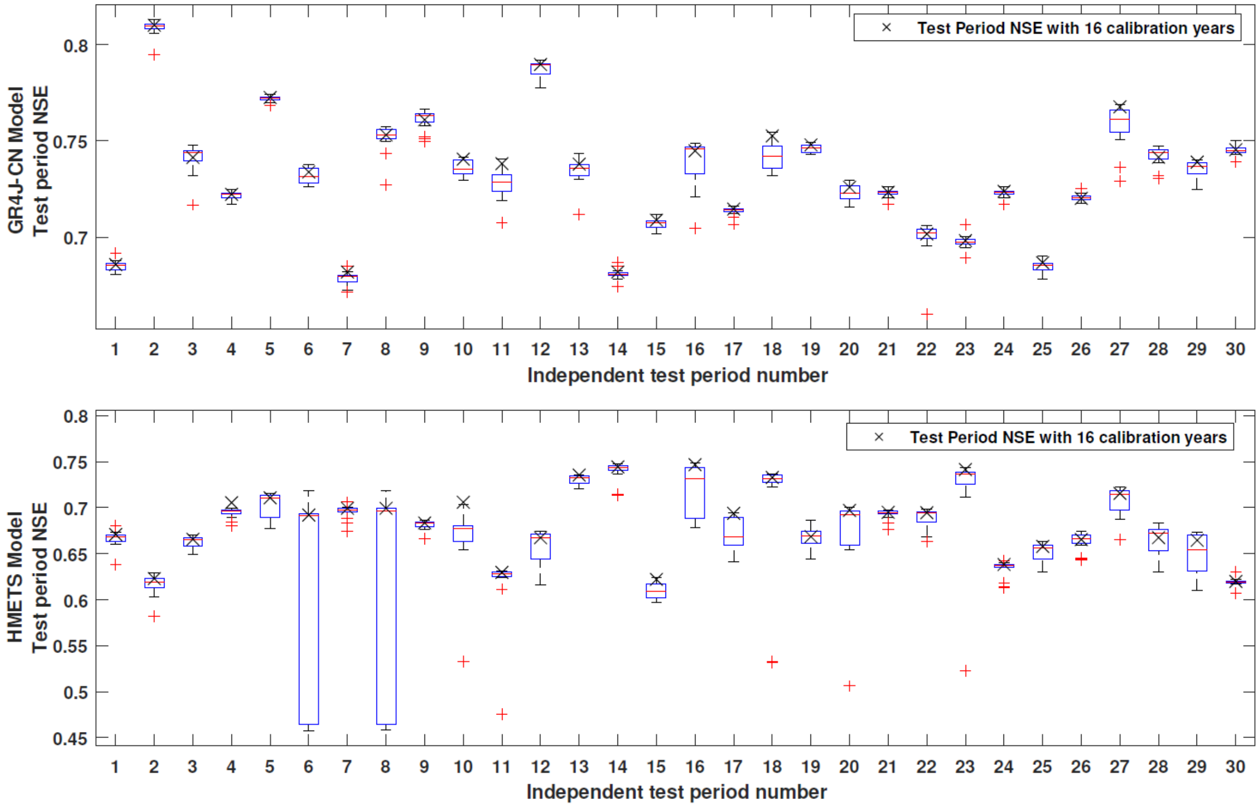

Figure 5. Test-period *NSE* values for the 30 independent test runs on the Clinch River catchment. The boxplots show the test-period *NSE* when calibrating on 15 of the 16 available years (16 possible combinations in total, composing each box plot). The red line denotes the median value of the dataset and the cross represents the test-period *NSE* obtained when calibrating on all 16 available years. Boxplot box edges represent the 25th and 75th quantiles, and the whiskers represent +/-2.7 standard deviations from the mean.

It is important to note that the boxplots show large variability between each other. This is caused by the fact that the independent test-periods are composed of different years from the dataset and are thus considered as independent cases. Some of the selected years end up being more difficult to simulate, resulting more variability within the *NSE* values.

To further explore the impact of the number of calibration years on the test-period results, the difference in the *NSE* values between a calibration using all 16 available years and the median of all calibrations using a subsample of *n* years is investigated and shown in Figure 6. In all cases, the *NSE* scores are the ones evaluated on the 30 independent testing periods.

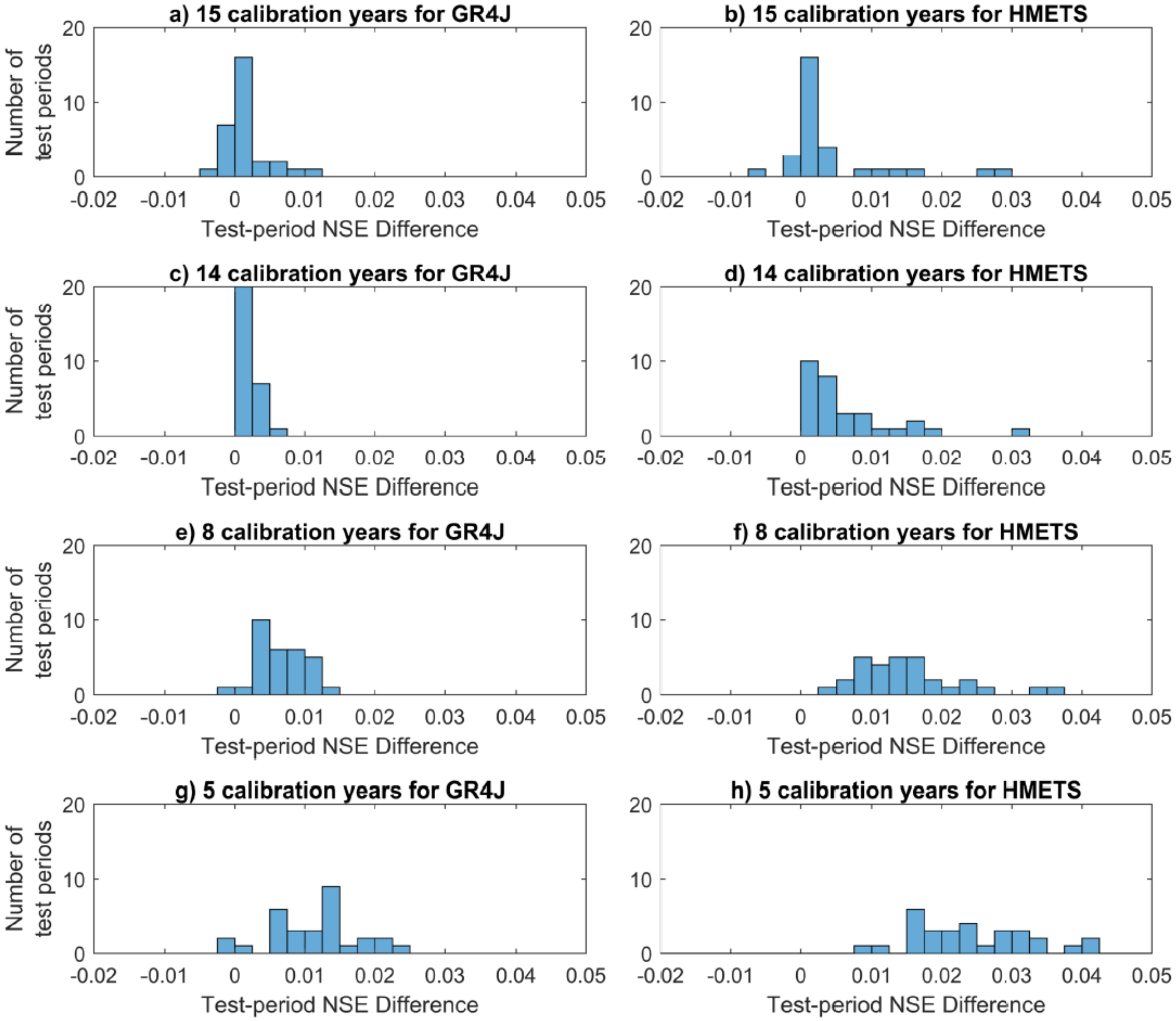

Figure 6. Histogram of the Clinch River catchment difference in the *NSE* values between the 30 independent test-periods when using 1) all 16 years for the calibration and 2) the median of calibrations using 15, 14, 8 and 5 years (top to bottom). A positive *NSE* difference indicates that calibrating on 16 years is better than on a smaller number of years. The histograms display the number of test-cases that fall in each *NSE* difference bin.

A clear trend can be seen in Figure 6, in which the shorter the calibration period, the lower the performance on the testing period as compared to when the 16 years are used for calibration. This is true for both models and all catchments, although only the Clinch River catchment test-case is shown in Figure 6. It is clear that the more years are used in calibration, the smaller the *NSE* differences are. Taken with the results from Figure 3, it is clear that adding more calibration years allows the model to perform better on the independent testing period.

The results presented in Figures 3-6 are not filtered based on performance, therefore one could argue that some of these results would not hold in an operational setting because if a calibration *NSE* is much higher than the validation period *NSE*, the parameter set could be rejected on the basis that the model was experiencing “overfitting” or that the parameter set was not robust. To put this hypothesis to the test, the same experiment as in Figure 6 was performed, but this time only cases where the validation *NSE* is at least equal to the calibration *NSE*. Therefore, the validation *NSE* would theoretically be accepted as the skill is at least as good as on the calibration period. Figure 7 presents the results.

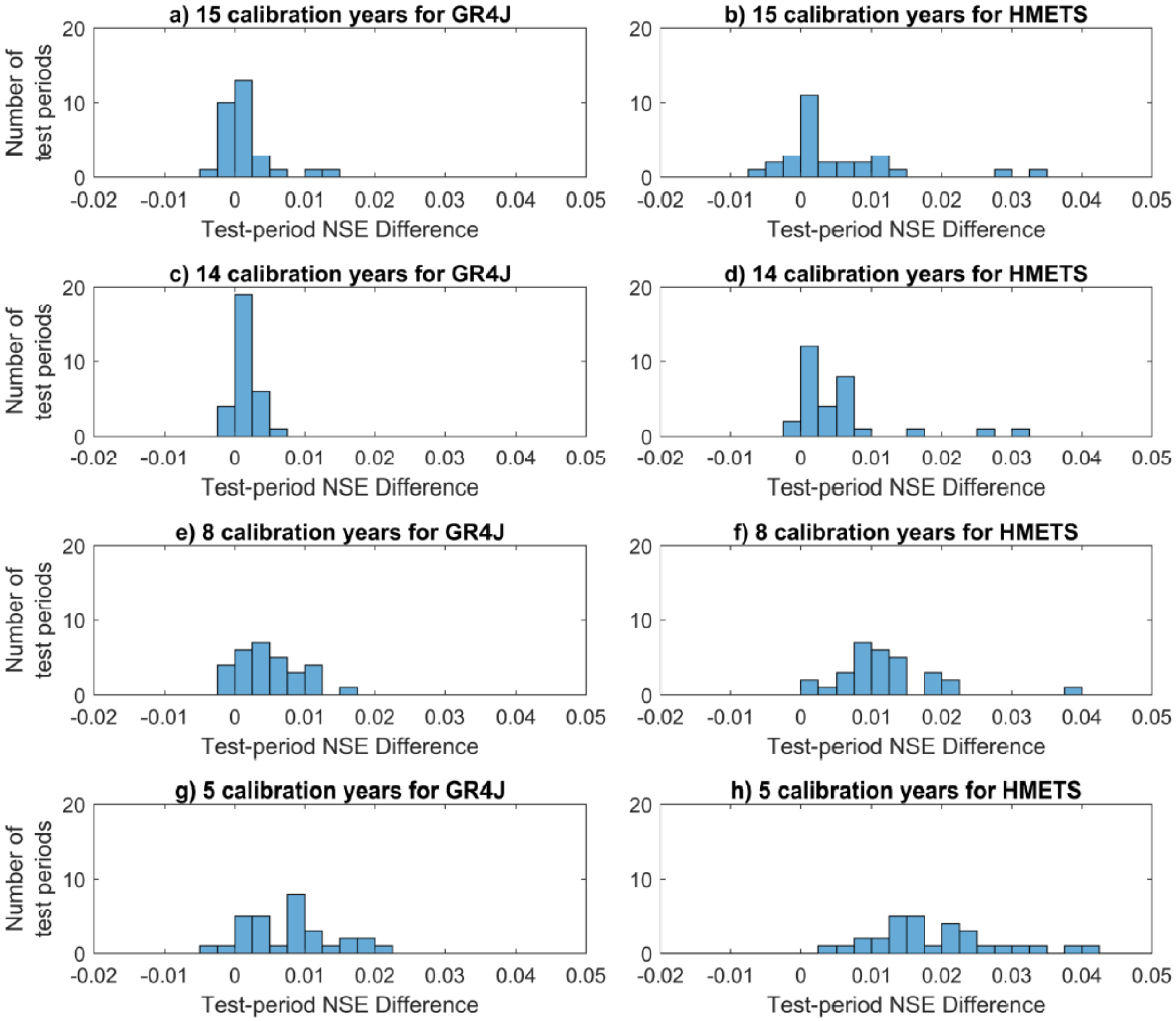

Figure 7. Same as Figure 6, but using only cases where the cross-validation *NSE* is at least equal to the calibration *NSE* are used to build the boxplots here.

Figure 7 clearly shows the same behavior as in Figure 6, indicating that the skill in calibration and validation is not necessarily correlated, but can be linked to the characteristics of the calibration and validation period time series. A statistical test was performed to verify the significance of these results. The Wilcoxon test on ranks was used as a non-parametric replacement test for the 1-sided Student *t*-test. This was self-imposed because the samples do not come from the same distribution given that each test has different test-period years. In all cases (filtered and unfiltered results from Figure 6 and Figure 7) the tests showed significant results. P-values ranged from 3x10^{-6} to 1x10^{-5} for the 15-out-of-16 years case, and all other results (14…1 out of 16 years) attained the lowest possible p-value limit for the Wilcoxon test with a sample size of 30, which is 1.73x10^{-6}. It is therefore clear that the differences, although small, are significant in this context, although it is obvious that an *NSE* difference of 0.001 might not be particularly useful in an operational context. Nonetheless, the fact that some combinations of 15-in-16 years are better than the 16-in-16 year case is still theoretical, because in practice there is no way of knowing which years are to be used and which ones are to be discarded in the calibration process.

Calibration and validation were analyzed from another angle. This time, the calibrated parameter sets issued from the 16-year calibration for each of the 30 independent test-cases are used to simulate streamflows on all of the combinations of calibration and validation periods, as described in Figure 2. In this graph, no use is made of the 8 independent years associated with each of the 30 independent test-case. This allowed investigating the *NSE* distributions on different periods and calibration time series lengths knowing that the parameter set is robust and performs well. Results are presented in Figure 8.

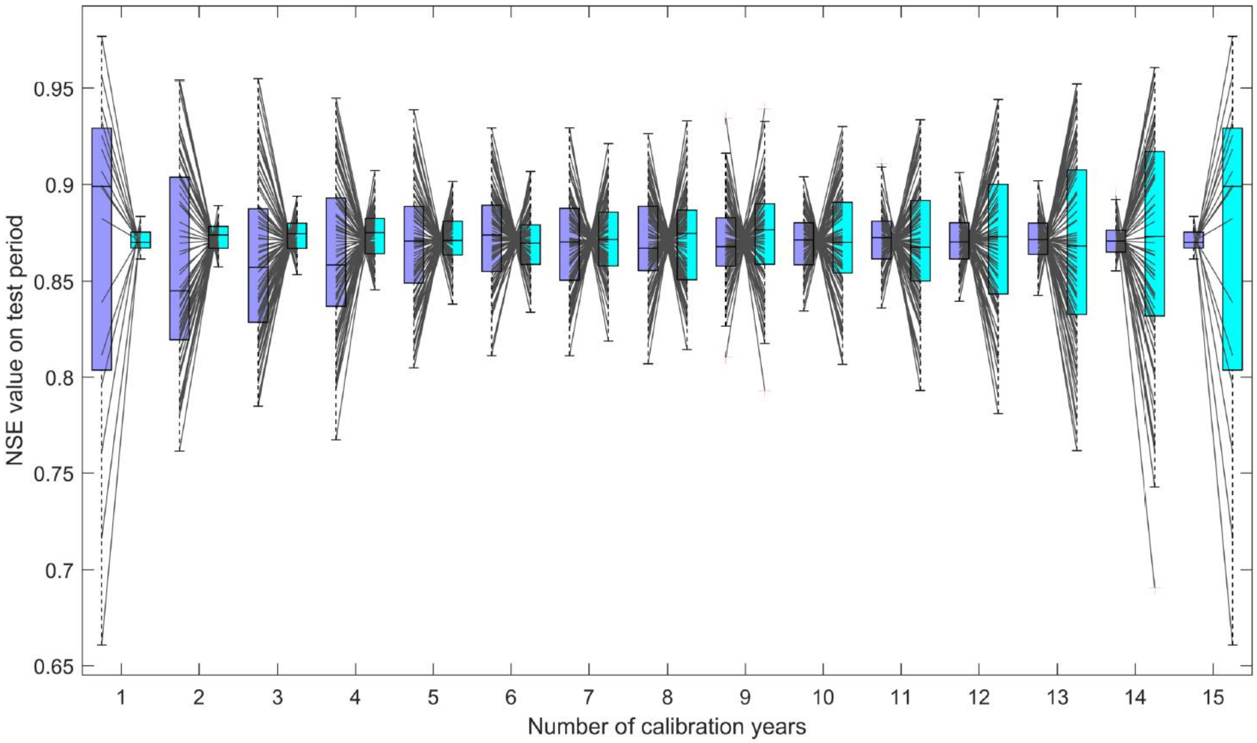

Figure 8. Calibration and validation *NSE* values for independent testing period 1, using *n* = 1 to 15 calibration years. Dark and pale blue box-and-whisker plots represent the bootstrapped calibration and validation *NSE* values respectively. The x-axis values refer to the *n* calibration years used for the bootstrap. Grey lines link the calibration-validation pairs for the *n* calibration years. Boxplot box edges represent the 25th and 75th quantiles, and the whiskers represent +/-2.7 standard deviations from the mean.

In Figure 8, it is possible to see the spread of *NSE* values that the hydrological model can return given randomly selected calibration years. The calibration-validation pairs, linked by dark lines in Figure 8, show that high-performing *NSE* calibration is linked to low-performing *NSE* in validation, and vice-versa. This is an artifact of conditioning the results to a unique and robust parameter set. If a hydrologist were to calibrate on 15 out of 16 years and keep a single validation year, the odds are approximately 50% that the *NSE* value obtained on the validation period would be below the calibration skill. In some instances, such as for the 15-year calibration set, there exist a few years that will dramatically underperform in validation, even though the parameter set is perfectly acceptable and robust.

A final test was conducted to evaluate the robustness of calibrating over the 16 years. In this test, the parameter set derived from the calibration on the full 16-year time-series was used to run the hydrological models on the same calibration years of evaluation *j* using *i* calibration years from methodological steps 4-7 (see Figure 2). The difference between the calibration and validation *NSE* values were computed for each of these cases. Furthermore, the same exercise was performed using the original calibration sets for each *i* and *j*. The idea was to measure the difference in calibration and validation *NSE* values using the original calibration sets and comparing those results to those obtained using a unique parameter set calibrated on the 16 years. Results are shown in Figure 9.

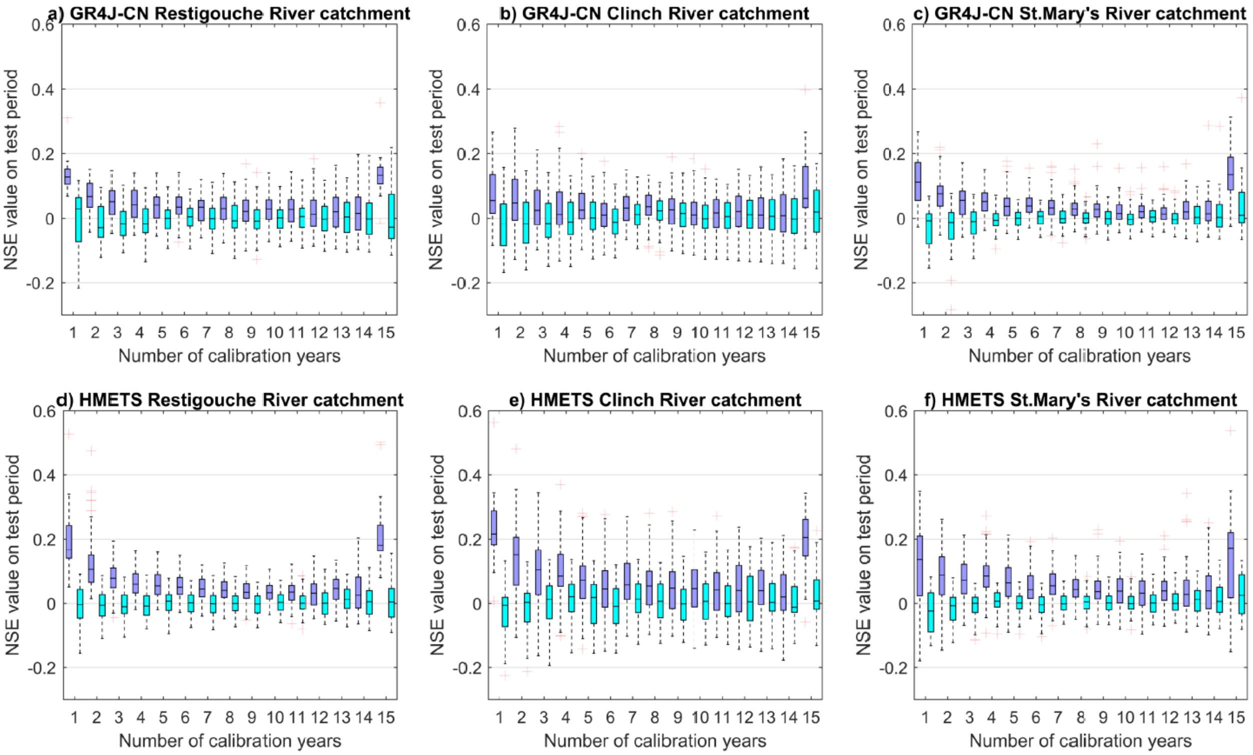

Figure 9. Difference between calibration and validation *NSE* values for one independent test period using the original parameter sets (dark blue) and the full time-series calibrated parameter set (light blue) for the 15 increments of numbers of calibration years. Boxplot box edges represent the 25th and 75th quantiles, and the whiskers represent +/-2.7 standard deviations from the mean.

It can be seen in Figure 9 that there is a bias when using the original parameters calibrated on *i* years, and that this bias shrinks significantly (and oscillates around 0) for most cases with the parameter set calibrated on the entire period. This shows that the full calibration provides more robust results than calibrating on split-samples and that the validation *NSE* score is not a reliable predictor of calibration parameter performance.

5. Discussion

5.1. The hydrological model validation sacrifice

Ever since hydrological models have existed, parameter calibration and validation have been intertwined and considered an inseparable pair. Understandably, a hydrologist will want the hydrological model to be robust enough to warrant its use in forecasting or other applications. A good validation skill comforts the user in the parameter set robustness for use in an operational setting. However, this comes at a cost on two levels.

First, the cross-validation is used purely as a parameter verification tool, and all the information contained within is withheld from the parameter set. Therefore, there might be some years in the cross-validation set that would be useful for future periods, but the parameter set would not be trained on those data and could thus not adequately react to the new inputs. This is a direct sacrifice of model performance for confidence in the model.

Second, it forces the hydrologist to use parameter sets that are perhaps not the best overall by sacrificing some skill in calibration to ensure calibration and validation are similar in terms of performance. In this case, a good parameter set could be rejected because of a larger-than-anticipated spread between the calibration and cross-validation *NSE*. As was shown in Figure 8, this can happen quite easily and be costly from an operational point of view. Figures 6 and 7 also show the same reasoning.

This study shows that the cost can be high in terms of overall performance on an independent test period. The resampling bootstrapping method implemented in this work shows that even on randomly selected years, it is almost always a good idea to use all available information when calibrating the hydrological model. Of course, it is possible to tailor a set of calibration years that will outperform the entire series (for example, by removing data years that are too dissimilar to the testing period) but this cannot be performed *a priori*. Therefore, in absence of information on the future years, the most conservative approach is to calibrate on the entire available time-series in order to ensure that the parameter sets contain as much information as possible. Figures 3, 4 and 5 demonstrate this point quite clearly.

5.2. The impact of the calibration objective function

In this study, only the *NSE* was used as the objective function during calibration. While it does have some drawbacks, as explained in section 3.2, it is generally recognized as providing satisfactory results. Ideally, this study would have included more objective functions in order to validate the conclusions on other metrics, but the complexity of performing the sheer number of calibrations made it prohibitive. Instead, the parameter sets obtained through calibration on the *NSE* metric were evaluated using two other metrics, namely the relative bias (%) and the Pearson correlation coefficient. Figures 10 and 11 show the results respectively for the relative bias and correlation for the three main catchments for a single test-case, although results are similar for all model-catchment pairs.

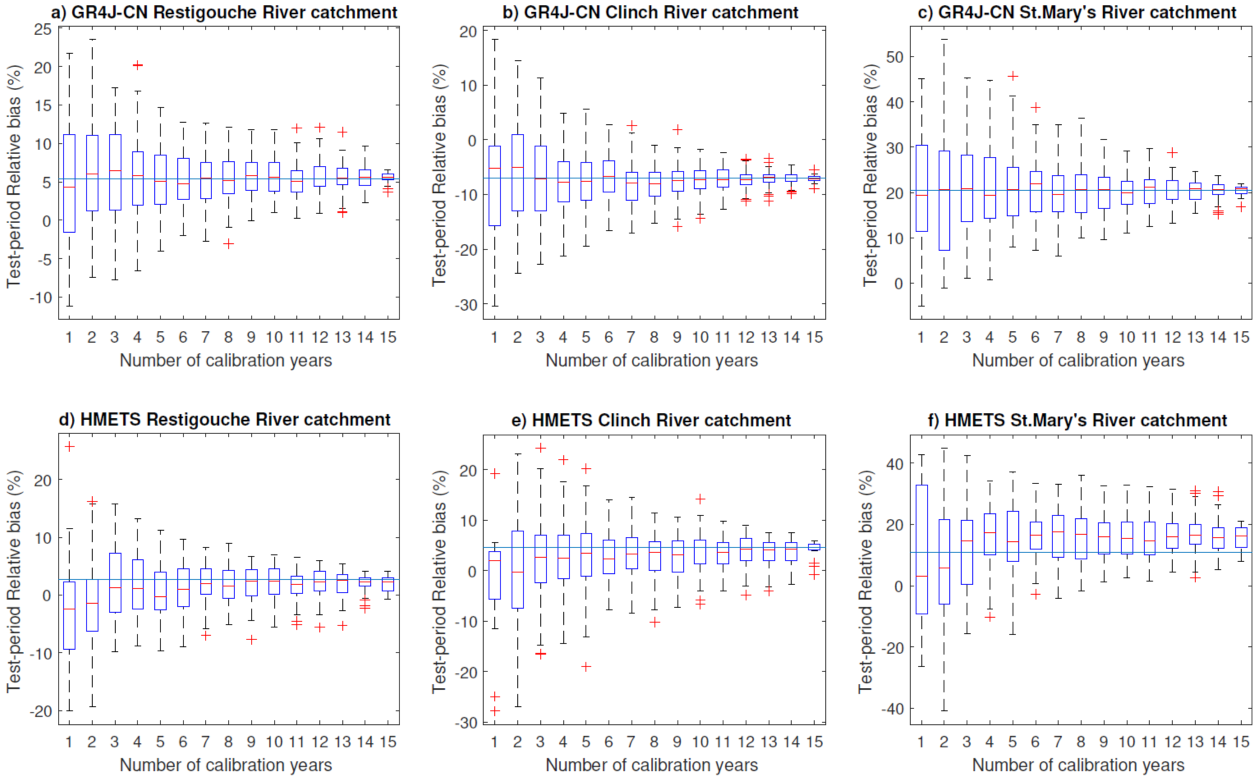

Figure 10. Test-period *Relative bias* values with increasing number of contributing calibration years on the *NSE* objective function for the GR4J-CN model (top row) and HMETS model (bottom row) for the three catchments. The horizontal line represents the test-period *Relative bias* when all 16 remaining years are used during calibration. Boxplot box edges represent the 25th and 75th quantiles, and the whiskers represent +/-2.7 standard deviations from the mean.

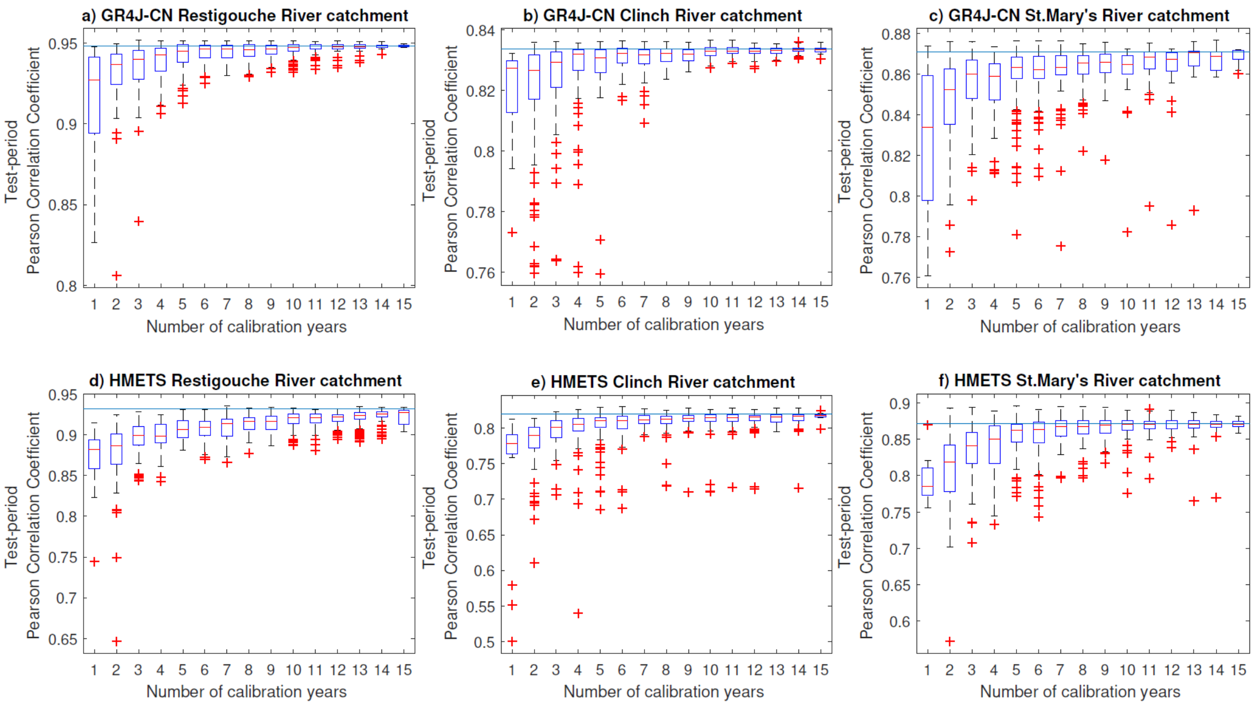

Figure 11. Test-period *Pearson Correlation Coefficient* values with increasing number of contributing calibration years on the NSE objective function for the GR4J-CN model (top row) and HMETS model (bottom row) for the three catchments. The horizontal line represents the test-period *Pearson Correlation Coefficient* when all 16 remaining years are used during calibration. Boxplot box edges represent the 25th and 75th quantiles, and the whiskers represent +/-2.7 standard deviations from the mean.

As can be seen in these figures, the bias and correlation values converge to a point as more and more calibration years are used. In Figure 10, the bias levels converged when more calibration years were added, but they did not necessarily converge towards zero. This seems to be caused by the *NSE* preferring correlation to bias during calibration. Furthermore, it intuitively seems more likely that there could be a larger difference in bias when selecting years randomly between the calibration and validation period. However, the results seem to indicate that if the calibration were to be performed on the bias instead of the *NSE* as the objective function, the same converging results would be found, but would converge closer to zero.

The correlations (Figure 11) follow the same patterns as *NSE* (Figure 5), again probably due to the fact that *NSE* gives more importance to the correlation than the bias in this study, as explained in Gupta et al. (2009). A comparative study using the Kling-Gupta Efficiency (*KGE*) metric (Gupta et al. 2009), which allows distinguishing more easily between the correlation, bias and variance, could be performed to assess these points further.

5.3. The sense of security provided by parameter validation: False advertising?

The results in Figures 3-5 and Figure 8 indicate that the validation of a calibrated hydrological model parameters is very sensitive to the characteristics of the underlying data in both the calibration and validation periods. Both hydrological models used in this study displayed the same behavior, which is an increase in independent testing skill as more calibration years are included. However, validation is often used to ensure that the model performs adequately on an independent period and that the model is not overfitting the data instead of representing the hydrological processes. The results in Figures 5 and 8 show that the range of *NSE* values calculated on the independent testing period can vary wildly depending on the properties of the selected years that are selected for calibration and testing. In many studies and operational settings, a much lower validation *NSE* would be addressed by other means such as changing the calibration/validation periods to ensure similar performances, and until a so-called robust parameter set is obtained. It can therefore be difficult to distinguish between overparameterization, model inadequacy and simply the quality of a certain time period that renders its modeling difficult.

In this context, and in light of the results obtained in this study, it seems clear that the optimal solution is to include all available years in the dataset for the calibration process. Consequently, the responsibility of the hydrologist is to ensure that there is not a structural problem with the meteorological or hydrometric data that could lead to the model attempting to fit the model on unrealistic data. By selecting all years for calibration, the parameter sets will contain the maximum amount of information for the study site, therefore maximizing its chances of performing adequately in future simulations and prediction. Of course, if calibrating the model on all years returns a low/poor *NSE* value, then there would be reason to suspect that the model is inadequate for the simulation conditions and would need to be investigated further, perhaps by changing the model altogether or inspecting the datasets for errors. The results in Figures 3 and 8 provide evidence that overparameterization cannot be distinguished from the seemingly poor results obtained from a randomly selected set of validation years. Ultimately, the *NSE* skill of a model calibrated on the entire available dataset should be the only criteria to determine if the model is adequate for simulating on the given catchment.

5.4. Is this concept generalizable?

This study used a contrasting set of three catchments with different hydrometeorological inputs and two hydrological models of different complexity. An analysis of calibrated parameter sets (results not shown) indicated that the GR4J-CN model displayed no equifinality when more than 12 years of data were included in the calibration set. Unlike GR4J-CN, HMETS' parameter sets spanned large sections of the parameter space (even when calibrating on 15 years), thus indicating presence of equifinality. In both cases, the optimal solution was to calibrate the hydrological model on all years, which resulted in superior performance on the independent testing periods. This evidence supports the idea that the parameter tuning, even when in presence of equifinality, integrates information from the learning set that can be applied afterwards. The fact that these conclusions hold for the six test-cases lends credibility to the idea that these findings should be applicable to a wide range of models and catchments. To investigate the issue further, the experiments were performed on five independent catchments using a less restrictive methodology. Instead of setting a maximum of 100 combinations for each number of calibration years, only 64 were used in order to maximize available computing resources. Furthermore, only 10 independent test-cases were performed instead of 30 as used in the full-scale experiment. Nonetheless, the results of this test show that the results also hold for these five catchments which gives more weight to the idea that the concept is generalizable. Table 2 shows the characteristics of the five supplementary catchments, and Figure 12 shows the same analysis as Figure 5 applied to these verification catchments.

Table 2. The five supplementary catchments' characteristics and mean annual hydrometeorological variables.

	Indian Creek, CA	Yadkin River, NC	Trinity River, CA	Sinnemahoning Creek, PA	Neches River, TX
Catchment characteristics:					
Drainage area (km^2)	1913	5903	7386	1773	2964
Centroid Latitude (degrees)	40.1°N	35.9°N	41.1°N	41.3°N	31.9°N
Centroid Longitude (degrees)	120.9°W	80.4°W	123.7°W	78.1°W	95.4°W
Mean hydrometeorological variables:					
Annual total precipitation (mm)	806	1188	1355	999	1008
Rainfall	599	1135	1231	810	990
Snowfall	207	53	124	189	18
Annual daily temperature (°C)					
Minimum	0.3	7.1	3.7	1.8	12.2
Maximum	17.0	20.4	19.3	14.5	25.0
Mean	8.7	13.7	11.5	8.1	18.6
Annual daily streamflow (m^{3}s^{-1})	17	82	162	31	17

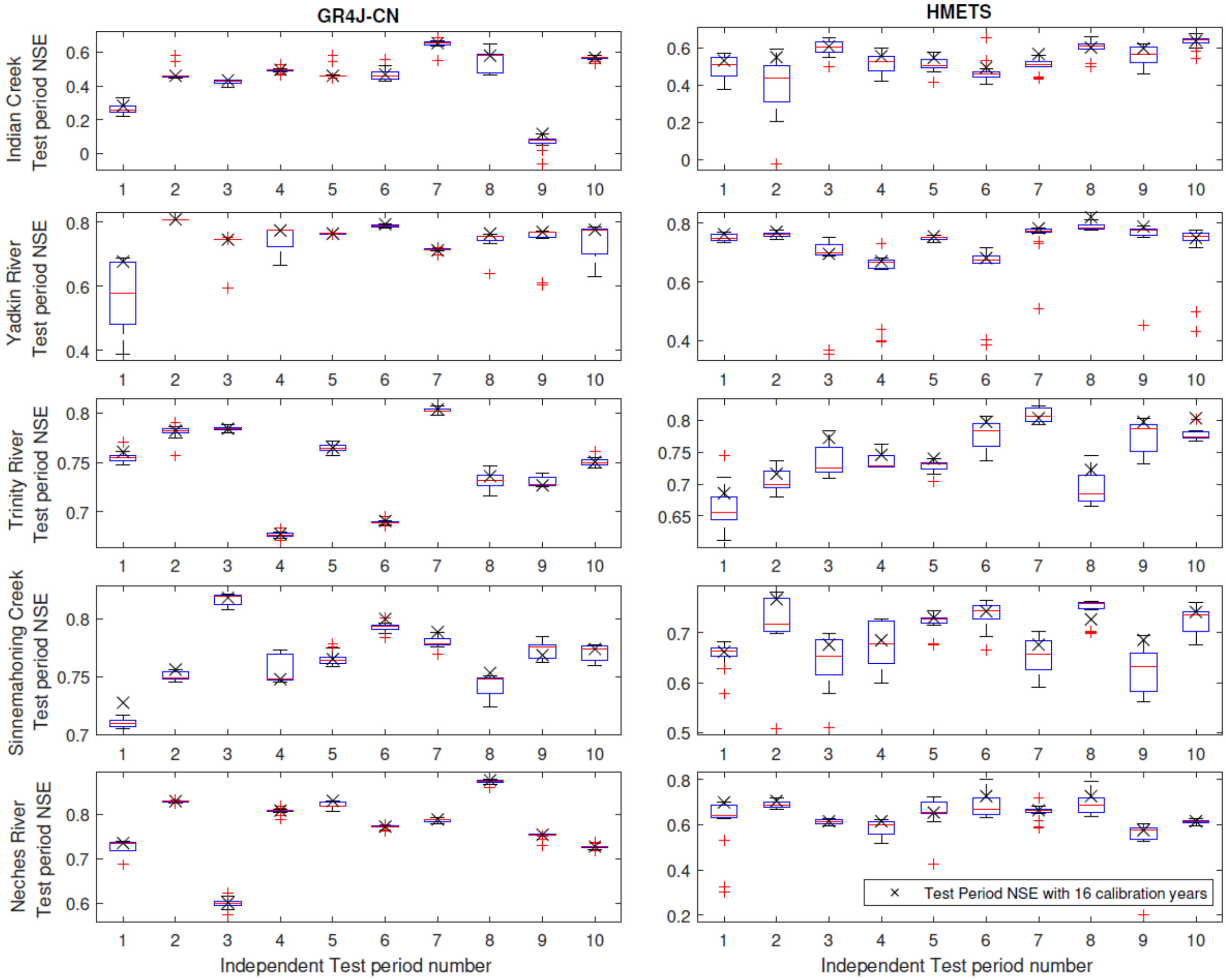

Figure 12. Test-period *NSE* values for the 10 independent test runs on the five verification catchments. The boxplots show the test-period *NSE* when calibrating on 15 of the 16 available years (16 possible combinations in total, composing each box plot). The red line denotes the median value of the dataset and the cross represents the test-period *NSE* obtained when calibrating on all 16 available years. Boxplot box edges represent the 25th and 75th quantiles, and the whiskers represent +/-2.7 standard deviations from the mean.

The Wilcoxon statistical test was once again applied on the new model-catchment pair results on the independent test period, and results are shown in Table 3. In all but two cases, the differences between using 15 and 16 calibration years was significant. The difference was significant for all model-catchment pairs when using 14 calibration years. It is important to note that for these cases, only 10 samples (10 random independent test periods) were used; therefore, the power of the statistical test is reduced and the lowest possible p-value is 0.002.

Table 3. P-values of *NSE* scores on the independent test period for the 5 verification catchments, for 14- and 15-in-16 calibration years as compared to the 16-in-16 year calibration case.

Catchment	p-values			
	GR4J-CN		HMETS	
	15 years (of 16)	14 years (of 16)	15 years (of 16)	14 years (of 16)
(1) Indian Creek, CA	0.002	0.002	0.020	0.002
(2) Yadkin River, NC	0.232	0.002	0.027	0.002
(3) Trinity River, CA	0.027	0.002	0.004	0.002
(4) Sinnemahoning Creek, PA	0.242	0.002	0.105	0.002
(5) Neches River, TX	0.004	0.014	0.020	0.002

These results, along with the results in Figures 10 and 11, seem to agree with the idea that the concept is generalizable at least to a certain extent. However, it is possible that the full-series calibration is not optimal for some other model-catchment setups, such as for catchments showing signs of non-stationarity as will be discussed in Section 5.7. While we are confident that in most cases the optimal methodology is to calibrate on all available years, more work needs to be done to prove its generalizability on a wider scale.

5.5. The compromise solution

It is understandable that modifying such an ingrained habit as using a split-sample calibration and validation is a tough sell, and that some, if not most, researchers and practitioners will prefer to keep the validation step as a robustness evaluation tool. However, it is also clear from the results analyzed in this work, that there is a cost to this solution. In an attempt to satisfy both sides of the equation, it is proposed to preserve the validation step (if so wanted by the operator), but that if the skill in validation is acceptable, then the calibration should be performed once again over the entire time series. In this way, the operator will know that the model is able to perform well on both independent periods, indicating that the processes are well simulated by the model. Then, by recalibrating over the entire dataset, the model will have been trained on more data, leading it to contain more information and making it more robust for future use.

This hybrid method can be seen as a compromise solution which would lead to the same conclusion as calibrating on all years and then judging the overall *NSE* score. In both cases, the entire time-series is used in calibration but the way that the model is deemed to be acceptable or not is performed in a separate step. The only drawback of the proposed solution is the required extra time to perform a split-sample calibration before the full calibration.

5.6. The issue of model complexity

There is a trend in hydrological science toward increasingly more complex process-based and/or distributed hydrological models with some models now implementing land surface schemes with complex formulations such as the Richards nonlinear differential equations of water movement in the non-saturated portion of the soil column (Paniconi and Putti 2015). In such models, calibration over the full dataset length would not be possible without access to massive parallel computing facility. In these cases, the selection of a calibration subsample would still be necessary but should not follow typical splitting strategies but rather focus on strategic approaches such as proposed by Singh and Bárdossy (2012) or Razavi and Tolson (2013). However, if the eventuality that a full calibration would be possible to the user, the additional information included within the calibration would very likely increase the performance of future simulations. Furthermore, a final analysis was performed to evaluate the effects of the length of the available time series. In this test, 51 years of data were used from the Trinity River catchment in California, ranging from 1948-1998. A single test run was performed while keeping the first year as the warm-up period, 15 random years as the independent test period and performing the calibration on random subsets of 1-34 calibration years. A maximum of 500 combinations per case was used to explore the larger dimension of the problem. The following evaluation of the *NSE* on the independent test period for these trials was compared to that of calibrating on all 35 available years. Results are presented in Figure 13.

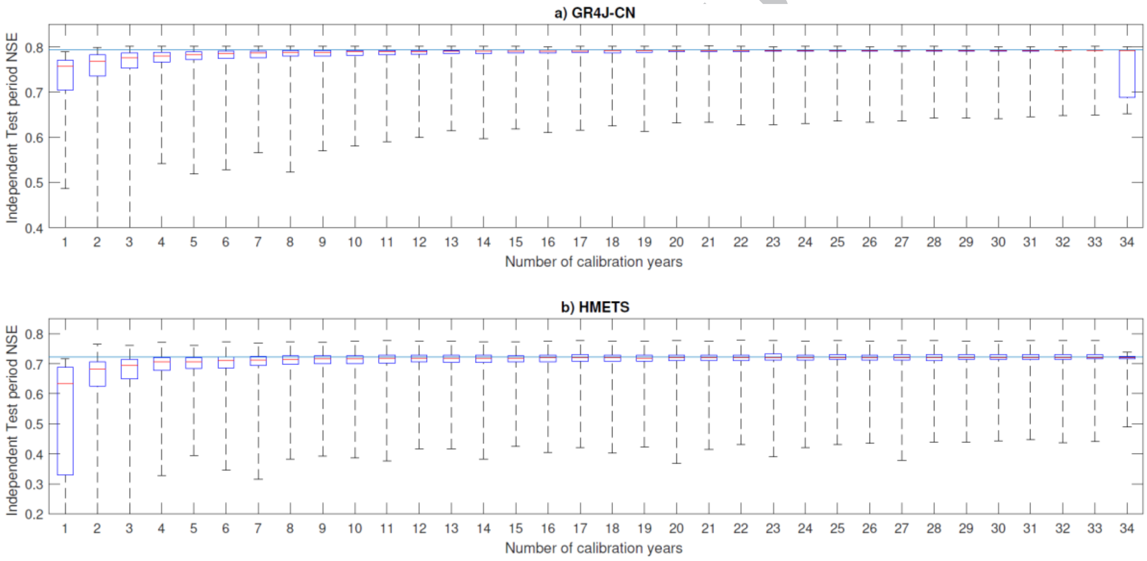

Figure 13. Boxplots of independent test-period *NSE* scores when calibrating on random subsets of 1-34 calibration year combinations for the a) GR4J and b) HMETS models. The horizontal line represents the independent test-period NSE when calibrating on the 35 available calibration years. Boxplot box edges represent the 25th and 75th quantiles, and the whiskers represent +/-2.7 standard deviations from the mean.

It is clear that the length of the available dataset plays a crucial role in the effects of split-sample calibration and validation. Clearly, from Figure 13, there is a limit to how much information a parameter set can contain. When there is sufficient information to represent the entirety of the time-series, then it seems that there is no gain in calibrating on all years, although there is a case to be made regarding the robustness of the parameter set. There are some combinations of 33 and 34 years that vastly underperform the 35-year case. Therefore, it is still recommended to use all available years during the calibration phase. Notice that the last boxplot in panel a) (34 out of 35 years) seems different than the others. This is because it contains only 35 points, whereas the other boxplots contain 500 points each and thus the threshold for outlier detection is wider.

5.7. The question of stationarity

In this study, the effect of calibration and validation is investigated on three catchments that did not show signs of non-stationarity, i.e. the mean annual streamflow did not contain a trend over a 25-year period. This allowed randomly sampling from the database to generate calibration and validation sets. This raises the question as to how the method would fare on a catchment that is subject to non-stationarity. Obviously, in this scenario, the independent test period would need to be in the most recent years and those years could not be randomly selected from the entire time series. Furthermore, the calibration and validation would have to make use of the remaining years, perhaps on a distinctly different climate. However, in an operational or research setting, there are two options to hedge one's bets.

First, the odd/even year split-sample method has been proposed to calibrate a hydrological model on non-stationary conditions. This method allows calibrating on the odd years and cross-validating on the even years (or vice-versa), thus including the same trend in the calibration and validation. While this gives confidence to the modeler (as explained in Section 5.3), the reality is that the time series does not contain all of the information that is available and that the future simulations will be performed with a model that under-exploited. Furthermore, the drawback to this method is that it takes essentially as much time to perform this type of calibration, as the whole period still has to be simulated to implement such split-sample techniques.

Second, the logical conclusion would be to perform a calibration on the entire time-series to include the information contained in the even years. As stated in Section 5.5, it would be possible to perform the validation and, given an acceptable *NSE*, recalibrate the parameters using the entire time series. In light of the results obtained in this study, it is posited that the method should perform as well as it does in stationary conditions, but this remains to be validated and should be explored in future works. Some of these issues are discussed in Thirel et al. (2015).

6. Conclusion

In this study, an experimental approach is presented to investigate the added benefit of foregoing the traditional validation step in hydrological model calibration. The results presented above point towards validation being futile at best, detrimental at worst, and deceiving in all cases. The main hypothesis that calibrating a hydrological model on a certain period, and then validating the model's parameter set on another period, was shown to be flawed due to numerous possibilities of false-negatives. When a validation *NSE* is inferior to a calibration *NSE*, typically the model is recalibrated or the calibration and validation years are inverted or shifted around. The end-result is a model whose parameters include only information from the calibration period, sacrificing the information contained in the rest of the time-series.

The method proposed in this study is to calibrate over the entire time-series and forego the validation step. If the calibration skill is acceptable, then the hypothesis is made that the model is able to simulate on that period. A hybrid method that conserves confidence in the model's robustness was proposed in which the model is recalibrated over the entire period if the typical calibration/validation step returns acceptable results.

The results are consistent over the three contrasting catchments and two models, which provides some evidence to the full time-series calibration to be the optimal strategy, although some calibration/validation combinations do outperform the full calibration skill in a small proportion. Statistical testing showed that the method was robust, and verification on 5 independent catchments showed similar results. Furthermore, the method translated well to other metrics such as the relative bias and the Pearson correlation coefficient.

Of course, the problem lies in the fact that the capacity of the model to perform better on an independent testing period cannot be evaluated a priori, therefore the optimal strategy is the one which is statistically more likely to occur given random inputs. In this study, in 30 independent runs and over six test-cases, the optimal strategy was convincingly to calibrate on the entire time-series and forgo validation. The same was found for the smaller verification run, with 10 independent runs over 10 model-catchment pairs.

Finally, it is important to note that this work has some limitations that should be addressed in future work. For example, non-stationary time-series should also benefit from this methodology, but this remains to be validated. Furthermore, the generalizability of the method to other catchments and models should be investigated, although it is expected that the same results will be found. For example, models operating on sub-daily time-steps could be tested using this methodology to see if the same behavior is observed. The fact that information is left out of the parameter set should manifest itself even further in more process-based hydrological models, but this also needs to be tested.

Acknowledgements

The authors would like to extend their gratitude to the anonymous reviewers and associate editor whose insightful comments were instrumental in shaping the paper in its current form. The authors would like to thank the École de technologie supérieure (ETS) and the Natural Sciences and Engineering Research Council of Canada (NSERC) for partial funding of this research.

References

Andreassian, V., A. Hall, N. Chahinian, and J. Schaake, 2006: Large sample basin experiments for hydrological model parameterization: results of the Model Parameter Experiment (MOPEX), 346 p.

Arnold, J. G., and Coauthors, 2012: SWAT: Model use, calibration, and validation. Transactions of the ASABE, 55, 1491. doi: 10.13031/2013.42256

Arsenault, R., M. Latraverse, and T. Duchesne, 2016a: An efficient method to correct under-dispersion in ensemble streamflow prediction of inflow volumes for reservoir optimization. Water Resources Management, 30, 4363-4380. doi: 10.1007/s11269-016-1425-4

Arsenault, R., G. R. C. Essou, and F. P. Brissette, 2017: Improving hydrological model simulations with combined multi-input and multimodel averaging frameworks. Journal of Hydrologic Engineering, 22, 04016066. doi: 10.1061/(ASCE)HE.1943-5584.0001489

Arsenault, R., A. Poulin, P. Côté, and F. Brissette, 2014: Comparison of stochastic optimization algorithms in hydrological model calibration. Journal of Hydrologic Engineering, 19, 1374-1384. doi: 10.1061/(ASCE)HE.1943-5584.0000938

Arsenault, R., R. Bazile, C. Ouellet Dallaire, and F. Brissette, 2016b: CANOPEX: A Canadian hydrometeorological watershed database. Hydrological Processes, 30, 2734-2736. doi: 10.1002/hyp.10880

Band, L. E., P. Patterson, R. Nemani, and S. W. Running, 1993: Forest ecosystem processes at the watershed scale: incorporating hillslope hydrology. Agricultural and Forest Meteorology, 63, 93-126. doi: 10.1016/0168-1923(93)90024-C

Bennett, N. D., and Coauthors, 2013: Characterising performance of environmental models. Environmental Modelling & Software, 40, 1-20. doi: 10.1016/j.envsoft.2012.09.011

Bergström, S., B. Carlsson, M. Gardelin, G. Lindström, A. Pettersson, and M. Rummukainen, 2001: Climate change impacts on runoff in Sweden - assessments by global climate models, dynamical downscaling and hydrological modelling. Climate Research, 16, 101-112.

Biondi, D., G. Freni, V. Iacobellis, G. Mascaro, and A. Montanari, 2012: Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. Physics and Chemistry of the Earth, Parts A/B/C, 42-44, 70-76. doi: 10.1016/j.pce.2011.07.037

Boyle, D. P., H. V. Gupta, and S. Sorroshian, 2000: Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. Water Resources Research, 36, 3663-3674. doi: 10.1029/2000WR900207

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen, 2004: An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. Journal of Hydrology, 298, 242-266. doi: 10.1016/j.jhydrol.2004.03.042

Chen, J., F. P. Brissette, and R. Leconte, 2011a: Uncertainty of downscaling method in quantifying the impact of climate change on hydrology. Journal of Hydrology, 401, 190-202. doi: 10.1016/j.jhydrol.2011.02.020

Chen, J., F. P. Brissette, A. Poulin, and R. Leconte, 2011b: Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed. Water Resources Research, 47. doi: 10.1029/2011WR010602

Daggupati, P., and Coauthors, 2015: A recommended calibration and validation strategy for hydrologic and water quality models. Transactions of the ASABE, 58, 1705. doi: 10.13031/trans.58.10712

Day, G. N., 1985: Extended Streamflow Forecasting Using NWSRFS. Journal of Water Resources Planning and Management, 111, 157-170. doi: 10.1061/(ASCE)0733-9496(1985)111:2(157)

Duan, Q., S. Sorooshian, and V. K. Gupta, 1994: Optimal use of the SCE-UA global optimization method for calibrating watershed models. Journal of Hydrology, 158, 265-284. doi: 10.1016/0022-1694(94)90057-4

Duan, Q., and Coauthors, 2006: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. Journal of Hydrology, 320, 3-17. doi: 10.1016/j.jhydrol.2005.07.031

Essou, G. R. C., R. Arsenault, and F. P. Brissette, 2016: Comparison of climate datasets for lumped hydrological modeling over the continental United States. Journal of Hydrology, 537, 334-345. doi: 10.1016/j.jhydrol.2016.03.063

Fatichi, S., and Coauthors, 2016: An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. Journal of Hydrology, 537, 45-60. doi: 10.1016/j.jhydrol.2016.03.026

Gaborit, É., S. Ricard, S. Lachance-Cloutier, F. Anctil, and R. Turcotte, 2015: Comparing global and local calibration schemes from a differential split-sample test perspective. Canadian Journal of Earth Sciences, 52, 990-999. doi: 10.1139/cjes-2015-0015

Garavaglia, F., M. Le Lay, F. Gottardi, R. Garçon, J. Gailhard, E. Paquet, and T. Mathevet, 2017: Impact of model structure on flow simulation and hydrological realism: from a lumped to a semi-distributed approach. Hydrology and Earth System Sciences, 21, 3937. doi: 10.5194/hess-21-3937-2017

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

Gaur, S., P. K. Paul, R. Singh, A. Mishra, P. K. Gupta, and R. P. Singh, 2017: Operational testing of Satellite based Hydrological Model (SHM). EGU General Assembly Conference Abstracts, 2543.

Gharari, S., M. Hrachowitz, F. Fenicia, and H. H. G. Savenije, 2013: An approach to identify time consistent model parameters: sub-period calibration. Hydrology and Earth System Sciences, 17, 149-161. doi: 10.5194/hess-17-149-2013

Gowda, P. H., D. J. Mulla, E. D. Desmond, A. D. Ward, and D. N. Moriasi, 2012: ADAPT: Model use, calibration, and validation. Transactions of the ASABE, 55, 1345. doi: 10.13031/2013.42246

Guo, C., G. Pleiss, Y. Sun, and K. Q. Weinberger, 2017: On calibration of modern neural networks. arXiv preprint arXiv:1706.04599.

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

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

Hansen, N., and A. Ostermeier, 1996: Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation. Proceedings of IEEE International Conference on Evolutionary Computation, 312-317.

——, 2001: Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation, 9, 159-195. doi: 10.1162/106365601750190398

Huard, D., and A. Mailhot, 2008: Calibration of hydrological model GR2M using Bayesian uncertainty analysis. Water Resources Research, 44. doi: 10.1029/2007WR005949

Hutchinson, M. F., D. W. McKenney, K. Lawrence, J. H. Pedlar, R. F. Hopkinson, E. Milewska, and P. Papadopol, 2009: Development and testing of Canada-wide interpolated spatial models of daily minimum–maximum temperature and precipitation for 1961–2003. Journal of Applied Meteorology and Climatology, 48, 725-741. doi: 10.1175/2008jamc1979.1

Jain, S. K., and K. P. Sudheer, 2008: Fitting of Hydrologic Models: A Close Look at the Nash-Sutcliffe Index. Journal of Hydrologic Engineering, 13, 981-986. doi: 10.1061/(ASCE)1084-0699(2008)13:10(981)

Jakeman, A. J., and G. M. Hornberger, 1993: How much complexity is warranted in a rainfall- runoff model? Water Resources Research, 29, 2637-2649. doi: 10.1029/93WR00877

Jiang, T., Y. D. Chen, C.-Y. Xu, X. Chen, X. Chen, and V. P. Singh, 2007: Comparison of hydrological impacts of climate change simulated by six hydrological models in the Dongjiang Basin, South China. Journal of Hydrology, 336, 316-333. doi: 10.1016/j.jhydrol.2007.01.010

Juston, J., J. Seibert, and P. O. Johansson, 2009: Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment. Hydrological Processes, 23, 3093-3109. doi: 10.1002/hyp.7421

Kendall, M. G., 1975: Rank correlation methods. 4th Edition ed. Charles Griffin.

KlemeŠ, V., 1986: Operational testing of hydrological simulation models. Hydrological Sciences Journal, 31, 13-24. doi: 10.1080/02626668609491024

Larabi, S., A. St-Hilaire, and F. Chebana, 2018: A new concept to calibrate and evaluate a hydrological model based on functional data analysis. Journal of Water Management Modeling. doi: 10.14796/JWMM.C442

Legates, D. R., and G. J. McCabe, 1999: Evaluating the use of "goodness- of- fit" Measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35, 233-241. doi: 10.1029/1998WR900018

Liu, D., S. Guo, Z. Wang, P. Liu, X. Yu, Q. Zhao, and H. Zou, 2018: Statistics for sample splitting for the calibration and validation of hydrological models. Stochastic Environmental Research and Risk Assessment. doi: 10.1007/s00477-018-1539-8

Martel, J.-L., K. Demeester, F. Brissette, A. Poulin, and R. Arsenault, 2017: HMETS—A simple and efficient hydrology model for teaching hydrological modelling, flow forecasting and climate change impacts. International Journal of Engineering Education, 33, 1307-1316.

Maurer, E. P., A. W. Wood, J. C. Adam, D. P. Lettenmaier, and B. Nijssen, 2002: A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. Journal of Climate, 15, 3237-3251. doi: 10.1175/1520-0442(2002)015<3237:Althbd>2.0.Co;2

McCuen, R. H., Z. Knight, and A. G. Cutter, 2006: Evaluation of the Nash-Sutcliffe efficiency index. Journal of Hydrologic Engineering, 11, 597-602. doi: 10.1061/(ASCE)1084-0699(2006)11:6(597)

Middelkoop, H., and Coauthors, 2001: Impact of climate change on hydrological regimes and water resources management in the Rhine Basin. Climatic Change, 49, 105-128. doi: 10.1023/a:1010784727448

Minville, M., F. Brissette, and R. Leconte, 2008: Uncertainty of the impact of climate change on the hydrology of a nordic watershed. Journal of Hydrology, 358, 70-83. doi: 10.1016/j.jhydrol.2008.05.033

Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, 2007: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50, 885. doi: 10.13031/2013.23153

Moriasi, D. N., and Coauthors, 2015: Hydrologic and water quality models: key calibration and validation topics. Transactions of the ASABE, 58, 1609. doi: 10.13031/trans.58.11075

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

Newman, A. J., and Coauthors, 2015: 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. doi: 10.5194/hess-19-209-2015

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

Paniconi, C., and M. Putti, 2015: Physically based modeling in catchment hydrology at 50: Survey and outlook. Water Resources Research, 51, 7090-7129. doi: 10.1002/2015WR017780

Perrin, C., C. Michel, and V. Andréassian, 2003: Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology, 279, 275-289. doi: 10.1016/S0022-1694(03)00225-7

Perrin, C., L. Oudin, V. Andreassian, C. Rojas-Serna, C. Michel, and T. Mathevet, 2007: Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models. Hydrological Sciences Journal, 52, 131-151. doi: 10.1623/hysj.52.1.131

Poissant, D., R. Arsenault, and F. Brissette, 2017: Impact of parameter set dimensionality and calibration procedures on streamflow prediction at ungauged catchments. Journal of Hydrology: Regional Studies, 12, 220-237. doi: 10.1016/j.ejrh.2017.05.005

Razavi, S., and B. A. Tolson, 2013: An efficient framework for hydrologic model calibration on long data periods. Water Resources Research, 49, 8418-8431. doi: 10.1002/2012WR013442

Refsgaard, J. C., H. J. Henriksen, W. G. Harrar, H. Scholten, and A. Kassahun, 2005: Quality assurance in model based water management – review of existing practice and outline of new approaches. Environmental Modelling & Software, 20, 1201-1215. doi: 10.1016/j.envsoft.2004.07.006

Singh, S. L., and A. Bárdossy, 2012: Calibration of hydrological models on hydrologically unusual events. Advances in Water Resources, 38, 81-91. doi: 10.1016/j.advwatres.2011.12.006

Singh, V. P., and D. A. Woolhiser, 2002: Mathematical Modeling of Watershed Hydrology. Journal of Hydrologic Engineering, 7, 270-292. doi: 10.1061/(ASCE)1084-0699(2002)7:4(270)

Thirel, G., V. Andréassian, and C. Perrin, 2015: On the need to test hydrological models under changing conditions. Hydrological Sciences Journal, 60, 1165-1173. doi: 10.1080/02626667.2015.1050027

Tolson, B. A., and C. A. Shoemaker, 2007: Cannonsville Reservoir watershed SWAT2000 model development, calibration and validation. Journal of Hydrology, 337, 68-86. doi: 10.1016/j.jhydrol.2007.01.017

Troin, M., R. Arsenault, and F. Brissette, 2015: Performance and uncertainty evaluation of snow models on snowmelt flow simulations over a nordic catchment (Mistassibi, Canada). Hydrology, 2, 289. doi: 10.3390/hydrology2040289

Troin, M., R. Arsenault, J.-L. Martel, and F. Brissette, 2018: Uncertainty of hydrological model components in climate change studies over two nordic Quebec catchments. Journal of Hydrometeorology, 19, 27-46. doi: 10.1175/jhm-d-17-0002.1

Valéry, A., 2010: Modélisation précipitations–débit sous influence nivale. Élaboration d'un module neige et évaluation sur 380 bassins versants. Ph.D. thesis, Agro Paris Tech, 417 pp. https://webgr.irstea.fr/wp-content/uploads/2012/07/2010-VALERY-THESE.pdf.

van der Spek, J. E., and M. Bakker, 2017: The influence of the length of the calibration period and observation frequency on predictive uncertainty in time series modeling of groundwater dynamics. Water Resources Research, 53, 2294-2311. doi: 10.1002/2016WR019704

Vehviläinen, B., 1992: Snow cover models in operational watershed forecasting. PhD Thesis, National Board of Waters and the Environment, Helsinki.

Velázquez, J. A., M. Troin, D. Caya, and F. Brissette, 2015: Evaluating the time-invariance hypothesis of climate model bias correction: Implications for hydrological impact studies. Journal of Hydrometeorology, 16, 2013-2026. doi: 10.1175/jhm-d-14-0159.1

Vrugt, J. A., H. V. Gupta, S. C. Dekker, S. Sorooshian, T. Wagener, and W. Bouten, 2006: Application of stochastic parameter optimization to the Sacramento Soil Moisture Accounting model. Journal of Hydrology, 325, 288-307. doi: 10.1016/j.jhydrol.2005.10.041

Wallner, M., U. Haberlandt, and J. Dietrich, 2012: Evaluation of different calibration strategies for large scale continuous hydrological modelling. Advances in Geosciences 31 (2012), 31, 67-74. doi: 10.5194/adgeo-31-67-2012

Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long- range experimental hydrologic forecasting for the eastern United States. Journal of Geophysical Research: Atmospheres, 107, ACL 6-1-ACL 6-15. doi: 10.1029/2001JD000659

- Calibration on full time series is shown to be more robust than split-sample methods
- 30 Bootstrapping tests on 6 cases provide evidence towards the method being optimal
- Verification on 10 independent catchment-model pairs support the conclusions
- Caveats of split-sampling on model performance are demonstrated
- Length of the calibration period is proportional to the parameter set robustness