

Water Resources Research®

RESEARCH ARTICLE

10.1029/2021WR031523

Key Points:

- A unique split-sample experiment is performed across 463 catchments to provide guidance on split sample decision-making in model calibration
- Calibrating models to the full available data period and skipping model validation entirely is the most robust choice
- Calibrating models to older data and then validating models on newer data, a very common approach in literature, is an inferior choice

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

H. Shen,
hongren.shen@uwaterloo.ca

Citation:

Shen, H., Tolson, B. A., & Mai, J. (2022). Time to update the split-sample approach in hydrological model calibration. *Water Resources Research*, 58, e2021WR031523. <https://doi.org/10.1029/2021WR031523>

Received 2 NOV 2021

Accepted 13 FEB 2022

Author Contributions:

Conceptualization: Hongren Shen, Bryan A. Tolson

Data curation: Hongren Shen, Julianne Mai

Formal analysis: Hongren Shen, Bryan A. Tolson, Julianne Mai

Funding acquisition: Bryan A. Tolson

Investigation: Hongren Shen, Bryan A. Tolson

Methodology: Hongren Shen, Bryan A. Tolson

Project Administration: Bryan A. Tolson




Resources: Hongren Shen

Software: Hongren Shen, Julianne Mai

© 2022. The Authors.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial License](https://creativecommons.org/licenses/by-nc/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Time to Update the Split-Sample Approach in Hydrological Model Calibration

Hongren Shen¹ , Bryan A. Tolson¹ , and Julianne Mai¹ 

¹Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, ON, Canada

Abstract Model calibration and validation are critical in hydrological model robustness assessment. Unfortunately, the commonly used split-sample test (SST) framework for data splitting requires modelers to make subjective decisions without clear guidelines. This large-sample SST assessment study empirically assesses how different data splitting methods influence post-validation model testing period performance, thereby identifying optimal data splitting methods under different conditions. This study investigates the performance of two lumped conceptual hydrological models calibrated and tested in 463 catchments across the United States using 50 different data splitting schemes. These schemes are established regarding the data availability, length and data recentness of continuous calibration sub-periods (CSPs). A full-period CSP is also included in the experiment, which skips model validation. The assessment approach is novel in multiple ways including how model building decisions are framed as a decision tree problem and viewing the model building process as a formal testing period classification problem, aiming to accurately predict model success/failure in the testing period. Results span different climate and catchment conditions across a 35-year period with available data, making conclusions quite generalizable. Calibrating to older data and then validating models on newer data produces inferior model testing period performance in every single analysis conducted and should be avoided. Calibrating to the full available data and skipping model validation entirely is the most robust split-sample decision. Experimental findings remain consistent no matter how model building factors (i.e., catchments, model types, data availability, and testing periods) are varied. Results strongly support revising the traditional split-sample approach in hydrological modeling.

Plain Language Summary Hydrological model calibration is a critical model building process that infers key model parameter values from observed system response data. Conventionally, this process requires the historical period to be split into a calibration period for tuning parameters and a validation period for testing model robustness (i.e., the split-sample). Unfortunately, there is a lack of empirical evidence supporting how exactly to define the split-sample. We designed an exhaustive and novel experiment comparing the range of possible split-sampling schemes, including calibrating to older/recent years, calibrating to a short/long period, and calibrating to the full period of available system response data. Each scheme was evaluated based on performance, assessed in three different ways, in numerous post-validation model testing periods for each of the 926 calibration case studies (two different hydrological models applied in 463 catchments). Results show that using older data for model calibration and then using newer data for validation, which is the typical practice in the literature, is an inferior choice and should be avoided. The results also show that calibrating to the full historical data and skipping model validation entirely is the most robust choice. Therefore, the split-sample approach applied in this community for decades should be revised.

1. Introduction

Advances in computing capabilities and data collection have inspired many hydrological models to be developed, utilized, and improved in the last half century (Beven, 1989, 2012; Devia et al., 2015; Savenije, 2009; Singh & Woolhiser, 2003). Hydrological models, which essentially are a set of mathematical equations based on simple physical laws that simulate sophisticated physics in hydrologic processes (Blöschl et al., 2013; Singh & Chow, 2016), have been extensively employed as tools to either advance the understanding of the hydrological cycle or facilitate decision-making for many purposes such as water resources management and planning, flood and drought forecasting, reservoir management, and climate change assessment (Beckers et al., 2009; Blöschl et al., 2013; Fowler et al., 2007; Hrachowitz et al., 2013; Mishra & Singh, 2011). However, when a hydrological model of a watershed is built for such applications, there are many impactful subjective decisions required, such

Supervision: Bryan A. Tolson

Validation: Hongren Shen

Visualization: Hongren Shen, Julianne Mai

Writing – original draft: Hongren Shen

Writing – review & editing: Hongren Shen, Bryan A. Tolson, Julianne Mai

as input data sets, model structure, spin-up/initialization strategy, parameter calibration and performance metrics (Melsen et al., 2019).

Model calibration is a process for either manually or automatically adjusting influential model parameters over a specific simulation period to obtain model outputs (e.g., streamflow at the catchment outlet) matching the corresponding observations as closely as possible (Arsenault et al., 2018; Beven, 2012; Duan et al., 1994; Legates & McCabe, 1999). Typically, calibrated hydrological model performance is also evaluated against observations that are not used in model calibration before the model is applied to support water resources management decisions. We adopt the word “validation” for this process and formally define validation as the quantitative and qualitative evaluation of model performance against new observations not used in calibration in order to ensure parameter transferability and model robustness (Arsenault et al., 2018; Biondi et al., 2012; Klemeš, 1986). Validation for hydrological models is not for testing scientific theory but a testing of whether models are acceptable for a given purpose (Refsgaard & Henriksen, 2004). Also, “evaluation” is often an alternative word used to mean the same thing as validation (Fowler, Peel et al., 2018; Fowler, Coxon et al., 2018). Therefore, validation and evaluation are used interchangeably throughout the paper.

Model performance in the validation period is conditioned by the choice of calibration period (Coron et al., 2012; Guo et al., 2020; Myers et al., 2021). Thus, the data splitting scheme is a key decision when building a model. Moreover, the length of calibration period is reported to have varied influences on hydrological modeling (Guo et al., 2018; Knoben et al., 2020). The information contained in a calibration period and the efficiency with which the information is extracted are key to model calibration (Sorooshian et al., 1983). Thus, some studies use a sufficiently long calibration period to include representative dry and wet conditions (e.g., see Gupta & Sorooshian, 1985; Yapo et al., 1996), while some studies suggest that models be calibrated on a sub-period of the full-period record that has representative hydrological dynamics to those expected to be the evaluation period (e.g., see Li et al., 2012).

This study focuses on the decision about how to split the available system response data between model calibration and model validation in order to achieve good quality model predictions in some post-validation model application (e.g., using the model in a decision-support context). If we restrict ourselves to a context where the watershed outlet streamflow is the prediction of interest and this is also the location of the only observations of system response, then, at the time the model is built, we can describe model simulations as covering three different time periods: the calibration, validation and model application period, where the application period could generally be considered to be some period in the future (e.g., after the model validation is completed and the model is deemed to be fit-for-purpose). In this context, all available system response data is split into a calibration and a validation sub-period.

Klemeš (1986) provided the formative framework for hydrological model validation. The split-sample test (SST), also called holdout method (Kohavi, 1995), is central to the four-level model performance validation framework proposed by Klemeš (1986). This framework consists of four different levels of model validation, including the SST and the differential split-sample test (DSST). The DSST method is a specific case of the SST, as it selects calibration and validation periods based on pre-defined or pre-screened climatic differences (Coron et al., 2012; Klemeš, 1986). In recent decades, many variations of the SST and DSST methods have evolved and been applied in many hydrological modeling studies (e.g., see Dakhloui et al., 2017; Essou et al., 2016; Coron et al., 2012). The original SST method (hereafter termed “SST-K”; Klemeš, 1986) splits a data record into two sub-periods, models are then calibrated over one sub-period and validated over the other sub-period and vice versa, thus requiring a “two-round” calibration and validation (i.e., performing two calibration plus validation experiments). Models are deemed as acceptable when the two-round model validation results are similar and both acceptable. Specifically, when the data record is sufficiently long, SST-K employs a data splitting scheme where the data record is split into two equal-length sub-periods for the two-round calibration and validation, that is, the first 50% for calibration and the last 50% for validation (denoted as C_{50}/V_{50}) and the first 50% for validation and the last 50% for calibration (denoted as V_{50}/C_{50}). When the data period length is insufficient, the two data splitting schemes are the first 70% for calibration and the last 30% for validation (denoted as C_{70}/V_{30}), and the last 70% for calibration and the first 30% for validation (denoted as V_{30}/C_{70}), respectively. However, there are three main drawbacks in the original SST-K method: (a) The “sufficiently long” data record is not adequately defined, which leaves it vague for the selection of data splitting schemes; (b) The 50:50 or 70:30 splitting schemes are suggested without empirical/numerical evidence provided to support selecting these splits; and (c) There is no guidance

on which parameter set should eventually be selected out of the two parameter sets that any SST-K method will produce.

Due at least in part to these shortcomings, a simplified SST variation has been widely adopted for deterministic hydrological modeling. The simplified SST method employs only one data splitting scheme from the SST-K method, defining a single calibration and single validation period, and has been the most commonly used data splitting method in hydrological modeling community (e.g., see Pool et al., 2018; Rakovec et al., 2019; Schlef et al., 2021). In the simplified SST approach used for model building, the data splitting scheme often does not follow the 50:50 or 70:30 guidance in Klemeš (1986). In fact, according to Daggupati et al. (2015) and Myers et al. (2021), the rationale for the selected data splitting scheme in hydrological model publications is rarely clarified. More interestingly, most studies tend to select calibration and validation data years chronologically, that is, the earlier years in the data record are used for calibration and the more recent years are retained for validation. Myers et al. (2021) summarized 25 papers on model calibration and validation for six hydrological models, in which 24 (96%) of them followed this data splitting approach but none of them clarified reasons. In our collective experience, we typically have applied this data splitting approach because it is most practical (convenient and computationally efficient) in continuous hydrological modeling (i.e., calibrating first and then validating at the end of the calibration period only requires initial conditions be specified once and then only the calibration period needs to be simulated during the iterative model calibration process). Doing it the other way, that is calibrating to later data and then validating to earlier data, forces the modeler to either (a) inefficiently simulate the entire validation plus calibration period during the iterative model calibration process or (b) somewhat inconveniently specify initial conditions for both a calibration period simulation and a validation period simulation (i.e., two different spin-up periods are required for calibration and validation), thus potentially introducing a discontinuity in model predictions if one was to then stitch together model outputs from the separate simulations. A key model benchmarking study for 531 catchments across the contiguous United States by Newman et al. (2017) did not follow the typical practice and selected instead (without any reported rationale) to calibrate to the 2000–2008 period and then validate to the earlier 1990–1999 period. As a result, a number of follow-up studies comparing to this benchmark have thus necessarily followed the same data splitting choice as Newman et al. (2017).

Model validation/evaluation studies proposing new or comparing alternative split sample approaches typically focus on evaluating similarity of model performance between the calibration and validation periods and do not assess performance in a third period that is independent of both the calibration and validation periods. Example studies in this category include Coron et al. (2012); Dakhlaoui et al. (2017, 2019); Essou et al. (2016); Fowler et al. (2016); Guo et al. (2018); Knoben et al. (2020); Li et al. (2012); Myers et al. (2021); Nicolle et al. (2021); Pool et al. (2018); Rakovec et al. (2019); Schlef et al. (2021); and Vaze et al. (2010). Splitting available data into either calibration or validation is a practical and understandable approach given limitations on the length of available system response data. However, empirical testing of split sample decisions or alternative model evaluation procedures during an independent model testing period, representing what model performance can be expected to be in some future application, would clearly provide a better assessment of the efficacy of the split sample approaches being studied.

In contrast to the above common two-period split-sample experimental design, there are only a few hydrological modeling studies, such as Arsenault et al. (2018) and Zheng et al. (2018), that utilized a third period, referred to as a model testing period, to evaluate alternative split-sample approaches. For example, in the extensive study by Arsenault et al. (2018), they demonstrated a Bootstrap-based DSST variation with a single discontinuous testing period. The remaining non-testing period years were randomly split into discontinuous years of calibration and validation periods. This sampling process was repeated multiple times (i.e., bootstrapping) to obtain many random combinations of discontinuous years for calibration. They assessed how calibration periods influence model performance using 239,940 calibration schemes based on randomly selected data years with lengths increased from 1 to 16 years. Models were also calibrated over the entire data period and validation was skipped to contrast with other calibration schemes. Based on model performance in an independent testing period, Arsenault et al. (2018) recommended using as many years as possible in the calibration step and to entirely disregard validation under certain conditions. Guo et al. (2018) and Singh and Bárdossy (2012) also reported that calibrating to all available data may be a robust strategy. The recommendation to skip model validation has huge implications and warrants a more extensive empirical assessment, especially since there are three assumptions in the experimental design of Arsenault et al. (2018) that can be improved to further generalize their key conclusions.

First, they used only three catchments in North America, which made it hard to exclude the influence of climatic and catchment characteristics; and thus, one may get significantly different results in another region. Second, calibrating to randomly selected (discontinuous) years is often not realistic in a practical model building process. Using discontinuous years for model calibration and validation requires that models be run over the full data record between the first and last calibration year to ensure model state variables are consistent in time (Arsenault et al., 2018; Essou et al., 2016), which increases computational burden in calibration. Thus, such a discontinuous calibration period may not be feasible for distributed hydrological modeling applications due to the much higher model complexity and larger computational burden than lumped models. We reviewed 48 publications reporting on the calibration of the Variable Infiltration Capacity (VIC) model, a distributed model, all published in 2018 that are cited in the bibliometric study by Addor & Melsen (2019). Based on our review, distributed hydrological modelers clearly resort to data splitting schemes with continuous calibration and validation periods since it is shown that all 48 of these studies used a continuous calibration period for VIC model calibration. Third, a key assumption in the experimental design used by Arsenault et al. (2018) is that they identified a static set of model testing years (i.e., all 1,332 different combinations of calibration and validation periods are evaluated for one fixed testing period in their experiment). Thus, it is unclear if their findings were conditional on this single testing period.

The purpose of model validation is to assess the *adequacy* of a model on the basis of the hydrological credibility of its outputs (Klemeš, 1986). Therefore, model validation procedures will sometimes function to identify inadequate models that should not be used in some future model application period. A robust validation procedure should therefore tend to successfully identify inadequate models. Unfortunately, the majority of model validation/evaluation methodological studies do not assess this aspect explicitly. For example, although Arsenault et al. (2018) evaluated which data splitting approaches led to the best testing period objective function values, they implicitly classified all calibration and validation results as successful since every one of their calibrated/validated model parameter sets always was evaluated in the model testing period. Their approach is not unique. Large-sample hydrological modeling studies that also do not explicitly characterize validation performance as adequate or inadequate include Bai et al. (2021); Essou et al. (2016); Fowler, Peel, et al. (2018); Fry et al. (2014); Gaborit et al. (2017); Guo et al. (2018); Mai et al. (2021); Mathevet et al. (2020); Newman et al. (2015, 2017); Rakovec et al. (2019); Smith et al. (2004, 2012); and Yang et al. (2019). We argue that the absence of explicit model failure criteria is a suboptimal approach to model building and that model failure handling at different steps in a model building process needs to be carefully considered, especially in the context of evaluating alternative model validation/evaluation strategies.

In calibration and validation, model performance is usually measured by quantitative metrics. A review on the performance metrics for environmental models is presented in Bennett et al. (2013). A model failure can be defined as an unacceptable simulation result when its corresponding model performance metric does not reach an acceptable level. There are two ways to define this acceptable level: (a) Absolute-level based criteria, which cuts off performance metrics into different ranges and arbitrarily define them as good or bad (see Guo et al., 2020; Moriasi et al., 2007, 2015; and Ritter & Muñoz-carpena, 2013); and (b) Reference models (also named benchmark, see Garrick et al., 1978; Knoben et al., 2020; Newman et al., 2015; and Schaeffli & Gupta, 2007), which are established based on mean observed flow.

Two large-sample studies that nicely assess explicit model failure instances in the context of model calibration/validation, but not in a post-validation model testing period, are Knoben et al. (2020) and Fowler et al. (2016). Knoben et al. (2020) computed a reference flow based on interannual mean/median discharge series on every calendar day over a specific reference period, and demonstrated this to be useful in characterizing whether a model is plausible (i.e., adequate) for a particular catchment. Fowler et al. (2016) is the first study we are aware of that framed model calibration results in the context of a confusion matrix where model calibration and evaluation results were categorized into four possible outcomes (both calibration and validation are good, both calibration and validation are poor, or calibration and validation have contradicting results). We note that Fowler et al. (2016) did not explicitly refer to their framing as a confusion matrix. We are unaware of past hydrological modeling studies using this confusion matrix-based classification framing to empirically evaluate the efficacy of alternate split sample decisions considering post-validation model testing periods.

In this study, we introduce a unique and comprehensive large-sample SST experimental design incorporating multiple post-validation model testing periods in order to empirically assess how best to perform a simplified

SST. In other words, we assess how to select a continuous sub-period for model calibration, thus leaving the remaining data for validation. Our experiments are conducted for 463 catchments, each with 35 years of available streamflow data, and two models in order to provide a reliable empirical assessment. We also highlight that the model build process includes decisions about how to handle if a model is deemed inadequate in the calibration or validation period. Experimental results are analyzed in multiple novel ways for more than a dozen different testing periods and all results point to the same general split-sample guidance (see details in Section 4): Calibrate to all data or at least calibrate to some of your most recent data, but do not calibrate to the oldest data and then validate on the newest data.

In Section 2, the large-sample of case study catchments, historical data and methodology are introduced. The key results and discussion is presented in Section 3 and Section 4, respectively. Finally, the conclusions are summarized in Section 5.

2. Data and Methodology

Section 2.1 introduces the novel SST experimental design, and then Section 2.2 introduces the catchment and data used in this study. Section 2.3 describes the methodology for analyzing the results.

2.1. Experimental Design for SST Assessment

This study applies multiple data splitting schemes for model calibration and validation. The calibration sub-periods (CSPs) are created based on the SST (Klemeš, 1986) and generalized split-sample test (Coron et al., 2012) frameworks. Unlike the philosophy in the DSST where the calibration period is created based on the pre-defined or pre-screened climatically contrasted conditions such as dry and wet years (Klemeš, 1986), this study only considers sub-periods defined by continuous years in CSP selection. We focus on continuous CSPs because: (a) we wanted to investigate the value of recent versus old calibration sub-periods, effectively precluding the use of discontinuous periods; (b) Arsenault et al. (2018) reported that calibrating to all data could be preferred to discontinuous sub-period calibrations; and (c) as discussed in Section 1, continuous CSPs are very common in the distributed model calibration literature.

Unlike previous studies, our SST assessment experiments define post-validation model testing periods. Such a three-period calibration, validation and testing scheme is stricter in SST assessment than the commonly used two-period approach. Multiple testing periods are created by simply pretending the model was built 5 years ago, thus leaving the five most-recent years as continuous model testing data, and then pretending the model was built 10 years ago, thus leaving the ten most-recent years as continuous model testing data, etc. As such, we utilize the terminology “model build year” in all of our experiments and different model build years generate different model testing periods. This rolling window approach to defining multiple model testing periods is, to the best of our knowledge, new in hydrological modeling, and extremely important since it avoids findings being specific to a single example model testing period (e.g., a single climatic condition). The available data prior to the model build year is split into many different continuous CSPs. Throughout this paper, available data refers to both the forcing data for the model and the system response data that model outputs will be compared to.

In this study, we use the calendar year for the SST experimental design such that a model simulation for a given year covers the 1 January–31 December period. In addition, data available prior to the model build year are used for model spin-up, calibration and validation. A model build year of 1990 implies the model was built instantaneously at midnight on 1 January 1990 and thus 1990 would be a year in the model testing period.

Figure 1 illustrates how our experimental design splits the 35 years of available data (1980–2014 based on our large sample of catchments described in Section 2.2) between the model spin-up, calibration, validation and model testing periods for five different model build years. The five model build years for building hydrological models (1990, 1995, 2000, 2005, and 2010) leave 10, 15, 20, 25, and 30 years of available data, respectively, for model spin-up, calibration and validation. For each model build year panel in Figure 1, CSPs are defined using sliding windows with varied lengths. Four representative lengths defined by the percentage of data available prior to the model build year are selected roughly as 30%, 50%, 70%, and 100%. These varied lengths of CSPs ensure the samples are composed of different information from short-period to the full-period (with a length indicated as 100%). The lengths 30%, 50%, and 70% of a data record are used here since they were proposed in the original

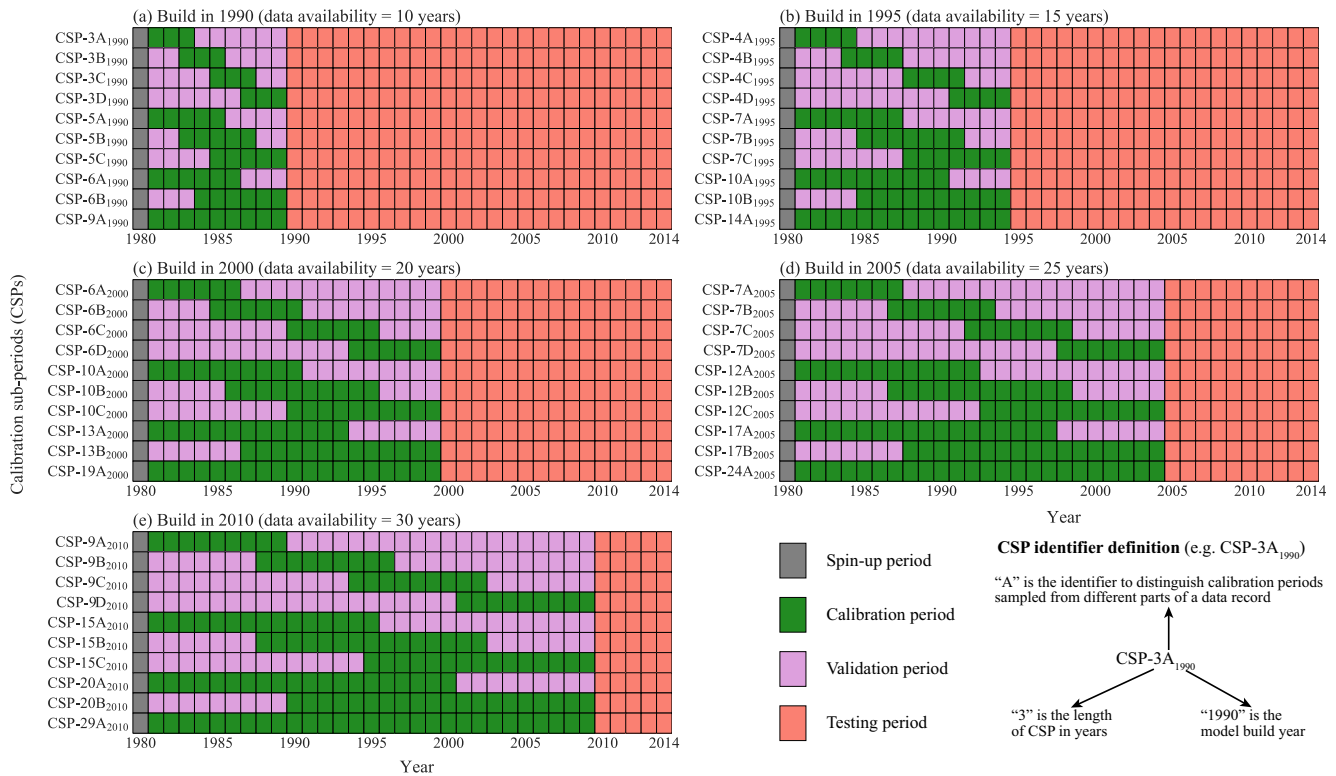


Figure 1. Experimental design for the split-sample test assessment. Calibration sub-periods (CSPs) are created for different model build years at (a) 1990, (b) 1995, (c) 2000, (d) 2005, and (e) 2010, with data availability for calibration then being 10, 15, 20, 25, and 30 years, respectively. Each CSP is assigned a unique identifier with a number denoting the CSP length in years, a letter corresponding to the unique calibration period, and a subscript indicating the model build year.

SST framework (Klemeš, 1986). Employing sliding windows allows CSPs of a different age (i.e., older vs. newer data) to be defined. For a given length of sliding window, there are multiple candidate CSPs. For example, given 9 years of available data for calibration and validation, using a 3-year sliding window can create up to seven CSPs. However, many of these CSPs overlap with one another and thus, calibration on all them can yield redundant information. This is a critical concern in a large-sample study like this one due to the high computational costs of excessive calibration experiments. Also, the autocorrelation in streamflow series may result in high correlation between consecutive data years (Kalra et al., 2008). Therefore, we require that the overlapping data years between two adjacent CSP samples be no more than 60% of the length of the sliding window. In addition, the CSPs of equal length are sometimes shifted slightly so that they are all symmetric over the available data (before the model build year). As shown in each panel of Figure 1, this CSP definition strategy creates 10 CSPs for each model build year.

With the various spin-up/calibration/validation/testing period configurations all defined, it is important to clarify exactly how the hydrological model simulations are conducted. Consider any row in any of the panels in Figure 1, the first year of available data (1980) is always used for model spin-up. There is no clear consensus on the optimal method for spinning up a model (Ajami et al., 2014). In general, spin-up behavior is found to be different with respect to catchments, models, state variables, and evaluation criteria. We use 1980 data recursively for three times to define a “three-year” spin-up period to initialize the hydrological models (i.e., force models with meteorological inputs in 1980 and repeatedly run these models in 1980 for three times with the end-of-day states on 31 December in the first 1980-run being the initial states on 1 January in the second 1980-run, and so forth), which is similar to how Lim et al. (2012) and Seck et al. (2015) built a multi-year spin-up in their model initialization studies. Running models with this yearly recursive forcing could eliminate interannual climate variability and lead models to an equilibrium state that is representative for the climatology of the 1-year forcing (Cosgrove et al., 2003). We evaluated results of this strategy across our case study area and found that the soil moisture content in our models were reaching a “practical” equilibrium state employed in Seck et al. (2015) after the 3-year simulation (soil moisture content at the end of the second and the third years are within 10%).

When calibrating a model, each time the model is simulated, the simulation period includes the 3-year spin-up period and terminates at the end of the CSP being evaluated. Although this can be inefficient when the validation period precedes the calibration period (e.g., see the fourth row in each panel of Figure 1), the benefit is that for all calibration experiments, the model initialization processes are completely consistent. Only the calibration period performance is assessed during calibration and model outputs for any validation period occurring prior to the calibration period are suppressed and thus not assessed. The best calibrated parameter set (identified using the calibration protocol in Section 2.4) is then used to simulate the model starting with the three-year spin-up period and ending at the end of 2014 (37-year simulation). This long time series of simulation results is then appropriately post-processed to compute the various calibration, validation and model testing period performance metrics.

Figure 1 shows testing periods for each model build year and they are referred to as full testing periods (i.e., the entire continuous period of years after the model build year). To even further generalize model testing regarding different climatic and hydrological conditions, each of these five full testing periods are augmented with two additional shorter length testing periods. Models are also tested in the first 3 years of the testing period and the first 5 years of the testing period. In total, there are 14 different testing periods for the five model build years (i.e., 5 model build years \times 3 testing periods $-$ 1 repeated testing period = 14). The 5-year and full-period testing periods are the same when models are built in 2010 and thus are only counted once. This spreads alternative testing periods across a 25-year period. Such a wide range of testing periods enable models to be tested in contrasting conditions and increase dissimilarities between a CSP and its corresponding testing periods, thus supporting more robust findings.

The 50 CSPs shown in Figure 1 are categorized into three classes, which hereafter are called as the *full-period* CSPs (no validation performed, such as CSP-9A₁₉₉₀ in Figure 1a), *recent* CSPs (calibration years immediately precede the model build year, such as CSP-3D₁₉₉₀, CSP-5C₁₉₉₀ and CSP-6B₁₉₉₀ in Figure 1a) and *older* CSPs (calibration years exclude the most recent years that immediately precede the model build year, such as CSP-3A₁₉₉₀, CSP-3B₁₉₉₀ and CSP-3C₁₉₉₀ in Figure 1a). The recent CSPs and older CSPs are also called *short-period* CSPs to be distinguished from the *full-period* CSPs.

In order to quantify just how old or new the data for the calibration period are, we use the term *recency* to describe how close CSPs are to the model build year (and hence, the start of the model testing period). Recency is computed as the ratio of two period lengths: the number of years between the CSP end date and 1980 over the number of years of available data prior to the model build year. For example, utilizing the CSP notation defined in Figure 1, considering 1990 as the model build year, CSP-3C₁₉₉₀ has a recency score of 8/10 or 80% whereas CSP-3D₁₉₉₀ has a recency score of 100%. The larger recency scores indicate more recent data years included in a CSP. We assign all CSPs into four recency bins/levels of 30%, 50%, 80%, and 100% even though precise recency scores for all of our CSPs are not exactly equal to these levels with minor rounding errors. For example, from Figure 1a panel, CSP-3C₁₉₉₀, CSP-5B₁₉₉₀, and CSP 6A₁₉₉₀ are all assigned a recency score of 80%.

2.2. Catchments and Data

The Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) data set provides long-term hydro-meteorological data of 671 catchments that are minimally impacted by human activities across the contiguous United States (CONUS; Addor et al., 2017; Newman et al., 2015). These data, available at the daily time step, include catchment-mean meteorological forcing from three datasets, Daymet, Maurer, and NLDAS, as well as daily observed streamflow for the catchment outlet from the United States Geological Survey (USGS). This data set serves as the candidate hydrological modeling inputs for our study. The CAMELS data set enables a large-sample study based on a wide range of hydroclimatic conditions, which facilitates robust statistical analysis of model performance and reduces the influence of case-specific studies, thereby further enabling robust hypothesis testing and statistically meaningful statements to be made using comparative hydrology (Gupta et al., 2014).

In this study, we perform a strict catchment filtering of the complete 671 catchments list. Commonly used catchment selection criteria on CAMELS datasets are based on specific catchment area ranges, catchment area discrepancies and water balance errors (Knoben et al., 2020; Kratzert et al., 2019; Newman et al., 2017). We require that catchment area discrepancies (calculated from the CAMELS derived catchment areas and the USGS reported drainage areas) be smaller than 10%, water balance errors be limited on Budyko curve (Budyko et al., 1974) that is similar to the CAMELS catchment filtering criterion used in Knoben et al. (2020), and the amount of missing

data be minimal. More specifically, consecutive missing data periods in a streamflow record must be less than four months in every year from 1980 to 2014 *and* all missing data for 1980–2014 is less than six months in total. These strict criteria are to minimize the negative impacts of outlier catchments on the controlled hydrological modeling experiments. After applying these criteria, 463 catchments are available with areas ranging from 4 km² to 25,800 km². Only 12 catchments in this list have missing data and since the amount of missing data is small (<1% of the 1980–2014 data), this will have negligible impacts on the hydrological modeling experiment.

The Daymet forcings are used in this study because of its longer availability period (1980–2014 compared to 1980–2008 for Maurer forcing) and the finer spatial resolutions (1 × 1 km compared to 12 × 12 km compared to Maurer and NLDAS forcings). Another reason for choosing Daymet is that Newman et al. (2015) and Addor et al. (2017) reported that Daymet forcings generated more accurate hydrological model simulation results.

The streamflow data record originally archived in the CAMELS data set contains many missing periods, especially in the latter part of the 1980–2014 time period. We therefore retrieved the latest streamflow data from the National Water Information System of the USGS to infill these missing periods.

The map for the spatial locations of all CAMELS catchments (including the 463 selected catchments and other filtered catchments) is presented in Figure S1 in Supporting Information S1. A table listing all 463 catchments selected is available in Table S1 in Supporting Information S1. The corresponding Daymet forcings and updated USGS streamflow data files for these catchments are all available online (see Data Availability Statement).

2.3. Hydrological Models

Two conceptual lumped hydrological models are applied in this study: the GR4J (du Génie Rural à 4 paramètres Journaliers) and Hydrological Model of École de technologie supérieure (HMETS) models. These two models are selected as representatives of different levels of model complexity (we calibrate six GR4J parameters and 21 HMETS parameters) to see how model complexity differences impact findings.

The GR4J model was originally developed as a four-parameter lumped model (Perrin et al., 2003), and has been extensively used in hydrological modeling worldwide (Mathevet et al., 2020; Oudin et al., 2018; Poncelet et al., 2017). In this study, it is coupled with a two-parameter snow accounting routine to consider snow processes, namely the CemaNeige degree-day snow model (Valéry, 2010), which is shown efficient and comparatively effective when associated with rainfall-runoff models at catchment scales (Valéry et al., 2014). Thus, there are six parameters in total for this version of GR4J in calibration (Note that it is named “GR6J” in Poncelet et al. (2017) and “GR4J-CN” in Arsenault et al. (2018), while here, we refer to this model as “GR4J”). GR4J employs two Unit Hydrographs for flow routing. Details of model structure and parameters of GR4J can be found in Perrin et al. (2003) and Valéry et al. (2014). GR4J calibration parameters and their ranges are provided in Table S2 in Supporting Information S1.

The HMETS introduced by Martel et al. (2017) is a more complex lumped model than GR4J, which considers more complicated hydrological processes and has up to 21 parameters for calibration, all of which are calibrated in our study. This model has been used in many hydrological modeling studies and has shown robust performance in previous studies (Arsenault et al., 2018; Chlumsky et al., 2021; Shen et al., 2018). HMETS employs two Unit Hydrographs to route the surface and delayed runoff. Details of model structure and parameters can be found in Martel et al. (2017). HMETS calibration parameters and their ranges are provided in Table S3 in Supporting Information S1.

In this study, both the GR4J and HMETS models are implemented in the Raven hydrological modeling framework (Craig et al., 2020). Raven is a robust and highly generalized object-oriented flexible modeling framework platform. It supports flexible customization in terms of a wide range of model structures, watershed discretization, process representations, forcing function estimation and interpolation methods and other numerical algorithms, which provides a standardized modeling platform and allows various types of hydrological modeling investigations, such as model structure sensitivity/uncertainty analysis (Chlumsky et al., 2021; Mai et al., 2022) and model inter-comparison (Mai et al., 2021). Raven conveniently unifies the format for both models' input and output files. Since all the inputs for GR4J and HMETS are in standardized Raven formats, these inputs form a useful CAMELS-based benchmark data set that are immediately available for use with any other Raven-configured

model structure (available online, see Data Availability Statement). Full details on the Raven framework can be found in Craig et al. (2020) and the Raven manual (Craig, 2020).

2.4. Calibration Protocol

In the proposed SST experiment, GR4J and HMETs are both calibrated in each of the 463 CAMELS catchments over the 50 CSPs introduced in Section 2.1. The dynamically dimensioned search (DDS) algorithm (Tolson & Shoemaker, 2007), which has been widely applied in hydrological model calibration studies (Chlumsky et al., 2021; Dembélé et al., 2020; Lahmers et al., 2019; Sharma et al., 2019; Spieler et al., 2020), is used to automatically calibrate model parameters. We utilize DDS as implemented in the optimization and calibration software toolkit OSTRICH (Matott, 2017). DDS is a neighborhood search algorithm and is based on a user-specified budget of model evaluations to find good quality calibration solutions (Tolson & Shoemaker, 2007). Given the different model complexities, we set the budget of model evaluations as 1,000 for GR4J and 3,000 for HMETs, and repeat 20 independent optimization trials with different randomly generated initial parameter sets in each CSP calibration. The best model parameter set out of the 20 optimization trials is then selected as the final calibrated parameter set and thus, only this parameter set is utilized to generate simulated hydrographs for the model validation period and model testing period. This approach (best of 20 optimization trials) is used to reduce the influence of optimization algorithm choice on results, as we believe the calibrated parameter set is very likely to be quite close to the globally optimal solution (i.e., a negligibly lower objective function than the globally optimal objective function value). Note that 20 trials were deemed to adequately balance the goal of closely approximating the global optimum against the need to also minimize the extreme computational burden associated with solving such a huge number of calibration problems. Accordingly, the total number of model calibration problems solved with DDS is 926,000 (2 models \times 50 CSPs \times 20 trials \times 463 catchments), and the total number of model test period hydrographs assessed is 129,640 (2 models \times [40 CSPs \times 3 testing periods + 10 CSPs \times 2 testing periods] \times 463 catchments).

The models are calibrated, validated and tested using the Kling-Gupta efficiency (KGE) metric (Gupta et al., 2009), which is a weighted combination of the three constitutive components (i.e., correlation, variability bias and mean bias) decomposed from the Nash-Sutcliffe efficiency (NSE; Nash & Sutcliffe, 1970) formula and is expressed as

$$\text{KGE} = 1 - \sqrt{(r - 1)^2 + \left(\frac{\sigma_{\text{sim}}}{\sigma_{\text{obs}}} - 1\right)^2 + \left(\frac{\mu_{\text{sim}}}{\mu_{\text{obs}}} - 1\right)^2} \quad (1)$$

where r is the linear correlation between observed and simulated flows, σ_{obs} and σ_{sim} are the standard deviation in observations and simulations, respectively, and μ_{obs} and μ_{sim} are the observation mean and simulation mean, respectively.

The KGE value ranges from $-\infty$ to 1.0, and KGE = 1.0 indicates the perfect agreement between simulations and observations. This metric has been demonstrated to be superior in estimating the variability in flows, especially for flow regimes with high seasonality, than the NSE (Gupta et al., 2009), and it is increasingly used in hydrological modeling studies. The selection of calibration and evaluation performance metric may be a subjective choice; however, every quantitative performance metric has its own pros and cons (Bennett et al., 2013). Although other performance metrics choices could be adopted in our framework, the comparison of different performance metrics is out of the scope of our study.

2.5. SST Comparative Performance Assessment

This section presents the methodology applied to compare how well one split-sample decision performs relative to other split-sample decisions. There are a myriad of ways to approach this comparison and so to ensure robust conclusions; we compare results in three different ways. How to do such a comparison depends on the modeler's subjective assessment of what is important and what constitutes a model failure. When comparing alternative CSPs across a large sample of catchments, we believe the following aspects of CSP performance as measured for the post-validation model testing period are example aspects of interest:

1. Frequency that one CSP is better than another CSP in terms of the objective function metric computed in the model testing period
2. Central tendency of the objective function metric as computed in the model testing period;
3. Frequency that a CSP correctly classifies model testing period failure (inadequacy) and success (adequacy)

With these general objectives in mind, we present three different assessment strategies in the subsections below (2.5.2–2.5.4). However, since all strategies depend on an explicit approach to model failure identification and handling, we first address this topic in Section 2.5.1 below.

2.5.1. Model Failure Handling

Model failure here is equivalent to how Klemeš (1986) describes model inadequacy: failure or inadequacy implies the model should not be utilized to support water resources decision-making. Here, we formalize the basis for the model failure/success determination using the reference climatology (flow). The reference climatology (flow) is established by calculating the mean value of observed streamflow on the reference period at a specific time scale (e.g., daily scale). Longer time scales, such as monthly means (Newman et al., 2015), provide smoother reference climatology, while shorter time scales capture more variability in hydrological regimes. We thus employ a daily scale mean flow to account for a stable seasonality in flow regimes every year. In this study, the reference period is dependent on whether we are in the model calibration/validation stage or the model testing phase of our experimental design. For example, when calculating reference flow for a calibration *or* a validation period, only the spin-up and calibration data years are available, while when calculating reference flow for a testing period, all data years prior to the model build year (spin-up, calibration and validation) can be utilized. Also note that reference flow series consist of 366 data points in leap years, in which the leap-day data point is generated from historical data on each 29 February.

Following Knoben et al. (2020), the KGE calculated using the reference flow as the predicted flow is denoted as *reference KGE*, and this reference KGE can be used to distinguish plausible/improbable model results. As such we identify a model simulation result as success or failure based on the KGE (calculated from simulations and observations) and its corresponding reference KGE (calculated from observations only). If a KGE value beats its reference KGE, the corresponding modeling result (i.e., parameter set) can be deemed a success; otherwise, it should be deemed as a failure. In our study, we evaluate success/failure of the calibration result (calibration period KGE vs. the reference KGE using the spin-up and calibration period observations), the validation result (validation period KGE vs. the reference KGE using the spin-up and calibration period observations), and the model testing result (testing period KGE vs. the reference KGE using the spin-up, calibration and validation period observations). Note the importance of our validation and testing period reference KGEs being independent of the corresponding observed data in these periods.

Model failure can be diagnosed in model calibration, in model validation or in model testing. In practice, a modeler must decide how to handle a model failure at the calibration or validation stage of the model building process. One option is to ignore the failure and proceed to use the model anyway (i.e., assuming a model can never fail in validation). An example is in Arsenault et al. (2018), which as described earlier is inconsistent with the philosophy of model validation. Another approach is to simply toss out the model completely and replace it with nothing in validation (e.g., see Guo et al., 2020; Knoben et al., 2020; Moriasi et al., 2007). The third approach is to throw away the model and instead identify a new model/new model parameter set or use the reference flow to try and predict conditions in the model testing period (practical and easy to implement in our study). The fourth approach, would be to make an optimal decision and choose between ignoring failure and simply using the reference flows, but this would require a large sample suite of experimental results evaluating both decision alternatives for performance in a model testing period.

In this study, we typically handle failures by replacing the model with the testing period reference flow that is available as of the model build year. The only analysis where we deviate from this approach is in our decision tree analysis (see Section 2.5.3 below) for comparing CSPs, where we utilize optimal decision-making to choose between ignoring failure and using the reference flow. The useful aspect of either failure handling approach is that both always yield a prediction in each catchment for the model testing period and thus generate a consistent sample size of 463 testing period outputs even if different CSPs have different rates of failure.

2.5.2. Frequency of Each Short-Period CSP Being Better Than Its Corresponding Full-Period CSP

In this simple analysis, we directly compare a pair of CSPs together by determining the frequency one does better than the other across all 463 catchments for a given testing period (computed for all 14 testing periods). Since we explicitly wanted to evaluate the hypothesis that all data should be used for calibration, the analysis here creates nine pairs (i.e., nine short-period CSPs each vs. their corresponding full-period CSP) for each model, model build year and testing period combination. For each pair, the frequency a short-period CSP has a better KGE in the model testing period than the full-period CSP is computed. This frequency is reported as a proportion and each proportion is calculated with a statistical sample size of 463. A relative frequency or proportion equal to 0.5 indicates that each CSP choice in the pair performs equally well. We further use a large sample 95% confidence interval for a proportion that shows only proportions smaller than 0.455 and larger than 0.545 are significantly different from 0.5. See Section 3.1 for the results.

2.5.3. Decision Tree Analysis

We use a decision tree as a first attempt to focus on assessing the CSP choice that optimizes the expected value of the objective function metric in the model testing period. The decision tree is a classic decision-making tool to help make sequential decisions under uncertainty. It is a well-used tool in water resources management decision-making (e.g., see Lund (1991); and Ray et al. (2019)). Decision trees are also a very common data mining approach used for classification and prediction (Nefeslioglu et al., 2010). To the best of our knowledge, this is the first time a decision tree analysis has been used to assess alternative split sample decisions in hydrologic model calibration.

The model building process is a sequence in time of various decisions and chance events. The chance or uncertain events in a model building and application context are whether our calibration period, validation period and model testing period predictive performance levels are each deemed to be a success or failure. In our context, we consider three explicit model build decisions in this order: (a) How to split available model build data between calibration and validation (10 options as shown in Figure 1); (b) How to handle a model failure in the calibration period (two options: discard the model and use reference flows in model testing period or ignore failure and proceed to model validation); (c) How to handle a model failure in the validation period (two options: discard the model and use reference flows in model testing period or ignore model failure and use it in model testing period).

Note that we already made another decision that is not considered explicitly in our decision tree analysis. That was the decision about how to define a model failure. Other subjective model build decisions could also fit into a decision tree framework, but our scope is to focus only on the above three.

Combining the above decisions and chance events yields the following sequence of decisions/events during the model building and model testing process defining our decision tree:

1. Split-sample decision
2. Chance calibration outcome (failure or success)
3. Decision on calibration failure handling (conditional on previous decisions and chance events)
4. Chance validation outcome (failure or success, also conditional)
5. Decision on validation failure handling (conditional)
6. Chance testing period outcome (failure or success, also conditional)

With the above introduction, it will be useful to refer to the example decision tree in Figure 3 (ignoring for now the results reported as numbers in the decision tree). A decision tree is generally composed of three types of nodes: decision, chance, and terminal nodes (Kami & Jakubczyk, 2018). Chance nodes are represented as circles and there are at least two chance outcomes (two branches) following each node, with all branches having an assigned conditional probability of occurrence. Chance nodes can be followed by either another chance node or a decision node. Decision nodes are represented as squares, and they can be followed by either another chance node or a decision node. Terminal nodes, denoted as triangles, each represent an outcome associated with the set of decisions/chance events leading to that node (i.e., a specific path from the start of the tree on the left to a terminal node). Key to the analysis is the assignment of an outcome/payoff associated with each terminal state of nature. Given any decision tree structure, the aforementioned chance outcome probabilities and outcomes/payoffs at the end of each path through the tree are required inputs in order to conduct the decision tree analysis to identify the optimal decisions at every decision node.

The set of experiments detailed in Figure 1 is post-processed to generate most of the decision tree inputs. In addition to finishing all Figure 1 experiments, all the relevant reference KGE values must be computed from observed streamflows as described in Section 2.5.1 above. With all our post-processed results and reference KGEs, we generate 2 models \times 14 model testing periods = 28 decision trees. Each one of these trees can be analyzed to determine the optimal CSP and the optimal calibration and optimal validation failure handling approach.

A decision tree identifies the set of decisions that maximizes the expected value of outcome/payoff. Although the KGE for the model testing period is the natural outcome of interest in our experiment, when we calculate an average KGE across multiple catchments and assign that as the payoff (i.e., to be maximized), that average is subject to extreme negative outliers that can disproportionately impact the average. Therefore, we instead assign the terminal outcomes equal to a simple metric, namely the average *KGE score*. A KGE score for a single catchment is expressed as:

$$KGE_s = \max(KGE_t, KGE_{truncate}) \quad (2)$$

where KGE_s is KGE score, KGE_t is the calculated KGE value in the testing period, and $KGE_{truncate}$ is a truncation threshold, below which KGE values are all regarded as equally bad. Since it is reported that a $KGE = 1 - \sqrt{2} (\approx -0.414)$ means the simulations are equal to using mean annual flow as a predictor (Knoben et al., 2019), we set a more conservative truncate threshold $KGE_{truncate} = -1.0$ in this study, thereby treating KGE values smaller than -1.0 as all equally poor. This value of -1.0 makes KGE scores symmetric around 0 and eliminates the large impact of outliers on the decision tree analysis. In this study, model testing KGE values that are truncated by this threshold account for about 5% in total, which is a minor part to the results.

Given a full set of payoffs and probabilities in our decision tree, the optimal decisions are the ones that maximize the expected value of the KGE score. These optimal decisions, or the optimal path through the tree, are determined via *rollback* calculations that start at the terminal nodes and move backwards through the tree. At a chance node, the expected KGE score is a simple expected value calculation using the payoff values (average KGE scores) at the end of each branch and the probabilities on each branch. At any decision node, the optimal decision is the one that has a maximum expected KGE score. The resulting optimal decisions regarding model failure handling and CSP selection can then serve as a guide for future modelers whose objective when building their model is solely on maximizing the expected performance in some future model application period. In results Section 3.2, we detail a few example rollback calculations, the results of which are encapsulated in Figure 3.

2.5.4. Multi-Objective CSP Assessment Considering Median KGE and Classification Accuracy in Testing Period

In this next assessment approach, we consider two aspects of model testing period CSP performance as model building objectives that we would like to simultaneously optimize. Objective one is to maximize the testing period median KGE, and objective two is to maximize the frequency that CSP performance in calibration and validation correctly classifies or predicts model testing period failure and success. This is a multi-objective decision-making problem to choose the best CSP and with our results from the experimental design described in Figure 1, there are 2 models \times 14 model testing periods = 28 decision different cases where we can evaluate CSP efficacy this way.

In this multi-objective assessment, model failures are never ignored. A failure in model calibration or a failure in model validation both trigger a decision to deem the model inadequate and replace it with reference flows that apply in the model testing period. Note that unlike our approach with decision trees, we used the median (across 463 catchments) to measure the central tendency of model testing period KGE results. The median is simpler than the KGE scores we use for our decision tree analyses and stable in the presence of outliers.

Our second objective requires that the model building process is framed as a classification problem which can be assessed using a confusion matrix. A confusion matrix is a classic way to assess the results of a classifier against known states of nature (Fawcett, 2006). Here, we view the model calibration plus model validation process (i.e., a CSP choice) as a binary classifier, indicating the calibrated model is either adequate or inadequate (success or failure) for testing period application. For the testing period, we can actually assess if the calibrated model is truly adequate or truly inadequate. While a confusion matrix has been used before for matching spatial patterns in flooding (Hosseiny et al., 2020) and to help identify the appropriate model complexity level (Schöninger et al., 2015), to the best of our knowledge, this is the first time a confusion matrix analysis has been used to assess alternative split-sample decisions for their ability to classify adequate versus inadequate calibrated models.

As a classification problem, we follow the confusion matrix convention to define “positive” as the not-normal class, which is a model failure, while “negative” represents normality, which is a model success in this study. Model testing results for a CSP (or more generally, an SST) are then classified into four categories for a given catchment and calibrated model:

1. True Positive (TP): the CSP is a failure, indicating the model is expected to be inadequate in the testing period and the model testing result is an actual failure, proving the model is inadequate in the testing period. In short, the CSP correctly predicted the model is inadequate for the testing period.
2. True Negative (TN): the CSP is a success, indicating the model is expected to be adequate in the testing period and the model testing result is an actual success, proving the model is adequate in the testing period. In short, the CSP correctly predicted the model is adequate for the testing period.
3. False Negative (FN): the CSP is a success, indicating the model is expected to be adequate in the testing period and the model testing result is an actual failure, proving the model is inadequate in the testing period. In short, the CSP incorrectly predicted the model is adequate for the testing period.
4. False Positive (FP): the CSP is a failure, indicating the model is expected to be inadequate in the testing period and the model testing result is an actual success, proving the model is adequate in the testing period. In short, the CSP incorrectly predicted the model is inadequate for the testing period.

A confusion matrix simply reports the frequency in each category in a two-by-two matrix, and then these quantities can be used to compute various classifier performance indices (e.g., see Fawcett, 2006; Hosseiny et al., 2020). In the context of CSP assessment, accuracy is a single metric of overall classifier performance and the equation is defined as below:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (3)$$

where TP, TN, FN, and FP are the counts of catchments classified into these four categories. The sum of these four counts is 463. For brevity, we utilize accuracy as our only measure of CSP classification performance. This effectively assumes that false positives and false negatives are equally important to avoid.

In this multi-objective decision problem, tradeoffs between median KGE and classification accuracy are assessed to identify which CSPs are preferred. Preferred CSPs are identified based on the Pareto front, which is formed from non-dominated CSP results. For a given model and model build year, a CSP_{x1} is said to dominate another CSP_{x2} if,

1. $\text{CSP}_{x1}(i) \geq \text{CSP}_{x2}(j)$, for all indices $i \in \{\text{median KGE, accuracy}\}$, and
2. $\text{CSP}_{x1}(i) > \text{CSP}_{x2}(j)$, for at least one index $j \in \{\text{median KGE, accuracy}\}$

Dominated CSPs are clearly an inferior choice, and a rational decision-maker would then use subjective value judgments to choose a CSP from among the non-dominated CSPs.

Multi-objective results are aggregated over multiple testing periods in order to report the relative frequency each CSP is non-dominated and the relative frequency each CSP dominates other CSPs. See Section 3.3 for the results.

3. Results

The SST experiments summarized in Figure 1 were conducted as described in Sections 2.1–2.4. Results were generated for two hydrologic models, GR4J and HMETs, with 10 CSPs per model build year and five model build years. Comparative performance of the CSPs is based on model testing period performance and thus, we do not explicitly present or compare calibration or validation performance. All results are derived from various post-processing analyses of 926,000 DDS optimization trials and 129,640 model testing period hydrographs.

3.1. Short-Period CSP Performance: Frequency They Beat the Full-Period CSP Benchmark

Employing the full-period CSPs (CSPs using 100% of the available calibration data) as a benchmark, the frequency a short-period CSP has a better KGE in the model testing period than the full-period CSP is computed. Figure 2 displays these results in the three testing periods for the two different models and show that at the 0.05

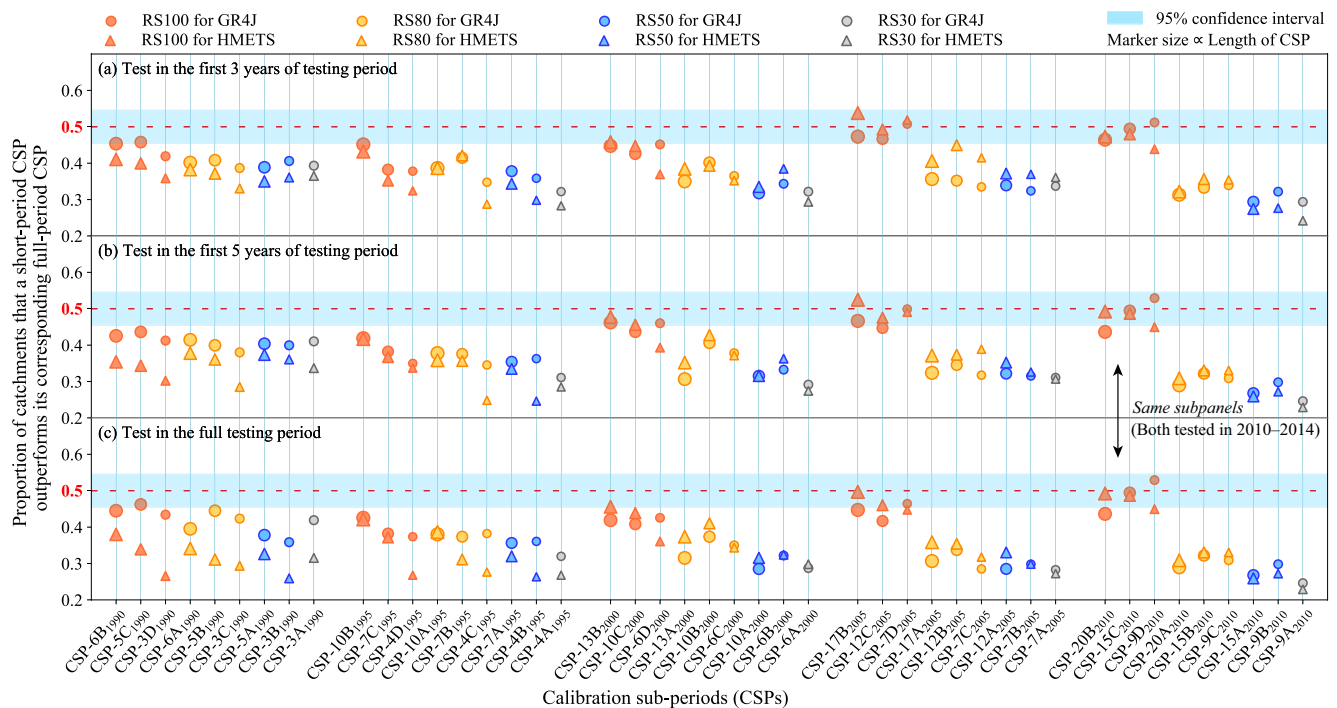


Figure 2. Proportion of 463 catchments that the short-period calibration sub-periods (CSPs) outperform their corresponding full-period CSP in all testing periods. The x -axis is grouped by model build years, then firstly sorted by recency scores (descending order, denoted as different colored markers) and second sorted from long-period to short-period CSPs (descending order, denoted as decreasing marker sizes). Recency score is represented by four different colors, and “RS100” in the legend means a recency score of 100%. The GR4J and Hydrological Model of École de technologie supérieure (HMETs) results are represented by circles and triangles, respectively. The marker sizes are in proportion to the lengths of CSPs. The red solid line is the proportion threshold at 0.5, below which implies that full-period CSPs outperform short-period CSPs in more than half of the catchments. The light blue shaded region ranging from 0.455 to 0.545 indicates the region where proportions are not significantly different than 0.5 (using a 0.05 significance level). Note that the definition of CSP identifiers is provided in Figure 1.

significance level, short-period CSPs are worse than full period CSPs for 88% of the 252 pairwise comparison proportions while 12% of these proportions show no significant difference from 0.5 (short-period CSPs perform as well as the and the full period CSPs). Calibrating to the full period is clearly a very robust strategy, for either model. Figure 2 shows a variety of additional patterns discussed in each paragraph below.

First, all proportions for older CSPs (recency score 30%–80%, represented as yellow, blue and gray markers in Figure 2) of both GR4J and HMETs are smaller than 0.5, ranging from 0.25 to 0.44 (GR4J) and 0.23–0.45 (HMETs). Furthermore, all 168 of these proportions for older CSPs, for both models, are statistically different than 0.5 at the significance level of 0.05 (i.e., below the lower boundary 0.455), thus indicating full-period CSPs always significantly outperform these older CSPs. This is even true for the two older CSPs in the two most recent build years with the longest available data period.

Second, the proportions for recent CSPs (recency score 100% but length of CSP is smaller than 100%, represented as red markers in Figure 2) exhibit a tendency to be conditioned by the calibration data availability and the model type. When there are at least 20 years of available calibration data (i.e., model building in 2000, 2005, and 2010), most of the recent CSPs for the two models are not significantly different from 0.5 at the significance level of 0.05 (i.e., 29 of 48 proportions or 60% of the data points). None of recent CSPs in these model build years significantly outperform the full-period CSPs (i.e., 0 of 48 proportions have values greater than 0.545). However, when the available calibration data period is less than 20 years (i.e., model building in 1990 and 1995), these two percentages for both models change substantially (i.e., 2 of 36 or 6% of proportions are not significantly different than 0.5, whereas the number of full-period CSPs performing significantly better than recent CSPs are now 34 of 36 or 94% of the proportions). Overall, Figure 2 shows that recent CSPs can in some instances perform very comparably to the full-period CSPs, and the best recent CSP choice would appear to be the longest recent CSP that covers the final 70% of the available calibration data period (e.g., CSP-6B₁₉₉₀, CSP-10B₁₉₉₅, etc.), as shorter-length recent CSPs are not as reliable (e.g., CSP-3D₁₉₉₀, CSP 4D₁₉₉₅, etc.). Alternatively, provided the

available calibration data is at least 20 years, calibrating to the final 50% of the available calibration data period (e.g., CSP-10C₂₀₀₀, CSP-12C₂₀₀₅, and CSP-15C₂₀₁₀) with either model generally works as well as calibrating to the full-period with only very limited exceptions.

3.2. Decision Tree Analysis: Optimal Decisions for Model Failure Handling and CSP Selection

Applying reference KGE to discriminate success/failure in model simulations and the decision tree to identify different model building paths, we can make optimal decisions for both model failure handling in calibration and validation and the best CSP selection provided the purpose of model building is to maximize the expected value of the outcome (model testing period performance), as described in Section 2.5.3. Figure 3 shows an example partial decision tree (1 out of 28 decision trees) for GR4J in model build year 2005 and the testing period in 2005–2007, with only three of 10 CSP branches shown. The other seven branches are skipped in Figure 3 for brevity but are taken into account in the analysis.

As introduced in Section 2.5.3, the decision tree is to identify decisions in model building that maximize the expected values of outcomes (expected KGE scores in model testing periods). Here we show an example of performing the decision tree analysis based on Figure 3 configuration. First, following the sequence of calibration, validation and model testing, the best calibrated GR4J model simulations for each single catchment on the three example CSPs are classified into different decision tree branches based on the model success/failure identification, which relies on the reference KGE for each period. Synthesizing the entire suite of 463-catchment results in this step, we obtain the preliminary expected KGE scores for model testing periods, which are the black bold numbers next to the triangles in Figure 3. Note that we report all possible outcomes in this step. For example, when calibrating to CSP-7D₂₀₀₅, 24 catchments are categorized as validation failure when their calibration is identified as success. We then report model testing period outcomes for these 24 catchments in two possible failure handling ways: one is ignoring the validation failure and apply the GR4J to testing periods (denoted as outcomes TS₂ and TF₂); the other one is discarding the model and using reference flow for testing periods (denoted as TA₃).

Second, we identify the optimal decisions on model failures in calibration and validation via a rollback calculation. Taking the above-mentioned example, it is easy to find out if ignoring the validation failure would be the better choice for the 24 catchments by comparing the testing outcomes of the two failure handling approaches. Rolling every terminal node back to the very first chance node of model calibration, we obtain two critical results: the first are the optimal model failure handling decisions when there are failures in calibration or validation (highlighted as bold red branches in Figure 3), and the second is the overall expected KGE score for a CSP based on the optimal model failure handling decisions. For CSP-7D₂₀₀₅, the expected KGE score is 0.454 over the 463-catchment sample analysis. Furthermore, we compare the expected KGE scores for each CSP choice in the decision tree and show that CSP-7D₂₀₀₅, having an expected KGE score 0.454, ranks the best out of the 10 CSPs and hence is the optimal choice in this decision tree. The full-period CSP (CSP-24A₂₀₀₅) ranks the fourth best out of the 10 CSPs with an expected KGE score of 0.429. Also note that there is only one decision node in the full-period CSP sub-tree, since full-period CSPs naturally ignore the validation phase in a model building path. Overall, the decision tree allows a transparent and easy way to interpret how model failures in calibration and validation can be properly handled for maximizing the expected KGE score in model testing period.

To further assess CSP choices regarding the length and recentness, the expected KGE scores of each CSP derived from the 28 decision trees are averaged and reported. Figure 4 displays the heatmaps of the averaged expected KGE score of each CSP for the two models. CSPs are grouped by model build years and aggregated with respect to their lengths and recency scores. Note that boxes for CSPs in model build year 2010 contain only two testing samples, while other CSPs contain three testing samples.

Figure 4 shows that the expected KGE scores vary with model (compare the upper and lower panels in Figure 4) and the HMETs model performs better than GR4J consistently in all CSPs. This indicates that HMETs can be a better model choice than GR4J when the model building goal is to maximize model testing period performance. Moreover, it can be seen that the best (bold values highlighted in Figure 4) out of 10 CSPs in each panel is consistently one of the recent CSPs with recency scores = 100% and in five of 10 cases, the full-period CSP choice is optimal on average. The differences among short-period recent CSPs and the full-period CSP (the right-most

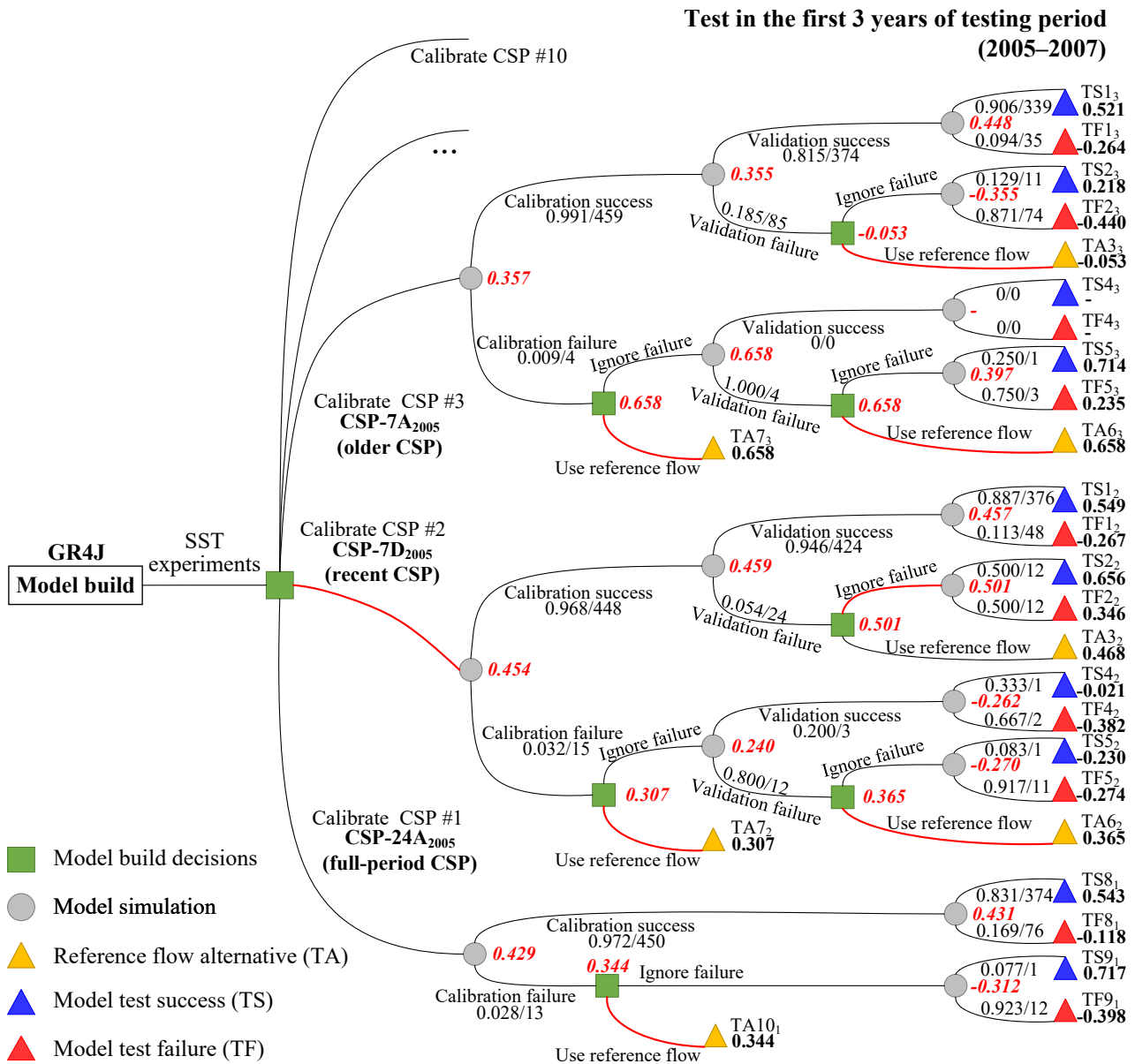


Figure 3. Example of a decision tree for GR4J on three calibration sub-periods (CSPs) (i.e., CSP-24A₂₀₀₅, CSP-7D₂₀₀₅, and CSP-7A₂₀₀₅) tested in the period of 2005–2007 and synthesized from the 463 catchment samples. The green boxes are the decision nodes for making decisions on CSPs and calibration/validation failure handling. The gray circles are the chance nodes for model calibration/validation/testing outcomes. All the three different colored triangles are the terminal nodes for all possible model building paths based on different model failure handling approaches (either ignore failures or discard models and use reference flow as an alternative). Yellow triangles indicate model testing results using reference flow as an alternative (outcomes denoted as TA). Blue triangles indicate model testing results that are identified as success (outcomes denoted as TS). Red triangles indicate model testing results that are identified as failure (outcomes denoted as TF). The number (from 1 to 10) that follow “TA,” “TS,” and “TF” is to discriminate outcomes associated with different model building paths. And the subsequent subscript number (from 1 to 3) discriminates the three CSPs in this example. The two black numbers separated by a slash indicate “proportion/number of catchments” identified for each branch. The black bold numbers next to the triangles are expected KGE scores for model testing period in different model building paths. The red italic bold numbers next to the gray circles and green boxes are the expected KGE scores in rollback calculation, which are computed based on the optimal decision on model failure handling. The red bold branches highlight the optimal paths of a model building regarding choice of the CSP and decisions on model failure handling in calibration and validation.

column in each panel in Figure 4) are minor. In contrast, the differences among various recency scores along the x-axis are much more substantial and hence, recency appears to be more important than length of CSP.

The decision tree analysis makes optimal decisions when there is a calibration failure or a validation failure. In most cases, the example decision tree in Figure 3 shows the optimal decision is to discard the model after it fails and instead use the reference flow, however there is one case where the optimal decision is to ignore validation

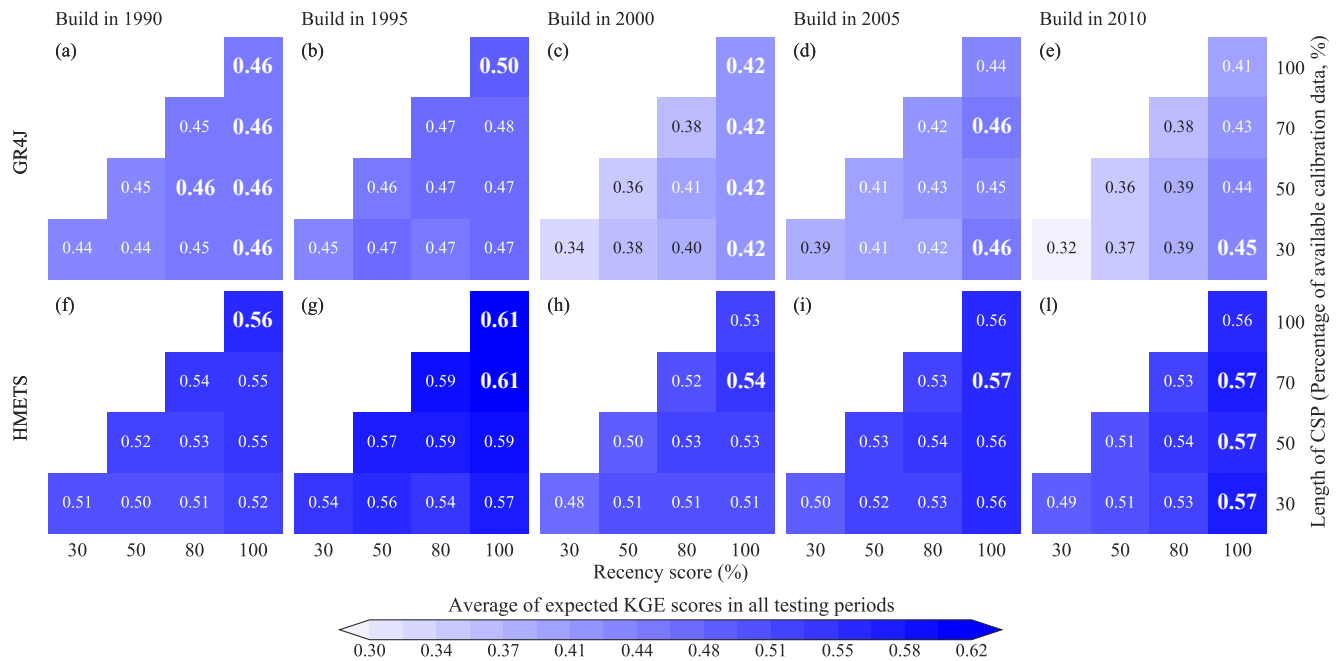


Figure 4. Heatmaps of expected Kling-Gupta efficiency scores of Calibration sub-periods (CSPs) averaged over all testing periods for GR4J and Hydrological Model of École de technologie supérieure (HMETS) based on the decision tree analysis (14 decision trees per model). CSPs are classified into different classes regarding the length of CSP (percentage of available calibration data) and recency score. Each colored box represents the average over all three testing periods, and the largest value (using the averages rounded to two decimal places) in each model build year group is highlighted in larger and bold font.

failure and use the model in the testing period. This optimal decision is not known in practice when building a model and so the results in Figure 4 which optimize failure handling should be considered idealized, eliminating the influence of the failure handling decision. In contrast, the multi-objective analysis in the next section is a more realistic assessment of CSP performance where we fix the model failure handling strategy to use reference flows when calibration or validation is a failure.

3.3. Multi-Objective CSP Assessment: Maximizing Both Median KGE and Accuracy in Testing

Unlike the decision tree analysis that only aims at maximizing expected values of model testing outcomes (Section 3.2), we perform another independent CSP assessment with two model building objectives being optimized simultaneously. The two objectives are maximizing the median KGE and maximizing the classification accuracy, both for the model testing periods. Note that model failures are never ignored in this assessment, meaning any model failures identified in model calibration and/or validation will trigger a decision to discard the model and use reference flow for model testing period instead.

Tradeoffs between the median KGE and accuracy of CSPs are assessed for all testing periods. Figure 5 presents an example tradeoff analysis between the median KGE and accuracy for the first 3 years in testing period. CSPs in Figure 5 are grouped by model build years and represented by different markers regarding their lengths and recency scores in each panel. In this example, note that markers located at the upper-right corner would be the preferred solutions for this multi-objective analysis, called the non-dominated solutions. These non-dominated solutions are highlighted by the Pareto front in Figures 5a and 5i, where in each case there are two non-dominated solutions. In all eight other subplots, there is only one CSP that is non-dominated (meaning it is superior to all others). It can be seen that full-period CSPs (represented as the largest red markers in Figure 5) are identified as non-dominated solutions for 9 out of 10 instances. Such a high frequency is also observed in other testing periods (not repeatedly shown here for brevity). Moreover, when comparing the results of GR4J and HMETS (the upper and the lower panels in Figure 5), it is observed that the HMETS performances are better than GR4J's. In this example, the median KGE of GR4J ranges from 0.51 to 0.63, whereas it ranges from 0.56 to 0.67 for HMETS. The accuracy ranges from 0.73 to 0.90 for GR4J and from 0.79 to 0.96 for HMETS. Similar relative model performance results are also observed in other testing periods (not shown here).

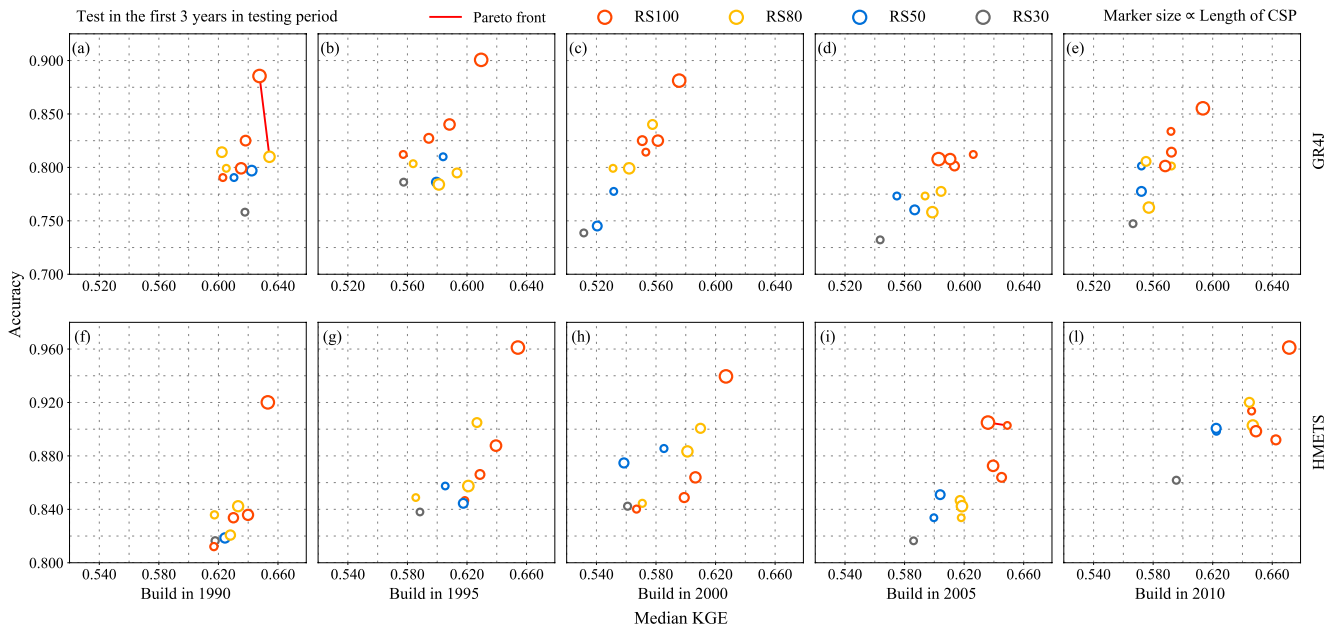


Figure 5. The Pareto solutions in the two-dimensional space regarding median Kling-Gupta efficiency (KGE) and accuracy metric of different Calibration sub-period (CSP) classes in the first 3 years of testing period. The first row of plots are results for GR4J and the second row of plots are for Hydrological Model of École de technologie supérieure (HMETs). The solutions lying in the upper-right of each panel with high values in both median KGE and accuracy metric are dominating solutions in the lower-left of the corresponding panel. The full-period, recent and older CSPs are indicated by red, blue, and gray outlined circles, respectively. The marker sizes are in proportion to the lengths of CSPs. The red solid line indicates the Pareto front, which is the set of all non-dominated solutions. Note that there is no Pareto front drawn in plots (b), (f), and (g) due to the sole non-dominated solution in each of the plots.

The accuracy differences between Pareto front solutions and the dominated solutions in Figure 5 are noteworthy (e.g., they are at least 0.04–0.08 in various subpanels). Considering the error rate ($1 - \text{accuracy}$) instead of accuracy, these differences translate into the dominated CSPs having classification error rates that are often double the error rates achieved by the non-dominated, full-period CSPs. Focusing on the older CSPs (yellow, blue and gray circles), we can see just how much more inferior these results are compared to the non-dominated solutions (accuracies lower than non-dominated solutions by more than 0.10 in a few cases, median KGE values often lower by 0.05 KGE units). Similar differences in error rate magnitudes are observed for the other testing periods (results not shown).

A key multi-objective assessment result is the frequency a CSP is a non-dominated solution in all testing periods (0/3–3/3) for a model build year. Another informative metric across the three testing periods is the frequency of each CSP dominating other CSPs (9 pairwise comparisons \times 3 testing periods = 27 and hence this ranges from 0/27 to 27/27 when models are built in 1990, 1995, 2000, and 2005, while there are only two testing periods when models are built in 2010, hence this ranges from 0/18 to 18/18). Thus, combining the results from Figure 5 (one testing period) with the tradeoff analyses from the other two testing periods (which are not shown individually), we produce Figure 6 which aggregates all these frequencies.

Figure 6 shows that full-period CSPs have a 1.0 frequency (tallest gray bars) of being non-dominated solutions in 9 of 10 subpanels regardless of model type and model build year. In the only other case, the full-period CSP is non-dominated in two of the three testing periods in Figure 6d. In 8 out of 10 panels, referring to the secondary y-axis, full-period CSP instances have the highest frequency of dominating other CSPs (ranging from 0.963 to 1.0, represented as the largest red markers in Figure 6). These results show the two models with full-period CSPs in most model building and application instances are able to simultaneously produce optimal testing results (median KGE in a large catchment sample) and maximize the frequency of correctly classifying model testing period success/failure.

Figure 6 also reveals just how poorly the older CSPs perform. Not a single older CSP was non-dominated in any of the 28 cases of tradeoff analyses. Furthermore, the older CSPs cluster to the right in each panel, in particular

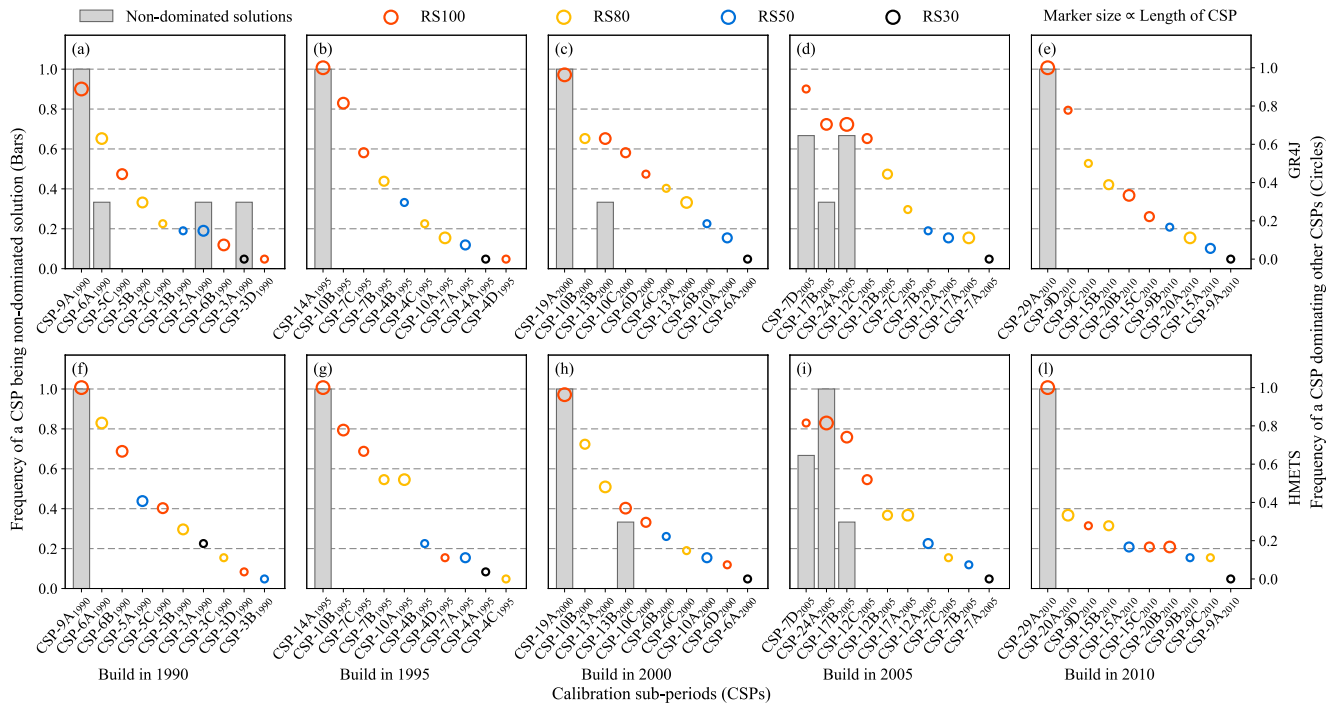


Figure 6. Summary of tradeoff between the median Kling-Gupta efficiency and accuracy over the 463 catchments on all three testing periods. Simulation results with failures in these model building processes are constantly rejected and the reference flow is used as the alternative. The first row of subplots are results for the GR4J model and the second row of subplots are for the Hydrological Model of École de technologie supérieure (HMETs) model. Gray bars indicate the relative frequency of each Calibration sub-periods (CSP) being non-dominated solutions (out of three testing periods), and the best value is 1.0. The circles show the relative frequency of each CSP dominating other CSPs (out of the total pairwise comparisons, which is 18 for build year 2010 and 27 for other build years) with the same model build year, and the best value is 1.0. The x-axis is sorted by the values corresponding to the secondary y-axis from the largest to the smallest. The full-period, recent and older CSPs are indicated by red, blue and gray circles, respectively. The marker sizes are in proportion to the lengths of CSPs. Note that the definition of CSP identifiers is provided in Figure 1.

for the CSPs with a recency score of 50% or 30% meaning the frequency they dominate another CSP are very low (typical around 0.2).

4. Discussion

This study presents results for a massive split-sample testing experiment for hydrological modeling across a large-sample of catchments. We analyzed the results of 926,000 model calibration experiments and 129,640 post-validation model testing instances generated using two hydrological models applied in 463 catchments across the CONUS in Section 3. We believe this to be the most extensive split-sample testing assessment completed to date considering the large sample size and the fact that unlike most split-sample or validation strategy studies, we also have independent model testing periods in addition to calibration and validation periods.

4.1. Guidance for Split-Sample Decision-Making and Implications for Modelers

Our exhaustive experimental design focused on considering only continuous calibration periods with the validation period at either the start or end (or both) of the calibration period, because our literature review revealed this to be common practice for spatially distributed hydrological models. Furthermore, our assessment was from the perspective of a modeler seeking a deterministic calibration result (i.e., a single parameter set). The literature also reveals this deterministic perspective to be common for spatially distributed model calibration. Thus, our results are conditional on the above assumptions.

SST Recommendation #1: Calibrating models to older data and then validating models on newer data produces inferior model testing period performance in every single analysis conducted and should be avoided.

Results in Figure 2 supporting this could not indicate any more clearly that when building a model to predict future streamflows for some 3-year to 25-year period into the future, the full-period CSP is superior to any older CSP that does not contain the most recent available data. We tried all combinations of using 2 models \times 6 older CSPs \times 14 different testing period configurations, and all of them were worse than the corresponding full-period CSPs. Results were nearly as strong in Figure 4 summarizing the decision trees showing only 1 in 10 instances where a slightly older CSP (2 years older than most recent data, see the definition of CSP-5B₁₉₉₀ in Figure 1) is tied on average with all the most recent CSPs. Results in our multi-objective analysis considering median KGE and classification accuracy also could not have been stronger, as not a single older CSP was non-dominated in any of the 28 cases of tradeoff analyses.

This has substantial implications given the preponderance of past studies who use the newest data to validate their models (as discussed in Section 1; Myers et al., 2021 report 24/25 papers they reviewed followed this practice). Indeed, across our past modeling studies initiated prior to discovering our results presented here, when validating models with a continuous calibration subperiod, we have done the opposite of recommendation #1 and followed our calibration period with a validation period. An important and well utilized model benchmarking paper for CAMELS catchments by Newman et al. (2017) is consistent with the above recommendation, in which the distributed model VIC was calibrated in the newest data period while validated in older data period. The challenge this recommendation creates is how to handle initial conditions at the start of calibration period (t_c) and at the start of the validation period ($t_v - \Delta t$, where Δt denotes the time interval between the start of validation and the start of calibration periods). See Section 2.1 for more discussion.

Relating our Recommendation #1 to the literature, studies such as Anctil et al. (2004), Melsen et al. (2014), Perrin et al. (2007) and Xia et al. (2004) suggested using calibration periods covering only 5 months to 8 years (compared to their available calibration data period, ranging from 18 to 39 years) are sufficient for model building. These are very short periods relative to their available calibration data period (<50%). It should be noted that limitations in these studies are similar to those mentioned in the previous paragraph that they performed model calibration on very limited sample size (i.e., from 1 to 12 catchments) and none evaluated findings for a post-validation model testing period. In contrast, our results show that using short-period CSPs (e.g., specifically using only 30% or 70% of the data) is not a wise choice in model building and that if a short-period CSP must be used, modelers need to utilize the newest data years in calibration if they want to avoid inferior model testing/application period predictions.

SST Recommendation #2: Calibrating models to the full available data period and skipping temporal model validation entirely is the most robust choice and eliminates additional subjective decisions.

Given Recommendation #1 is to be followed, justifying Recommendation #2 from empirical results only requires focusing on the results in Section 3 for the most recent CSPs with lengths 100%, 70%, 50%, and 30%. In Figure 2, 87% (88% for GR4J and 86% for HMETs) of short-period recent CSPs (with lengths 70%, 50%, and 30%) are significantly ($\alpha = 0.05$) worse than the full-period CSPs, while none of those short-period CSPs are significantly better than the full-period ones. Thus, there is a very strong advantage to the 100% CSP in Figure 2 results over other CSPs. For Figure 4, counting the bold optimal KGE scores, there is a very slight advantage for the 70% CSP (count = 7 in 10) over the 100% CSP (count = 5 in 10), while the 50% and 30% CSPs are no better than the 100% CSP. Thus, there is a very slight advantage to the 70% CSP in Figure 4 results over the 100% CSP. Fortunately, our most robust and multi-objective assessment in Figure 6 (focused on median KGE and failure/success classification accuracy) shows the 100% CSP is vastly preferred over the others. For example, the 100% CSP is non-dominated in 27/28 tradeoff analyses while all other CSPs are non-dominated in only two or fewer tradeoff analyses. Therefore, based on the overall empirical results, the 100% CSP is recommended.

Recommendation #2 is also justified because following it eliminates two subjective decisions facing modelers who otherwise would have validated their model. First of all, calibrating to all data means there is no decision to make about which data to assign to calibration (e.g., results above make it unclear if a most recent CSP that covers a length of 60%, 70% or more would be preferred). Second of all, calibrating to all data obviates the need to deal with the inconvenient initial condition problem discussed regarding Recommendation #1.

Relating our Recommendation #2 to the literature, studies such as Arsenault et al. (2018); Guo et al. (2018) and Singh and Bárdossy (2012) report that calibrating hydrological models on the full data period generally yields robust model performance. However, discontinuous calibration periods are utilized in the modeling experiment performed by Arsenault et al. (2018) and Singh and Bárdossy (2012), which, as discussed in Section 1, is out of the scope of this study and less common than the continuous CSPs that we utilize. Also, Guo et al. (2018) and Singh and Bárdossy (2012) only perform a two-period assessment in their split-sample model building studies, that is, model calibration and validation only, without any independent model testing periods. Most importantly, all three studies utilize a very small sample size (i.e., three catchments or less), leaving their findings very case-specific and thus, their conclusions are not generalizable. Considering our sample size of 463 catchments and the other key features in our experimental design for SST assessment, our finding regarding the efficacy of calibrating to all available data is very robust and quite generalizable.

4.2. Study Limitations and Future Work

Our SST assessment framework is designed to evaluate model performance in the years immediately following when a model is built (i.e., following both the calibration and validation periods). This design exactly matches the operational hydrological model development context (e.g., streamflow forecasting), and it also works well in the context of various water management studies evaluating near-term changes to the watershed. More generally, our study identifies optimal continuous calibration period split-sample decisions relevant for those who want to build their models to predict overall historical period system behavior with the intention to apply (extrapolate) these models to an independent time period (e.g., a future period). Therefore, our recommendations can also conditionally apply in the context of model building for the purpose of climate change impact assessment. Such example climate change studies fitting their models to all their baseline (historical) period data (or to a continuous subset of historical data thought to be representative of the entire historical period) include Poulin et al. (2011), Schnorbus and Cannon (2014), and Tarek et al. (2020).

However, our recommendations do not apply to climate change impact assessment studies focused on carefully assessing and ensuring parameter transferability under contrasting climates. Such studies require models to be calibrated and validated in climatically contrasting sub-periods (e.g., either dry or wet), thereby evaluating how contrasting climatology impacts hydrological model performance (Bérubé et al., 2022; Coron et al., 2012; Dakhlaoui et al., 2017, 2019; Fowler et al., 2016). Model building in these studies focus on calibrating/validating models in a “specific” climate condition with calibration and validation periods being split by contrasting climates (i.e., the DSST proposed by Klemesš, 1986). Future studies should try to adapt our large sample SST evaluation framework to directly compare how our SST recommendations hold up against the split-sample decision-making approaches commonly employed when contrasting climates are thought to be critical (see a somewhat related effort by Nicolle et al., 2021).

In this study, we aimed at empirically testing alternative choices for selecting a continuous calibration sub-period from a period of available data for model building and hence our focus was on temporal validation only. Future work should also assess if there are any spatial patterns across our large sample of catchments as there may be regions where the results are less (or more) striking. In temporal validation, a special case of the DSST that uses odd years for calibration and even years for validation, or vice versa, is a potentially advantageous approach as demonstrated in Essou et al. (2016) and Xu (2021) that should be investigated within our experimental design in future work. Although such an approach could overcome non-stationarity issues with historical period climate and simultaneously provide the ability to perform validation, this approach would have a computational burden equal to the full-period CSP but would use less information for calibration. We have started work on both of these follow-up investigations.

We only used one model calibration objective function in this work: KGE of daily discharge. The integrated KGE metric, having the same constitutive components to the NSE (Gupta et al., 2009), although now widely employed in the hydrological modeling community, is unable to equally consider the significance of different limbs of a hydrograph. Furthermore, the recent work in Clark et al. (2021) could be used in future work to account for the uncertainty in KGE in our experimental design. Given our reliance on performance across a sample of 463 catchments, we do not believe our findings will be sensitive to accounting for KGE uncertainty. Fundamentally

different additional calibration objective functions can and should be evaluated by our experimental design, such as calibrating models to hydrological signatures (e.g., see Shafii & Tolson, 2015).

Perhaps the most obvious remaining open research question is to try and determine the physical reasons behind our findings. There are a few possible reasons why it is observed that full-period and recent CSPs are the most robust model building decisions (considering testing period performance) and in particular, why older CSPs are inferior. One reason we can eliminate from consideration is problems with model initialization as model initialization was very carefully designed and assessed to show it was appropriate in the context of our study (see details in Section 2.1). We speculate that our finding on data recency being so critical to calibration success, could be due to some combination of: (a) relatively poor quality in older forcing and/or streamflow data; (b) non-stationary climate (e.g., climate variability and/or climate change, even gradual, may result in noticeable differences in recent data compared to those older ones in a long-term accumulation); (c) non-stationary watershed conditions due to anthropogenic influence; and (d) the inherent autocorrelation in streamflows, such that data immediately preceding the testing period are related to the testing period data and so calibrating to the newest data is advantageous. New SST experiments where the oldest data are discarded completely (not used for calibration, validation or model testing) could help answer this question but a more in-depth look at the nature and sources of the forcing and streamflow data, as well as a careful trend analysis of time series and even spatio-temporal data is likely also necessary. While answering this question is important, the answer will not change the fact that our empirical model testing period results show that calibrating to older data and then validating to the newest data is an absolutely inferior strategy if one plans to use models for some purpose in the post-validation time period.

5. Conclusions

In this study, a novel and comprehensive split-sample test (SST) experimental assessment is established and applied to two conceptual hydrological models in 463 catchments across the United States, and the KGE is used as the calibration objective and model testing metric. Novel aspects in our SST assessment framework include defining multiple post-validation model testing periods with a rolling window approach to define model build year, the framing of the way model validation failures are handled, the assessment analysis that views model building decisions as a decision tree, and finally, the assessment analysis framing the calibration-validation exercise as a formal classification problem to bin models as either a success or failure. We evaluated 10 different continuous CSPs for model calibration (varying data period length and recency) across five different model build year scenarios to ensure results are robust across all kinds of testing period conditions. Model performance in testing periods were assessed from three independent aspects: frequency of each short-period CSP being better than its corresponding full-period CSP; central tendency of the objective function metric as computed in model testing period; and frequency that a CSP correctly classifies model testing period failure and success.

Overall, our extensive empirical results evaluating model testing period performance strongly supported two fundamental and generalizable recommendations for modelers facing the common decision about how to split their available data over time in order to define a continuous calibration subperiod. First, calibrating models to older data and then validating models on newer data produces inferior model testing period performance in every single analysis conducted and should be avoided. This is exactly the opposite approach to what is typically done in hydrological modeling studies. Second, calibrating a model to the full available data period and skipping temporal model validation entirely is the most robust choice. We provide, by far, the most convincing empirical evidence to date to support skipping model validation.

Data Availability Statement

The Raven formatted Daymet forcing and the USGS gauge streamflow data as described in Section 2.2 and used for all experiments in this study are available on Zenodo (<https://doi.org/10.5281/zenodo.5915374>). The reference KGE and KGE for each catchment-model combination in calibration, validation, and testing periods used for modeling results analysis in Section 3 are also available on Zenodo (<https://doi.org/10.5281/zenodo.5915374>). The Raven Hydrologic Modeling Framework v3.0.4 used in this study is available at <http://raven.uwaterloo.ca/Downloads.html>. The DDS algorithm and Ostrich software v17.12.19 used in this study are available at <http://www.civil.uwaterloo.ca/envmodeling/Ostrich.html>.

Acknowledgments

The authors are very grateful for the constructive comments provided by the Editor, Dr. Hoshin Gupta, and two anonymous reviewers. The authors would like to acknowledge the primary funding from the Canada First Research Excellence Fund provided to the Global Water Futures (GWF) Project and the Integrated Modeling Program for Canada (IMPC) to H. Shen, and the secondary source of support from Dr. Tolson's Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery for the first author as well. The work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET; www.sharcnet.ca) and Compute/Calcul Canada (www.compute.canada.ca).

References

- Addor, N., & Melsen, L. A. (2019). Legacy, rather than adequacy, drives the selection of hydrological models. *Water Resources Research*, 55(1), 378–390. <https://doi.org/10.1029/2018WR022958>
- Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, 21(10), 5293–5313. <https://doi.org/10.5194/hess-21-5293-2017>
- Ajami, H., McCabe, M. F., Evans, J. P., & Stisen, S. (2014). Assessing the impact of model spin-up on surface water-groundwater interactions using an integrated hydrologic model. *Water Resources Research*, 50(3), 2636–2656. <https://doi.org/10.1002/2013wr014258>
- Ancil, F., Perrin, C., & Andréassian, V. (2004). Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models. *Environmental Modelling & Software*, 19(4), 357–368. [https://doi.org/10.1016/S1364-8152\(03\)00135-X](https://doi.org/10.1016/S1364-8152(03)00135-X)
- Arsenault, R., Brissette, F., & Martel, J. L. (2018). The hazards of split-sample validation in hydrological model calibration. *Journal of Hydrology*, 566(September), 346–362. <https://doi.org/10.1016/j.jhydrol.2018.09.027>
- Bai, P., Liu, X., & Xie, J. (2021). Simulating runoff under changing climatic conditions: A comparison of the long short-term memory network with two conceptual hydrologic models. *Journal of Hydrology*, 592(November 2020), 125779. <https://doi.org/10.1016/j.jhydrol.2020.125779>
- Beckers, J., Smerdon, B., & Wilson, M. (2009). *Review of hydrologic models for forest management and climate change applications in British Columbia and Alberta*. Forrex Forum for Research and Extension. Retrieved from http://epe.lac-bac.gc.ca/100/200/300/forrex/forrex_series/FS25.pdf
- Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., et al. (2013). Characterising performance of environmental models. *Environmental Modelling & Software*, 40, 1–20. <https://doi.org/10.1016/j.envsoft.2012.09.011>
- Bérubé, S., Brissette, F., & Arsenault, R. (2022). Optimal hydrological model calibration strategy for climate change impact studies. *Journal of Hydrologic Engineering*, 27(3), 1–13. [https://doi.org/10.1061/\(asce\)he.1943-5584.0002148](https://doi.org/10.1061/(asce)he.1943-5584.0002148)
- Beven, K. (1989). Changing ideas in hydrology—The case of physically-based models. *Journal of Hydrology*, 105(1–2), 157–172. [https://doi.org/10.1016/0022-1694\(89\)90101-7](https://doi.org/10.1016/0022-1694(89)90101-7)
- Beven, K. (2012). *Rainfall-runoff modelling*. Wiley. <https://doi.org/10.1002/9781119951001>
- Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., & Montanari, A. (2012). Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. *Physics and Chemistry of the Earth*, 42(44), 70–76. <https://doi.org/10.1016/j.pce.2011.07.037>
- Blöschl, G., Sivapalan, M., Savenije, H., Wagener, T., & Viglione, A. (2013). *Runoff prediction in ungauged basins: Synthesis across processes, places and scales*. Cambridge University Press.
- Budyko, M. I., Miller, D. H., & Miller, D. H. (1974). *Climate and Life* (Vol. 508). Academic press.
- Chlumsky, R., Mai, J., Craig, J. R., & Tolson, B. A. (2021). Simultaneous calibration of hydrologic model structure and parameters using a blended model. *Water Resources Research*, 57, e2020WR029229. <https://doi.org/10.1029/2020WR029229>
- Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., et al. (2021). The abuse of popular performance metrics in hydrologic modeling. *Water Resources Research*, 57(9), 1–16. <https://doi.org/10.1029/2020WR029001>
- Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., & Hendrickx, F. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resources Research*, 48(5), 1–17. <https://doi.org/10.1029/2011WR011721>
- Cosgrove, B. A., Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., et al. (2003). *Land surface model spin-up behavior in the North American land data assimilation system (NLDAS)*, (Vol. 108). <https://doi.org/10.1029/2002JD003316>
- Craig, J. R. (2020). *Raven user's and developer's manual*. University of Waterloo. Retrieved from <http://raven.uwaterloo.ca/>
- Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., et al. (2020). Flexible watershed simulation with the Raven hydrological modelling framework. *Environmental Modelling & Software*, 129, 104728. <https://doi.org/10.1016/j.envsoft.2020.104728>
- Daggupati, P., Pai, N., Ale, S., Douglas-Mankin, K. R., Zeckoski, R. W., Jeong, J., et al. (2015). A recommended calibration and validation strategy for hydrologic and water quality models. *Transactions of the ASABE*, 58(6), 1705–1719. <https://doi.org/10.13031/trans.58.10712>
- Dakhlou, H., Ruelland, D., & Trambly, Y. (2019). A bootstrap-based differential split-sample test to assess the transferability of conceptual rainfall-runoff models under past and future climate variability. *Journal of Hydrology*, 575(May), 470–486. <https://doi.org/10.1016/j.jhydrol.2019.05.056>
- Dakhlou, H., Ruelland, D., Trambly, Y., & Bargaoui, Z. (2017). Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern Tunisia. *Journal of Hydrology*, 550, 201–217. <https://doi.org/10.1016/j.jhydrol.2017.04.032>
- Dembélé, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., & Schaeffli, B. (2020). Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite data sets. *Water Resources Research*, 56(1), 1–26. <https://doi.org/10.1029/2019WR026085>
- Devia, G. K., Ganasri, B. P., & Dwarakish, G. S. (2015). A review on hydrological models. *Aquatic Procedia*, 4(Icwrcoe), 1001–1007. <https://doi.org/10.1016/j.aqpro.2015.02.126>
- Duan, Q., Sorooshian, S., & Gupta, V. K. (1994). Optimal use of the SCE-UA global optimization method for calibrating watershed models. *Journal of Hydrology*, 158(3–4), 265–284. [https://doi.org/10.1016/0022-1694\(94\)90057-4](https://doi.org/10.1016/0022-1694(94)90057-4)
- Essou, G. R. C. C., Arsenault, R., & Brissette, F. P. (2016). Comparison of climate datasets for lumped hydrological modeling over the continental United States. *Journal of Hydrology*, 537, 334–345. <https://doi.org/10.1016/j.jhydrol.2016.03.063>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Fowler, H. J., Blenkinsop, S., & Tebaldi, C. (2007). Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, 27(12), 1547–1578. <https://doi.org/10.1002/joc.1556>
- Fowler, K., Coxon, G., Freer, J., Peel, M., Wagener, T., Western, A., et al. (2018). Simulating runoff under changing climatic conditions: A framework for model improvement. *Water Resources Research*, 54(12), 9812–9832. <https://doi.org/10.1029/2018WR023989>
- Fowler, K., Peel, M., Western, A., & Zhang, L. (2018). Improved rainfall-runoff calibration for drying climate: Choice of objective function. *Water Resources Research*, 54(5), 3392–3408. <https://doi.org/10.1029/2017WR022466>
- Fowler, K. J. A., Peel, M. C., Western, A. W., Zhang, L., & Peterson, T. J. (2016). Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models. *Water Resources Research*, 52(3), 1820–1846. <https://doi.org/10.1002/2015wr018068>
- Fry, L. M., Gronewold, A. D., Fortin, V., Buan, S., Clites, A. H., Luukkonen, C., et al. (2014). The great lakes runoff intercomparison project phase 1: Lake Michigan (GRIP-M). *Journal of Hydrology*, 519(PD), 3448–3465. <https://doi.org/10.1016/j.jhydrol.2014.07.021>
- Gaborit, É., Fortin, V., Tolson, B., Fry, L., Hunter, T., & Gronewold, A. D. (2017). Great lakes runoff inter-comparison project, phase 2: Lake Ontario (GRIP-O). *Journal of Great Lakes Research*, 43(2), 217–227. <https://doi.org/10.1016/j.jglr.2016.10.004>
- Garrick, M., Cunneane, C., & Nash, J. E. (1978). A criterion of efficiency for rainfall-runoff models. *Journal of Hydrology*, 36(36), 375375–381381. [https://doi.org/10.1016/0022-1694\(78\)90155-5](https://doi.org/10.1016/0022-1694(78)90155-5)

- Guo, D., Johnson, F., & Marshall, L. (2018). Assessing the potential robustness of conceptual rainfall-runoff models under a changing climate. *Water Resources Research*, 54(7), 5030–5049. <https://doi.org/10.1029/2018WR022636>
- Guo, D., Zheng, F., Gupta, H., & Maier, H. R. (2020). On the robustness of conceptual rainfall-runoff models to calibration and evaluation data set splits selection: A large sample investigation. *Water Resources Research*, 56(3), 1–21. <https://doi.org/10.1029/2019WR026752>
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>
- Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, 18(2), 463–477. <https://doi.org/10.5194/hess-18-463-2014>
- Gupta, V. K., & Sorooshian, S. (1985). The relationship between data and the precision of parameter estimates of hydrologic models. *Journal of Hydrology*, 81(1–2), 57–77. [https://doi.org/10.1016/0022-1694\(85\)90167-2](https://doi.org/10.1016/0022-1694(85)90167-2)
- Hosseiny, H., Nazari, F., Smith, V., & Nataraj, C. (2020). A framework for modeling flood depth using a hybrid of hydraulics and machine learning. *Scientific Reports*, 10(1), 1–14. <https://doi.org/10.1038/s41598-020-65232-5>
- Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of predictions in ungauged basins (PUB)—A review. *Hydrological Sciences Journal*, 58(6), 1198–1255. <https://doi.org/10.1080/02626667.2013.803183>
- Kalra, A., Piechota, T. C., Davies, R., & Tootle, G. A. (2008). Changes in US streamflow and western US snowpack. *Journal of Hydrologic Engineering*, 13(3), 156–163. [https://doi.org/10.1061/\(asce\)1084-0699\(2008\)13:3\(156\)](https://doi.org/10.1061/(asce)1084-0699(2008)13:3(156))
- Kami, B., & Jakubczyk, M. (2018). A framework for sensitivity analysis of decision trees. 135–159. <https://doi.org/10.1007/s10100-017-0479-6>
- Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, 31(1), 13–24. <https://doi.org/10.1080/02626668609491024>
- Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., & Woods, R. A. (2020). A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. *Water Resources Research*, 56(9), 1–23. <https://doi.org/10.1029/2019WR025975>
- Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing nash-sutcliffe and kling-gupta efficiency scores. *Hydrology and Earth System Sciences*, 23(10), 4323–4331. <https://doi.org/10.5194/hess-23-4323-2019>
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International joint conference of artificial intelligence*.
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23(12), 5089–5110. <https://doi.org/10.5194/hess-23-5089-2019>
- Lahmers, T. M., Gupta, H., Castro, C. L., Gochis, D. J., Yates, D., Dugger, A., et al. (2019). Enhancing the structure of the WRF-hydro hydrologic model for semiarid environments. *Journal of Hydrometeorology*, 20(4), 691–714. <https://doi.org/10.1175/JHM-D-18-0064.1>
- Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1), 233–241. <https://doi.org/10.1029/1998WR900018>
- Li, C. Z., Zhang, L., Wang, H., Zhang, Y. Q., Yu, F. L., & Yan, D. H. (2012). The transferability of hydrological models under nonstationary climatic conditions. *Hydrology and Earth System Sciences*, 16(4), 1239–1254. <https://doi.org/10.5194/hess-16-1239-2012>
- Lim, Y.-J., Hong, J., & Lee, T.-Y. (2012). Spin-up behavior of soil moisture content over East Asia in a land surface model. *Meteorology and Atmospheric Physics*, 118(3), 151–161. <https://doi.org/10.1007/s00703-012-0212-x>
- Lund, J. R. (1991). Random variables versus uncertain values: Stochastic modeling and design. *Journal of Water Resources Planning and Management*, 117(2), 179–194. [https://doi.org/10.1061/\(asce\)0733-9496\(1991\)117:2\(179\)](https://doi.org/10.1061/(asce)0733-9496(1991)117:2(179))
- Mai, J., Craig, J. R., Tolson, B. A., & Arsenault, R. (2022). The sensitivity of simulated streamflow to individual hydrologic processes across North America. *Nature Communications*, 13(1), 1–11. <https://doi.org/10.1038/s41467-022-28010-7>
- Mai, J., Tolson, B. A., Shen, H., Gaborit, É., Fortin, V., Gasset, N., et al. (2021). The great lakes runoff intercomparison project phase 3: Lake Erie (GRIP-E). *Journal of Hydrologic Engineering*, 26(9), 05021020. [https://doi.org/10.1061/\(asce\)hse.1943-5584.0002097](https://doi.org/10.1061/(asce)hse.1943-5584.0002097)
- Martel, J. L., Demeester, K., Brissette, F., Poulin, A., & Arsenault, R. (2017). HMETs-A simple and efficient hydrology model for teaching hydrological modelling, flow forecasting and climate change impacts. *International Journal of Engineering Education*, 33(4), 1307–1316.
- Mathevet, T., Gupta, H., Perrin, C., Andréassian, V., & Le Moine, N. (2020). Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds. *Journal of Hydrology*, 585(October 2019), 124698. <https://doi.org/10.1016/j.jhydrol.2020.124698>
- Matott, L. S. (2017). *OSTRICH—an optimization software toolkit for research involving computational heuristics documentation and user’s guide*. State University of New York. Retrieved from www.eng.buffalo.edu/~lsmatott/Ostrich/OstrichMain.html
- Melsen, L. A., Teuling, A. J., Torfs, P. J. J. F., Zappa, M., Mizukami, N., Mendoza, P. A., et al. (2019). Subjective modeling decisions can significantly impact the simulation of flood and drought events. *Journal of Hydrology*, 568(September 2017), 1093–1104. <https://doi.org/10.1016/j.jhydrol.2018.11.046>
- Melsen, L. A., Teuling, A. J., Van Berkum, S. W., Torfs, P., & Uijlenhoet, R. (2014). Catchments as simple dynamical systems: A case study on methods and data requirements for parameter identification. *Water Resources Research*, 50(7), 5577–5596. <https://doi.org/10.1002/2013wr014720>
- Mishra, A. K., & Singh, V. P. (2011). Drought modeling—a review. *Journal of Hydrology*, 403(1–2), 157–175. <https://doi.org/10.1016/j.jhydrol.2011.03.049>
- Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3), 885–900. <https://doi.org/10.13031/2013.23153>
- Moriasi, D. N., Gitau, M. W., Pai, N., & Daggupati, P. (2015). Hydrologic and water quality models: Performance measures and evaluation criteria. *Transactions of the ASABE*, 58(6), 1763–1785. <https://doi.org/10.13031/trans.58.10715>
- Myers, D. T., Ficklin, D. L., Robeson, S. M., Neupane, R. P., Botero-Acosta, A., & Avellaneda, P. M. (2021). Choosing an arbitrary calibration period for hydrologic models: How much does it influence water balance simulations? *Hydrological Processes*, 35(2), 1–17. <https://doi.org/10.1002/hyp.14045>
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—a discussion of principles. *Journal of Hydrology*, 10(3), 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
- Nefeslioglu, H. A., Sezer, E., Gokceoglu, C., Bozkir, A. S., & Duman, T. Y. (2010). Assessment of landslide susceptibility by decision trees in the metropolitan area of istanbul, Turkey. *Mathematical Problems in Engineering*, 2010, 1–15. <https://doi.org/10.1155/2010/901095>
- Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydro-meteorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, 19(1), 209–223. <https://doi.org/10.5194/hess-19-209-2015>
- Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., & Nearing, G. (2017). Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, 18(8), 2215–2225. <https://doi.org/10.1175/jhm-d-16-0284.1>

- Nicollé, P., Andréassian, V., Gaspard, P. R., Perrin, C., Thirel, G., Coron, L., & Santos, L. (2021). *Technical note—RAT: A robustness assessment test for calibrated and uncalibrated hydrological models* key words key points 1 introduction (pp. 1–22).
- Oudin, L., Salavati, B., Furusho-Percot, C., Ribstein, P., & Saadi, M. (2018). Hydrological impacts of urbanization at the catchment scale. *Journal of Hydrology*, 559, 774–786. <https://doi.org/10.1016/j.jhydrol.2018.02.064>
- Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, 279(1–4), 275–289. [https://doi.org/10.1016/S0022-1694\(03\)00225-7](https://doi.org/10.1016/S0022-1694(03)00225-7)
- Perrin, C., Oudin, L., Andréassian, V., Rojas-Serna, C., Michel, C., & Mathevet, T. (2007). Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models. *Hydrological Sciences Journal*, 52(1), 131–151. <https://doi.org/10.1623/hysj.52.1.131>
- Poncelet, C., Merz, R., Merz, B., Parajka, J., Oudin, L., Andréassian, V., & Perrin, C. (2017). Process-based interpretation of conceptual hydrological model performance using a multinational catchment set. *Water Resources Research*, 53(8), 7247–7268. <https://doi.org/10.1002/2016WR019991>
- Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: Towards a non-parametric variant of the kling-gupta efficiency. *Hydrological Sciences Journal*, 63(13–14), 1941–1953. <https://doi.org/10.1080/02626667.2018.1552002>
- Poulin, A., Brissette, F., Leconte, R., Arsenault, R., & Malo, J. S. (2011). Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin. *Journal of Hydrology*, 409(3–4), 626–636. <https://doi.org/10.1016/j.jhydrol.2011.08.057>
- Rakovec, O., Mizukami, N., Kumar, R., Newman, A. J., Thober, S., Wood, A. W., et al. (2019). Diagnostic evaluation of large-domain hydrologic models calibrated across the contiguous United States. *Journal of Geophysical Research: Atmospheres*, 124(24), 13991–14007. <https://doi.org/10.1029/2019JD030767>
- Ray, P. A., Taner, M. Ü., Schlef, K. E., Wi, S., Khan, H. F., Freeman, S. S. G., & Brown, C. M. (2019). Growth of the decision tree: Advances in bottom-up climate change risk management. *Journal of the American Water Resources Association*, 55(4), 920–937. <https://doi.org/10.1111/1752-1688.12701>
- Refsgaard, J. C., & Henriksen, H. J. (2004). Modelling guidelines - terminology and guiding principles. *Advances in Water Resources*, 27(1), 71–82. <https://doi.org/10.1016/j.advwatres.2003.08.006>
- Ritter, A., & Muñoz-Carpena, R. (2013). Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology*, 480, 33–45. <https://doi.org/10.1016/j.jhydrol.2012.12.004>
- Savenije, H. H. G. (2009). HESS opinions: “The art of hydrology”. *Hydrology and Earth System Sciences*, 13(2), 157–161. <https://doi.org/10.5194/hess-13-157-2009>
- Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes: An International Journal*, 21(15), 2075–2080.
- Schlef, K. E., François, B., & Brown, C. (2021). Comparing flood projection approaches across hydro-climatologically diverse United States river basins. *Water Resources Research*, 57(1), 1–21. <https://doi.org/10.1029/2019wr025861>
- Schnorbus, M. A., & Cannon, A. J. (2014). Statistical emulation of streamflow projections from a distributed hydrological model: Application to CMIP3 and CMIP5 climate projections for British Columbia, Canada. *Water Resources Research*, 50(11), 8907–8926. <https://doi.org/10.1002/2014wr015279>
- Schöniger, A., Illman, W. A., Wöhling, T., & Nowak, W. (2015). Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. *Journal of Hydrology*, 531, 96–110.
- Seck, A., Welty, C., & Maxwell, R. M. (2015). Spin-up behavior and effects of initial conditions for an integrated hydrologic model. *Water Resources Research*, 51(4), 2188–2210. <https://doi.org/10.1002/2014wr016371>
- Shafii, M., & Tolson, B. A. (2015). Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resources Research*, 51(5), 3796–3814. <https://doi.org/10.1002/2014wr016520>
- Sharma, S., Siddique, R., Reed, S., Ahnert, P., & Mejia, A. (2019). Hydrological model diversity enhances streamflow forecast skill at short-to medium-range timescales. *Water Resources Research*, 55(2), 1510–1530. <https://doi.org/10.1029/2018WR023197>
- Shen, M., Chen, J., Zhuan, M., Chen, H., Xu, C.-Y., & Xiong, L. (2018). Estimating uncertainty and its temporal variation related to global climate models in quantifying climate change impacts on hydrology. *Journal of Hydrology*, 556, 10–24. <https://doi.org/10.1016/j.jhydrol.2017.11.004>
- Singh, S. K., & Bárdossy, A. (2012). Calibration of hydrological models on hydrologically unusual events. *Advances in Water Resources*, 38, 81–91. <https://doi.org/10.1016/j.advwatres.2011.12.006>
- Singh, V. P., & Chow, V. T. (2016). *Handbook of applied hydrology* (2nd ed.). McGraw-Hill Education.
- Singh, V. P., & Woolhiser, D. A. (2003). Mathematical modeling of watershed hydrology. *Perspectives in Civil Engineering: Commemorating the 150th Anniversary of the American Society of Civil Engineers*, 7(4), 345–367. [https://doi.org/10.1061/\(asce\)1084-0699](https://doi.org/10.1061/(asce)1084-0699)
- Smith, M. B., Koren, V., Zhang, Z., Zhang, Y., Reed, S. M., Cui, Z., et al. (2012). Results of the DMIP 2 Oklahoma experiments. *Journal of Hydrology*, 418–419, 17–48. <https://doi.org/10.1016/j.jhydrol.2011.08.056>
- Smith, M. B., Seo, D. J., Koren, V. I., Reed, S. M., Zhang, Z., Duan, Q., et al. (2004). The distributed model intercomparison project (DMIP): Motivation and experiment design. *Journal of Hydrology*, 298(1–4), 4–26. <https://doi.org/10.1016/j.jhydrol.2004.03.040>
- Sorooshian, S., Gupta, V. K., & Fulton, J. L. (1983). Evaluation of Maximum Likelihood Parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility. *Water Resources Research*, 19, 251–259. <https://doi.org/10.1029/WR019i001p00251>
- Spieler, D., Mai, J., Craig, J. R., Tolson, B. A., & Schütze, N. (2020). Automatic model structure identification for conceptual hydrologic models. *Water Resources Research*, 56(9), e2019WR027009. <https://doi.org/10.1029/2019WR027009>
- Tarek, M., Brissette, F. P., & Arsenault, R. (2020). Large-scale analysis of global gridded precipitation and temperature datasets for climate change impact studies. *Journal of Hydrometeorology*, 21(11), 2623–2640. <https://doi.org/10.1175/JHM-D-20-0100.1>
- Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, 43(1), 1–16. <https://doi.org/10.1029/2005WR004723>
- Valéry, A. (2010). *Modélisation précipitations débit sous influence nivale: Elaboration d'un module neige et évaluation sur 380 bassins versants*. Doctoral dissertation, Doctorat Hydrobiologie, Institut des Sciences et Industries du Vivant et de l'Environnement AgroParisTech.
- Valéry, A., Andréassian, V., & Perrin, C. (2014). “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>
- Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J.-M., Viney, N. R., & Teng, J. (2010). Climate non-stationarity—validity of calibrated rainfall-runoff models for use in climate change studies. *Journal of Hydrology*, 394(3–4), 447–457. <https://doi.org/10.1016/j.jhydrol.2010.09.018>
- Xia, Y., Yang, Z. L., Jackson, C., Stoffa, P. L., & Sen, M. K. (2004). Impacts of data length on optimal parameter and uncertainty estimation of a land surface model. *Journal of Geophysical Research-D: Atmospheres*, 109(7), 1–13. <https://doi.org/10.1029/2003JD004419>
- Xu, C. (2021). Issues influencing accuracy of hydrological modeling in a changing environment. *Water Science and Engineering*, 14(2), 167–170. <https://doi.org/10.1016/j.wse.2021.06.005>

- Yang, Y., Pan, M., Beck, H. E., Fisher, C. K., Beighley, R. E., Kao, S. C., et al. (2019). In quest of calibration density and consistency in hydrologic modeling: Distributed parameter calibration against streamflow characteristics. *Water Resources Research*, 55(9), 7784–7803. <https://doi.org/10.1029/2018WR024178>
- Yapo, P. O., Gupta, H. V., & Sorooshian, S. (1996). Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. *Journal of Hydrology*, 181(1–4), 23–48. [https://doi.org/10.1016/0022-1694\(95\)02918-4](https://doi.org/10.1016/0022-1694(95)02918-4)
- Zheng, F., Maier, H. R., Wu, W., Dandy, G. C., Gupta, H. V., & Zhang, T. (2018). On lack of robustness in hydrological model development due to absence of guidelines for selecting calibration and evaluation data: Demonstration for data-driven models. *Water Resources Research*, 54(2), 1013–1030. <https://doi.org/10.1002/2017WR021470>

Reference From the Supporting Information

- Mai, J., Craig, J. R., & Tolson, B. A. (2020). Simultaneously determining global sensitivities of model parameters and model structure. *Hydrology and Earth System Sciences*, 24(12), 5835–5858.