Rethinking the Split-Sample Approach in Hydrological Model Calibration

by

Hongren Shen

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Civil Engineering (Water)

Waterloo, Ontario, Canada, 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Prasad Daggupati Associate Professor, School of Engineering, University of Guelph
Supervisor	Bryan A. Tolson Professor, Department of Civil and Environmental Engineering, University of Waterloo
Internal Member	James R. Craig Associate Professor, Department of Civil and Environmental Engineering, University of Waterloo
Internal Member	Bruce MacVicar Associate Professor, Department of Civil and Environmental Engineering, University of Waterloo
Internal-External Member	David L. Rudolph Professor, Department of Earth and Environmental Sciences, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Hongren Shen was the sole author of Chapters 1, 2, 4, 5, and 6 of this thesis, which were under the supervision of Dr. Bryan A. Tolson.

Chapter 3 of this thesis consists of a published paper by Shen et al. (2022), which was co-authored with Dr. Bryan A. Tolson and Dr. Juliane Mai. Hongren Shen's contribution to this paper includes conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, visualization, original draft, and review and editing. Dr. Bryan A. Tolson's contribution to this paper includes conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, supervision, review and editing. Dr. Juliane Mai's contribution to this paper includes data curation, formal analysis, software, visualization, and review and editing.

Shen, H., Tolson, B. A., & Mai, J. (2022). Time to Update the Split-Sample Approach in Hydrological Model Calibration. *Water Resources Research*, 58(3), 1–26. https://doi.org/10.1029/2021WR031523

Abstract

Hydrological models, which have become increasingly complex in the last half century due to the advances in computing capabilities and data collection, have been extensively utilized to facilitate decision-making in water resources management. Such computer-based models generally contain considerable parameters that cannot be directly measured, and hence calibration and validation are required to ensure model transferability and robustness in model building (development). The most widely used method used for assessing model transferability in time is the split-sample test (SST) framework, which has even been a paradigm in the hydrological modeling community for decades.

However, there is no clear guidance or empirical/numerical evidence that supports how a dataset should be split into the calibration and validation subsets. The SST decisions usually appear to be unclear and even subjective in literature. Even though past studies have spared tremendous efforts to investigate possible ways to improve model performance by adopting various data splitting methods; however, such *problem of data splitting* still remain as a challenge and no consensus has achieved on which splitting method may be optimal in hydrological modeling community. One of the key reasons is lacking a robust evaluation framework to objectively compare different data splitting methods in the "out-of-sample" model application period. To mitigate these gaps, this thesis aims at assessing different data splitting methods using the large-sample hydrology approach to identify optimal data splitting methods under different conditions, as well as exploring alternative validation methods to improve model robustness that is usually done by the SST method.

First, the thesis introduces a unique and comprehensive evaluation framework to compare different data splitting methods. This evaluation framework defines different model build years, as such models can be built in various data availability scenarios. Years after the model build year are retained as model testing period, which acts as an "out-of-sample" data beyond the model building period and matches how models are applied in operational use. The evaluation framework allows to incorporate various data splitting methods into comparison, as the comparison of model performance is performed in the common testing period no matter how calibration and validation data are split in model building period. Moreover, a reference climatology, which is purely observation data-based, is applied to benchmark our model simulations. Model inadequacy is properly handled by considering the possible decisions modelers may make when faced with bad model simulations. As such, the model building can be more robust and realistic. Example approaches which cover a wide range of aspects modelers may care about in practice are provided to assess large-sample modeling results.

Two large-sample modeling experiments are performed in the proposed evaluation framework to compare different data splitting methods. In the first experiment, two conceptual hydrological models are applied in 463 catchments across the United States to evaluate 50 different continuous calibration sub-periods (CSPs) for model calibration (varying data period length and recency) across five different model build year scenarios, which ensures robust results across three testing period conditions. Model performance in testing periods are 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. The second experiment assesses 44 representative continuous and discontinuous data splitting methods using a conceptual hydrological model in 463 catchments across the United States. These data splitting methods consist of all the ways hydrological model calibration split-sampling is currently done when only a single split sample is evaluated and one method found in data-driven modeling. This results in over 0.4 million model calibration-validation and 1.7 million model testing exercises for an extensive analysis. Model performance in testing periods are assessed in similar ways in the first experiment except that all model optimization trials are utilized to draw even more robust conclusions.

Three SST recommendations are made based on the strong empirical evidence. 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. Calibrating a model to the full available data period and skipping temporal model validation entirely is the most robust choice. It is recommended that hydrological modelers rebuild models after their validation experiments, but prior to operational use of the model, by calibrating models to all available data.

Last but not least, alternative model validation methods are further tested to enhance model robustness based on the above large-sample modeling results. A proxy validation is adopted to replace the traditional validation period in the SST method by using Split Kling-Gupta Efficiency (KGE) and Split Reference KGE in calibration to identify unacceptable models. The proxy validation is demonstrated to have some promise to enhance model robustness when all data are used in calibration.

Acknowledgements

I am profoundly grateful to all those who have supported me and contributed to my academic growth during this impressive journey.

I would like to thank Dr. Bryan A. Tolson for his mentorship and support in these years, especially his great support on me during the two-year pandemic when I was physically stuck in China. Dr. Tolson's insight, guidance, patience, constructive critiques, and critical thought have been the driving forces behind the evolution of this work.

I would like to thank my thesis and comprehensive exam committee members, Dr. James R. Craig, Dr. Bruce MacVicar, Dr. David L. Rudolph, Dr. Prasad Daggupati, and Dr. John Quilty, for their expertise, evaluation, and suggestions that significantly elevated the rigor and quality of this research.

I would like to thank my former colleague Dr. Juliane Mai. Her excellent expertise, professional advice, and rigorous attitude toward research deeply influenced me in our collaboration. Dr. Mai made great contributions to my first paper. Her guidance and support on me in our collaboration of GRIP-E and GRIP-GL were hugely significant to my academic journey.

I extend my appreciation to my colleagues and peers in the Hydrology Research Group. The discussion, camaraderie, shared experiences have enriched the boring campus life. I would especially like to thank Dr. Ming Han for guiding me to use, test, and maintain BasinMaker, which has hugely contributed to my research.

The financial support from the Canada First Research Excellence Fund provided to the Global Water Futures (GWF) Project and the Integrated Modeling Program for Canada (IMPC) and the University of Waterloo International Doctoral Student Awards are gratefully acknowledged. Partial financial support during the PhD was provided by the Ontario Ministry of Northern Development, Mines, Natural Resources and Forestry and the Environment and Climate Change Canada G&C Program, which are hereby acknowledged.

This research was enabled in part by support provided by the Digital Research Alliance of Canada (alliancecan.ca). This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada.

To family and friends, your constant belief in me was the biggest source of motivation in these years. I am immensely grateful for the unwavering encouragement, support and love from my parents. I am also thankful for a special family member, my corgi Yuanxin, who witnessed my graduation from master and soon the PhD. He makes everything around me beautiful.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Research objectives and scope

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; Vijay P. 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, climate change assessment, etc. (Beckers et al., 2009; Blöschl et al., 2013; Fowler et al., 2007; Hrachowitz et al., 2013; Mishra & Singh, 2011).

Hydrological modeling generally requires a model building (or development) process for the historical period by using specific model inputs (e.g., meteorological forcings and basin geo-spatial characteristics) and system response data (e.g., observed streamflow at basin outlet) to select appropriate model structures and model parameters (Blöschl et al., 2013; KlemeŠ, 1986; Mai et al., 2020; Singh & Chow, 2016), which include two phases in practice: model calibration and validation. Model calibration is a process for either manually or automatically adjusting influential model parameters over a specific simulation period to obtain model outputs matching the corresponding observations as closely as possible. Model validation in this research is 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).

Among those model calibration challenges listed in Mai (2023), the *problem of data splitting* is the focus of this thesis, which refers to the procedure of choosing appropriate data for model calibration and validation. In hydrological modeling, models are generally calibrated and validated in the framework of the split-sample test (SST) proposed by KlemeŠ (1986). Data splitting methods that have been applied in hydrological modeling may be classified by how calibration data are organized in time, including temporally continuous and discontinuous splitting. Continuous data splitting is the most widely used in hydrological modeling but it varies with which part of the dataset is used in calibration (e.g., see Coron et al., 2012; Guo et al., 2018, 2020; Knoben et al., 2020; Mai, Craig, et al., 2022; Mai, Shen, et al., 2022; Newman et al., 2015, 2017; and Rakovec et al., 2019). Discontinuous

data splitting is less dominant in hydrological modeling literature, but it is claimed to be beneficial as sampling calibration and validation data from the entire dataset (e.g., using odd and even years) may retain important statistical features (e.g., trend) in the subsets (e.g., see Arsenault et al., 2018; Essou et al., 2016; and Zheng et al., 2022). Nevertheless, there is no consensus on the data splitting method that can be applied for all (at least most) SST practice, and it is reported that different data splitting may have substantial influence on model robustness (Coron et al., 2012; Daggupati et al., 2015; Guo et al., 2020; KlemeŠ, 1986; Shen et al., 2022a).

In this thesis, different data splitting methods are assessed in a novel, large-sample hydrology approach to identify optimal data splitting methods under different conditions. Four main objectives of this thesis are (a) introducing a new evaluation framework for a more objective comparison of different data splitting methods; (b) comparing all the ways hydrological model calibration split-sampling is currently done and alternative model validation methods for the split-sample test; (c) explicitly handling model inadequacy issues (e.g., models failing validation) when comparing different data splitting methods in order to achieve more statistically robust conclusions; and (d) providing a comprehensive methodology for assessing massive hydrological modeling results across a large-sample catchments. We limit our scope to addressing the single-site temporal calibration and validation problem. Multi-site model calibration and spatial model validation are not considered in this thesis.

1.2 Key contributions

Key contributions of this thesis include:

1. A unique evaluation framework for comparing different data splitting methods for hydrological modeling. The evaluation framework consists of several unique aspects such as multiple model building scenarios and it enables model tested in a third "out-of-sample" period which mimics how models are applied in operational use.
2. Based on strong, large-sample empirical evidence, three split-sample test (SST) recommendations for future hydrological modelers are as follows:

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.

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.

SST recommendation #3: Modelers should rebuild models after any validation experiments, but prior to operational use of the model, by calibrating models to all available data.

3. An evaluation of an example alternative model validation method that can be applied when SST recommendation #2 is followed and model calibration uses all available data. The alternative validation method is a proxy validation approach considering model performance in individual years and it shows some promise. Our evaluation framework also allows more alternative validation methods to be explored in the future.
4. Several comparative performance metrics and methodologies for assessing large-sample model building and testing results are demonstrated. Examples include pairwise comparison between different SST decisions, central tendency of model performance, ability of SST decisions functioning as binary classifiers to correctly predict model failures, and multi-objective analyses considering different aspects simultaneously. Such methodologies can be further applied in the future SST studies along with the evaluation framework.

1.3 Thesis structure

The chapters of this thesis are organized as follows: Chapter 2 presents literature review about the hydrological modeling, the split-sample test in model calibration and validation, and the large-sample hydrology. Chapter 3 presents an empirical large-sample study on assessing continuous data splitting in model building and testing, which is a mirror of a manuscript published on *Water Resources Research*. Chapter 4 further assesses continuous and discontinuous data splitting, which is based on a prepared manuscript that is expected to be submitted as of the thesis defense date (October 2). Chapter 5 explores alternative validation methods to enhance model robustness when all data are used in calibration, which is based on a manuscript in preparation. Chapter 6 ends this thesis with major findings, limitations and future work.

Chapter 2
Literature Review

This chapter reviews literature about the evolution of hydrological models, the general hydrological model development (building) process in practice, the split-sample test (SST) approach used in hydrological modeling, the model inadequacy issue in hydrological modeling, and the large-sample hydrology approach applied in hydrological modeling. At the end of this chapter, key research gaps identified from literature are provided.

2.1 Overview of hydrological models

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 hydrological 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, climate change assessment, etc. (Beckers et al., 2009; Blöschl et al., 2013; Fowler et al., 2007; Hrachowitz et al., 2013; Mishra & Singh, 2011).

The evolution of hydrological models has a long history and can be traced to nearly 170 years ago when the rational method was developed by Irish engineer Thomas James Mulvaney (Beven, 2012). The rational method relates storm runoff peak to rainfall intensity. This method assumes that a portion of rainfall was considered to account for abstractions not contributing to runoff. It also assumes that rainfall is of constant intensity and it is uniformly distributed over the watershed, thus making it feasible only for small watersheds. The introduction of the unit hydrograph (UH) by Sherman (1932) provided a method to translate effective rainfall into runoff rates over time (e.g., a hydrograph). In the UH method, methods for estimating the effective rainfall, from total rainfall, are required. Horton (1933) developed a theory of infiltration, which helped estimate how much of the rainfall became runoff and thus better predict the shape of the hydrograph.

For determining the total amount of runoff due to a single rainfall event, the Soil Conservation Service (SCS) of the U.S. Department of Agriculture developed the SCS curve number method in 1956 (SCS, 1956). This method accounts for the effects of soil types, antecedent condition, and land use, and has been widely used on agricultural, forested, urban, and other types of watersheds. This method

assumes the ratio of the actual runoff to the potential runoff is equal to the ratio of the actual retention to the potential retention. Nash (1958) derived the theory of the instantaneous UH (IUH) by using a cascade of n linear reservoirs each of lag time k to represent a watershed. Dooge (1959) developed the generalized UH theory of which most conceptual IUH models are special cases.

In the 1960s, with the creation of the digital computer, the first generation of hydrological models that was able to simulate the entire watershed response over time was developed by Crawford and Linsley (1966), known as the Stanford Watershed Model (SWM). This model later evolved into the Hydrological Simulation Program FORTRAN (HSPF) (Bicknell et al., 1997) and was widely used in hydrological consulting. Since then, many models have been developed and improved and are still in current use, including the HSPF model, Sacramento models from the USA (Sorooshian et al., 1993), the Xinanjiang (XAJ) model from China (Zhao, 1992), the HBV model from Sweden (Lindström et al., 1997), the Tank model from Japan (Sugawara, 1974), the UBC model from Canada (Quick & Pipes, 1977), and the AWBM model from Australia (Boughton, 2004). These models were called explicit soil moisture accounting (ESMA) models by O'Connell (2012), as they all employ a collection of storage (bucket) units to represent different hydrological processes and describe fluxes between or within those storage units.

In the late 1960s, Freeze and Harlan (1969) presented a blueprint for developing physically based models that are directly based on equations describing all the surface and subsurface flow processes in the catchment. This kind of model is called a distributed hydrological model, which defines parameter values in each modeling element. Distributed hydrological models have been widely used in recent decades (Abbott & Refsgaard, 2012; Devia et al., 2015; Smith et al., 2004, 2012), because these models can formulate important hydrological processes and assess the impact of climate and land cover changes on hydrological response. Moreover, the increase in computer power makes it easier for programming and large-scale computing. Examples of distributed hydrological models include the Systeme Hydrologique Europeen (SHE) model (Abbott et al., 1986), which later evolved into MIKE SHE and SHETRA versions, and the Institute of Hydrology Distributed Model (IHDM) in the UK (K Beven et al., 1987), the THALES model in the Australia (Grayson, 1996), the Gridded Surface/Subsurface Hydrological Analysis (GSSHA) model in the USA (Downer & Ogden, 2006), and the earth system simulator HydroGeoSphere in Canada (Brunner & Simmons, 2012). These models were essentially based on the blueprint proposed by Freeze and Harlan (1969) or on simplifications of it. There is also a branch of models that attempt to maintain the distributed description of catchment response by applying a form of distribution function to represent the spatial variability of runoff generation. The distribution may be based on a statistical description in the Probability Distributed

Moisture (PDM) model, a simple function form in the XAJ model and Variable Infiltration Capacity (VIC) model (Gao et al., 2009), GIS-derived hydrological response units, and some simplified physically based index such as the topography-based TOPMODEL (Keith Beven, 1997). These models have the advantages of describing the non-linearity of runoff generation process using the simplified distribution function and avoid introducing too many parameters compared to the fully physically based distributed models.

However, it should be noted that the explicit soil moisture accounting models or distributed models described above can only be valid for specific spatio-temporal scales. Considerable uncertainties may be associated with their watershed discretization, model structures and parameters, given their fixed choices of process and these process algorithms are usually only validated against data from tightly controlled field or laboratory scale experiments before applied in the real world (e.g., at the basin scale) (Craig et al., 2020). Thus, it is recognized that a solution to cope with the above issues is to develop flexible hydrological modeling frameworks in recent decades. Flexible models or modular-based models support building a wide range of models with different model complexities. Flexible models also support place-based customizations in the same model by choosing different processes representations on different climates and landscapes. Model complexity in such flexible modeling frameworks can be easily adjusted to achieve higher predictive skill and better system understanding (Craig et al., 2020). Various flexible modeling frameworks have been developed in recent decades, such as the Precipitation-Runoff Modeling System (PRMS; Leavesley & Stannard, 1995; Markstrom et al., 2008), the cold regions hydrological model (CRHM) platform (Pomeroy et al., 2007), the Imperial College Rainfall-Runoff Modeling Toolbox (RRMT; Wheater et al., 2008), the Framework for Understanding Structural Errors (FUSE; Clark et al., 2008), the SUMMA land surface model (Clark et al., 2015), the Modular Assessment of Rainfall-Runoff Models Toolbox (Knoben, Freer, Fowler, et al., 2019), and the Raven hydrological modeling framework (Craig et al., 2020).

These flexible modeling frameworks generally provide a standardized platform for users to choose, customize and/or build new models, even though their specific functionalities may differ with one another. Taking the Raven hydrological modeling framework as an example, it is a highly generalized object-oriented and open-source hydrological modeling framework, which encapsulates more than 100 compatible process algorithms to emulate approaches widely used in hydrological models. Raven 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 and uncertainty assessment

(Chlumsky et al., 2021; Juliane Mai, Craig, et al., 2022), model inter-comparison (Ahmed et al., 2023; Mai et al., 2021; Mai, Shen, et al., 2022), hypotheses tests about different aspects in hydrological modeling (Shen et al., 2022a; Taheri et al., 2023). Also, Raven conveniently unifies the format for both models' input and output files, as such it is efficient to use benchmark datasets for different Raven-configured model structures. Moreover, additional process algorithm modules are easy and straightforward to be added into the Raven framework without incompatibility issues.

2.2 Overview of the deterministic hydrological model building process

Hydrologic modeling entails simulating state variables and water fluxes for various hydrological compartments. Hydrological modeling typically refers to continuous simulation (as opposed to event-based simulation), which contains algorithms to maintain a continuous water balance for the catchment, as such the antecedent conditions for each rainfall-runoff event being simulated are estimated. Continuous simulation allows modelers to track the soil moisture continuously throughout the period for which storm events are to be evaluated (Beven, 2012; Singh & Chow, 2016). In addition, continuous modeling can provide a continuous estimate for the water fluxes at various spatial and temporal scales, and for processes and fluxes that are difficult to observe or measure in field and/or laboratory, such as lateral subsurface flow and evapotranspiration. Drooger and Perry (2008) stated that continuous modeling improves the understanding of how hydrological processes interact and enables the development and evaluation of scenario analysis in water management. To conduct effective and efficient hydrologic modeling, modelers need to be clear about how hydrological processes function in the real world and how they can be approximated in models.

Hydrological modeling generally requires a model building (or development) process in historical period by using specific model inputs (e.g., meteorological forcings and basin geo-spatial characteristics) and system response data (e.g., observed streamflow at basin interior inlet or outlet) to select appropriate model structures and model parameters (Blöschl et al., 2013; KlemeŠ, 1986; Mai et al., 2020; Singh & Chow, 2016). After the model is successfully developed, it can be further deployed to the future application period (i.e., "out-of-sample" period) to support different purposes of decision-making in water resources management, such as facilitating planning and design, monitoring and predicting floods and droughts, and assessing climate change impact (e.g., see Blöschl et al., 2013; Clark et al., 2016; Fowler et al., 2007; Hrachowitz et al., 2013; Mishra & Singh, 2011; Nohara et al., 2006; and Singh, 2018). It should be noted that in the recent study by Maier et al. (2023), validation (evaluation) is an independent phase of model building (development). This, however, is contrasting to the model building definition in this thesis as well as studies in machine learning field. In practice, an inadequate model result, either in calibration or validation, is a signal that the model should not be

used in further applications. This usually leads to recalibrating and revalidating the model until an acceptable model result is achieved. Thus, model calibration and validation can be an iterative processess, and it is more appropriate to consider validation as a phase of model building.

Hydrological model building can generally be viewed to include two phases in practice: model calibration and validation. 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, Coxon, et al., 2018; Fowler, Peel, et al., 2018). Validation and evaluation are used interchangeably throughout the thesis.

In addition, the general procedures for setting up hydrological modeling in a catchment are summarized as follows (Bedient et al., 2008):

(1) Select and pre-process data based on study objectives, catchment characteristics, data availability, and project budget.

(2) Choose an appropriate hydrological model (or build a model using modeling framework such as Raven), which should be targeted on the key processes or mechanisms the modeler would like to simulate.

(3) Obtain all necessary input data, including meteorological forcing (i.e., precipitation, temperature, radiation, wind speed, atmospheric pressure, humidity, etc.), streamflow data, topography, land cover and land use, catchment characteristics, channel geometry, lake and reservoir features, and other aspects of data that are required in the modeling.

(4) Choose calibration objectives quantifying the degree of agreement between the model predictions and the observed data, as well as any calibration constraints, and the calibration protocol, i.e., manual or auto-calibration. Then calibrate the model.

(5) Validate model using data under different conditions (temporally) or at different in-situ measurement locations (spatially).

(6) Evaluate usefulness of the model and comment on needed changes or modifications. Then apply the model to support decision-making in water resources management if the model is adequate in evaluation.

The model building process is fraught with modeling decisions with respect to six different sources of uncertainty specified by Vrugt and Sadegh (2013), including parameters, input data, initial state, model structure, output, and state variables. Modelers' decisions can be subjective, and these can then affect hydrological modeling results (Melsen et al., 2019). Melsen (2022) interviewed researchers from different institutions to survey 83 motivations of certain hydrological modeling decisions and found that modeling decisions can be specific to research teams and modelers' experiences. In the above outlined six model building steps, model calibration may be one of the most laborious procedures, as it often intertwines with other steps and decisions along the way before a calibration can be carried out or can be deemed as successful. Calibration is usually an iterative process before the experimental setup can be used reliably for a given model and dataset (Mai, 2023). These steps and decisions are summarized as a "calibration life cycle" by Mai (2023) and ten strategies were provided to aid modelers to perform a successful model calibration, such as how to deal with data, model parameters, objective functions, calibration algorithms, etc.

2.3 Overview of split-sample test (SST)

KlemeŠ (1986) provides 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) (see more in Section 2.3.2). This framework consists of four different levels of model validation, including the SST and the differential split-sample test (DSST). Basic idea of SST is to divide the dataset into two non-overlapping subsets: one subset is used for model calibration and the other one is retained for validation (KlemeŠ, 1986). The DSST method is a specific case of the SST (see more in Section 2.3.2), as it selects calibration and validation periods based on pre-defined or pre-screened climatic differences (Coron et al., 2012; KlemeŠ, 1986). The SST has been in standard use, even a paradigm, in hydrological model building in the past half century (Andréassian et al., 2009; Daggupati et al., 2015). In recent decades, many variations of the SST and DSST methods have evolved and been applied in many hydrological modeling studies (e.g., see Dakhlaoui et al., 2017; Essou et al., 2016; and Coron et al., 2012).

The procedure of choosing appropriate data for model calibration and validation is referred to as the *problem of data splitting*. Data splitting is one of the key challenges in the SST framework in model building. However, the data splitting method originally proposed in the SST framework did not have clear guidance on deterministic hydrological modeling and was lacking in empirical/numerical foundations. Even though the SST method has long been extensively applied in hydrological model building, there is no consensus on the data splitting method that can be applied for all (or most) model building practices, and it is reported that different data splitting may have substantial influence on model robustness (Coron et al., 2012; Daggupati et al., 2015; Guo et al., 2020; KlemeŠ, 1986).

In this section, we firstly overview the common data splitting methods, including those using continuous series and discontinuous series for calibration in Section 2.3.1 and 2.3.2, respectively. We then introduce an unconventional model building practice in Section 2.3.3 that some modelers calibrate their models with all available data (i.e., skipping temporal validation), which is also the key motivation of our study. In Section 2.3.4, we compare several recent evaluation frameworks applied for comparing different data splitting methods.

2.3.1 Continuous data splitting

The split-sample approach in hydrological modeling mostly utilizes *temporally continuous* calibration sub-periods (CSPs) (e.g., see Coron et al., 2012; Guo et al., 2018, 2020; Knoben et al., 2020; Mai, Craig, et al., 2022; Mai, Shen, et al., 2022; Newman et al., 2015, 2017; and Rakovec et al., 2019). A key reason for this may be the relatively low computational cost associated with simulated such continuous sub-periods, especially for distributed modeling.

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, i.e., 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: (1) The "sufficiently long" data record is not adequately defined, which leaves it vague for the selection of data splitting schemes; (2) The 50:50 or 70:30

splitting schemes are suggested without empirical/numerical evidence provided to support selecting these splits; and (3) 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; and 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, i.e., 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 the collective experience of the contributors to Chapter 3, they 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 (1) inefficiently simulate the entire validation plus calibration period during the iterative model calibration process or (2) 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 basins 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 1999–2008 period and then validate to the earlier 1988–1999 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).

2.3.2 Discontinuous data splitting

Unlike those using a continuous calibration sub-period for model calibration, another important category of data splitting in SST is using *temporally discontinuous* data for model calibration

and validation. Discontinuous data splitting is less dominant in hydrological modeling in literature, but it has received increasing attention in recent years due in part to: (a) Discontinuous data splits (e.g., see odd/even years method used in Essou et al. (2016)) could retain statistical features (e.g., trend) of the full record into its subsets by sampling data across the entire time series; and (b) Using discontinuous splits in model building naturally leaves a part of data for validation. This second point is important as (Chen et al., 2022) note that skipping validation entirely would go against the split-sample approach conventions that a validation phase must follow with the calibration

Those studies adopting discontinuous data splits can be distinctive regarding the various data splitting methods utilized to select calibration and validation subsets. The most used discontinuous data splitting method is the simple random sampling approach, which randomly samples data with a uniform distribution and thus, each sample has equal probability to be selected. Random sampling is efficient and easy to implement, but it could suffer from high variance or bias when data are not uniformly distributed, e.g., streamflow. In hydrological modeling, this method was employed by Arsenault et al. (2018) to select 1,333 out of 65,534 possible (2%) random annual-based splits for model calibration at each of three catchments in the North America, and they repeated the random sampling of splits for 30 times. This sampling, however, left substantial splits not being considered, which may significantly influence the representativeness of splits used in hydrological modeling if variability in the dataset is not negligible (Reitermanov, 2010).

Differing with random sampling approach to data splitting, some methods are deterministic, i.e., only one data split will be created for a dataset. These methods could alleviate the computational burden of model calibration. An example is the systematic sampling method (Zhang & Berardi, 2001), which is designated for ordered datasets such as time series. This method is simple and efficient in use. One of its typical implementations in practice is the odd/even years method, which splits datasets into odd-year and even-year subsets. Each of them is used for calibration and validation and vice versa. Odd/even years method has been adopted in many hydrological modeling studies (Arsenault et al., 2017; Jie Chen et al., 2013; Essou et al., 2016). Another typical example, which employs more sophisticated algorithms, is distance-based Kennard–Stone sampling (CADEX) (Kennard & Stone, 1969) and its improved version DUPLEX (Snee, 1977). Zheng et al. (2022) further proposed a modified version of the DUPLEX algorithm (i.e., MDUPLEX) and showed its advantage of achieving statistically consistent subsets.

Lastly, another typical data splitting method, which selects calibration and validation data according to climatic differences (e.g., dry/wet and hot/cold), is defined as another level for model validation in KlemeŠ (1986) and is called the differential split-sample test (DSST). DSST is a special

version of the SST and is widely used for analyzing model performance change under diverse hydro-climatic conditions (e.g., see Bai et al., 2021; Coron et al., 2012; Dakhlaoui et al., 2017; Fowler et al., 2018; Fowler et al., 2016; Gaborit et al., 2015; Motavita et al., 2019; and Seiller et al., 2012). In these applications, DSST is generally adopted for hydrological modeling under changing climatic conditions (e.g., climate change impact studies) that models are calibrated and validated in climatically contrasted periods to explore how much performance would loss due to inappropriate parameter transfer and hence assess model application limits (Andréassian et al., 2009; Coron et al., 2012). Note that DSST is out of our scope in this thesis, since the focus on exploring the optimal data split in model building for the prediction objective in the future, where hydro-climatic conditions are assumed to be unknown.

According to those above-mentioned studies on discontinuous data splits, it is also worth noting the drawbacks of utilizing discontinuous data splits in hydrological modeling. The random sampling method is more suitable for testing the sensitivity of using different splits in hydrological modeling than for operational use, because the computational cost is usually not affordable in practice. Although a randomly created data split can be optimal in some cases, it is clear that such an approach does not generalize well to different locations when practical calibration studies require modellers to select a single data split, rather than multiple data splits. The random and deterministic method have the same computational burden as calibrating models to the full-period dataset, i.e., models need to be simulated across the entire duration of model calibration and validation for the continuity in hydrological model state variables updating (e.g., see Arsenault et al., 2018; Dakhlaoui et al., 2019; Essou et al., 2016; and Zheng et al., 2022). Those sophisticated splitting methods, such as the MDUPLEX, though deterministic, require more computational expenses to split the dataset (Zheng et al., 2022).

The above-mentioned discontinuous splitting methods are generally employed from the machine learning field, where discontinuous splits are most used in their model development (May et al., 2010; Reitermanov, 2010). Different to hydrological modeling, machine learning applications generally have no strict requirements on the time dependence in data samples, while temporal order is critical in hydrological time series. A detailed description on discontinuous data splitting methods used in machine learning can be found in literature, e.g., see Chen et al. (2022); May et al. (2010); Reitermanov (2010); and Sharma (2017).

2.3.3 To split data or not? That is the question

Maier et al. (2023) stated three general principles to split available data into calibration and evaluation (i.e., validation), including that calibration and evaluation data should be different, all patterns/events relevant to the modeling purposes should be included in model calibration, and all

patterns/events should be also included in model evaluation. These statements establish the importance of not losing information content in the calibration data compared to the entire available dataset, as such models may be more effectively "learn" the input-output mappings.

Instead of trying data splitting methods with different levels of complexity to meet the above-mentioned general principles during data portioning, some studies found that calibrating models to the full-period dataset is an optimal strategy. A typical example is the study by Arsenault et al. (2018). In their extensive modeling experiments, they demonstrated a Bootstrap-based data splitting method that they firstly randomly chose a part of data years as a test period, which is independent of the calibration and validation data. The remaining non-test 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 the independent test 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 (S. K. Singh & Bárdossy, 2012) also reported that calibrating to all available data may be a robust strategy. Zheng et al. (2023) employed models calibrated to all available data as one of their benchmarks in comparison of two other data splitting algorithms in 163 Australian catchments, and found there is no reason to avoid using all data for model calibration.

The recommendation to skip model validation has huge implications, as it gets rid of any complicated data splitting, thereby largely simplifying hydrological model building, while it also meets the general principles outlined by Maier et al. (2023). However, it 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. A representative sample of 48 calibration publications reporting on the calibration of VIC model, a distributed model, all published in 2018 that are cited in the bibliometric study by Addor & Melsen (2019) was carefully reviewed. Based on this analysis of the literature, 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 test years (i.e., all 1,332 different combinations of calibration and validation periods are evaluated for one fixed test period in their experiment). Thus, it is unclear if their findings were conditional on this single test period.

2.3.4 Evaluation frameworks for comparing different data splitting

Model validation/evaluation studies proposing new or comparing multiple validation approaches, e.g. alternative split sample configurations, 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 and thus representative of some future model application period. Example studies in this category include Coron et al., 2012; Dakhlaoui et al., 2017; Dakhlaoui et al., 2019; Essou et al., 2016; Fowler et al., 2016; Guo et al., 2018; Knoben et al., 2020; Li et al., 2012; Moriasi et al., 2015; Myers et al., 2021; Nicolle et al., 2021; Pool et al., 2018; Rakovec et al., 2019; Schlef et al., 2021; Vaze et al., 2010; and Zheng et al., 2022. 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/evaluation decisions being studied.

Moreover, a robust evaluation framework for testing the efficacy of different SST decisions in hydrological model building should encapsulate most influential aspects of how to perform and evaluate the SST experiments (e.g., controlled modeling experiment and typical methodologies for experiment assessment). The controlled experiment should be designed to remove (or minimize) the influence of extraneous variables not considered but would potentially impact the experiment results. Typical methodologies for evaluating model performances should consider the purpose of building the model and critical aspects that modelers would assess in practice.

Three recent typical evaluation frameworks for comparing different data splitting methods were carefully reviewed. Table 2-1 summarizes how these evaluation frameworks were designed with

respect to five aspects, including the data splits, experiment data, model configuration, calibration protocol, and performance assessment. Details on these evaluation frameworks can be found in Arsenault et al. (2018), Coron et al. (2012, and Zheng et al. (2022). Hereafter we refer to these frameworks as Arsenault's, Coron's, and Zheng's, respectively.

We also note why the other three frameworks are less desirable in the comparative assessment of different data splitting methods as follows.

Arsenault's framework, to our knowledge, was the first one employing "three-period" model evaluation approach in hydrological modeling for comparing different data splits on independent years not utilized in calibration and validation. However, we need to note Arsenault's third period (called test period in Arsenault et al. (2018)) is randomly sampled from the entire period, in which data years may also be allocated to calibration and validation. Although Arsenault's test period is independent to calibration and validation, it is an "in-sample" sub-period in model building and could undermine its representativeness for testing model credibility in the post-validation conditions.

Zheng's framework may be less rigorous with respect to the data years they utilized to compare model performances of different data splits. Zheng et al. (2022) assessed and reported model performance in their calibration, evaluation (which is the same as the validation in this study), and the entire period (i.e., calibration plus evaluation). Two critical issues going with this methodology are that: (a) Different data splits have different calibration/evaluation periods and thus, they are not comparable. Even though the entire period can be a common period for all data splits, it is not rigorous to compare performance of those data splits over the full period, which has been partially used in calibration. This undermines independence requirement for model evaluation; and (b) Doing so in (a) is essentially conducting an "in-sample" evaluation, which is undesirable if modelers aim at using the model in "out-of-sample" hydrology (KlemeŠ, 1986).

Coron's framework aimed at generating all possible equal-length (e.g., 5 years and 10 years) calibration and validation sub-periods to assess model transferability under similar or contrasted conditions. Such a framework is useful for evaluating the sensitivity of model robustness under various conditions. Coron's framework can be helpful if modeler needs to assess model performance when its parameters are transferred to similar or contrasted conditions, and this somewhat relates to the effort by Zheng et al. (2022), which explored model parameters transferred to a statistically similar period. However, the scope in this study (see in Section 1.1) is to identify the optimal data splits to predict system behavior in a post-validation period.

Table 2-1. Summary of four evaluation frameworks with respect to the key aspects of how they compared different data splitting methods.

Key Aspects Reported in Literature		Coron et al. (2012)	Arsenault et al. (2018)	Zheng et al. (2022)
Data Splits	Candidate data splits	GSST method, 125 splits per catchment on average	Randomly sampled years for test, calibration, and validation. 239,940 SST splits in total	MDUPLEX splits, 3 continuous splits, and all data for calibration.
Experiment Data	Catchment selection	216 catchments in southeast Australia	3 catchments in North America	163 catchments in Australia
	Data availability	1974–2006 (33 years)	1986–2010 (25 years)	25–70 years in different catchments
	Data variables	Daily rainfall, potential evapotranspiration, and streamflow	Daily precipitation, temperature, and streamflow	Daily precipitation, potential evapotranspiration, and streamflow
Model Configuration	Model selection	GR4J, MORDOR6 and SIMHYD	GR4J-CN and HMETS	GR4J, AWBM, and CMD
	Watershed discretization	Lumped	Lumped	Lumped
	Model building period	1974–2006	1986–2010	**Not reported**
Calibration Protocol	Model initialization	**Not reported**	One year (1986)	Reported as "at the beginning of the entire period" but not clarified the exact period
	Simulation year type	**Not reported**	**Not reported**	**Not reported**
	Calibration algorithm	A simple steepest descent local search procedure	CMA-ES	Shuffled Complex Evolution global optimization procedure
	Calibration criteria	RMSE and BIAS	NSE	KGE
	Optimization efforts	**Not reported**	Budget of 10,000 model evaluations	**Not reported**
Performance Assessment	Model failure handling	**Not explicitly considered.** Model deficiency is reported.	**Not reported**	**Not reported**
	Evaluation period	1974–2006 validation periods	"In-sample" test periods	Evaluation period and all data period (calibration and evaluation)
	Evaluation criteria	MRC and NSE	NSE	RE, KGE over all data, KGE over calibration, KGE over evaluation (i.e., validation in this paper), and ΔKGE

Table 2-1 (continued)

Key Aspects Reported in Literature	Coron et al. (2012)	Arsenault et al. (2018)	Zheng et al. (2022)
Evaluation aspects	Performance loss, change in climatic characteristics	Test-period NSE and NSE difference between models calibrated to all data and calibrated to short-period data	KGE over all data, RE, ΔKGE, skewness of runoff, catchment size

Definitions of key abbreviations in Table 2-1

GSST: Generalized split-sample test
RMSE: Root-mean-squared error
BIAS: Relative errors in discharge
MRC: Model robustness criteria
NSE: Nash-Sutcliffe efficiency
CMA-ES: Covariance-Matrix Adaptation Evolution Strategy
KGE: Kling-Gupta Efficiency
RE: Percentage relative error
ΔKGE: Difference between KGE over calibration and evaluation
CAMELS: Catchment Attributes and Meteorology for Large-sample Studies
USGS: United States Geological Survey
CSP: calibration sub-period
DDS: Dynamically dimensioned search

2.4 Overview of model inadequacy issue in hydrological modeling

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; K. 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; Smith et al., 2012; and Yang et al., 2019. The absence of explicit model failure criteria is a suboptimal approach to model building and 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: (1) 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; and Ritter & Muñoz-Carpena, 2013); and (2) Reference models (also named benchmark, see Garrick et al., 1978; Knoben et al., 2020; Newman et al., 2015; and Schaefli & 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). 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 practice, a modeler must decide how to handle a model failure at the calibration or validation stage of the model building process. This is especially true in studies evaluating alternative data splitting methods. Generally, there are four strategies to handle model failures: (a) Assuming a model can never fail in calibration and validation, thus skipping handling failures and proceed to use all model calibration and validation results for further analysis, which does not explicitly handle failures and is commonly seen in modeling studies (e.g., see Arsenault et al., 2018). As such, it may lead to unreliable conclusions due to failure contamination; (b) Simply tossing out the failed model completely and replace it with nothing in validation (e.g., see Guo et al., 2020 and Knoben et al., 2020), which will reduce sample size in evaluation as some models (catchments) are disregarded. This is less desirable since catchments with failed models are informative as they imply if our SST decisions identify inferior models; (c) Discarding failed models and then identifying a new model/new model parameter set to predict conditions in the model testing period providing an alternative adequate prediction in testing periods and hence addresses the sample size issue in (b); and (d) Optimizing the choice between option (a) and (c) based on the posterior model testing results evaluated from large-sample model building and testing results.

2.5 Overview of large-sample hydrology

Large-sample hydrology is an approach making use of datasets containing large catchment samples to draw statistically robust conclusions in hydrological modeling (Gupta et al., 2014). Since hydrological processes are dependent on spatial and temporal scales (Sivapalan & Blöschl, 2015; Skøien et al., 2003) and are related to catchment characteristics such as climate (Budyko et al., 1974), soil (Vereecken et al., 2015), geology (Kuentz et al., 2017), vegetation (Stephenson, 1990), land covers (Carlson & Arthur, 2000), and topography (Condon & Maxwell, 2015), utilizing a large sample of catchments instead of learning from place specific cases can provide broader information across a variety of different hydro-climatic environments and spatio-temporal scales (Gupta et al., 2014).

Large-sample hydrology is an important branch of comparative hydrology, which is to understand how various factors and conditions influence hydrological behavior, identify similarities and differences between different hydrological systems, and interpret these in terms of underlying

climate-landscape-human controls (Addor et al., 2020; Falkenmark & Chapman, 1989; Hrachowitz et al., 2013; Kovács, 1984; Thompson et al., 2011). Gupta et al. (2014) highlighted several typical benefits of the large-sample hydrology approach, including improving understanding of hydrology, achieving statistical robustness in analysis, facilitating catchment classification, regionalization, generalization and transposability, and enhancing the estimation of uncertainty in model prediction. Large-sample hydrology has been the foundation of many hydrological studies including those dedicated to catchment classification (Knoben et al., 2018), sensitivity analysis (Mai, Craig, et al., 2022), model evaluation and benchmarking (Towler et al., 2023), model uncertainty assessment (Knoben et al., 2020), model inter-comparisons (Mai, Shen, et al., 2022), streamflow prediction (Hrachowitz et al., 2013) and climate change impacts assessment (Melsen et al., 2018).

The large-sample hydrology approach highly relies on the availability of large sample of catchment datasets. Addor et al. (2020) highlighted two key requirements of a large-sample hydrology datasets that it must contain streamflow observations and basic identifiers for stream gauges along with these data. The number of catchment samples may vary from tens for regional comparisons to thousands for continental and global scales investigations (Addor et al., 2020; Burn & Whitfield, 2018; Fowler, Coxon, et al., 2018; Gudmundsson et al., 2019; Mai, Craig, et al., 2022). Addor et al. (2020) and Gupta et al. (2014) listed multiple available large-sample catchment datasets, including the Model Parameter Estimation Experiment project (MOPEX; Duan et al., 2006), which contains 438 catchments across the United States with long-term hydrometeorological observations to 2003 as well as attributes for catchments representing different hydroclimatic conditions. Later developed typical large-sample hydrology datasets at the national scale are the Catchment Attributes and Meteorology for Large-sample Studies dataset (CAMELS) in the US (Addor et al., 2017; Newman et al., 2015), Chile (Alvarez-Garreton et al., 2018), Brazil (Chagas et al., 2020), Great Britain (Coxon et al., 2020), Australia (Fowler et al., 2021), Austria (Klingler et al., 2021). Similar datasets are also available at regional scale and larger continental scales such as the Great Lakes Region (Mai et al., 2021; Mai, Shen, et al., 2022), the Arctic Region (https://www.r-arcticnet.sr.unh.edu/v4.0/index.html), North America (Arsenault et al., 2020) and Europe (Kuentz et al., 2017). Recently, the CAMELS series datasets were further aggregated and standardized as one dataset called Caravan, which includes meteorological forcing, streamflow, and static catchment attributes for 6830 catchments (Kratzert et al., 2023).

Past comparative data splitting studies do not normally use large sample hydrology datasets (e.g., see Arsenault et al. 2018; and Myers et al., 2021) and thus bring into question the robustness and generality of their conclusions.

2.6 Research gaps

In our collective review on those aspects presented in Section 2.1 to Section 2.5, we identify key reasons why the problem of data splitting still remains unsolved in hydrological modeling, as well as the data splitting research gaps we would like to mitigate in this thesis:

1. Data splitting is generally performed only once in most modeling studies without checking other possible splits, which may result from both modelers' preference or experience and the constraints of computational capacity (Melsen, 2022; Melsen et al., 2019). This, in turn, leads to arbitrary selection of calibration and validation in hydrological modeling (Myers et al., 2021).
2. Poor or inadequate models that would be considered to fail the validation check in practice are not properly handled in data splitting studies, which is not realistic for operational model building and may have significant influence on model performance assessment and further decision-making.
3. Studies on comparing different data splitting methods require a robust evaluation framework to provide rigorous experimental methodology, but most evaluation frameworks seen in literature are not ideal. In particular, past data splitting studies have either (typically), not considered a more realistic post-validation model testing period beyond calibration and validation or they have considered a single deterministic example testing period.
4. Most studies focus on interpreting results based on small samples of catchments, which makes it difficult to generalize the conclusions across different conditions. However, the large-sample hydrology approach can provide more robust empirical/numerical evidence to support conclusion drawn and generalization.

Chapter 3
Time to Update the Split-Sample Approach in Hydrological Model Calibration

This chapter is a mirror of the following published article by Shen et al. (2022a). Most of the literature review content in the article (e.g., the introduction and methodology sections) is adapted to Chapter 2, and only a shortened version of the introduction goes with this chapter. Other sections such as results and conclusions are all consistent with the article. All references are unified at the end of the thesis.

Shen, H., Tolson, B. A., & Mai, J. (2022a). Time to Update the Split-Sample Approach in Hydrological Model Calibration. *Water Resources Research*, 58(3), 1–26. https://doi.org/10.1029/2021WR031523

Summary

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 models to older data and then validating 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.

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

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; and 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). In addition, 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. Most studies tend to select calibration and validation data years chronologically, i.e., 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. More discussion on the SST and model inadequacy issue in hydrological modeling can be referred to Chapter 2.

This chapter 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.

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 3.3): 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 3.2, the large sample of case study catchments, historical data and methodology are introduced. The key results and discussion are presented in Section 3.3 and Section 3.4, respectively. Finally, the conclusions are summarized in Section 3.5.

3.2 Data and methodology

Section 3.2.1 introduces the novel SST experimental design, and then Section 3.2.2 introduces the catchment and data used in this study. Section 3.2.3 and Section 3.2.4 describes the hydrological models and model calibration protocol, respectively. Section 3.2.5 describes the methodology for analyzing the results.

3.2.1 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 (GSST) (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: (1) we wanted to investigate the value of recent versus old calibration sub-periods, effectively precluding the use of discontinuous periods; (2) Arsenault et al. (2018) reported that calibrating to all data could be preferred to discontinuous sub-period calibrations; and (3) as discussed in Section 2.3.1, 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 scheme (i.e., calibration, validation and testing) is stricter in SST assessment than the commonly used two-period approach. Multiple testing periods are created by simply pretending the model was built five years ago, thus leaving the five most-recent years as continuous model testing data, and then pretending the model was built ten 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 as different model build years to 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 thesis, 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 3-1 illustrates how our experimental design splits the 35 years of available data (1980–2014 based on our large sample of catchments described in Section 3.2.2) 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 3-1, 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 SST framework (KlemeŠ, 1986). Employing sliding windows allows CSPs of a different age (i.e., older versus newer data) to be defined. For a given length of sliding window, there are multiple candidate CSPs. For example, given nine years of available data for calibration and validation, using a 3-yr sliding window can create up to seven CSPs. However, many of these CSPs overlap with one another and thus, calibration on all of 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 3-1, 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 3-1, 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 one-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 three-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 three-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 3-1), 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 (i.e., these validation years are not used for criterion computation during the calibration). The best calibrated parameter set (identified using the calibration protocol in Section 3.2.4) 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 3-1 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 3-1 are categorized into three classes, which hereafter are called as the *full-period* CSPs (no validation performed, such as CSP-9A_{1990} in Figure 3-1a), *recent* CSPs (calibration years immediately precede the model build year, such as CSP-3D_{1990}, CSP-5C_{1990} and CSP-6B_{1990} in Figure 3-1a), and *older* CSPs (calibration years exclude the most recent years that immediately precede the model build year, such as CSP-3A_{1990}, CSP-3B_{1990} and CSP-3C_{1990} in Figure 3-1a). 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 3-1, considering 1990 as the model build year, CSP-3C_{1990} has a recency score of 8/10 or 80% whereas CSP-3D_{1990} 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 3-1a panel, CSP-3C$_{1990}$, CSP-5B$_{1990}$ and CSP 6A$_{1990}$ are all assigned a recency score of 80%.

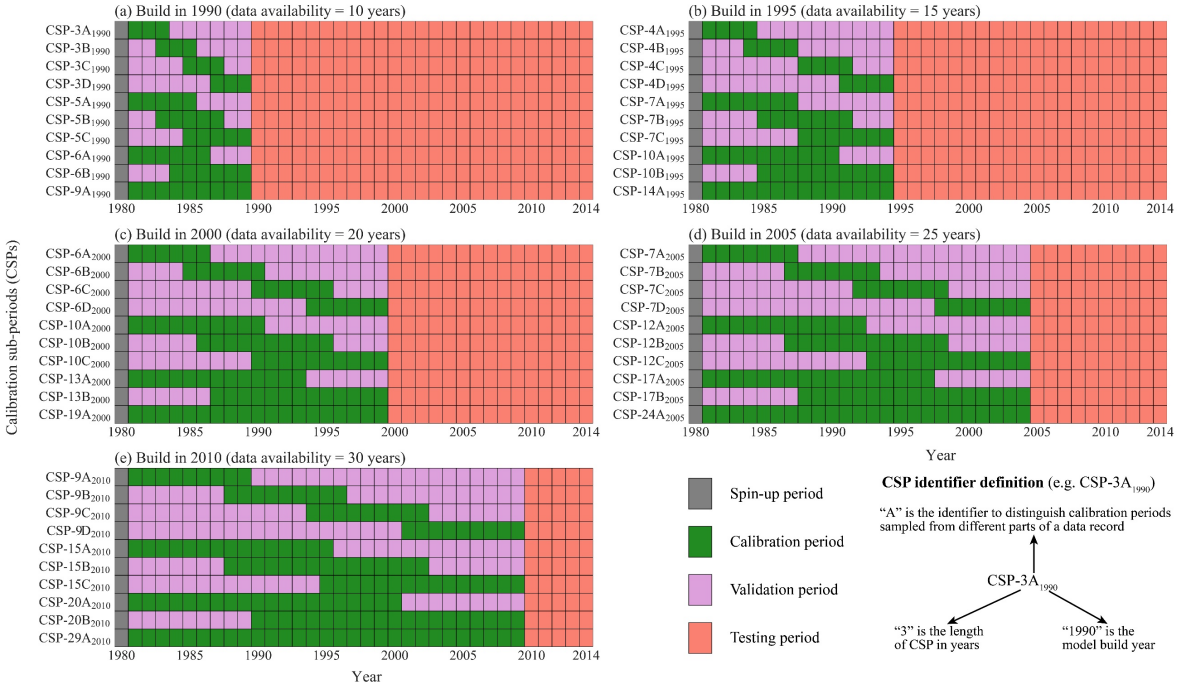

Figure 3-1. Experimental design for the split-sample test assessment. Calibration sub-periods (CSPs) are created for different model build years at (a) 1990, (b) 1995, (c) 2000, (d) 2005, and (e) 2010, with data availability for calibration then being 10, 15, 20, 25, and 30 years, respectively. Each 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.

3.2.2 Catchments and data

The Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) dataset 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 dataset serves as the candidate hydrological modeling inputs for our study. The CAMELS dataset 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 (H. V 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 1980–2014 *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 km^2 to 25,800 km^2. 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 km × 1 km compared to 12 km × 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 dataset 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 A1-1 in Appendices A-1. The detailed information of the 463 catchments and the corresponding Daymet forcings and updated USGS streamflow data files for these catchments are all available online (see Appendices A-2)

3.2.3 Hydrological models

Two conceptual lumped hydrological models are applied in this study: the GR4J (which stands for modèle du Génie Rural à 4 paramètres Journaliers) and HMETS (which stands for Hydrological

Model of École de technologie supérieure) 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 A3-1 in Appendices A-3.

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 A4-1 in Appendices A-4.

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, Craig, 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 dataset that are immediately available for use with any other Raven-configured model structure (available online, see Appendices A-2). Full details on the Raven framework can be found in Craig et al. (2020) and the Raven manual (Craig, 2023).

3.2.4 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 3.2.1. 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 (note the number of trials 20 is an adequate balance of minimizing uncertainty in calibration and maximizing feasibility in running such a huge number of models). 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 × 50 CSPs × 20 trials × 463 catchments), and the total number of model test period hydrographs assessed is 129,640 (2 models × (40 CSPs × 3 testing periods + 10 CSPs × 2 testing periods) × 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

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\mu_{sim}}{\mu_{obs}} - 1\right)^2} \tag{3-1}$$

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 -∞ 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.

3.2.5 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 basins, 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 (3.2.5.2–3.2.5.4). However, since all strategies depend on an explicit approach to model failure identification and handling, we first address this topic in Section 3.2.5.1 below.

3.2.5.1 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 independent on whether we are in the model calibration/validation stage or the model testing phase of our experimental design. All data years from 1980 to the year before the model build year define the five different reference periods utilized here. 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 Feb 29th.

Following Knoben et al. (2020), the KGE calculated using the reference flow as the predicted flow is denoted as *reference KGE*. Note that the KGE is computed by comparing the model simulated flow and observed flow, while the simulated flow is replaced with the reference flow in reference KGE calculation. The reference KGE can be used to distinguish plausible/implausible 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 versus the reference KGE for the calibration period), the validation result (validation period KGE versus the reference KGE for the validation period), and the model testing result (testing period KGE versus the reference KGE for the testing period). Note the importance of our testing period reference KGE being independent of the corresponding observed data in testing period. The best calibrated hydrographs, KGE and reference KGE of different sub-periods are available online (see Appendices A-2).

Model failure handling strategies applied in practice are introduced in Section 2.4. 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 3.2.5.3 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 basin 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.

3.2.5.2 Frequency of each short-period CSP being better than its corresponding full-period CSP

In this simple analysis, we directly compare a pair CSPs together by determining the frequency one does better than the other across all 463 basins 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 versus 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 only proportions smaller than 0.455 and larger than 0.545 are significantly different from 0.5. See Section 3.3.1 for the results.

3.2.5.3 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 3-1); (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 have 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 3-3 (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 3-1 is post-processed to generate most of the decision tree inputs. In addition to finishing all Figure 3-1 experiments, all the relevant reference KGE values must be computed from observed streamflows as described in Section 3.2.5.1 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 basin is expressed as:

$$\mathrm{KGE}_S = \max(\mathrm{KGE}_t, \mathrm{KGE}_{truncate}) \tag{3-2}$$

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 = $1-\sqrt{2}$ (≈ -0.414) means the simulations are equal to using mean annual flow as a predictor (Knoben, Freer, & Woods, 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 3.3.2, we detail a few example rollback calculations, the results of which are encapsulated in Figure 3-3.

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 3-1, there are 2 models × 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 (Schöniger 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 basin 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; and 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:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \tag{3-3}$$

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. $\text{CSP}_{x1}(i) \geq \text{CSP}_{x_2}(i)$, for all indices $i \in \{\text{median KGE, accuracy}\}$, and
2. $\text{CSP}_{x1}(j) > \text{CSP}_{x_2}(j)$, 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 judgements 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 3.3.3 for the results.

3.3 Results

3.3.1 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 3-2 displays these results in the three testing periods for the two different models and show that at the 0.05 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 3-2 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 3-2) of both GR4J and HMETS are smaller than 0.5, ranging from 0.25 to 0.44

(GR4J) and 0.23 to 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 calibration data period, i.e., CSP-17A$_{2005}$ and CSP-20A$_{2010}$.

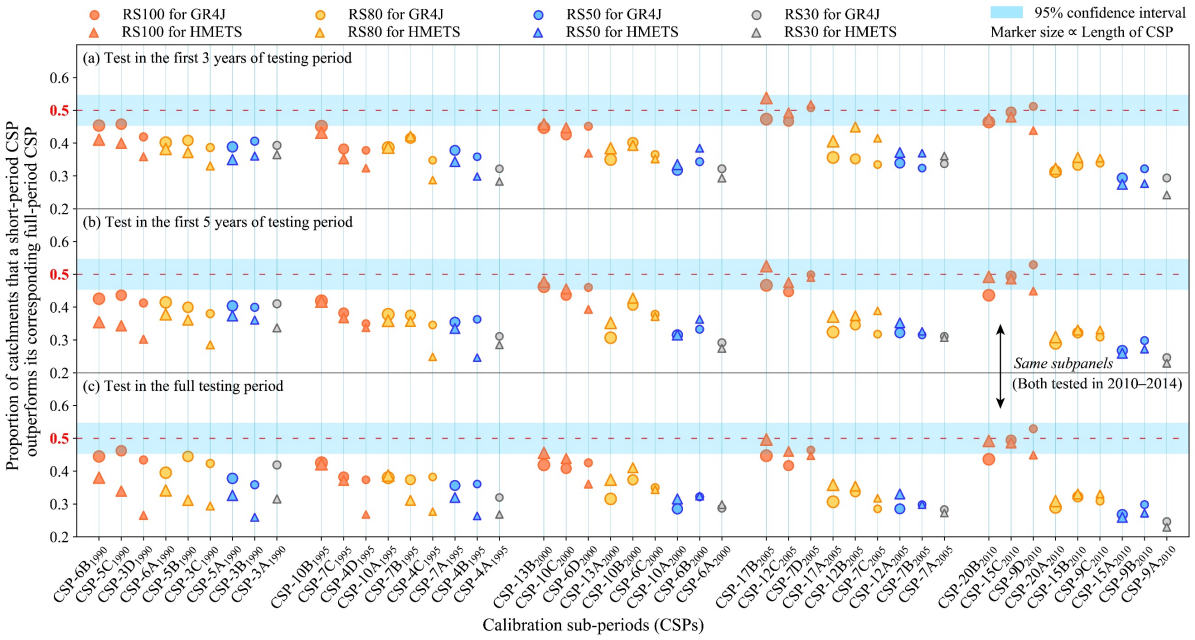

Figure 3-2. Proportion of 463 catchments that the short-period calibration sub-periods (CSPs) outperform their corresponding full-period CSP in all testing periods. The *x*-axis is grouped by model build years, then firstly sorted by recency scores (descending order, denoted as different colored markers) and secondly 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 HMETS results are represented by circles and triangles, respectively. Marker sizes are in proportion to the lengths of CSPs represented by the percentages of calibration data availability. 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 3-1.

Second, the proportions for recent CSPs (recency score 100% but length of CSP is smaller than 100%, represented as red markers in Figure 3-2) 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 3-2 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-6B_{1990}, CSP-10B_{1995}, etc.), as shorter-length recent CSPs are not as reliable (e.g., CSP-3D_{1990}, CSP 4D_{1995}, 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-10C_{2000}, CSP-12C_{2005} and CSP-15C_{2010}) with either model generally works as well as calibrating to the full-period with only very limited exceptions.

3.3.2 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 3.2.5.3. Figure 3-3 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 ten CSP branches shown. The other seven branches are skipped in Figure 3-3 for brevity but are taken into account in analysis.

As introduced in Section 3.2.5.3, 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 the Figure 3-3 configuration. Firstly, 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. For the clarity of this presentation, calibration and validation results are not reported in Figure 3-3. 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 3-3. Note that we report all possible outcomes in this step. For example, when calibrating to CSP-7D$_{2005}$, 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$_2$ and TF2$_2$); the other one is discarding the model and using reference flow for testing periods (denoted as TA3$_2$).

Secondly, 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 3-3), and the second is the overall expected KGE score for a CSP based on the optimal model failure handling decisions. For CSP-7D$_{2005}$, 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-7D$_{2005}$, having an expected KGE score 0.454, ranks the best out of the ten CSPs and hence is the optimal choice in this decision tree. The full-period CSP (CSP-24A$_{2005}$) 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 3-4 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.

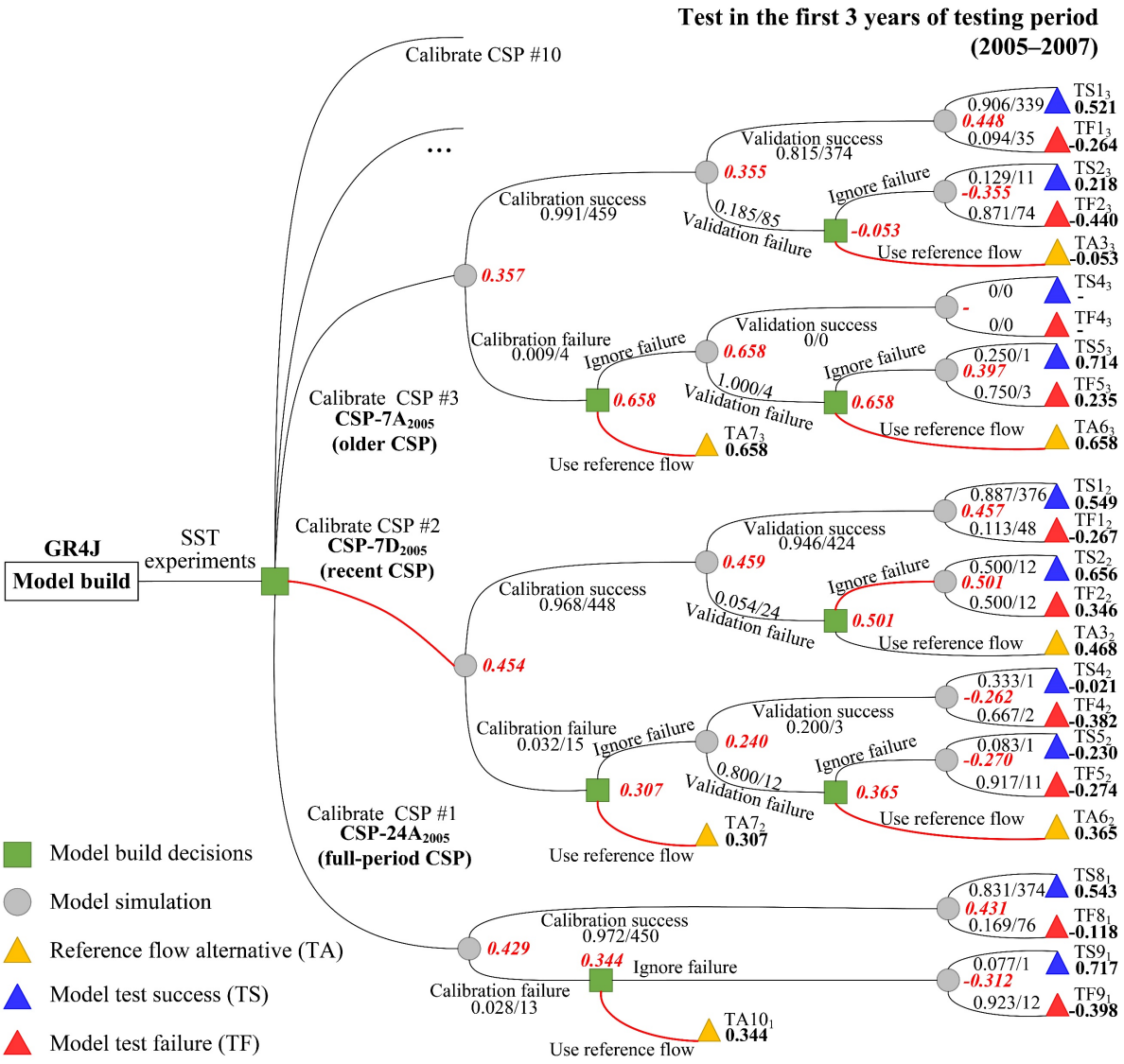

Figure 3-3. Example of a decision tree for GR4J on three calibration sub-periods (CSPs, i.e., CSP-24A$_{2005}$, CSP-7D$_{2005}$ and CSP-7A$_{2005}$) tested in the period of 2005–2007 and synthesized from the 463 catchment samples. The green boxes are the decision nodes for making decisions on CSPs and 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.

Figure 3-4 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. Moreover, it can be seen that the best (bold values highlighted in Figure 3-4) out of ten CSPs in each panel is consistently one of the recent CSPs with recency scores = 100% and in five of ten 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 column in each panel in Figure 3-4) 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 3-3 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 failure and use the model in the testing period performance. This optimal decision is not known in practice when building a model and so the results in Figure 3-4 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 handling strategy to use reference flows when calibration or validation is a failure.

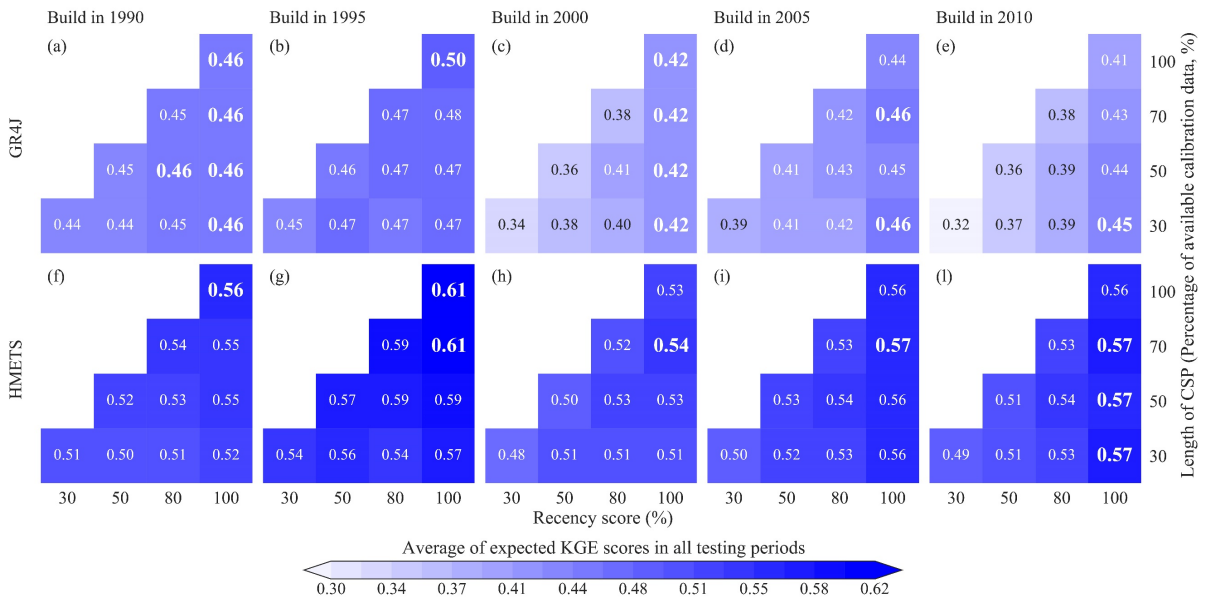

Figure 3-4. Heatmaps of expected KGE scores of calibration sub-periods (CSPs) averaged over all testing periods for GR4J and HMETS based on the decision tree analysis (14 decision trees per model). CSPs are classified into different classes regarding the length of CSP (percentage of available calibration 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.

3.3.3 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 3.3.2), 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 3-5 presents an example tradeoff analysis between the median KGE and accuracy for the first 3 years testing period. CSPs in Figure 3-5 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 (maximizing both objectives) for this

multi-objective analysis, called the non-dominated solutions. These non-dominated solutions are highlighted by the Pareto front in Figure 3-5d and Figure 3-5e, where in each case there are multiple non-dominated solutions. In all 8 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 3-5) are identified as non-dominated solutions for 8 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 3-5), 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 0.83 to 0.91 for GR4J and from 0.85 to 0.96 for HMETS. Similar relative model performance results are also observed in other testing periods (not shown here).

The accuracy differences between Pareto front solutions and the dominated solutions in Figure 3-5 are noteworthy for HMETS (e.g., they are up to 0.08 in various subpanels). Considering the error rate (1 − accuracy) instead of accuracy, these differences translate into the dominated CSPs having classification error rates that can sometimes be 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 0.08 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 to 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 3-5 (one testing period) with the tradeoff analyses from the other two testing periods (which are not shown individually), we produce Figure 3-6 which aggregates all these frequencies.

Figure 3-6 shows that full-period CSPs have the highest 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 one of the three testing periods in Figure 3-6d. In 7 out of 10 panels, referring to the secondary *y*-axis, full-period CSP instances have the highest frequency of dominating other CSPs (ranging from 0.667 to 1.0, represented as the largest red markers

in Figure 3-6). 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 3-6 also reveals just how poorly the older CSPs perform. Not a single older CSP was non-dominated in any of the 14 cases of tradeoff analyses for the HMETS model. For the GR4J model, the older CSPs can be non-dominated but the full-period CSPs are most frequently non-dominated (in 11 of 14 tradeoff analyses). Furthermore, the older CSPs cluster to the right in each panel, in particular for the CSPs with a recency score of 50% or 30% meaning the frequency they dominate another CSP are very low (typically around 0.2).

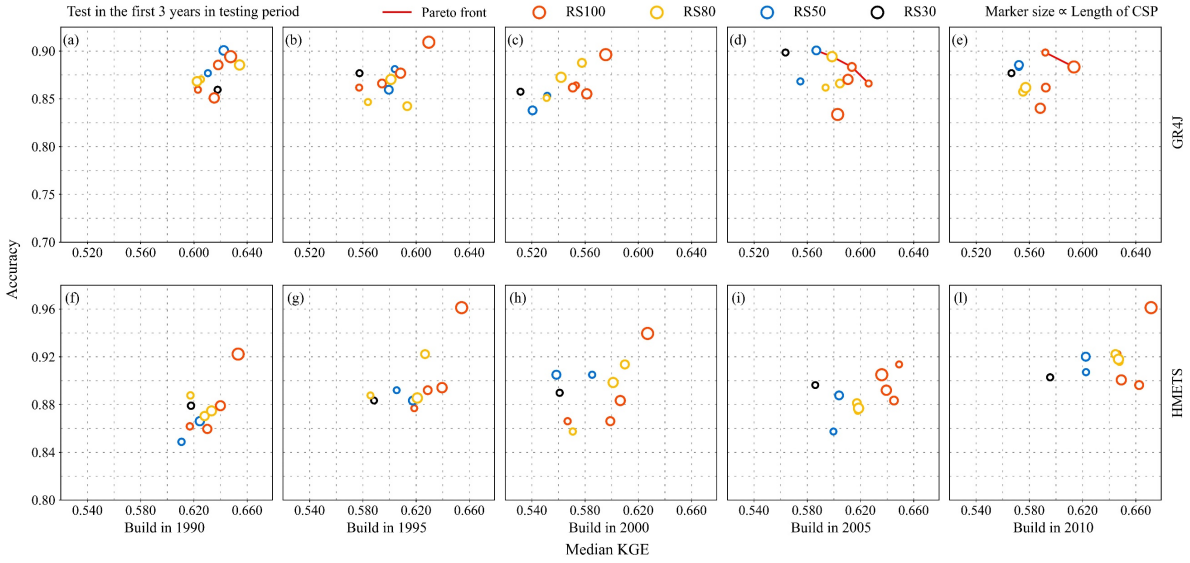

Figure 3-5. The Pareto solutions in the two-dimensional space regarding median KGE and accuracy metric of different calibration sub-period (CSP) classes in the first three years of testing period. The first row of plots are results for GR4J and the second row of plots are for HMETS. The solutions lying in the upper-right panel with high values in both median KGE and accuracy metric are dominating solutions in their lower-left positions. 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 CSP. 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 except (d) and (e) due to the sole non-dominated solution in each of the plots.

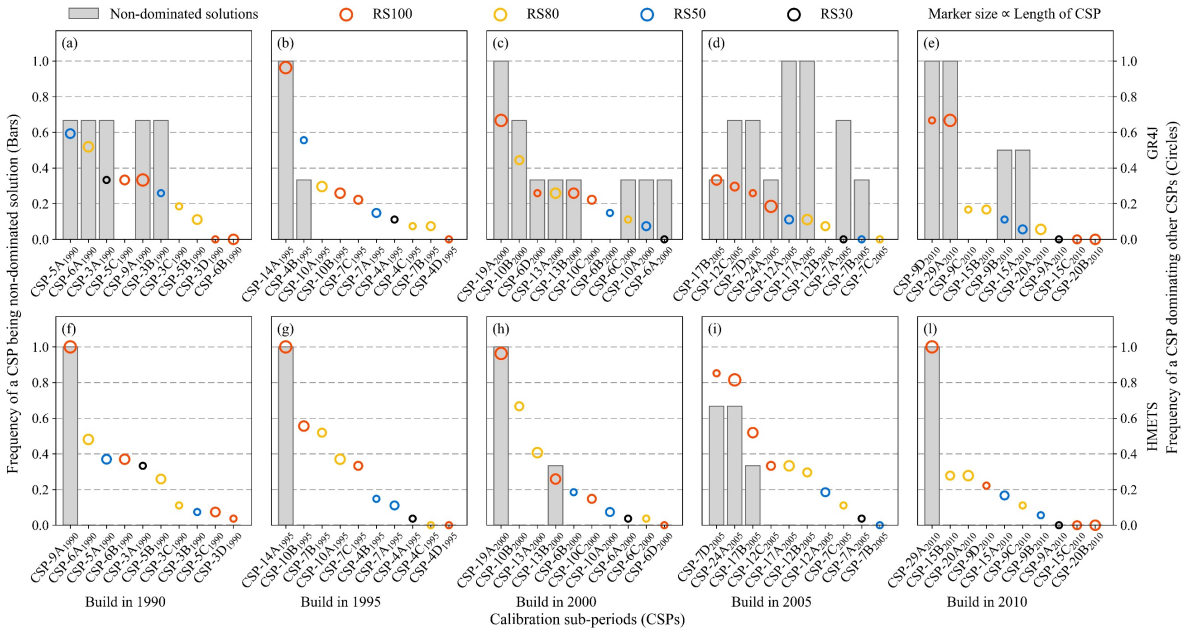

Figure 3-6. Summary of tradeoff between the median KGE and accuracy over the 463 catchments on all three testing periods. Simulation results with failures in these model building processes are constantly rejected and the reference flow is used as the alternative. The first row of subplots are results for the GR4J model and the second row of subplots are for the HMETS model. Gray bars indicate the relative frequency of each calibration sub-period (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 size is in proportion to the length of CSP. Note that the definition of CSP identifiers is provided in Figure 3-1.

3.4 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 3.3. 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.

3.4.1 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 3-2 supporting this could not indicate any more clearly that when building a model to predict future streamflows for some 3-yr to 25-yr 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-5B$_{1990}$ in Figure 3-1) is tied on average with all the most recent CSPs. Results in our multi-objective analysis considering median KGE and classification accuracy were also extremely strong, as not a single older CSP was non-dominated in any of the 14 cases of tradeoff analyses of HMETS model, and for the GR4J model, the older CSPs can be non-dominated but the full-period CSPs are most frequently non-dominated (in 11 of 14 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 3.1, Myers et al. (2021) reported 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_1) and at

the start of the validation period (t_1 - Δt, where Δt denotes the time interval between the start of validation and the start of calibration periods). See Section 3.2.1 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 3.3 for the most recent CSPs with lengths 100%, 70%, 50% and 30%. In Figure 3-2, 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 3-2 results over other CSPs. For Figure 3-4, 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 3-4 results over the 100% CSP. Fortunately, our most robust and multi-objective assessment in Figure 3-6 (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 24/28 tradeoff analyses while all other CSPs are non-dominated in at most 7 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 2.3.3, 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, i.e., 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.

3.4.2 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 & 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; Dakhlaoui et al., 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 differential split sample test (DSST) proposed by KlemeŠ (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 3.2.1). We speculate that our finding on data recency being so critical to calibration success, could be due to some combination of: (1) relatively poor quality in older forcing and/or streamflow data; (2) 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); (3) non-stationary watershed conditions due to anthropogenic influence; and (4) 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.

3.5 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 50 different continuous calibration sub-periods (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.

Chapter 4
Can Hydrologists Benefit from Using Discontinuous Data in Model Calibration?

This chapter is a replicate of the following manuscript that is currently in preparation. Most of the literature review content in the article (e.g., the introduction and methodology sections) is adapted to Chapter 2, and only a shortened version of the introduction goes with this chapter. Other contents such as results and conclusions are all consistent with the manuscript. All references are unified at the end of the thesis.

Shen, H. & Tolson, B. A. (2023). Can Hydrologists Benefit from Using Discontinuous Data in Model Calibration? (manuscript in preparation)

Summary

Hydrological model calibration and validation are critical steps in the split-sample test (SST) for model building (development). However, model robustness can be heavily influenced by how the data available is partitioned for these procedures. Choosing temporally continuous data for calibration is the dominant approach in hydrological modeling community, while sampling data from different segments of data series for calibration and validation may be a promising strategy for model building, as it can retain similar information for both subsets. This large-sample SST assessment study empirically compares both continuous and discontinuous data splitting methods in model building under different conditions and assesses how they impact model performance in the post-validation model testing periods. A conceptual model is calibrated and validated in 463 catchments across the United States based on 44 SST approaches such as the continuous calibration sub-period (CSP), discontinuous calibration sub-period (DCSP), modified DUPLEX (MDUPLEX), and full-period CSP). Discontinuous splits are generated in a deterministic way by either systematic sampling or the MDUPLEX algorithm, which is designed to create statistically similar subsets. The SST assessment is novel in multiple aspects including all SST decisions are compared in the model testing period, which defines a common "out-of-sample" period for an objective comparison; model inadequacy (failure) is properly defined and handled in the assessment; and the accuracy of models correctly predicting testing success/failure and model performance in testing period are optimized simultaneously. The large-sample empirical results strongly support calibrating models to the full-period CSP and skipping temporal validation entirely, while calibrating to discontinuous data offer no clear advantages over full-period CSP. It is recommended that hydrological modelers rebuild models after their validation

experiments, but prior to operational use of the model, by calibrating models to all available data. This step will not invalidate the validation already done but makes use of all available data for fitting the model.

4.1 Introduction

Hydrological modeling generally requires a model building (or development) process in historical period by using specific model inputs (e.g., meteorological forcings and basin geo-spatial characteristics) and system response data (e.g., observed streamflow at basin interior inlet or outlet) to select appropriate model structures and model parameters (Blöschl et al., 2013; KlemeŠ, 1986; Mai et al., 2020; Singh & Chow, 2016). After the model is successfully developed, it can be further deployed to the future application period (i.e., "out-of-sample" period) to support different purposes of decision-making in water resources management, such as facilitating planning and design, monitoring and predicting floods and droughts, and assessing climate change impact (e.g., see Blöschl et al., 2013; Clark et al., 2016; Fowler et al., 2007; Hrachowitz et al., 2013; Mishra & Singh, 2011; Nohara et al., 2006; and Singh, 2018).

Hydrological model building can generally be viewed as two phases in practice: model calibration and validation. Among those challenges fraught with model calibration listed in Mai (2023), we underline the *problem of data splitting*, which refers to the procedure of choosing appropriate data for model calibration and validation. Unfortunately, there is no consensus on the data splitting method that can be applied for all (at least most) model building practices, and it is reported that different data splitting may have substantial influence on model robustness (Coron et al., 2012; Daggupati et al., 2015; Guo et al., 2020; KlemeŠ, 1986; Shen et al., 2022a). More discussion about the problem of data splitting is presented in Section 2.3.

In Chapter 3 (Shen et al., 2022a), we provided two recommendations for the split-sample test (SST) practices in hydrological modeling:

"**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; and

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".

Two recent large-sample modeling studies adopted the above SST recommendation #2 and both showed supportive results: (a) Wasko et al. (2023) calibrated hydrological models to all available data across 467 catchments in Australia and showed the strength of calibrating models to the full-period

dataset targeting streamflow extremes; and (b) Zheng et al. (2023) employed models calibrated to all available data as one of their benchmarks as a comparison of two other data splitting algorithms in 163 Australian catchments, and their results show there is no reason to avoid using all data for model calibration.

However, Chapter 3 (Shen et al., 2022a) only assessed temporally continuous data splits in their evaluation framework, while another important category of data splitting in SST, i.e., using *temporally discontinuous* data for model calibration and validation, has not been assessed as a contrasting experiment to those continuous data splits and full-period dataset. As stated in some past studies, using discontinuous data splits in hydrological modeling may have two key advantages: (a) Discontinuous data splits (e.g., see odd/even years method used in Essou et al. (2016)) could retain statistical features (e.g., trend) of the full record into its subsets by sampling data across the entire time series; and (b) Using discontinuous splits in model building naturally leaves a part of data for validation. It is a concern in hydrological modeling community that calibrating models to the full-period dataset and skipping validation entirely would go against the long standing split-sample approach convention that a validation phase must follow model calibration (Chen et al., 2022). Using discontinuous splits would satisfy the need for a validation. A review of using discontinuous data splits in the SST method can be found in Section 2.3.2.

In this study, we adopt the evaluation framework proposed in Chapter 3 (Shen et al., 2022a) to investigate the potential of using discontinuous data splitting in hydrological model building. We limit our scope to addressing the *single-site temporal* calibration and validation problem. Multi-site model calibration and spatial model validation are not considered. The two research questions are described as follows.

1. For the streamflow prediction objective, can we benefit from calibrating hydrological models to discontinuous data splits? And can any of these alternative split samples be more robust than calibrating models to all data available?
2. Are conclusions in (1) robust if model performance is assessed in different ways, such as using multiple calibration trials versus using the best-calibrated trial only?

Answering the first question mitigates the gap in SST recommendation made in Chapter 3 (Shen et al., 2022a), which did not consider discontinuous splits in the controlled experiments. Our initial hypothesis is that utilizing discontinuous splits in model building is unlikely to change the SST recommendations that calibrating to all data is the most robust choice for streamflow prediction in testing periods. A possible reason is that any split samples lose information compared to the full-period

choice. If a model is less likely to be prone to the overfitting issue, using all data in calibration would be a better choice.

In this study, we adopt the dynamically dimensioned search (DDS) algorithm for model calibration. Due to the stochastic nature of the DDS algorithm, it is generally required to perform multiple optimization trials each with independent populations (Tolson & Shoemaker, 2007). Chapter 3 (Shen et al., 2022a) utilized the best-calibrated trial out of 20 optimization trials to eliminate variability of optimization algorithm performance but that approach reduces the sample size of the experiments. In contrast, considering all trials could reveal the central tendency of all optimization trials and may be more rigorous for our hypothesis testing.

Overall, our contributions by addressing the above two questions in this chapter are:

1. To reaffirm the SST recommendation made in Chapter 3 (Shen et al., 2022a) that calibrating models to all available data is still the most robust choice considering alternative split samples (i.e., discontinuous splits) based on more robust assessment approach (i.e., using all 20 optimization trial results).
2. To make the best practice recommendation for operational modeling based on all the ways modelers could possibly split their dataset, including continuous and discontinuous data splitting, to enhance model performance in the model application periods.

In Section 4.2, we present the controlled SST experiments adapted from Chapter 3 (Shen et al., 2022a) as well as the large-sample catchments, modeling data, model calibration protocol and methodologies for performance assessment. The key results and discussion are presented in Section 4.3 and Section 4.4, respectively. Finally, the conclusions are summarized in Section 4.5.

4.2 Data and methodology

This section describes the key methodologies adopted in the split-sample test (SST) controlled experiment and model performance assessment. Section 4.2.1 introduces the unique SST experimental design for applying continuous and discontinuous splits for hydrological model building. Section 4.2.2 introduces catchment and data used in the SST experiment. Section 4.2.3 describes the hydrological model employed in the experiment and model calibration protocol. Section 4.2.4 presents approaches to assessing the relative performance of various split-sample decisions.

4.2.1 Experimental design

This study aims at assessing performance of different SST decisions (interchangeably used with split samples, CSP choices, and data splits in this thesis) defined by both continuous and

discontinuous splitting methods in hydrological model building. The differential split-sample test (DSST), which is widely used for evaluating model robustness loss due to different hydro-climatic conditions in model calibration and validation periods, is not considered as a candidate data splitting method in this study, because our goal is to identify optimal splits for operational hydrological model building and aid model applications in the future period (e.g., streamflow prediction), in which hydro-climatic conditions are usually unknown. The DSST assumption that climatic characteristics of calibration and validation periods are pre-known and play a critical role in the model transferability is out of our current scope. In addition, randomly sampling segments of data series (e.g., years, months, or days) for calibration (e.g., see Arsenault et al., 2018; Kim & Kaluarachchi, 2009; Li et al., 2010; and Perrin et al., 2007) is out of our scope, since we aim at exploring deterministic data splitting methods that are easy to generalize in different cases and convenient to test (random splits require replicates to better generalize performance). Nevertheless, the random sampling method can be valuable in some applications such as Zheng et al. (2023).

We employ the SST evaluation framework proposed by Shen et al. (2022a) in this study, which was presented in Chapter 3. Several key concepts such as model build year, calibration and validation sub-periods, model testing period (i.e., post-validation model application period) can be found in Section 3.2.1. We here underline that *validation* is the process of checking if model performance is adequate in a period/location the model was not calibrated to, and it occurs *before* the model is deployed to answer questions or management problems it was built to support, while *testing* period is not another substitution for validation period. Testing period occurs *after* the model development and is the process of checking how well the model actually performed in the model application period (i.e., testing period). As such, our testing period performance is not an indication of how well the model might do in operation, it is an assessment of how well the model actually did. Therefore, different SST decisions are adopted for model building in the "historical" period prior to the model build year and are then tested in "future" testing period years (after the model build year), which mimics how model building works in operational model development and allows a fair comparison in testing period (see more discussion in Section 2.3.4).

Figure 4-1 illustrates the SST experiments designed with 35 years of available data (1980–2014, which will be described in Section 4.2.2). Five model build years (i.e., 1990, 1995, 2000, 2005, and 2010) are used for building the hydrological model, which leaves 10, 15, 20, 25, and 30 years of available model development data, respectively, for model spin-up, calibration and validation. For each model build year panel in Figure 4-1, totally four categories of SST decisions are defined: continuous calibration sub-period (CSP) (Shen et al., 2022a), discontinuous calibration sub-period (DCSP), a

modified version of DUPLEX (MDUPLEX) (Zheng et al., 2022), and full-period CSP (Shen et al., 2022a). Model build years 1990 and 2010 are two representative scenarios with most limited (i.e., 10 years) and longest (i.e., 30 years) data availability in model building, respectively. To keep SST decisions representative in different build year scenarios and computationally feasible, the above four categories of SST decisions are all applied in 1990 and 2010, while only the DCSP and full-period CSP are applied in the other three build years.

Continuous CSP and full-period CSP are denoted as CSP-x%y_z (where x% is the percentage of available model development data used for calibration, y is an identifier to distinguish different sub-periods with same x, which is skipped for the full-period CSP for brevity, and z is the model build year). We here employ 18 short-period CSPs only for 1990 and 2010 shown in Figure 4-1a and Figure 4-1e. Six of the 18 short-period CSPs use recent data for calibration (i.e., recent CSPs) and the other 12 use older data for calibration (i.e., older CSPs)). The five full-period CSPs are also from Chapter 3 (Shen et al., 2022a) for comparison with other discontinuous splits. More details about these continuous short-period CSPs and full-period CSPs can be found in Shen et al. (2022a) or Section 3.2.1.

Discontinuous splits adopted in this experiment are generated by two methods: (a) Using systematic sampling to produce DCSP splits with data partitioned at the annual scale; and (b) Adopting the MDUPLEX algrithm to split data at the daily scale. These two discontinuous splitting methods are described as follows.

1. Discontinuous splits generated by systematic sampling are denoted as DCSP-x%y_z (where x, y, and z are similarly defined as continuous CSP identifiers). For this 35-year data series, We firstly define a sampling frame with length L-year (L = 2, 3, 4) and thus, the period with a length of N years will be divided into N/L non-overlapping frames (Note that the frame at the tail of the data series may cover years less then L). We then select every kth year (k = 1 to L) in each frame as validation years and the remaining years are for calibration. The resultant DCSP splits have typical lengths 50%, 67% and 75% of data available in model building. For a typical length, multiple DCSP variations can be produced. For example, a 4-year sampling frame will create four different splits denoted as DCSP-75%A, DCSP-75%B, DCSP-75%C and DCSP-75%D in Figure 4-1. We believe these variations are not distinguishable in practice. Thus, we define them as an *equivalent splits* group for DCSP-75%. Each equivalent split is applied independently in hydrological model building but their results are assessed as a group together. In total, DCSP splits in our assessment have three different groups: DCSP-50%, DCSP-67% and DCSP-75%.
2. Discontinuous splits generated by the MDUPLEX algorithm are denoted as MDUPLEX-x%$_z$ (where x and z are similarly defined as full-period CSP identifiers). The MDUPLEX algorithm

proposed by Zheng et al. (2022) is a modified version of the DUPLEX (Snee, 1977), both of which are to generate subsets of the measured data (hydrographs here) with similar statistical properties based on distances (e.g., Euclidian) between all possible pairs of data points (e.g., observed daily streamflow). Both algorithms produce deterministic split or subset for a fixed input data. However, changes in input data, such as different lengths of data or changes in data values, may lead to different splitting results. The key difference between DUPLEX and MDUPLEX is that DUPLEX could produce two subsets with biased statistical properties when it is used to create two subsets with different proportions of data, while the MDUPLEX enables the creation of subsets with unequal sample size while maintaining similar statistical properties (Zheng et al., 2022). In this study, we apply MDUPLEX algorithm for generating three different calibration sub-periods with typical lengths of 50%, 67%, and 75% of data available in model building. Although MDUPLEX algorithm produces deterministic splits, it depends on variable(s) used for calculating the Euclidian distance. In this study, we only use observed daily streamflow at each flow gauge as the input variable for MDUPLEX and thus, the resultant calibration/validation sub-periods are specific to gauges and model build years. In total, there are 2,778 different MDUPLEX splits produced (i.e., 463 catchments × 2 build years × 3 typical lengths, where catchment data are described in Section 4.2.2). More details about the DUPLEX and MDUPLEX algrithms can be found in Snee (1977) and Zheng et al. (2022).

In total, we adopt 44 different SST decisions in this experiment, including 18 CSP decisions, 15 DCSP decisions (same-length equivalent splits are viewed as one decision), 6 MDUPLEX decisions, and 5 full-period CSP decisions. Note that the discontinuous splitting methods sample data across the full coverage of model building period. Thus, DCSP and MDUPLEX splits can be considered to be equal to recent CSPs and full-period CSPs with respect to the data recentness (i.e., DCSP, MDUPLEX, recent CSPs and full-period CSPs all have a 100% recency score as defined in Chapter 3 (Shen et al., 2022a) indicating the most-recent data are used in calibration, while older CSPs have recency scores such as 30%, 50% and 80% indicating different proportion of recent data are used in calibration).

Based on the aforementioned 44 SST decisions, we can build and test hydrological models (the model and calibration protocol will be introduced in Section 4.2.3). Here, we introduce how models are developed (i.e., initialized, calibrated, validated) and deployed (i.e., tested). Unlike Chapter 3 (Shen et al., 2022a) where models were simulated in calendar years (period of 1 January to 31 December), we build models in hydrological years (period of 1 October to the next 30 September) in this study. This is to minimize the impact of splitting a single snow season in half when modeling discontinuous

years. Consider any row in any of the panels in Figure 4-1, the first hydrological year of available data (1 October 1980 to 30 September 1981) is always used for model spin-up. We use 1980 data recursively for three times to define a "three-year" spin-up period to initialize the hydrological model (i.e., force models with meteorological inputs in 1980 and repeatedly run the model in 1980 for three times with the end-of-day states on 30 September 1981 in the first 1980-run being the initial states on 1 October 1980 in the second 1980-run, and so forth). This initialization approach is adjusted relative to Shen et al. (2022a) in Chapter 3. Note that in testing phase, model simulation always starts from 1 January to 31 December, as testing periods serve as common periods to compare different SST decisions and hence are not influenced by either hydrological or calendar year configuration.

The actual model run period in model optimization starts at the beginning of the 3-year spin-up period to either the end of the calibration period (for continuous CSP splits) or the end of model building period (for DCSP and MDUPLEX splits). And then only the simulations in calibration period will be used for performance criteria calculations in optimization. The best calibrated parameter set in each optimization trial (identified using the calibration protocol in Section 4.2.3) is used to simulate the model starting with the three-year spin-up period and ending at the end of 2014 (37-year simulation). This entire set of simulation result time series is then appropriately post-processed to compute the various calibration, validation and model testing period performance metrics. This ensures the model initialization processes and states updating are completely consistent across different sub-periods.

The full testing period for each model build year is shown as red blocks in each panel of Figure 4-1. Each of those five full testing periods are augmented with four additional shorter length testing periods (i.e., the first three and first five years of the entire testing period, the last three and last five years of the entire testing period, and the entire testing period. Example of 1990 is illustrated in Figure 4-1a) to consider different hydro-climatic conditions and enlarge dissimilarities between each split and its testing periods, thereby enabling a more robust hypothesis test. Note that testing models in last three or last five years of testing period (i.e., "Testing3" and "Testing4" in Figure 1a) naturally produces a gap between model building and model testing (e.g., at least a 20-year gap period in 1990 in Figure 4-1a), thus allowing us to assess if our findings by immediately applying models after calibration and validation may drastically change when the model is not immediately tested after model building. This actually explores how climate change in the gap period, if any, may impact the model testing performance after waiting for a while it has been built. Building models in 1990 is the best scenario for this check as the gap period is long enough (i.e., 20 years). In total, there are 23 different testing periods

for the five model build years (i.e., 5 build years × 5 testing periods per build year - 2 repeated testing period in 2010).

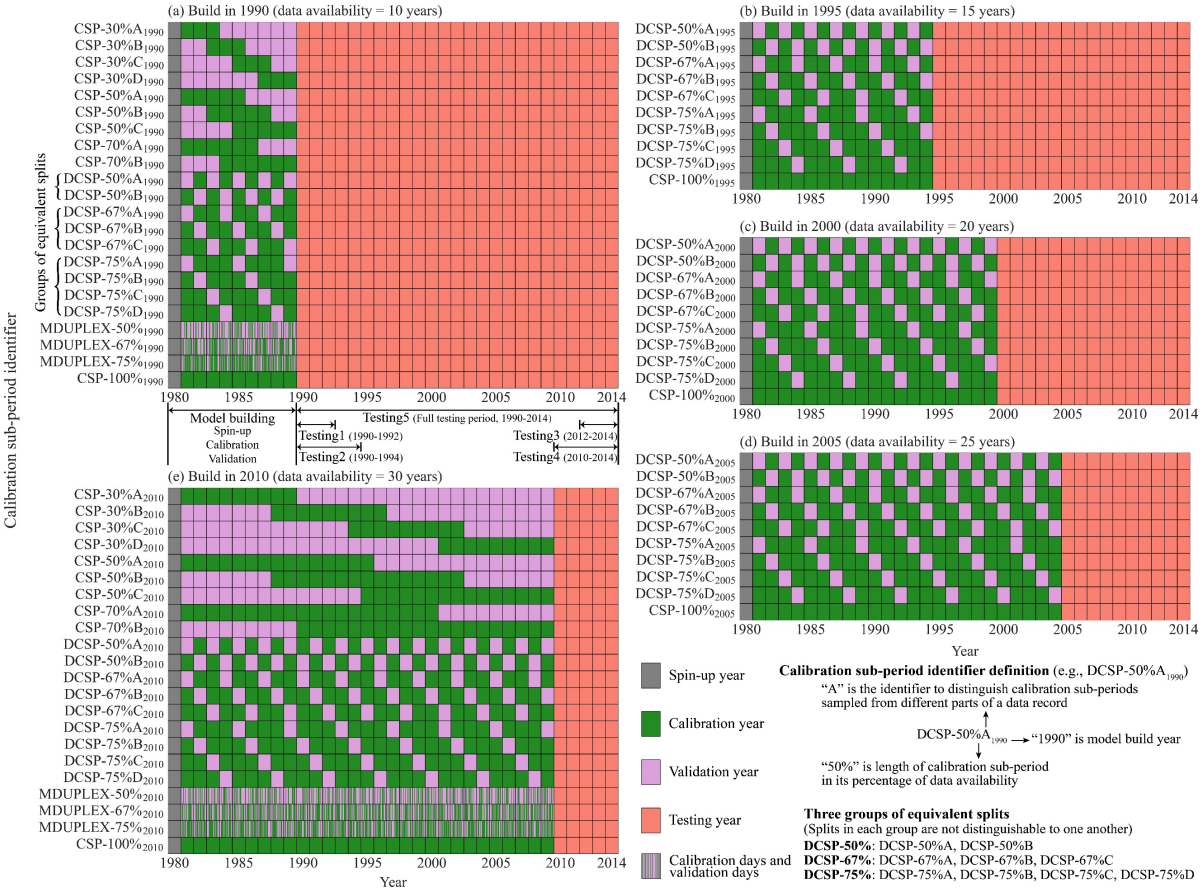

Figure 4-1. Experimental design for split-sample test (SST) assessment with model built in (a) 1990, (b) 1995, (c) 2000, (d) 2005, and (e) 2010. Four categories of SST decisions are adopted: continuous calibration sub-period (CSP), discontinuous calibration sub-period (DCSP), MDUPLEX, and full-period CSP. All these four categories of SST decisions are adopted in (a) 1990 and (e) 2010, while only DCSP and full-period CSP are adopted in the remaining three build years. Continuous splits and full-period CSP are denoted as CSP-x%y_z (where x% is the percentage of calibration data in available data for model building, y is an identifier to distinguish different sub-periods with same x, which is skipped for the full-period CSP for brevity, and z is the model build year, which may be skipped hereafter if its meaning is clear in the context). DCSP splits in each of the panels are represented as their equivalent splits denoted as DCSP-x%y_z (where x, y, and z are similarly defined as continuous CSP identifiers). In total, there are three groups of DCSP splits in the results assessment: DCSP-50%, DCSP-67% and DCSP-75%. MDUPLEX splits are denoted as MDUPLEX-x%$_z$ (where x and z are similarly defined as continuous CSP identifiers and y is not used here since MDUPLEX splits is deterministic for a fixed dataset). Each row of MDUPLEX splits in (a) and (e) is to conceptually show

the split is produced at the daily scale, which differs from other annual scale-based splits in other rows but note that MDUPLEX splits are gauge and build year specific. The year information of all these identifiers subscript may be ignored hereafter if its meaning is clear in the specific context.

4.2.2 Catchments and data

The Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) data set are used in this study, which provides 671 catchments that are minimally impacted by human activities across the contiguous United States (CONUS; Addor et al., 2017; Newman et al., 2015). We used the 463-catchment subset of CAMELS deliberately selected and processed in Chapter 3 (Shen et al., 2022a). The Daymet forcings (archived in the CAMELS data set) and the observed daily streamflow data at these 463 gauges (originally archived in the CAMELS and missing data infilled using newer streamflow data from the National Water Information System of the USGS by Shen et al. (2022b)) are used in this study, both spanning from 1 January 1980 to 31 December 2014. More details on gauge selection can be found in Shen et al. (2022a) or Section 3.2.2.

The map for the spatial locations of all CAMELS catchments (including the 463 selected catchments and other filtered catchments) is presented in Figure A1-1 in Appendices A-1. The detailed information of the 463 catchments and the corresponding Daymet forcings and updated USGS streamflow data files for these catchments are all available online (see Appendices A-2).

4.2.3 Hydrological models and calibration protocol

We employ a conceptual lumped hydrological model, i.e., the HMETS model (which stands for Hydrological Model of École de technologie supérieure), to test the SST alternatives described in Section 4.2.1. The HMETS model is a lumped hydrological model using two buckets to simulate water recharge/discharge in vadose zone and saturated zone (Martel et al., 2017). HMETS has been proved to perform well against two other lumped models over 320 catchments in the US from the Model Parameters Estimation Experiment (MOPEX) database (Martel et al., 2017). Chlumsky et al. (2021) compared 111 different models including HMETS across 12 MOPEX catchments and showed that the HMETS model tends to rank within the top quartile for almost all catchments. Chapter 3 (Shen et al., 2022a) also showed HMETS generally outperformed the 6-parameter GR4J model (which stands for du Génie Rural à 4 paramètre Journaliers; Perrin et al., 2003; Valéry et al., 2014) when comparing expected or median Kling-Gupta Efficiency (KGE) in testing periods across 463 catchments (see Figure 3-4 and Figure 3-5 in Section 3.3).

Key hydrological processes simulated in HMETS are snow accumulation, melting and refreezing, evapotranspiration, infiltration, and flow routing. It requires daily precipitation, minimum

and maximum air temperature. The HMETS has 21 parameters and all of them are calibrated in this study. Details of HMETS model structure and parameters can be found in Martel et al. (2017). Definition of parameters and their ranges can be referred to Chapter 3 (Shen et al., 2022a). In this study, the HMETS is implemented in the Raven hydrological modeling framework, which is a robust and highly generalized object-oriented flexible modeling framework platform (Craig et al., 2020). Raven provides many widely used numerical algorithms for hydrological processes, and it accepts unified model input files, which makes it an ideal tool for various types of modeling investigation such as multi-model intercomparison (e.g., see Mai, Shen, et al., 2022). Details on the Raven framework can be found in Craig et al. (2020) and the Raven manual (Craig, 2023).

In this study, HMETS is calibrated to the 44 splits at each of the 463 CAMELS catchments. The dynamically dimensioned search (DDS) algorithm (Tolson & Shoemaker, 2007), which has been widely applied in HMETS calibration studies (Chlumsky et al., 2021; Mai, Shen, et al., 2022; Martel et al., 2017; Shen et al., 2022a), is employed in this calibration experiment. DDS is implemented in the optimization and calibration software toolkit OSTRICH (Matott, 2017). We utilize a budget of 3,000 model evaluations per optimization trial. HMETS with each of the 44 splits is calibrated to 20 independent optimization trials, and different randomly generated initial parameter sets are sampled in each trial to minimize the influence of initial conditions on DDS. Note that since the three DCSP splits (i.e., DCSP-50%, DCSP-67%, and DCSP-75%) corresponds to multiple equivalent splits as introduced in Section 4.2.1, we require that the total optimization trails performed for a group of equivalent splits be at least 20, thus leaving a minimal trial number of 10, 7, and 5 for DCSP-50% (2 equivalent splits), DCSP-67% (3 equivalent splits), and DCSP-75% (4 equivalent splits), respectively. This requirement effectively reduces computational costs for DCSP splits and maintains the sample size of optimization trials consistent with other splits in performance assessment. Note that each split has 20 independent optimization trials at each catchment except that DCSP-67% split has 21 trials. This one extra trial of DCSP-67% split will be further post-processed (either maintaining it or removing it) in the assessments presented in Section 4.2.4.

Accordingly, the total number of model calibration problems solved with DDS is 409,755 (39 splits × 20 trials × 463 catchments + 5 DCSP-67% splits × 21 trials × 463 catchments), and the total number of model testing hydrographs assessed is 1,751,529 (409,755 trials × 5 testing periods – repeated testing periods in build year 2010 [(15 splits × 20 trials × 463 catchments + 1 DCSP-67% split × 21 trials × 463 catchments) × 2 testing periods]).

The model is calibrated, validated and tested using the 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 (see Equation (3-1)). The KGE value ranges from -∞ to 1.0, and KGE = 1.0 indicates the perfect agreement between simulations and observations. We continue using KGE in this study, as such our findings can be compared with the previous study in Shen et al. (2022a) (see Chapter 3). Other alternative criteria may be adopted in this evaluation framework according to some specific modeling objectives (e.g., low flows), but an extensive user survey by Gauch et al. (2022) suggested that KGE is the most important metric when rating the overall hydrograph and high flows.

4.2.4 SST comparative performance assessment

This section introduces methodology applied for assessing HMETS model performance of the 44 SST decisions based on the large-sample calibration, validation, and testing results. When assessing those alternative splits across large-sample catchments, we believe the performance in model testing periods instead of validation should be of interest for an objective comparison. Shen et al. (2022a) demonstrated three typical ways to approach this comparison (see in Chapter 3) including pairwise comparisons between short-period CSPs and their corresponding full-period CSP, central tendency of KGE in testing periods and frequency of a CSP correctly classifying model testing period failure and success, which are further adapted to this study. A key adjustment in this study is using all model optimization trials in the assessment. Considering multiple replicates per split-sample decision (in a given model build year and catchment) enables different assessments relative to Chapter 3 and should enhance robustness of conclusions. Four assessments applied are outlined as follows:

1. Rank of splits at each catchment based on raw KGE (i.e., KGE before failure handling) in testing periods.
2. Pairwise comparison of any two splits at each catchment in testing period (i.e., KGE after failure handling).
3. Ability of splits to function as accurate binary classifiers for predicting model testing states (model failure versus model success) from model building (calibration/validation) states.
4. A multi-objective analysis to simultaneously optimize median KGE (used in Assessment 2) and classifier performance metrics (used in Assessment 3) to identify the optimal SST decision.

We apply two model failure handling strategies in these assessments: (a) As most comparative SST modeling studies do, we skip handling failures and proceed to use all model calibration and validation results for performance analysis in testing periods, and (b) We identify model failures in model calibration and validation phases and discard those failed models. In model failure identification, we employ a reference model as the lowest acceptable level, which is based on reference climatology (i.e., reference flow) established by calculating the mean value of observed streamflow on the reference

period at the daily scale (Knoben et al., 2020). As introduced in Chapter 3 (Shen et al., 2022a), reference period for model calibration, validation and testing periods is kept constant as all data years prior to the model build year (i.e., spin-up, calibration and validation). More details on calculating reference KGE, which is the observation-based metric resulted from the reference flow and observed flow, are described in Shen et al. (2022a) (see in Section 3.2.5.1).

Assessment 1 applies failure handling strategy (a), which does not handle model failures and hence reproduces a typical analysis approach, is an exploratory analysis to show our initial assessment on different splits excluding the influence of handling model failures.

Assessments 2 and 4 both adopt failure handling strategy (b) to minimize the impact of inadequate model building instances. Since these two assessments both focus on the performance metric KGE, we require that any failed model building instance use the reference flow for predicting hydrographs in testing periods. Thus, the sample size in testing periods will not be reduced due to the failure handling.

Assessment 3, however, by definition of the binary classifier (see in Section 4.2.4.3), applies strategy (a) such that all raw simulations in testing periods (no matter the failure/success state assessed in model building) are used to check testing period states. Also note that in this assessment, more practical options of the threshold level for failure identification are considered.

It is also important to note that assessments 1 and 2 require the trial sample size of each split to be equal. We therefore remove 1 out of 21 trials of DCSP-67% splits for each model (i.e., per catchment per model build year) following the rule of thumb that we remove the trial whose testing KGE is closest to the median of those 21-trial KGE values, which may minimize the impacts on the KGE distribution. Assessments 3 and 4 are not sensitive to this equal-sample requirement, thus we will maintain using all 21 trials for DCSP-67% splits in these two analyses.

4.2.4.1 Rank splits by raw KGE in model testing period

This is an exploratory analysis on the raw model testing results with model failure handling strategy (a) described in the leading parts of Section 4.2.4, which excludes the influence of handling model failures. For a given model build year and a testing period, we rank the testing KGE of different splits at each of the 463 catchments. Note that it is required that the optimization trial sample size of each split be equal. Thus, 1 out of 21 trials of DCSP-67% splits is removed from the analysis (see Section 4.2.4). Then, we have 320 trials (16 splits × 20 trials) in 1990 and 2010 and 80 trials (4 splits × 20 trials) in 1995, 2000 and 2005 per catchment in the ranking analysis. Afterwards, we count the number of trials of each split ranking in the best/worst 20% at each catchment and aggregated these

counts across all 463 catchments. It is expected that a good split will have a large count number in the best 20% and a small count number in the worst 20% in the ranking. See Section 4.3.1 for the results.

4.2.4.2 Pairwise comparison of splits in model testing period

A pairwise comparison quantifies the difference between any two splits in the testing period KGE. Chapter 3 (Shen et al., 2022a) assessed pairwise splits with their best calibrated trials in model testing periods, however, as we assess all calibration trials, the pairwise comparison method used in Chapter 3 (Shen et al., 2022a) is not suitable. In this study, we evaluate whether the medians of 20-trial testing KGEs of any two splits show significant statistical differences. In this assessment, model failures are handled with strategy (b) described in Section 4.2.4. It is also required that the trial sample size of each split be equal. Thus, we remove 1 out of 21 trials of DCSP-67% splits from the analysis (see Section 4.2.4).

We employ the nonparametric Wilcoxon rank-sum test (also named as Mann-Whitney U rank test)to assess whether one group tends to produce larger observations than the second group and is presented as a test for difference in group medians (Helsel et al., 2020; Mann & Whitney, 1947; Wilcoxon, 1992), in order to evaluate whether there is any significant difference in the testing KGE medians of any two splits at each of the 463 catchments. Given we have 20-trial model testing results of any two splits X and Y (X and Y denote two SST decisions represented in Figure 4-1) per catchment per model build year per testing year, the null hypothesis can be expressed as the probability of an x value (i.e., testing KGE of X) being larger or smaller than any given y value (i.e., testing KGE of Y) is 0.5. The alternative hypothesis is the above probability is not 0.5. If the null hypothesis is rejected, medians of the 20-trial of X and Y are checked to determine which one is larger. Details of the Wilcoxon rank-sum test computation can be found in Helsel et al. (2020). This test has been implemented in the open-source SciPy library in Python (Virtanen et al., 2020).

In total, we perform 240 pairwise tests for model build years 1990 and 2010 (16 trials × 15 trials) and 12 pairwise tests for model built in 1995, 2000 and 2005 (4 trials × 3 trials) per testing period at each of the 463 catchments all excluding same-split pairs. We then count the number of catchments if the pairwise comparison shows significance (significance level $\alpha = 0.05$ used in this study). This leads to two different counts: One is count of catchments showing X median significantly larger than Y median ($M_{X>Y}$), and the other is count of catchments showing Y median significantly larger than X median ($M_{Y>X}$). The difference of these two counts, i.e., the net count ($M_{X>Y} - M_{Y>X}$), represents how much better the performance of split X is relative to split Y (negative values show split Y is better than split X). This net count can be transformed into net percentage by dividing the total count of catchments, represented in 16 × 16 (1990 and 2010) or 4 × 4 (1995, 2000 and 2005) matrices. A positive net

percentage shows split X being better than split Y over the 463 catchments. See Section 4.3.2 for the results.

4.2.4.3 Performance of splits as binary classifiers

In this assessment, we view each SST decision as a binary classifier which yields only two types of states, i.e., model is adequate (success) or inadequate (failure) (Shen et al., 2022a). Binary classifiers predict whether the model building is adequate (success) or inadequate (failure) for testing period prediction. Further, we can actually assess whether the model building is truly adequate or truly inadequate in the testing period. Note that we use 21 trials for DCSP-67% splits, because the impact of the unequal sample size is negligible in this assessment.

A two-by-two confusion matrix is a classic way to assess the results of a binary classifier against known states of nature, and it has been adopted to assess different SST decisions for their ability to classify model states in building and testing phases (Shen et al., 2022a). Four possible classes of our hydrological model building (i.e., calibration plus validation) and testing states are defined and are consistent with Shen et al. (2022a) (see in Section 3.2.5.4). Note that we define "positive" as the not-normal class, which is a model failure, while "negative" represents normality, which is a model success.

We further employ two metrics to interpret these four classes in our hydrological modeling context that we modelers would care about (a) how accurate a model can predict the hydrograph in both building and testing and (b) how often the model may be failed and what the consequences are. The two metrics are as follows:

1. Classification accuracy. Accuracy is the is the ratio of all correctly classified model "building-testing" instances divided by the total count of instances (see Equation (3-3)). Accuracy reduces performance to a single metric, and the value 1.0 indicates the classifier is perfect.
2. Fractions FN and FP as expressed in Equation (4-1). FN and FP instances are both undesirable in practice. The cost of FN instances is that a model deemed as a success after model building is actually a failure in the testing period,. A FP instance is when model building process assesses a model as a failure but in the testing period, the model is actually a success. The cost here, using a strict definition of failure, is that after model building, modelers incorrectly believe they are left without an acceptable model to use in the model application period. . The tradeoff between FNF and FPF reveals relative costs of different splits if they fail in model building and testing, and the ideal values of both fractions are 0.

$$FNF = \frac{FN}{TP+FP+FN+TN}, FPF = \frac{FP}{TP+FP+FN+TN} \quad (4\text{-}1)$$

where TP, TN, FN, and FP are the counts of catchments classified into these four categories defined in Section 3.2.5.4.

It should be noted that in our confusion matrix assessment, the success/failure state of testing periods are dependent on the threshold chosen for classification. This is unlike the conventional applications of confusion matrix, such as in medical diagnosis, a patient being healthy or unhealthy is a binary state. In our hydrological modeling cases, model performance depicted by KGE may range from -∞ to 1 instead of binary states (e.g., 0 and 1) and thus, we need to define a KGE threshold (i.e., an acceptable level), above which the testing results can be deemed as adequate (a success) and otherwise inadequate (a failure).

If this KGE threshold varies, the testing states and the classification results may change accordingly. However, the magnitude of the KGE threshold influencing the model building and testing states is not fully investigated yet. Chapter 3 (Shen et al., 2022a) used reference KGE to distinguish model success and failures in building and testing phases. However, other alternative thresholds such as a constant level may also be feasible (e.g., see Knoben, Freer, & Woods, 2019; Moriasi et al., 2015; and Ritter & Muñoz-Carpena, 2013). Therefore, we employ variable KGE thresholds in this study to explore if different thresholds used in model failure handling would substantially change the optimal SST decisions.

Rationale of using different KGE thresholds to benchmark modeling results over different phases (i.e., calibration, validation, and testing) are that: (a) In practical model building, it is generally expected a degradation would appear in model validation relative to calibration performance, while a significant performance degradation between calibration and validation may raise flags of overfitting (Arsenault et al., 2018); and (b) Past large-sample studies utilizing reference KGE as the lower bound of model simulations generally showed this benchmark is easy to achieve in most catchments (e.g., see Knoben et al., 2020; Shen et al., 2022a; and Towler et al., 2023). For example, the medians of reference KGE over the 463 catchments in build year 1990 are all around 0.2, 0.2, and 0 over the calibration, validation, and testing periods, respectively. And these medians are very close to zero in build year 2010 over the calibration, validation, and testing periods. Modelers may expect a higher level the model can achieve over reference KGE in practice (e.g., see Mai et al. (2022) where they thoughtfully argue that in general, a KGE less than 0.48 would be considered poor). Therefore, we propose the following two approaches to classifying model performance in building and testing phases:

1. Reference KGE-based thresholds. We add an additional performance expectation (i.e., Δ in Eq.4) in model calibration during the model failure identification, which is in accord with model building in practice. And the three KGE thresholds are defined as follows:

$$\begin{aligned} \text{Threshold}_{\text{cal}} &= \text{KGE}_{\text{cal}}^{\text{ref}} + \Delta \\ \text{Threshold}_{\text{val}} &= \text{KGE}_{\text{val}}^{\text{ref}} \\ \text{Threshold}_{\text{test}} &= \text{KGE}_{\text{test}}^{\text{ref}} \end{aligned} \tag{4-2}$$

where subscripts "cal", "val", and "test" indicate calibration, validation, and testing period, respectively, and KGE with superscript "ref" denotes reference KGE. Δ indicates how much calibration performance in KGE units is expected to exceed the calibration period reference KGE. In this assessment, we test Δ with values 0, 0.1, 0.2, 0.3 and 0.4. We also limit the lower and upper bounds of calibration thresholds to -0.41 and 0.8, respectively. The lower bound threshold -0.41 stands for the KGE by using mean observed flow in prediction (Knoben, Freer, & Woods, 2019). The maximum reference KGE values of our 44 different splits over the 463 catchments are around 0.8, which is thus set as the upper bound of calibration thresholds.

2. Constant KGE-based thresholds. As described in (1), we vary threshold in calibration by adding to its reference KGE, but thresholds in validation and testing are still the corresponding reference KGE without any adjustment. In this assessment, we use constant KGE levels (i.e., 0, 0.1, 0.2, 0.3 and 0.4) as thresholds, whilst adding up with a fixed Δ (i.e., 0.2) in calibration.

$$\begin{aligned} \text{Threshold}_{\text{cal}} &= \text{KGE}_{\text{const}} + \Delta \\ \text{Threshold}_{\text{val}} &= \text{KGE}_{\text{const}} \\ \text{Threshold}_{\text{test}} &= \text{KGE}_{\text{const}} \end{aligned} \tag{4-3}$$

where subscripts "cal", "val", and "test" indicate calibration, validation, and testing period, respectively, and KGE with subscript "const" denotes a constant KGE (either 0, 0.1, 0.2, 0.3 or 0.4). Δ is a fixed value of 0.2 KGE units. Although various Δ values can be assessed, we skip reporting more here for brevity, as changes in this variable are not expected to influence the overall results pattern between the 44 split-sample decisions.

4.2.4.4 Multi-objective assessment of splits considering median KGE and binary classification metrics in testing period

In this assessment, we frame the model building processes as a multi-objective decision-making problem to simultaneously optimize two objectives: Objective one is to maximize the testing period KGE quantified as median KGE (consistent with median KGE used in Section 4.2.4.2), and

objective two is to maximize the performance of splits functioning as binary classifiers quantified as classification accuracy introduced in Section 4.2.4.3. Two different tradeoffs are considered:

1. Tradeoff between median KGE and classification accuracy using all optimization trials (20 or 21) of each split-sample decision at *each* of the 463 catchments. Each SST decision contains 20 or 21 data points for both median KGE and accuracy calculation.
2. Tradeoff between median KGE and classification accuracy using all optimization trials (20 or 21) of each SST decision *aggregated across all* 463 catchments. Each SST decision contains 9260 or 9723 (20 or 21 × 463) data points for median KGE and accuracy calculation.

Tradeoff 1 reveals how each split may perform differently at a single catchment. Single gauge-based tradeoff results need to be further aggregated across 463 gauges to show the overall performance of different splits. We define a simple metric percent distance to depict the degree of each solution approaching the optimal values in their tradeoff analysis. Percent distance (PD) is formulated as follows and each variable can be referred to Figure 4-6a in Section 4.3.4. The percent distance ranges from 0 to 100%, and 100% is the perfect value that denotes the solution on average ranks the highest frequency to be the optimal one among all solutions.

$$PD_{\text{median KGE}} = \frac{x_p - x_n}{x_i - x_n} \times 100\% = \frac{d_1}{d_1 + d_2} \times 100\% \tag{4-4}$$

$$PD_{\text{accuracy}} = \frac{y_p - y_n}{y_i - y_n} \times 100\% = \frac{d_3}{d_3 + d_4} \times 100\% \tag{4-5}$$

where (x_p, y_p) is a sample solution P in the coordinates of median KGE versus accuracy, p indicates any of solutions in the tradeoff.

In Figure 4-6a, we can identify two feature points: ideal point I (x_i, y_i) and nadir point N (x_n, y_n). The ideal point has the optimal (largest) values of both median KGE and accuracy identified from all sample solutions, while the nadir point is the worst point with lowest values of both metrics. However, it should be noted that median KGE of a split in some catchments can be even lower than -0.41, which is the KGE value when using mean observed flow to make predictions (Knoben, Freer, & Woods, 2019). For these catchments, we limit the nadir point to be -0.41 with respect to the median KGE, while the worst accuracy can be 0. Thus, coordinates the of two feature points are as follows:

$$x_i = \max\left(x_1, x_2, ..., x_p, ...\right),\ y_i = \max\left(y_1, y_2, ..., y_p, ...\right) \tag{4-6}$$

$$x_n = \max\left\{\min\left(x_1, x_2, ..., x_p, ...\right), -0.41\right\},\ y_i = \max\left\{\min\left(y_1, y_2, ..., y_p, ...\right), 0\right\} \tag{4-7}$$

Note that solutions with median KGE lower than nadir point, the PD is set as 0. In this analysis, we have 1389 tradeoffs (463 catchments × 3 testing periods) in 2010 and 2315 tradeoffs (463 catchments × 5 testing periods) in each of the other four build years. We further aggregate the percent distance metrics by averaging them across catchments and testing periods, and then evaluate the averaged percent distances of different splits. See Section 4.3.4 for tradeoff 1 results.

Tradeoff 2 performs only one tradeoff per build year per testing period, which directly assesses the overall performance of splits across all gauges and is consistent with the method used in Chapter 3 (Shen et al., 2022a). We also assessed tradeoff 2 in this study and the results will be reported in the Appendices A-6.

4.3 Results

4.3.1 Ranking of splits: Frequency of them being the best/worst splits in model testing

Ranking the raw model testing results of each data split enables us to assess their relative performance. Figure 4-2 displays the count of calibration trials that rank in the best/worst 20% in the first 5 years of testing period. The results for the other four testing periods show similar patterns and hence are not presented here for brevity.

Figure 4-2 shows an overall pattern that the full-period CSP presents comparably good rankings at the best 20% tail (top row panels in Figure 4-2) except the three recent CSPs outperform others notably in 2010 (Figure 4-2e1), while full-period CSP consistently rank the least at the worst 20% tail in all build years (bottom row panels in Figure 4-2), which indicates its superiority in this comparison. The three recent CSPs (i.e., CSP-30%D_{2010}, CSP-50%C_{2010} and CSP-70%B_{2010}) possess the largest quartiles in best-rank counts (Figure 4-2e1), and they could even be triple the count of their corresponding older CSPs (e.g., CSP-30%D_{2010} versus CSP-30%A_{2010}). However, it is also worth noting that the full-period CSP has smallest worst-rank count in 2010 (Figure 4-2e2), where those recent CSPs can be up to 4.5 time as many as the count of the full-period CSP, which indicates recent CSPs results can be more extreme in the two tails of ranking.

Figure 4-2 also shows the ranking results, especially at the best 20% tail, vary with model build years (i.e., data availability). The three recent CSPs dominate in the best 20% ranking in 2010 (Figure 4-2e1) when model is built with 30 years of data available, but this prevalence is not seen in 1990 (Figure 4-2a1) when model is built with only 10 years of data available. DCSP and MDUPLEX splits generally are comparable to the full-period CSPs in the best 20% rankings. However, for the worst 20% ranking results, the DCSP and MDUPLEX splits are nearly double (1.2–2.8 times) the count of

full-period CSP on average in 2010 (Figure 4-2e2). These differences are even amplified in 1990 (Figure 4-2a2), where the worst 20% counts of DCSP and MDUPLEX splits are on average 2.4 times (1.3–4 times) as many as the count of full-period CSP.

The full-period CSPs show overall robust performance in all five model build years and all five testing periods, but considering model failures are not handled in this initial assessment, further assessment in next sections is required.

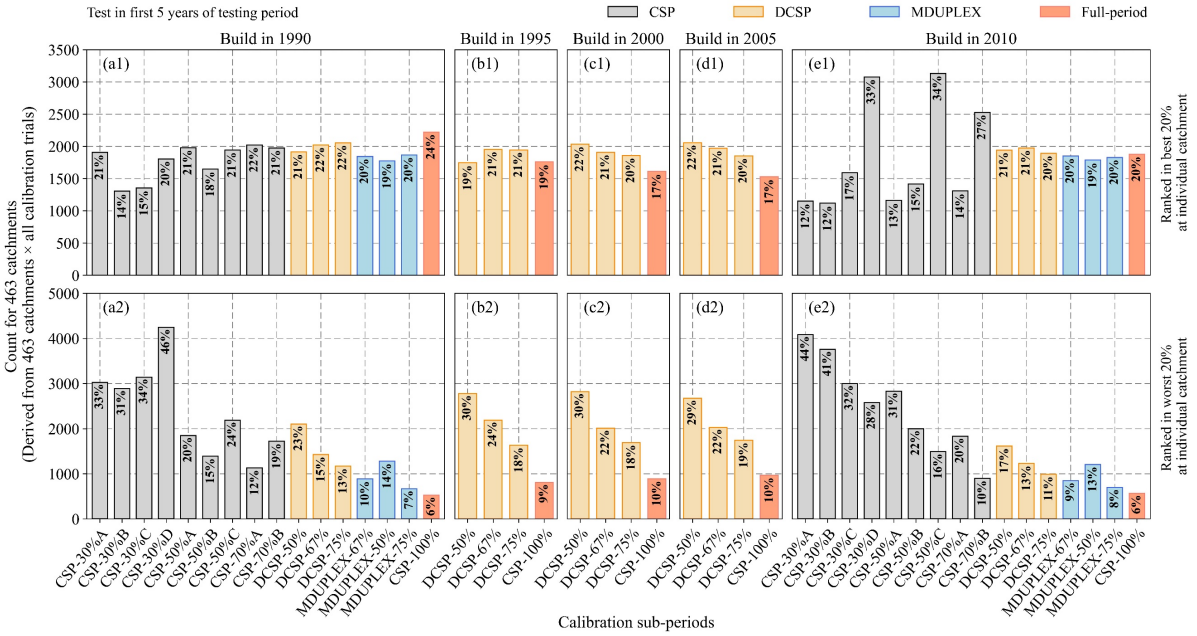

Figure 4-2. Count of each data split ranking (a1, b1, c1, d1, e1) in the best 20% and (a2, b2, c2, d2, e2) in the worst 20% of results in five model build years during the first 5 years of testing period. Note that the raw model testing KGE values are used in this ranking, i.e., no failure handling strategy applied. Each ranking analysis contains 320 (16 splits × 20 trials) KGE samples for models built in 1990 and 2010 and 80 (4 splits × 20 trials) KGE samples for models built in 1995, 2000 and 2005. Bars denote the total count of how many times a split being the best/worst 20% in 463 catchments. Number in each bar is the proportion of the best/worst 20% trials in total trials of each data split. Bar colors are to distinguish the four splitting categories: continuous calibration sub-period (CSP), discontinuous calibration sub-period (DCSP), modified DUPLEX (MDUPLEX), and full-period CSP. The data split identifiers (*x*-axis) are defined in Figure 4-1.

4.3.2 Wilcoxon rank-sum test: Pairwise comparison of any two splits' KGE medians in testing period

In this assessment, we perform the Wilcoxon rank-sum test between any two data splits at individual catchments using all their optimization trials (i.e., 20 trials). We then aggregate the test results for all catchments reported as the metric net percentage (see Section 4.2.4.2). In this assessment, model failures in calibration/validation will trigger a decision to discard the model in testing and instead use reference flows for testing periods. Figure 4-3 demonstrates how the Wilcoxon rank-sum test is performed at a single catchment and then aggregated across the 463 catchments to quantify how often a split X significantly outperforms another split Y.

Figure 4-3a shows the empirical cumulative distribution functions (ECDFs) of the 20-trial's KGE at the example gauge 01013500. It is not always straightforward to directly compare the differences between two ECDFs if they seem similar, while the Wilcoxon rank-sum test can aid in quantifying those differences with respect to the medians. Figure 4-3b presents the test results at gauge 01013500, and such test results are aggregated across all 463 gauges in Figure 4-3c. It is noted that in Figure 4-3c, any two blocks that are symmetric to the diagonal line indicate count of X being significantly better than split Y ($C_{X>Y}$) and count of Y being significantly better than X ($C_{Y>X}$) in the pairwise comparisons of (X, Y) and (Y, X), respectively. For example, the count of CSP-30A% significantly outperforming CSP-100% is 76 gauges, while the count of CSP-100% significantly outperforming CSP-30%A is 247 gauges in Figure 4-3c and thus, no significant differences are tested between the two splits in the remaining 140 gauges. Figure 4-3d shows the metric net percentage (i.e., $(C_{X>Y} - C_{Y>X}) / 463 \times 100\%$) transformed from Figure 4-3c.

Figure 4-3d provides a comprehensive screening of the overall test results. Viewing X as the nine short-period CSPs and Y as the six DCSP and MDUPLEX splits, which leaves us 54 pairwise comparisons, we can obtain a few useful scores that short-period CSPs only slightly outperform discontinuous splits in 8/54 (15%) instances, while in 16/54 (30%) discontinuous splits outperform these CSPs at more than 25% gauges. It is notable that CSP-30%D (shortest recent CSP) appears to be the worst split in Figure 4-3d, as its net percentage values are smaller than -30% in 12 out of 15 instances. Recall that CSP-30%D ranks the 1st at the worst 20% tail in Figure 4-2a2, which somehow explains its inferiority in Figure 4-3d. However, it should also be noted that recent CSPs perform much better than the older CSPs and can be comparable to the full-period CSP with respect to the net percentage in model build year 2010 (see in Figure A5-1 in Appendices A5).

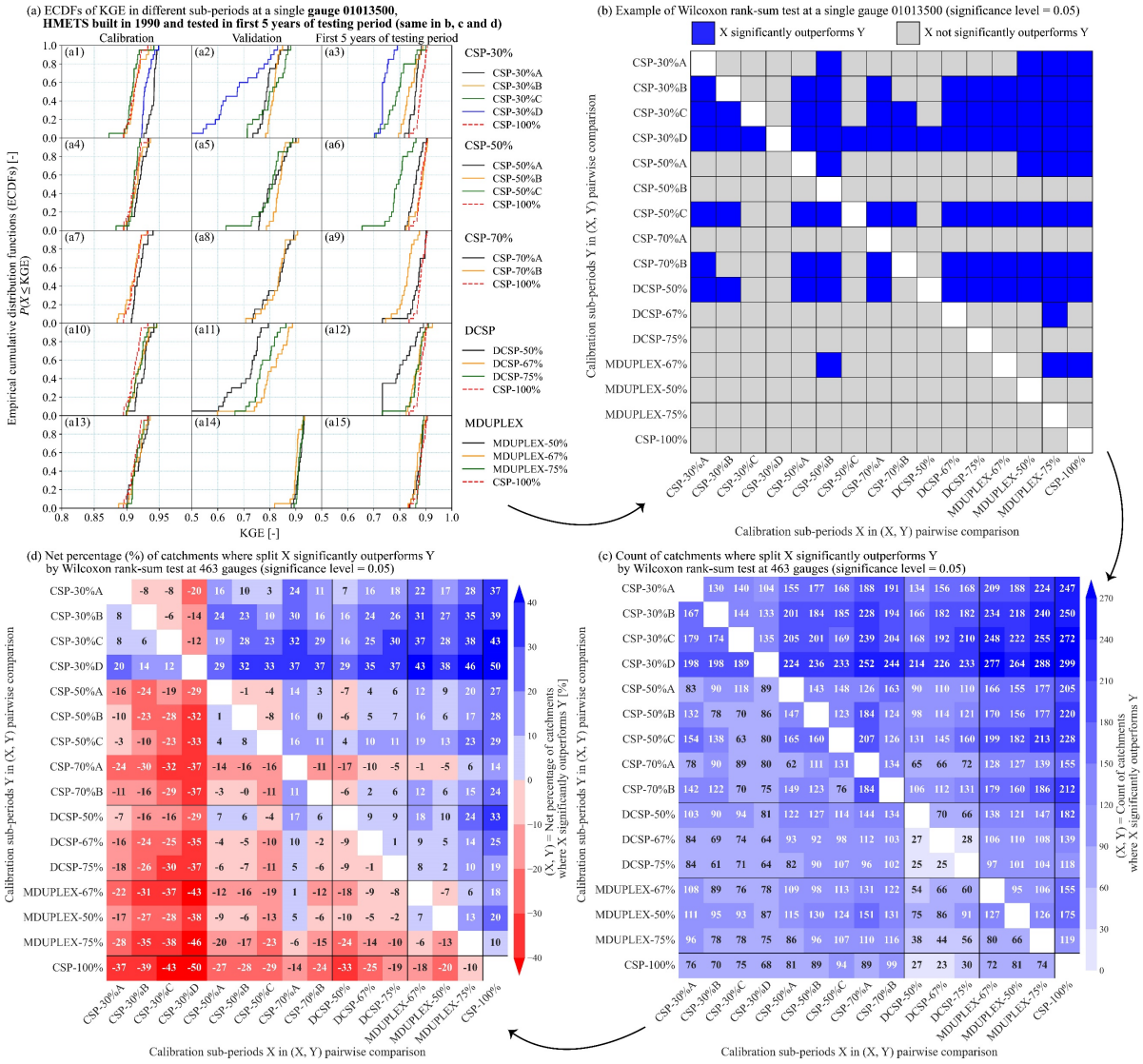

Figure 4-3. Demonstration of the pairwise comparison of 16 data splits based on the Wilcoxon rank-sum test (significance level $\alpha = 0.05$) in model build year 1990 during the first 5 years of model testing period. Note that model failures are handled in this analysis by discarding failed models in model building and instead using reference flow for testing periods prediction. The empirical cumulative distribution functions (ECDFs) for the calibration, validation, and testing period KGE are presented in (a) at an example gauge 01013500. The 16 splits in (a) constitute 240 pairs of splits (X, Y) for Wilcoxon rank-sum tests, and the test results are highlighted as two categories (i.e., split X significantly outperforms Y or not) in (b). Wilcoxon rank-sum tests across all 463 gauges are then aggregated in (c), showing the count of catchments where split X significantly outperforms Y. Note that any two blocks in (c) symmetric with respect to the diagonal line indicate count of X being significantly better than split Y ($C_{X>Y}$) and count of Y being significantly better than X ($C_{Y>X}$). These two counts are further

transformed to a single value metric net percentage (i.e., $(C_{X>Y} - C_{Y>X}) / 463 \times 100\%)$) in (d), which ranges from -100% to 100%. The data split identifiers are defined in Figure 4-1.

Viewing X as the full-period CSP and Y as any one of the 15 alternative splits, it can be seen the full-period CSP overall outperforms those 15 splits in 1990 (see the last column in Figure 4-3d). The average net percentage values for full-period CSP versus groups of older CSPs, recent CSPs, DCSP splits, and MDUPLEX splits are 31%, 34%, 26%, and 16%, respectively. When the model is built in 2010 with 30 years of data available during the same testing period (see the last column in Figure A5-1b in Appendices A-5), the full-period CSP maintain the superiority in terms of the above four average net percentage values being 44%, 5%, 13% and 9%. In short, this implies in 2010, full-period CSPs perform only slightly better than other splits except older CSPs, but such prevalence can be much amplified in 1990 (e.g., when building models with limited data for calibration and validation).

Similar results are observed in other testing periods that full-period CSP is superior to all other splits in all model build years (not presented for brevity). The recent CSPs perform comparably well with full-period CSP in 2010 but can be significantly worse in 1990 with a large decrease in net percentages. DCSP and MDULEX splits perform only slightly worse than full-period CSP in 2010, and their performance degradation in 1990 is minor compared to recent CSPs. Older CSPs, however, are overall the worst splits in both 2010 and 1990.

4.3.3 Split as binary classifier: Ability to correctly classify model failures in model building and testing

In Section 4.3.3.1, we present the classification accuracy metric derived from different failure handling thresholds. In Section 4.3.3.2, fractions of false negatives and false positives are separately displayed to show the consequences of incorrectly classified models in building and testing. Note that only results for model build years 1990 and 2010 are presented here, as they contain more data split samples (i.e., 16 data splits), and these two scenarios stands for the most data limit and sufficient conditions.

4.3.3.1 Classification accuracy: Frequency of model testing success/failure being correctly predicted

The constituent components of classification accuracy, i.e., true positives (TP) and true negatives (TN), stands for the benefits in model building and testing (i.e., the SST decisions work correctly). TP and TN both suggest model testing results are consistent with building results. Accuracy is tested with various thresholds defined by the two different approaches described in Section 4.2.4.3. Figure 4-4 shows the accuracy variation with different calibration thresholds, which is quantified as

median value of all the 463 catchments. Figure 4-4 only presents results in the first 5 years of the testing period, while other four different testing periods show similar patterns hence are skipped for brevity.

The most-left points in all panels in Figure 4-4 are the benchmark threshold scenarios, which denote the default reference KGE or constant KGE without additional performance expectation (i.e., Δ). Even though Δ being 0.4 in the reference KGE-based calibration threshold (the most-right points in each panel of the top two rows) may be overstrict for the calibration, while it can be helpful to show some extreme scenarios in this assessment.

Accuracy generally decreases with larger Δ added to the calibration threshold, i.e., stricter calibration is required. For the reference KGE-based threshold approach, testing thresholds are constantly the reference KGE and thus, testing states are fixed in all scenarios. And increasing calibration thresholds may lead to transitions from FN instances to TP instances, as well as TN instances to FP instances. The panels in the top two rows in Figure 4-4 shows decreasing trends in accuracy, implying more TN instances would be classified into FP instances.

For the constant KGE-based threshold approach, the calibration, validation, and testing thresholds are all perturbed simultaneously. Thus, negative predictions in testing may also transit into positive predictions, which is more complicated than the former approach. However, the overall trends in accuracy in these panels (bottom two rows in Figure 4-4) remain similar to the former one. Both approaches imply larger KGE thresholds used to constrain model building could result in an accuracy loss (0.05 to 0.1 in Figure 4-4), and the key consequence is the increasing cost of having more FP and FN instances, which needs to be considered in choosing the optimal splits.

Among all splits in Figure 4-4, the full-period CSP (CSP-100%) generally ranks at the top quartiles in all panels, implying full-period CSP is robust to achieve a high score in accuracy and appears to be the best one considering all panels and all threshold scenarios. The MDUPLEX splits are very close to the full-period CSP in all scenarios (Figure 4-4e1 to Figure 4-4e4). However, Figure 4-4 also shows different thresholds may impact the accuracy pattern notably. Typical examples are CSP-30%D and CSP-70%A, which are more variable in accuracy when tested with the two threshold approaches and in two representative build years, implying these splits can be less stable than CSP-100%.

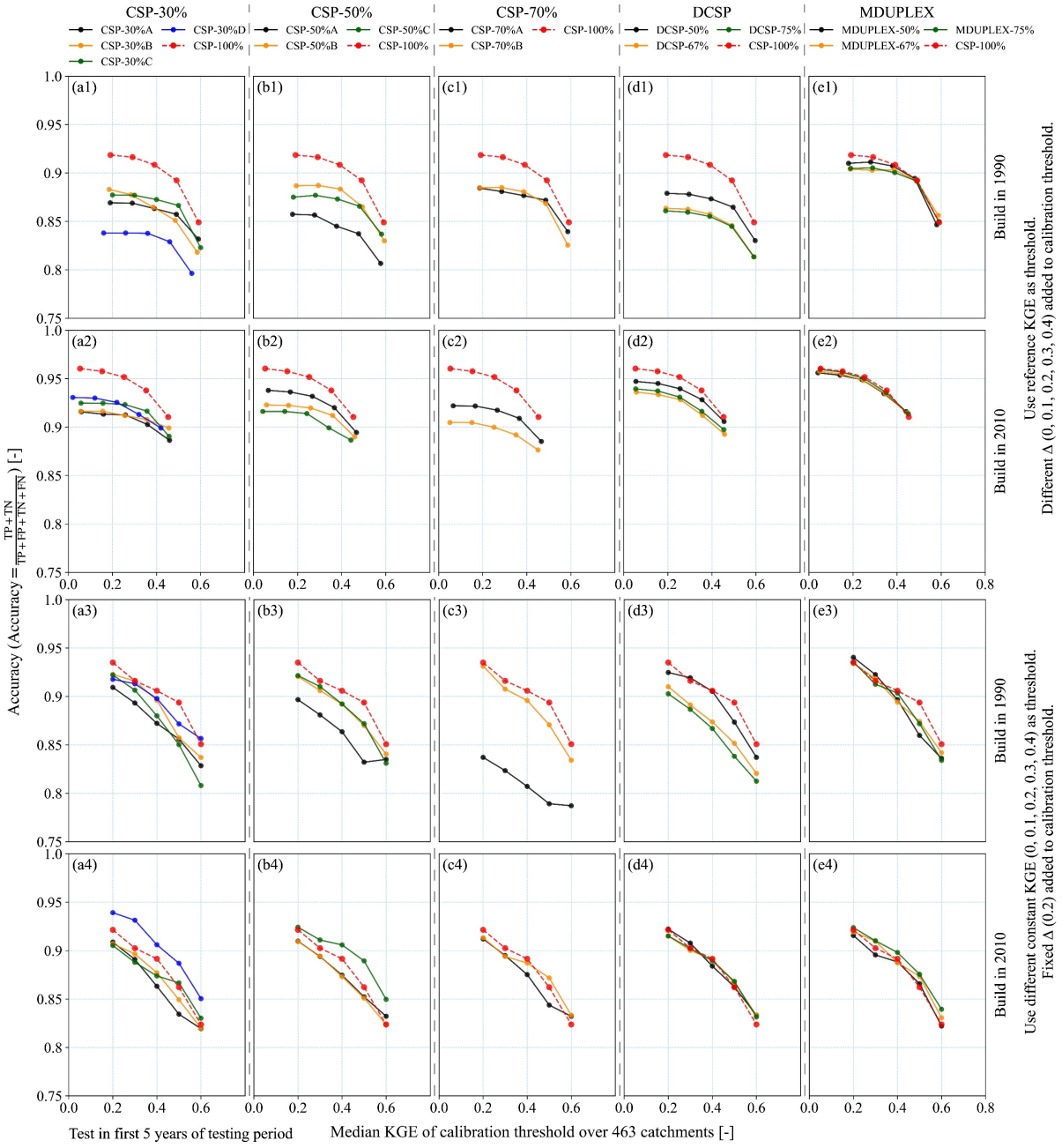

Figure 4-4. Classification accuracy variation with different KGE thresholds applied in the confusion matrix-based classification in the first five years of testing period. The *x*-axis of each panel denotes the median KGE of calibration threshold over all 463 catchments, and the *y*-axis denotes the accuracy metric. Panels in the top two rows (a1 to e1 for 1990 and a2 to e2 for 2010) display results using reference KGE as threshold, where calibration threshold is the reference KGE added with a variable value Δ (0, 0.1, 0.2, 0.3 and 0.4) and validation/testing thresholds are their corresponding reference KGE. Panels in the bottom two rows (a3 to e3 for 1990 and a4 to e4 for 2010) present results using different constant KGE as threshold (0, 0.1, 0.2, 0.3 and 0.4), where calibration threshold is the constant

KGE added with a fixed value $\Delta = 0.2$, and the validation and testing thresholds both are the constant KGE. Note that CSP-100% repeats in every panel to contrast with other splits. The data split identifiers are defined in Figure 4-1.

4.3.3.2 Fractions of FN and FP: Costs of incorrect classification in model building and testing

Mathematically, fraction of FN (FNF) plus fraction of FP (FPF) is equal to 1 minus accuracy. However, it is important to keep in mind that there can be different practical costs associated with FN and FP instances. Therefore, we further analyze these two separate components. Figure 4-5 shows FNF (x-axis) versus FPF (y-axis) with two typical Δ scenarios of the two threshold approaches in the first 5 years of the testing period. Other Δ scenarios and testing periods yield similar patterns hence are skipped here for brevity but are also considered in the analysis.

Figure 4-5 displays an important pattern that CSP-30%D, CSP-70%A and CSP-100% are constantly non-dominated solutions in 1990 (panels in top row of Figure 4-5). However, these non-dominated solutions are different with respect to the magnitudes of both axes. CSP-30%D and CSP-70%A in 1990 have the lowest FNF but the highest FPF (up to 18% in all instances), indicating modelers may need to make more efforts recalibrating models to achieve the acceptable level in practice. On the contrary, FPF of CSP-100% in 1990 is generally the smallest and can even be close to 0 when threshold is small, but the FNF usually ranks at the largest ones (6% to 10%). This indicates that full-period CSP may introduce higher risks of getting testing failures when the model building is deemed as a success. MDUPLEX splits are similar to CSP-100% in 1990 with respect to the magnitude of FNF (5% to 10%). Figure 4-5 also shows all splits in 2010 are much closer to one another with respect to FPF (ranges are within 5% in 2010, which however can be three times as large as those in 1990), this implies longer data available for model building may reduce the costs of model failures in building and testing.

Comparing the relative quantity of these two costs in 1990 (10 years of data available for model building), non-dominated solutions CSP-100% tend to have a large frequency in FN instances (5%–10% in Figure 4-5), which can be 2–3 times as large as CSP-30%D and CSP-70%A (2%–6% in Figure 4-5). However, the frequency of FP instance of CSP-30%D and CSP-70%A can be 2–70 times (14%–18% in Figure 4-5) as large as CSP-100% (0.4%–10% in Figure 4-5) in 1990. It can be seen that each of these non-dominated solutions in 1990 can be superior considering only one of the axes, while if one needs to take into account both costs that they may have in model building and testing, the full-period CSP may be the best to keep both costs at a low level.

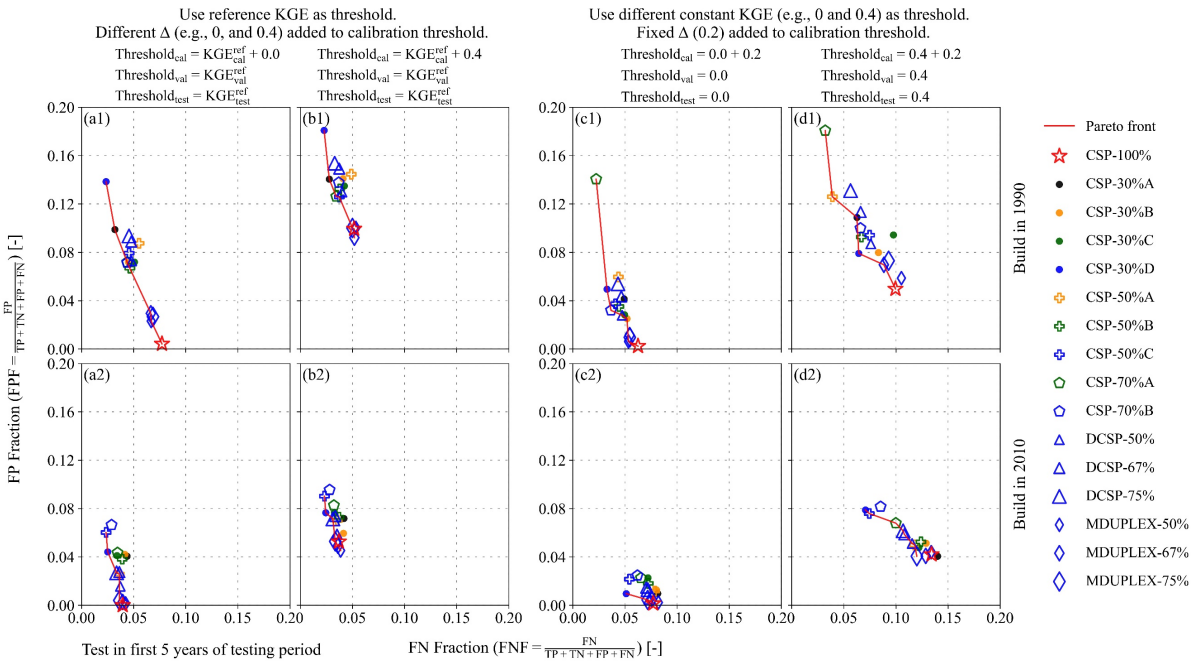

Figure 4-5. Tradeoff between the fraction of false negative (FNF) and fraction of false positive (FPF) in the first five years of testing period. Panels in the left two columns (a1 and b1 for 1990 and a2 and b2 for 2010) display results using reference KGE as threshold, where calibration threshold is the reference KGE added with a variable value Δ (0 and 0.4 as examples) and validation/testing thresholds are their corresponding reference KGE. Panels in the right two columns (c1 and d1 for 1990 and c2 and d2 for 2010) present results using different constant KGE as threshold (0 and 0.4 as examples), where calibration threshold is the constant KGE added with a fixed value Δ = 0.2, and the validation and testing thresholds both are the constant KGE. The data split identifiers are defined in Figure 4-1.

4.3.4 Multi-objective assessment of splits: Tradeoff between median KGE and classification accuracy

In this assessment, we frame a multi-objective problem to simultaneously optimize median KGE (see in Section 4.3.2) and classification accuracy (see in Section 4.3.3) in model testing periods. In this section, we present the averaged percent distances of all gauges and all testing periods. Note that the accuracy patterns of different splits in Section 4.3.3.1 generally do not vary with thresholds. Thus, we keep using reference KGE-based accuracy, i.e., the benchmark, in this assessment. Median KGE are calculated using all optimization trials of a split at a single gauge, and any failed models in calibration and validation will trigger the decision to use reference flows to predict testing hydrographs.

Figure 4-6a demonstrates how the percent distances of median KGE and accuracy are calculated from a single gauge tradeoff. Details can be found in Section 4.2.4.4. Figure 4-6b to 6f

display the percent distances of those two metrics averaged across 463 gauges and all testing periods in the five model build years. Note that percent distance metric is conditional to the sample size solutions in a tradeoff. Low percent distance values in 1995, 2000, 2005 (Figure 4-6c to 6e) are partially due to the definition of this metric that the lowest median KGE may be identified as the nadir point, thus leaving 0 percent distance value for the split. This suggests percent distance is only for comparing relative differences of splits in the same tradeoff. As such, percent distance values in build years 1995, 2000, and 2005 are not comparable with those in 1990 and 2010.

Figure 4-6 shows that the full-period CSP ranks the largest percent distances of both median KGE and accuracy in all model build years except that the percent distance of accuracy is the second largest in 2010. This indicates that full-period CSP on average approaches the best median KGE and accuracy in all 463 gauges. Note that the patterns of the percent distance of individual testing periods (not presented here) are consistent with the averaged one shown in Figure 4-6. Figure 4-6 also displays that the accuracy is better than median KGE regarding the percent distance values for all splits, and the range of accuracy is generally above 75%, while median KGE generally ranges from 34% to 68%. This implies that despite the variable KGE differences in testing periods, these splits can generally achieve a good accuracy score in all gauges, i.e., consistent model building and testing results. This is consistent with the accuracy results presented in Section 4.3.3.1.

Comparing full-period CSP with short-period CSPs in 1990 and 2010 (red and gray bars in Figure 4-6b and Figure 4-6f), it can be seen that performances of short-period CSPs are relatively not stable. The percent distances of median KGE for short-period CSPs are in ranges of 41%–61% and 38%–62% in 1990 and 2010, respectively. However, the recent CSPs (e.g., CSP-30%D) appear to be better in 2010 than in 1990. For example, CSP-30% in 1990 (Figure 4-6b) are the worst in 1990 with lowest percent distance values for both metrics, but these performances are notably improved in 2010.

Comparing full-period CSP with discontinuous splits (i.e., DCSP and MDUPLEX splits), it can be seen that MDUPLEX performances can be close to full-period CSP in 1990 and 2010 (Figure 4-6b and Figure 4-6f) but no significant advantage of MDUPLEX splits can be observed in these results. On the other hand, DCSP splits are worse than full-period CSP in 1990 with at most a 13% drop in percent distance of median KGE and a 15% drop in percent distance of accuracy, but DCSP splits performances are improved in 2010. In 1995, 2000, and 2005 (Figure 4-6c to 6e), DCSP splits are also seen to be worse than full-period CSP with respect to both metrics.

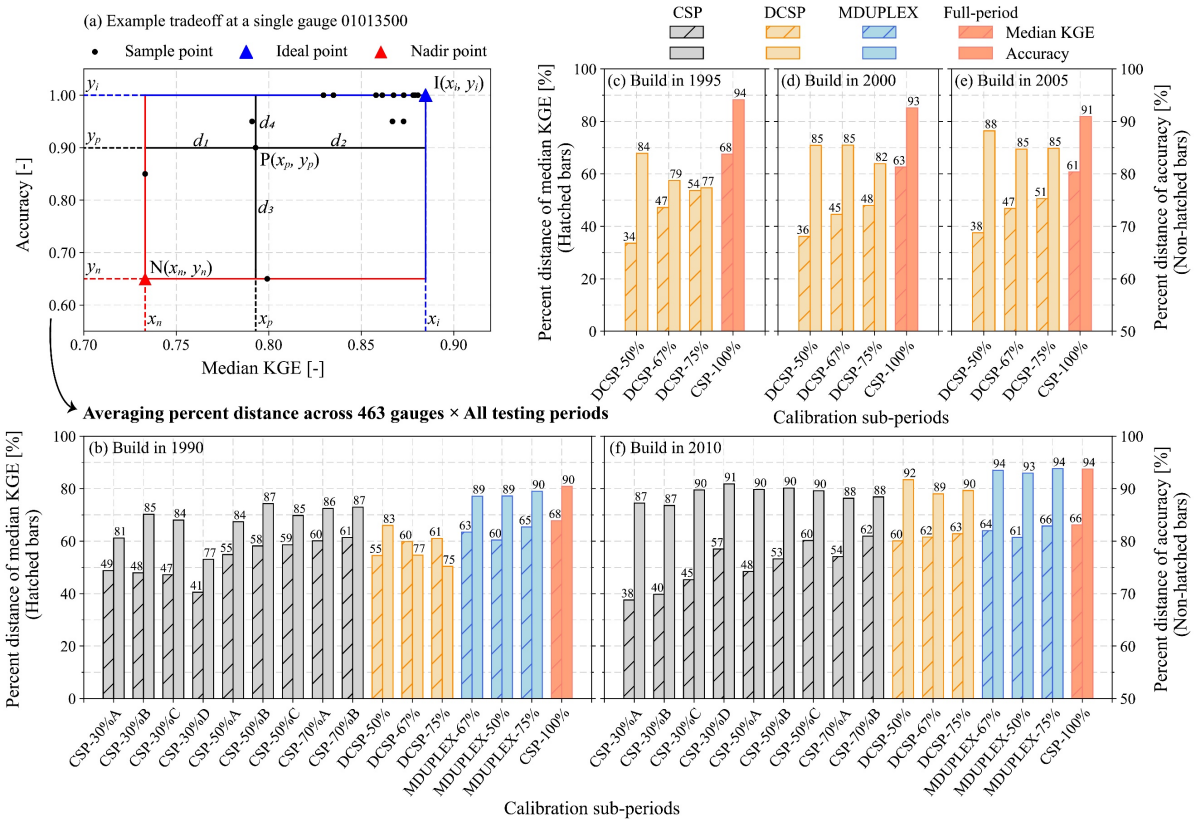

Figure 4-6. Demonstration of percent distance (PD) of median KGE and accuracy based on the multi-objective decision-making problem to simultaneously optimize median KGE and accuracy in model testing periods. An example tradeoff at a single gauge 01013500 is displayed in (a) with models built in 1990 and tested in the first 5 years of testing period. The ideal and nadir points in (a) represent the best and worst median KGE and accuracy all splits can achieve, respectively (see detailed definition in Section 4.2.4.4). The point P represents any split in this tradeoff. The percent distance of each split is calculated at single gauges and averaged across all 463 gauges and all testing periods, yielding results in (b) 1990, (c) 1995, (d) 2000, (e) 2005, and (f) 2010. Note that the accuracy is calculated with reference KGE being threshold and $\Delta = 0$ (see in Section 4.2.4.3). Model failures are handled when calculating median KGE that failed models use reference flow for testing period prediction instead. Also note that there are five different testing periods in 1990, 1995, 2000, and 2005, while there are only three unrepeated testing periods in 2010. The hatched bars on the primary *y*-axis denote percent distance of median KGE, while non-hatched bars on the secondary *y*-axis denote percent distance of accuracy. Different colors of bars represent the four categories of splits. The percent distance values are annotated on top of the bars. The data split identifiers are defined in Figure 4-1.

4.4 Discussion

This study adapts the split-sample test (SST) framework proposed by Shen et al. (2022a) (see in Chapter 3) to investigate whether hydrological modelers can benefit from calibrating their models to discontinuous data. In this study, we tested 44 calibration splits (i.e., 18 continuous CSPs from Shen et al. (2022a), 6 MDUPLEX splits from Zheng et al. (2022), 15 DCSP splits, and 5 full-period CSPs) using a 21-parameter model (i.e., HMETS) in 463 catchments across the CONUS, and the total counts of our model calibration and model testing cases are 409,755 and 1,751,529, respectively. We assessed all the ways hydrological model calibration split-sampling is currently done when only a single split sample is evaluated and one method found in data-driven modeling (i.e., MDUPLEX).

We utilized all model optimization trials in performance assessments. Our assessments included ranking model testing KGE without handling failures in model building, comparing medians of testing KGE with model failures handled by the reference climatology, viewing each SST decision as a binary classifier for model building and testing states, and framing maximizing the medians of testing KGE and maximizing the binary classifier accuracy as a multi-objective problem. In this section, we further discuss the implications of our massive empirical results.

4.4.1 SST recommendations for hydrological modelers

For practical hydrological model building that seeks a deterministic calibrated model for use in a future application period, assuming the model is calibrated under only a single SST and the calibration effort and formulation are not unreasonably poor, calibrating to all data is the most robust choice for our array of metrics than any of the split-sample decisions (either continuous or discontinuous calibration sub-periods) we looked here. It is also important to note that we do not recommend doing such a "three-period" (i.e., calibration, validation, and testing) splitting for model building. In contrast, our "three-period" experiments are designed for application only in data splitting comparison studies where a third "out-of-sample" testing period is needed. Practical hydrological model building without a third period (e.g., either split dataset or not) can follow the recommendations we made here.

Our results in Section 4.3 reaffirm the validity of the two SST recommendations made in Shen et al. (2022a) (see in Section 4.1 or Section 3.4.1) even when 26 representative discontinuous splits (i.e., DCSP and MDUPLEX splits) are incorporated into the evaluation framework. The conclusions do not change despite some adjustment made in the experimental design and model performance assessment between Chapter 3 (Shen et al., 2022a) and this study. For example, we calibrated models in hydrological years and assessed performance of all optimization trials in this study, while Chapter 3

(Shen et al. (2022a)) calibrated models in calendar years and only assessed performance of the best-calibrated trial.

Using all available data to calibrate models and skipping temporal validation entirely was empirically demonstrated to be the most robust choice for model building, given the objective is to apply models for streamflow prediction in the post-validation model testing period. Rationale may be its retaining of all information in the calibration, thus allowing hydrological models to better "learn" different processes and patterns mapping from the input data to the output system response data, which is in accord with the data splitting principles highlighted by Maier et al. (2023). In addition, hydrological models that are established based on physical constraints may be less likely prone to overfitting issues than those data-driven models with much higher degrees of freedom (i.e., more parameters) (Sungmin et al., 2020).

Our analyses here empirically assessed the value of discontinuous data splits (i.e., DCSP and MDUPLEX splits). Our results show calibrating models to those discontinuous splits offer no advantage relative to calibrating to the full-period dataset in terms of (a) KGE in testing period (see Figure 4-2, Figure 4-3, and Figure 4-6), (b) the ability of each split functioning as binary classifier to correctly predict testing period states (see Figure 4-4, Figure 4-5, and Figure 4-6), (c) computational costs that calibrating models to discontinuous splits and all data available both require running models through the entire model building period for every iteration in optimization (see discussion in Section 4.2.1), and (d) the efforts required on the implementation of splitting algorithms. Note that the MDUPLEX algorithm requires heavier computation than CSP and DCSP, especially when the data record is long (e.g., model building in 2010 with 29 years available), since the algorithm needs to calculate specific metrics (e.g., distance) between any pair of data points and the computation increases notably with more data points being considered. In addition, MDUPLEX splitting results may change when the data record is extended with new data points. As a result, using full-period record for calibration avoids any of the inconvenience in model building while providing robust model performance in the testing (i.e., model application) periods. Hence, recommendation #1 in Chapter 3 (Shen et al., 2022a) holds when both continuous and discontinuous deterministic data splitting options are considered.

As stated by Maier et al. (2023), calibration and validation data should be different but each needs to contain all patterns/events. Two data splits that clearly satisfy this requirement are 1) building models on the modified DUPLEX (MDUPLEX) split which consists of two statistically similar subsets (Zheng et al., 2022), and 2) using all data for calibration and skipping validation entirely. Our empirical results support that building models on MDUPLEX splits can be very close to but no better than

calibrating to the full period in terms of many aspects such as accuracy and median KGE in testing period. Despite the advantage of MDUPLEX algorithm to preserve similar statistical features in its subsets, it still appears to be relatively inferior to full-period dataset, which keeps all information content in calibration.

In addition, the results also showed calibrating models to some data splits (e.g., CSP-30% and CSP-70%), which may contain large variance in calibration and validation (i.e., calibration and validation data are dissimilar), led to a lower false negative fraction (i.e., models identified as adequate in calibration and validation but failed in model testing period) than MDUPLEX and full-period splits, although the false negative fraction is very minor in the total count (see in Section 4.3.3 and Section 4.4.1). Arguably, this shows the value of validating models to a period dissimilar to calibration data, which is the underlying philosophy behind the differential split-sample test (DSST) method proposed by KlemeŠ (1986). DSST is a special version of the SST and is widely used for analyzing model performance change under diverse hydro-climatic conditions (e.g., see Bai et al., 2021; Coron et al., 2012; Dakhlaoui et al., 2017; Fowler et al., 2018; Fowler et al., 2016; Gaborit et al., 2015; Motavita et al., 2019; and Seiller et al., 2012). An important implication from many DSST studies is that model performance degradation may be observed when transfer model parameter sets to another period with contrasting conditions (e.g., from wet to dry period) (e.g., see Coron et al., 2012).

On the other hand, calibrating to the full period data always yields low level in false positive instances, which is superior to other options. In practice, FN instances may lead to model failure in the future period applications (e.g., wrong prediction), and FP instances may require modelers to put more efforts in rebuilding the model until it yields acceptable results in calibration and validation to ensure there is at least one model available for future period applications. However, this can be much more computationally expensive in practice. Costs of FN and FP instances need to be considered in model building, but it may be another problem of how to weigh the consequences in operational modeling, which is out of our main consideration in this study.

It is worth noting that some hydrological modeling studies reported using odd/even years (i.e., DCSP-50% in this study) for calibration and validation (or vice versa) could overcome non-stationarity conditions due to the changing environment (see e.g., Arsenault et al. (2017); Essou et al. (2016); Xu, 2021; and Yang et al. (2020)). In this study, DCSP-50% splits can sometimes be better than short-period CSPs in the multi-objective analysis (Figure 4-6f), which proves its applicability in some cases such as when data available are sufficiently long (e.g., 30-year data available in 2010). But it is also obvious that in all model build years in Figure 4-6, DCSP-50% splits are less robust than calibrating to the full-period data, and also worse than MDUPLEX splits. The reason may be that even though the

odd/even year method, as well as other DCSP splits, samples calibration/validation years over the entire data record, there is still a loss of information for calibration compared to the full-period data. Since DCSP samples data at the annual scale, it may lose a year with key information for model calibration. The MDUPLEX, however, selects data points at the daily scale and considers data information similarity in the algorithm (i.e., described by the Euclidian distance between data points), may be more potent to achieve two similar subsets and can be better than DCSP. However, when model building data is relatively sparse (e.g., in 1990), DCSP and MDUPLEX are both worse than full-period CSP (e.g., see the worst 20% ranking results in Figure 4-2 and multi-objective analysis in Figure 4-6).

We also need to note that performing "in-sample" or "out-of-sample" assessment is an on-going debate in hydrological modeling community. Some studies emphasized on "in-sample" analysis as a well-performed model in the "in-sample" period may imply the ability to perform well in the "out-of-sample" conditions (Chen et al., 2022; Maier et al., 2023). Typical application is the study by Zheng et al. (2022), which compared MDUPLEX splits with some traditional continuous data splits during the model building period (i.e., validation or calibration plus validation). We do not deny the value of "in-sample" analysis that it may provide a deep understanding of hydrological processes under conditions that are known to modelers. However, we note that "in-sample" analysis should not be overemphasized and key reasons are that (a) "in-sample" analysis is less desirable if the model building purposes are to support model applications in the future such as streamflow prediction (KlemeŠ, 1986); and (b) "in-sample" analysis does not provide an independent period (i.e., the data in this period should not be used in model calibration and validation) to compare different SST decisions. True "out-of-sample" data are never available for modelers, however, the "out-of-sample" period can be mimicked under a rigorously defined evaluation framework such as the one applied here and in Chapter 3 (Shen et al., 2022a). One of the key features in our evaluation framework is that we always assess model performance in the testing periods no matter how data are split into calibration and validation or not split in model building process. This feature allows us to compare different SST decisions in an "out-of-sample" period for more objective and realistic analyses, which distinguishes our evaluation framework from others.

4.4.2 Best practice in the split-sample test in hydrological modeling

Combining SST experiment results reported in Chapter 3 (Shen et al. (2022a)) and the new discontinuous splits results in this study, we reaffirmed the validity of two SST recommendations made in Chapter 3 (Shen et al., 2022a) in Section 4.4.1. Here, we further discuss the implications of our massive empirical results on hydrological modeling.

In typical model building practice, modelers generally compare model performance of calibration and validation periods. A significant performance drop in validation period may indicate the problem of overfitting and or data quality issues (Arsenault et al., 2018). This convention implicitly assumes the validation error is representative of the errors/performance in other periods (Wu et al., 2013), i.e., the performance drop between calibration and validation is extrapolated to represent possible performance loss when apply the model to another different period (e.g., streamflow prediction into the future). The testing period in our SST assessment framework, which always follows the model building period, actually reveals how models including those successfully validated perform in this model application period. A useful validation period should function to correctly identify models that will fail in the model application period, which identifies those bad models somehow behaving well in calibration.

Our experimental design contains four distinct categories of SST decisions (see in Section 4.2.1). Data splits produced by the MDUPLEX algorithms consist of statistically very similar calibration and validation sub-periods, while other CSP and DCSP splits may have a large variance between their calibration and validation sub-periods (i.e., very different data). Our results show that calibrating and validating to data with large differences may lead to a degraded accuracy but an improved (e.g., lower) fraction of false negative instances (see in Figure 4-4 and Figure 4-5). However, calibrating and validating to similar data (i.e., MDUPLEX splits) yields very close accuracy and similar fractions of false positive and false negative instances to the full-period CSP. This implies that validating models to some fundamentally different data (with new characteristics relative to calibration data) may be more helpful to identify those bad models somehow behaving well in calibration, compared to validating to data that is very similar to the calibration data.

Choosing a data splitting method for model building is influenced by many factors in practice, such as data accessibility and quality, default data use team and settings, and modeler's experience and judgement (Lieke A. Melsen, 2022). Due in part to these reasons, modelers may still insist on validating models, since the "calibration-validation" paradigm has been extensively adopted in hydrological community for over half a century. However, our large sample-based results here and in Shen at al. (2022a) all strongly suggest that it is time to update the split-sample test approach since calibrating models to all data yields the best overall models over the range of all possible deterministic data splits. Considering both of these factors, we make a third more practical SST recommendation.

SST recommendation #3: Modelers rebuild models after any validation experiments, but prior to operational use of the model, by calibrating models to all available data.

Recalibrating the model to the full-period dataset prior to model use in the model application period no matter the prior SST decisions being used to perform model validation is easy to implement presuming a second, potentially longer calibration exercise can be repeated. This general strategy is used in machine learning and deep learning fields for model development (Raschka, 2018) where after these data-driven models are tested on data not used in their training, they are retrained with the complete dataset before being deployed operationally.

Combining Chapter 3 (Shen et al., 2022a) and this study, we made three SST recommendations for practical hydrological model building. However, these recommendations were made within the scope of only considering *single-site* and *temporal* model validation problems. Our models and catchments applied in the experiments are also limited to conceptual and lumped models (GR4J and HMETS in Chapter 3 and HMETS only in this study) and (near) natural catchments (463 US catchments minimally impacted by human activities). Thus, it is also of significant importance to clarify the possibility of transferring our recommendations to those model building conditions we did not focus on in this study, such as basins beyond the extent of the US, basins with anthropogenic influence (e.g., hydraulic constructions in rivers and land use and land cover change), and models with higher levels of complexity (e.g., distributed models with more parameters).

The 463 CAMELS catchments with a wide range of spatial extents utilized in this study are generally representative for different flow regimes and climates in the US. We believe the three recommendations can be transferred to other similar catchments either within or beyond the US, because our conclusions are all drawn in statistically robust ways instead of specific catchments and thus are possible to be generalized to catchments out of our assessment samples. These recommendations can also be applied for managed catchments. However, modelers need to be cautious when modeling managed catchments with significant anthropogenic influences such as urbanization and hydraulic constructions. Hydrological models used in this study only consider processes in the natural system, while it is critical to consider human activities, such as land-use change, irrigation, water diversion, etc., in a managed catchment when simulating its hydrological processes. For such situations, modelers need to carefully select their hydrological models that can accurately depict the key processes in those catchments. A typical option is the Raven hydrological modeling framework which can take into account the water demand (e.g., irrigation and water treatment withdrawal) and flow diversions (Craig et al., 2020; Craig, 2023).

Distributed hydrological models generally employ more sophisticated math/physical equations to describe more complex hydrological processes and are distinct in terms of considering more input variables at finer spatial scales (e.g., grid-cells). As such, distributed modeling may require more input

data (e.g., water stage, soil moisture, evapotranspiration, snow depth, lake level, groundwater level, etc.) for to further constrain model parameters (Hunt et al., 2006; Hunt et al., 2013; Mei et al., 2023; Refsgaard, 1997). Such additional data used in distributed modeling have been demonstrated to be beneficial for improving model performance (Dembélé et al., 2020; Mei et al., 2023). Even though our recommendations are made from modeling results of two conceptual lumped models, it is likely that they can be transferred to distributed modeling, while employing distributed models in our experiments for rigorous tests in the future would be the most convincing way to prove it. The reason that those recommendations can be applicable for distributed modeling is that both conceptual and physically-based hydrological models are based on water balance (i.e., mass conservation) despite different levels in their complexities, which makes them internally constrained by physical laws and assumptions, and thus making these models less flexible to react to unusual combinations of processes (i.e., more difficult to get overfitted) comparing to those data-driven models with much higher flexibility in model training (i.e., having much more parameters from hundreds to thousands) (Ayzel and Heistermann, 2021; Sungmin et al., 2020). As a result, we do not recommend data-driven modelers (e.g., LSTM) skipping their model testing (which is equivalent to our validation).

Another relevant aspect in distributed modeling is to validate models spatially, either using internal points at subbasin level or outside proxy basins not used in calibration (KlemeŠ, 1986). Since hydro-climatic variability in space is more critical when dealing with spatial model validation problem, it is usually required to perform more complex validation test, such as the differential split-sample test (DSST) and proxy basin test (KlemeŠ, 1986). Also, spatial validation is usually related to multi-site problem. A typical example to justify the potential of transferring our recommendations into multi-site and spatial validation problems is the recent Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL; Mai, Shen, et al., 2022). The GRIP-GL study comprised 13 models (including machine learning based, lumped, subbasin-based, and grid-cell based) that are calibrated locally (single site-based), regionally (one of each of the six predefined regions-based) or globally (all basins-based). These models were validated in three ways: temporally, spatially, and spatio-temporally. They reported that regionally calibrated models show stronger spatial robustness than locally calibrated models. This may support applying our recommendations such as using all available data for calibration (regionally) and then it is expected that the spatial validation and spatio-temporal validation can be further improved. Regionally calibrated (multi-site-based) models provide seamless prediction of streamflow in space, thereby allowing an enhanced parameter transfer to ungauged locations if our recommendations contribute to such field by further investigation in the future.

4.4.3 Study Limitations and Future Work

As stated in Section 4.1, the scope of this study is addressing the problem of single-site and temporal model calibration and validation. Our recommendations may be feasible for multi-site model calibration and spatial validation problems (e.g., see Mai, Shen, et al., 2022) but further exploration would be required to test this. However, for single-site calibration problems involving a distributed/semi-distributed model, including basins with complex management activities to represent, we still would expect these recommendations to hold given that even more parameters must be estimated in calibration (versus a lumped model), and thus maximizing the data used for calibration would be even more important.

As discussed in Shen et al. (2022a) (see in Section 3.4.2), our SST recommendations do not apply to climate change impact assessment studies focused on parameter transferability under contrasting climates. This is still true in this study and the key reasons are (a) our data splits do not require climatic contrasting or climatic similar in calibration and validation and (b) our testing periods are set after model build years but are not limited by climates. As such, our model building and testing combinations could constitute a wide range of climatic conditions to support robust conclusion drawn. An assumption is that we do not know how testing (future) conditions may be in our experiments, and hence we assess all the ways hydrological model calibration split-sampling is currently done to explore the optimal choice in unknown testing conditions. However, if the future conditions are predicted with some confidence, such as provided by the global climate model (GCM) or regional climate model (RCM), and plus the intention is to assess how hydrological changes would be in the future period, we believe it is appropriate to build models over a historical period with similar conditions to the future, which can be referred to the DSST studies (e.g., see Bai et al., 2021; Coron et al., 2012; Fowler et al., 2016; and Motavita et al., 2019).

In this study, we only use a lumped conceptual model, i.e., HMETS model, to test our hypothesis. Applying other models such as distributed or semi-distributed hydrological models in our SST experiments may provide a broader spectrum of results regarding different model complexities. However, we note that computational expense is the key limiting factor. For example, we have over 0.4 million calibration experiments (see in Section 4.2.3), which may not be practically feasible for distributed modeling. Nonetheless, testing more complex models such as distributed models in our evaluation framework is important future work.

We also assume the model is calibrated to only a single SST and the calibration effort and formulation are not relatively poor. As results imply in this study, validating models to some different data can periodically be valuable to identify bad models. Our future work may focus on conducting

some validation while simultaneously calibrating the model to all available data. An existing example of this is the K-fold cross-validation method. The K-fold cross-validation partitions the full-period dataset into k non-overlapping equal length subsets (i.e., folds) (Kohavi, 1995). Each subset is successively used for model validation and the remaining (k-1) folds play the role of calibration. K-fold cross-validation ensures every fold can be adopted in both calibration and validation, thus making use of the information of the full dataset in model building. This could also reduce the chance of (un)lucky splitting in single split. We believe K-fold cross-validation can be an alternative approach to building hydrological models in data sparse regions. However, the drawback is the high computational expense of building k times of model, especially for distributed modeling. Future work may also be needed to answer the question of how to determine the final model based on the k model building results if our scope is for identifying a deterministic model.

4.5 Conclusions

In this study, 44 representative continuous and discontinuous split-sample test (SST) decisions are compared in the comprehensive evaluation framework proposed in Chapter 3 (Shen et al., 2022a) using a conceptual hydrological model in 463 catchments across the United States, and the extensive experiments results of over 0.4 million model calibration experiments and 1.7 million model testing assessments are thoroughly assessed. Similar to Chapter 3 (Shen et al., 2022a), we build models in five different scenarios with data available ranging from 10 to 30 years and test models in various model testing periods after building the model. Model testing period defines a common "out-of-sample" period for an objective comparison of different SST decisions. In this study, we evaluated 44 SST decisions including continuous calibration sub-period (CSP), discontinuous calibration sub-period (DCSP), MDUPLEX, and full-period CSP for model calibration and validation and assessed the results across 23 different testing periods in five model build year scenarios. All model calibration optimization trials (i.e., at least 20 trials) are utilized in the performance assessment to achieve robust conclusions. Model performance in testing periods were assessed from different aspects: First, an exploratory analysis on ranking the raw KGE in testing period was conducted to show result patterns when model inadequacy (failures) is not explicitly handled in calibration and validation. Then, the nonparametric statistical test Wilcoxon rank-sum test was performed on any pair of SST decisions in testing period to reveal if their medians (of all optimization trials) are significantly different. Afterwards, each SST decision was viewed as a binary classifier to bin models either a success or failure and its ability to correctly predict model testing states are assessed by the classification accuracy metric and its constitutive components. Finally, we framed model building processes as a multi-objective decision-making problem to simultaneously optimize both the accuracy and median KGE

and the tradeoff between accuracy and median KGE was analyzed for both individual gauges and all 463 gauges in aggregate.

Overall, our empirical results considering new discontinuous SST decisions in model building and testing assessment strongly support that the SST recommendations made in Chapter 3 (Shen et al., 2022a) (e.g., calibrating models to all available data is the most robust choice) is still valid even when representative discontinuous SST decisions are evaluated under the same framework. Strong evidence shows the superiority in calibrating to all data, while those discontinuous SST decisions have no clear advantage over the full-period CSP. We recommend that hydrological modelers continuing to validate their models rebuild their models after their validation experiments, but prior to operational use of the model, by calibrating models to all available data. Empirical results show that such an approach would be expected to improve multiple aspects of model performance in the model application period and the improvement will be more significant as the model building data available is more and more limited.

Chapter 5
Exploring Alternative Model Validation Methods for An Updated Split-Sample Test

This chapter is a replicate of the following manuscript that is currently in preparation. Most of the literature review content in the article (e.g., the introduction and methodology sections) is adapted to Chapter 2, and only a shortened version of the introduction goes with this chapter. Other contents such as results and conclusions are all consistent with the manuscript. All references are unified at the end of the thesis.

Shen, H. & Tolson, B. A. (2023). Exploring Alternative Model Validation Methods for An Updated Split-Sample Test (manuscript in prep)

Summary

Hydrological model validation has been traditionally viewed as a critical procedure in the split-sample test (SST) framework to ensure model robustness. However, a typical recommendation from recent SST studies is that calibrating models to all data available is the most robust choice. Using all data in calibration naturally skips temporal validation entirely and this can reduce the chance of identifying poor quality models. In this study, alternative validation methods are explored to try and improve the successful rate of identification of poor-quality models when all data are utilized in calibration. We propose three validation methods (VMs), including the traditional calibration and validation approach (VM1), the traditional calibration and a proxy validation approach which relies on an assessment of annual model performance statistics (Split KGE and Split Reference KGE) in calibration to identify unacceptable models (VM2), and skipping validation entirely (VM3). We employ six continuous calibration sub-periods (CSPs) that utilize recent data for calibration and older data for validation to test the three validation methods by post-processing their model building and testing results from Chapter 4. The experiments are conducted in 463 catchments across the US and were assessed in multiple aspects such as the accuracy of each CSP with a validation to correctly predict model failures in model building and testing and the costs of failing to do so. The accuracy and median KGE in the testing period of each validation method are also framed as a multi-objective problem to be optimized simultaneously. Our validation experiments tested on a large sample of catchments generally indicated using Split KGE and Split Reference KGE can be more effectively detect hydrographs that may be overfitted to specific years such as high flow year. The VM3 showed overall the best performance in model testing period with respect to the accuracy and KGE. However, VM2

improved both VM1 and VM3 by using Split KGE and Split Reference KGE in terms of a more balanced consequences when considering the false positive and false negative instances in model building and testing. Alternative validation methods such as VM2 enable model calibration based on all data while also providing a check against overfitting to further improve model robustness. Results suggest further exploration of more proxy validation methods like VM2.

5.1 Introduction

Hydrological models are developed to seek a better understanding of physical processes and facilitate decision-making for many purposes such as water resources management and planning, flood and drought forecasting, reservoir management, climate change assessment, etc. (e.g., see Beckers et al., 2009; Blöschl et al., 2013; Fowler et al., 2007; Hrachowitz et al., 2013; and Mishra & Singh, 2011). Such computer-based models have become increasingly complex in the last half century due to the advances in computing capabilities and data collection (Beven, 1989, 2012; Craig et al., 2020; Devia et al., 2015; Savenije, 2009; Singh & Woolhiser, 2003). Many of these hydrological models contain considerable parameters that are established for empirical equations and cannot be directly measured. Model calibration is therefore a critical procedure to identify the best parameter values by comparing the model simulated and observed system response data (Arsenault et al., 2018; Beven, 2012; Duan et al., 1994; Legates & McCabe, 1999; Mai, 2023). It can be even more important to validate the model adequacy and model robustness in data set that is not used in calibration. Such a validation procedure is to ensure model parameter transferability in time and/or in space (KlemeŠ, 1986).

Model calibration and validation are the central of a model building (development) process and are usually bound together in practice. The most-used model validation method in hydrological modeling community is the split-sample test (SST) framework (KlemeŠ, 1986), which partitions the available dataset into two mutually exclusive subsets and utilizes one set for model calibration and the other set is retained for validation. In the original SST method proposed by KlemeŠ (1986), the two subsets are used for a "two-round" calibration and validation experiments, i.e., one used in calibration and the other one used in validation and vice versa. However, such a "two-round" model calibration and validation suggestion has not been the dominant approach in model building in the following decades, while a simplified SST version that only one round of calibration and validation is adopted in model building has been widely adopted (e.g., see Pool et al., 2018; Rakovec et al., 2019; and Schlef et al., 2021). Both the original SST version and the simplified SST version test the calibrated model in an independent period that was not used in calibration to ensure parameter transferability and model robustness. However, the original SST method determines the model to be adequate if the "two-round" model validation results are similar and both acceptable (KlemeŠ, 1986). Differently, the simplified

SST method usually compares performance criteria of the calibration and validation periods and an important performance drop between calibration and validation may imply overfitting (also called overparameterization) or data quality problems (e.g., see Arsenault et al., 2018; Knoben et al., 2020; and Schoups et al., 2008). This convention implicitly assumes the validation error is representative of the errors/performance in other periods (Wu et al., 2013), i.e., the performance drop between calibration and validation is extrapolated to represent possible performance loss when apply the model to another different period (e.g., streamflow prediction period). Rationale of this assumption is that if state of the system (e.g., climates and landscapes) remain stationary in time, the calibration and validation performance would likely be similar; however, a performance loss between these two periods may imply how much model prediction skill could be lost when extrapolating the model to a future period due to factors such as climate change and anthropogenic influence (Mai, 2023).

Past studies have made tremendous efforts to investigate and discuss possible ways to improve hydrological model performance and model robustness with respect to data splitting between calibration and validation (e.g., see Arsenault et al., 2018; Coron et al., 2012; Daggupati et al., 2015; Dakhlaoui et al., 2019; Guo et al., 2018, 2020; Motavita et al., 2019; Razavi & Tolson, 2013; Shen et al., 2022a; and Zheng et al., 2022). Although much attention has been paid on this topic, unfortunately, the *problem of data splitting* still remain as a challenge and there are no consensus on which data splitting methods that can be applied for all (or most) model building practices in hydrological modeling community (Daggupati et al., 2015; Maier et al., 2023; Myers et al., 2021). It should be noted that some studies have recently demonstrated models can be more robust when all data are used for calibration and none are held back for temporal validation (e.g., see Arsenault et al., 2018; Shen et al., 2022a; and Wasko et al., 2023). More discussion on the SST method, the problem of data splitting, and the question if a data set should be split or not (i.e., skipping validation) are presented in Section 2.3.

Here, we underscore the three SST recommendations made in Chapter 3 (which is a mirror of Shen et al. (2022a)) and Chapter 4, which are listed as follows.

"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.

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.

SST recommendation #3: Modelers rebuild models after their validation experiments, but prior to operational use of the model, by calibrating models to all available data."

These recommendations can be quite generalizable as discussed in Chapter 3 and Chapter 4. Using all available data to calibrate models and skipping temporal validation entirely was empirically demonstrated to be the most robust choice for model building, given the objective is to apply models for streamflow prediction in the post-validation model testing period (see in Chapter 3 and Chapter 4). Rationale may be the retention of all information in the calibration, thus allowing hydrological models to better "learn" different processes and patterns mapping from the input data to the output system response data, which is in accord with the data splitting principles highlighted by Maier et al. (2023). In addition, hydrological models that are established based on physical constraints may be less likely prone to overfitting issues than those data-driven models with much higher degrees of freedom (i.e., more parameters) (Sungmin et al., 2020).

However, as the second phase of a model building process, model validation still shows some value in identifying model inadequacy as shown in our previous experiments. For model building cases using all data in calibration, we also saw a relatively higher fraction of false negative instances in model testing state classification (see in Section 4.4.1). Those false negative instances are when models were successfully built in the full-period dataset but actually failed in streamflow prediction in model testing period. Even though those fractions of false negative instances were minor (e.g., less than 10% of model building and testing instances), performing some sort of "validation" or overfitting check in model building may detect more model failures before applying the model operationally.

In this study, we explore possible alternative approaches to reduce the frequency of false negative instances when traditional validation is skipped in the model build process. The key motivation is from the uncertainty in the objective functions we employed. The evaluation of model performance in either calibration or validation is characterized by some quantitative performance metric (e.g., objective function), which typically calculate a single value for the entire period (e.g., calibration) to measure the closeness of the model and system response (e.g., see Bennett et al., 2013; and Moriasi et al., 2015). In the previous chapters, we employed 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). Both KGE and NSE have been extensively employed as objective functions in hydrological modeling, but the KGE 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).

However, some concerns are also related to the KGE metric. For example, some important features in a hydrograph, such as shape of rising limbs and recessions, and timing of peak flows, are all lumped into the single correlation component in the KGE metric (Knoben, Freer, & Woods, 2019).

Thus, it is suggested KGE components be separately evaluated for a better understanding of the overall KGE value. However, these components lose the initial physical meaning (Santos et al., 2018). The quantitative evaluation of uncertainty in KGE metric has also drawn attention recently (e.g., see Clark et al., 2021; and Vrugt & de Oliveira, 2022), and an important conclusion is that the KGE metric can be heavily influenced by just a few data points, which means a well-fitted hydrograph may be "overfitting" to some highly influential segments. Fowler, Peel, et al. (2018) tried to alleviate the "overfitting" to high flow years by calculating the KGE metric over individual years, which was called the Split KGE. They found the Split KGE metric-based calibration results are more balanced between dry and non-dry years. Similar methods have also been demonstrated to be able to improve model robustness when evaluating NSE at different time steps from daily to annual and decadal by Hartmann & Bárdossy (2005).

In this study, we post-process the model building and testing experiment results from Chapter 4 to further test alternative model "validation" methods. Calibrating models to all data is recommended in the previous chapters, but it naturally eliminates the traditional validation phase in model building. We therefore propose alternative validation methods such as reusing calibration data in a different way as a proxy validation to identify model failures. We ask the question:

For a streamflow prediction objective, how can hydrologists garner the substantial overall benefits of calibrating to all data while also better guarding against infrequent but undiagnosed overfitting?

We limit our scope to build models for supporting streamflow prediction in the future period, which is quite a common objective in operational model use. We also limit the scope to single-site and temporal validation problem, which is consistent with the previous chapters in this thesis. Also, further explorations can all be compatible with the massive experiments we have already conducted, and more approaches can be easily added into the new investigation in the future, if any.

In Section 5.2, we introduce the experimental design for testing alternative validation methods, the large-sample catchments data, model building data, and methodologies for performance assessment. The results and discussion are presented in Section 5.3. Finally, the conclusions are summarized in Section 5.4.

5.2 Data and methodology

This section describes data and key methodologies applied to assess alternative model validation methods. Section 5.2.1 introduces the experimental design for testing alternative validation methods in hydrological model building. Section 5.2.2 introduces catchment and data used in the validation experiments. Section 5.2.3 describes the hydrological model building data we employed for

the validation experiments. Section 5.2.4 presents methodologies to evaluate different validation methods in the model testing period.

5.2.1 Experimental design

In this study, we aim at exploring possible alternative model validation methods to identify model inadequacy in calibration, thereby enhancing the robustness of hydrological model building and improve model performance in post-validation model application periods. We employ model building and testing results based on six different split-sample test (SST) decisions from Chapter 4 and post-process these in different ways to evaluate three possible validation methods. These SST decisions all use recent years for calibration with different lengths (i.e., 30%, 50% and 70% of the available data in model building). The SST decisions employed in this study, as well as model building and testing results based on them are presented in Section 5.2.3.

This exploratory work purposefully avoids experimental assessments relying on the 100% CSP experiments from Chapter 4. This is because this chapter is focused on new proxy validation method development and the benchmarking any new methods against the 100% CSP results in Chapter 4 is left for future work after such methods are finalized.

Validation method 1 (VM1) follows the traditional calibration and validation routine that we firstly calibrate models to each of the six SST decisions, which are all continuous calibration sub-periods (CSPs) employed from Chapter 4. And then models are validated to the validation periods (pink blocks in Figure 5-1a1 and Figure 5-1a2). The VM1 repeats how models are generally built in practice in that it all data available is fully used in model building. Model calibration and validation are also benchmarked against the reference flows (see details in Section 5.2.3), which define the lower boundary of a plausible model. Model failure in calibration or validation will trigger the decision to disregard the model and the reference flow will be used in testing prediction instead. Also note that model validation performed in Chapter 3 and Chapter 4 both followed the VM1.

Validation method 2 (VM2) skips validating models to the validation period (while blocks in in Figure 5-1b1 and Figure 5-1b2) while providing a proxy validation. Instead of how we validate models in the VM1, we perform a validation-like analysis to evaluate if the calibration result is acceptable and we do so by reusing the calibration data in a new way. Firstly, we calibrate models to each of the six SST decisions, and calibration results are benchmarked against the reference flows, which is similar to VM1. Then, we reuse calibration data as the proxy validation. The proxy validation calculates Split Kling-Gupta Efficiency (KGE) (see in Section 5.2.3) of each individual year in the calibration period and the Split KGE values are compared to their corresponding reference flows at

each individual year. The rule of thumb to identify model failures in proxy validation is when the Split KGE value is worse than reference flows (measured as Split Reference KGE) in 50% or more years. Counting the number of inadequate years can effectively identify hydrographs "overfitting" to specific high flow years. Although validation period is required to be independent to calibration period, we here use calibration data in a new way, post-calibration, to detect unacceptable calibration results by characterising hydrographs "overfitting" to high influential years in calibration at the annual scale (Fowler, Peel, et al., 2018; Gharari et al., 2013). Doing so avoids using Split KGE information in both the calibration objective function and proxy validation.

Validation method 3 (VM3) is skipping validation entirely and in fact, unlike Chapter 3 and Chapter 4, discarding some of the available model building data. This fencepost benchmark is useful for two reasons. First, in comparison with VM1, we can evaluate if for a fixed calibration result, traditional validation improves or degrades model performance in the model testing period. Second, in comparison with VM2, we can now assess if proxy validation is helpful relative to the recommended approach from Chapters 3 and 4 to calibrate the full-period CSP (e.g., we essentially pretend the validation period data used in VM1 is not available for model building). We firstly calibrate models to each of the six SST decisions, and then models are benchmarked against reference flows. Adequate models will be directly used in testing prediction, while inadequate models will be disregarded and we use reference flow for testing prediction instead, which is similar to VM1 and VM2.

All models are calibrated and validated in hydrological year (period of 1 October to the next 30 September), while model testing is in calendar year (period of 1 January to 31 December) since testing periods serve as common periods to compare different SST decisions and hence are not influenced by either hydrological or calendar year configuration.

The first hydrological year of available data (1 October 1980 to 30 September 1981) is always used for model spin-up. We use 1980 data recursively for three times to define a "three-year" spin-up period to initialize the hydrological model (i.e., force models with meteorological inputs in 1980 and repeatedly run the model in 1980 for three times with the end-of-day states on 30 September 1981 in the first 1980-run being the initial states on 1 October 1980 in the second 1980-run, and so forth), which is consistent with how we initialize, calibrate, validate, and test models in Chapter 4. Although this leads to a gap (white blocks in Figure 5-1) between the spin-up period and model calibration period, we believe the impact of the initial state variables of calibration period in VM2 and VM3 are minor, because the model initialization in 1980 has achieved a "practical" equilibrium state and thus, running models in extra while-block years in Figure 5-1 will have minor impact on the following calibration

period except more computational cost, but it is affordable as we only post-processed results from Chapter 4.

The full testing period for each model build year is shown as red blocks in each panel of Figure 5-1. Each of those five full testing periods are augmented with four additional shorter length testing periods (i.e., the first three and first five years of the entire testing period, the last three and last five years of the entire testing period, and the entire testing period. This is also consistent with the testing period setting in Chapter 4.

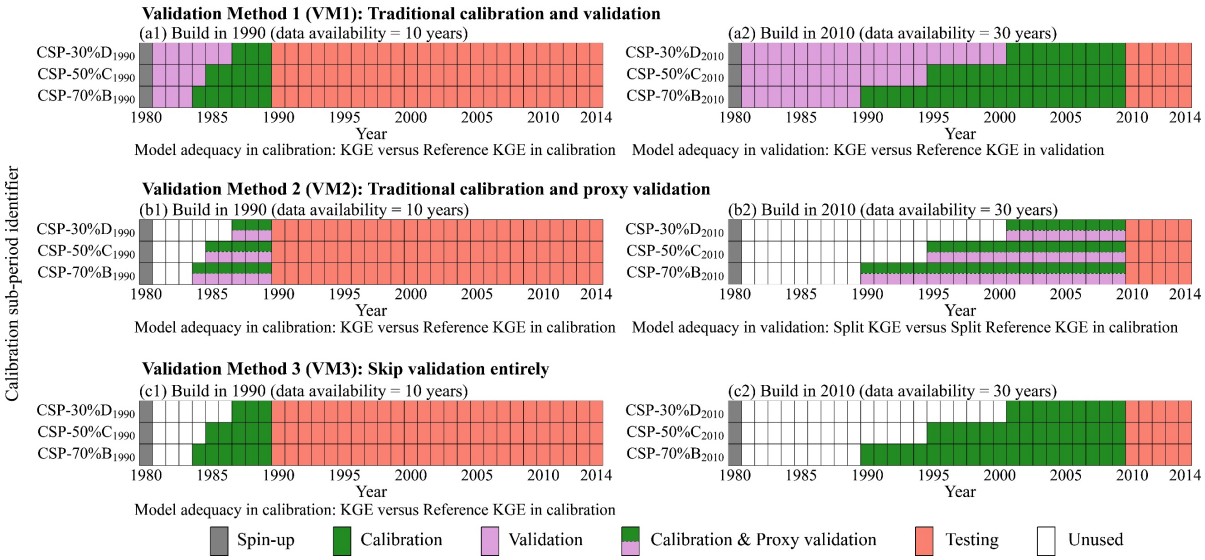

Figure 5-1. Experimental design of testing three validation methods (VMs), each including two model build year scenarios (left panel 1990 and right panel 2010). Six continuous calibration sub-periods (CSPs) are adopted in model building and testing. Continuous splits and full-period CSP are denoted as CSP-x%y_z (where x% is the percentage of calibration data in available data for model building, y is an identifier to distinguish different sub-periods with same x, which is skipped for the full-period CSP for brevity, and z is the model build year, which may be skipped hereafter if its meaning is clear in the context).

5.2.2 Catchments and data

The Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) data set are used in this study, which provides 671 catchments that are minimally impacted by human activities across the contiguous United States (CONUS; Addor et al., 2017; Newman et al., 2015). We used the 463-catchment subset of CAMELS deliberately selected and processed in Chapter 3 (Shen et al., 2022a), which is also consistent with Chapter 4. The Daymet forcings (archived in the CAMELS data set) and the observed daily streamflow data at these 463 gauges (originally archived in the CAMELS

and missing data infilled using newer streamflow data from the National Water Information System of the USGS by Shen et al. (2022b)) are used in this study, both spanning from 1 January 1980 to 31 December 2014. More details on gauge selection can be found in Shen et al. (2022a) or Section 3.2.2.

The map for the spatial locations of all CAMELS catchments (including the 463 selected catchments and other filtered catchments) is presented in Figure A1-1 in Appendices A-1. The detailed information of the 463 catchments and the corresponding Daymet forcings and updated USGS streamflow data files for these catchments are all available online (see Appendices A-2).

5.2.3 Hydrological model building data

We select six recent continuous calibration sub-period (CSPs) with calibration period lengths in 30%, 50%, and 70% of the data availability in model build year 1990 and 2010 (see in Figure 5-1), as well as their model building and testing results from Chapter 4. We only post-process the results to test different model validation methods.

The HMETS (which stands for Hydrological Model of École de technologie supérieure) is calibrated to each of the six SST decisions at the 463 CAMELS catchments. The dynamically dimensioned search (DDS) algorithm (Tolson & Shoemaker, 2007), which has been widely applied in HMETS calibration studies (Chlumsky et al., 2021; Mai, Shen, et al., 2022; Martel et al., 2017; Shen et al., 2022a), was employed in the calibration experiment. DDS was implemented in the optimization and calibration software toolkit OSTRICH (Matott, 2017). We utilized a budget of 3,000 model evaluations per optimization trial. HMETS with each of the SST decisions was calibrated to 20 independent optimization trials, and different randomly generated initial parameter sets were adopted to these independent trials to minimize the influence of initial conditions on DDS. The model is calibrated, validated and tested using the Kling-Gupta Efficiency (KGE) metric (Gupta et al., 2009).

Accordingly, the total number of model calibration and validation problems we assessed in this study is 166,680 (6 splits × 20 trials × 463 catchments × 3 validation methods), and the total number of model testing hydrographs assessed is 222,240 (3 splits × 20 trials × 463 catchments × 5 testing periods – 3 splits × 20 trials × 463 catchments × 3 testing periods). Note that there are only three non-repeated testing periods in 2010.

In VM2 in Section 5.2.1, we applied a new metric in the proposed proxy validation method, which is the Split KGE. The Split KGE metric considers the KGE value of each individual (hydrological) year and equally quantify the influence of each year (Fowler, Peel, et al., 2018). However, we avoid averaging Split KGE values of all years into the final value as Fowler, Peel, et al. (2018) did, since the KGE metric can be highly skewed with very negative values (e.g., -10 or even

smaller), hence the simple averaging may lose information when assessing the proxy validation results. Instead, we apply a fairly objective rule of thumb that defines model failure (inadequacy) in proxy validation to be when the Split KGE value is worse than reference flows (measured as Split Reference KGE) in 50% or more years. Since our goal of applying the proxy validation is to identify hydrographs may be overfitted to specific years (e.g., high flow), such a simple rule can effectively characterise the overfitting benchmarked against reference flows.

The reference flow is established by calculating the mean value of observed streamflow on the reference period at the daily scale (Knoben et al., 2020). Reference period for model calibration, validation and testing periods is constantly all data years prior to the model build year (i.e., spin-up, calibration and validation) (Shen et al., 2022a). More details on calculating reference KGE, which is the observation-based metric resulted from the reference flow and observed flow, are described in Shen et al. (2022a) (see in Section 3.2.5.1). Similarly, the Split Reference KGE is calculated based on individual hydrological year data.

5.2.4 Alternative validation methods assessment

This section introduces methodology applied for assessing performance of different validation methods. Since our objective and scope in this experiment are consistent with Chapter 4, we can employ the comparative performance assessment used in Section 4.2.4 to analyze how different validation methods influence model performance in the testing period. Note that we consider all 20 model optimization trials for each SST decision in the following assessments.

First, we perform a graphical analysis on hydrographs at all catchments to screen inconsistent model inadequacy results identified from (a) KGE and reference KGE in model building and (b) Split KGE and Split Reference KGE in model building. The inconsistency may highlight the significance of using Split KGE metric to measure more detailed goodness-of-fit variability in a hydrograph during calibration, thereby supporting the use of Split KGE for the proxy validation in VM2.

Second, we view each SST decision as a binary classifier which yields only two types of states, i.e., model is adequate (success) or inadequate (failure) (Shen et al., 2022a). Binary classifiers are used to assess whether the model building is adequate (success) or inadequate (failure) for testing period prediction. Further, we can actually assess whether the model building is truly adequate or truly inadequate in the testing period. Model failures are handled as described in Section 5.2.1. Four possible classes of our hydrological model building (i.e., calibration plus validation) and testing states are defined and are consistent with Shen et al. (2022a) (see in Section 3.2.5.4). Note that we define

"positive" as the not-normal class, which is a model failure, while "negative" represents normality, which is a model success.

We further employ two metrics, i.e., the classification accuracy and fractions of False Negative (FN) and False Positive (FP), to interpret these four classes in our hydrological modeling context that we modelers would care about (a) how accurate a model can predict the hydrograph in both building and testing and (b) how often the model may be failed and what the consequences are. More details of these two metrics are presented in Section 4.2.4.3. In this assessment, we applied reference KGE-based thresholds to add an additional performance expectation (i.e., 0 and 0.2) in model calibration during the model failure handling. The KGE thresholds are defined in Equation (4-2) in Section 4.2.4.3. Testing more thresholds may be possible but it is not a significant influential factor, as results in Section 4.3.3 showed different thresholds generally affect the metric magnitude but do not change result patterns.

Third, we frame the model building processes as a multi-objective decision-making problem to simultaneously optimize two objectives: Objective one is to maximize the testing period KGE quantified as median KGE, and objective two is to maximize the performance of splits functioning as binary classifiers quantified as classification accuracy. We consider the tradeoff analysis between median KGE and classification accuracy using all 20 optimization trials of each validation method per SST decision per gauge.

Each tradeoff contains three validation method solutions. Considering each solution is derived from trial samples of a single gauge, it is not distinguishable to compare how many times a validation method may be the non-dominated solution. Similar to Section 4.2.4.4, we adopt the percent distance (PD) metric to depict the degree of each solution approaching the optimal values in their tradeoff analysis. A graphical example of how percent distance is calculated at each tradeoff is displayed in Figure 4-6 in Section 4.3.4. The percent distance ranges from 0 to 100%, and 100% is the perfect value that denotes the solution on average ranks the highest frequency to be the optimal one among all solutions.

5.3 Results and discussion

5.3.1 Disproportional influence of high flows on KGE: Overfitting to high-flow years

Figure 5-2 displays an example hydrograph of the HMETS model built in 1990 on CSP-50%B_{1990}, as well as the performance metrics KGE, reference KGE, Split KGE and Split reference KGE. The two panels in Figure 5-2 displays the same hydrographs, while Figure 5-2a shows both calibration and validation periods, thus demonstrating the validation method 1 (VM1). Figure 5-2b

shows calibration period and proxy validation, which demonstrates the validation method 2 (VM2) and validation method 3 (VM3). The testing period in this example (i.e., the first five years of the testing period) is identified as model failure but not shown in Figure 5-2 for brevity.

Figure 5-2a shows the simulated hydrograph is acceptable based on VM1. The calibration and validation KGE values (0.78 and 0.56, respectively) both exceed the reference KGE values (0.66 and 0.53, respectively), denoting the model building is adequate (success). However, the testing results (not shown here) are identified as inadequate (failure), hence the model building and testing is false negative, demonstrating this model building does not correctly predict testing states. The VM2 and VM3 both skip the traditional validation period in Figure 5-2b. Similar to the VM1, VM3 result of this example is false negative, which is not preferred in practice. However, VM2 result is improved by the proxy validation.

The proxy validation results show large variability in the Split KGE, which ranges from -0.04 to 0.96. Based on the count of Split KGE exceeding Split Reference KGE, the proxy validation shows model building is inadequate, which is consistent with the testing results. Thus, the model building and testing based on VM2 is true positive. It can be seen that the simulated flow fits the observed flow well in the high flow year 1986 (Split KGE 0.96), however, the simulated flow in 1985, 1987, and 1989 are all worse than reference flows. This indicates the calibration KGE may be "overfitted" to individual years (e.g., 1986) and the KGE value cannot objectively represent the goodness-of-fit in every year. The Split KGE and Split Reference KGE can reveal the closeness of fitting in each year with equal weight. Aggregating Split KGE values to either mean or median may not be able to properly reveal model failures, since extreme values will affect the mean significantly, while simply ignoring very bad Split KGE (e.g., negative values) in median will lose critical information in model building that the model may be overfitted to individual years. The way we count number of years may overcome the shortages in mean and median.

Among all hydrographs of the six SST decisions we are assessing (9,620 hydrographs per SST decision per build year), we further check how different the model failure results can be when using KGE and Split KGE to evaluate the calibration and validation. We evaluate the percentages of hydrographs with adequate traditional validation results but inadequate proxy validation. These hydrographs are similar to the example in Figure 5-2. The results show such percentages are 6%–18% for the three SST decisions in 1990 and 4%–6% for the three SST decisions in 2010. Thus, the proxy validation may have some promise to improve model building.

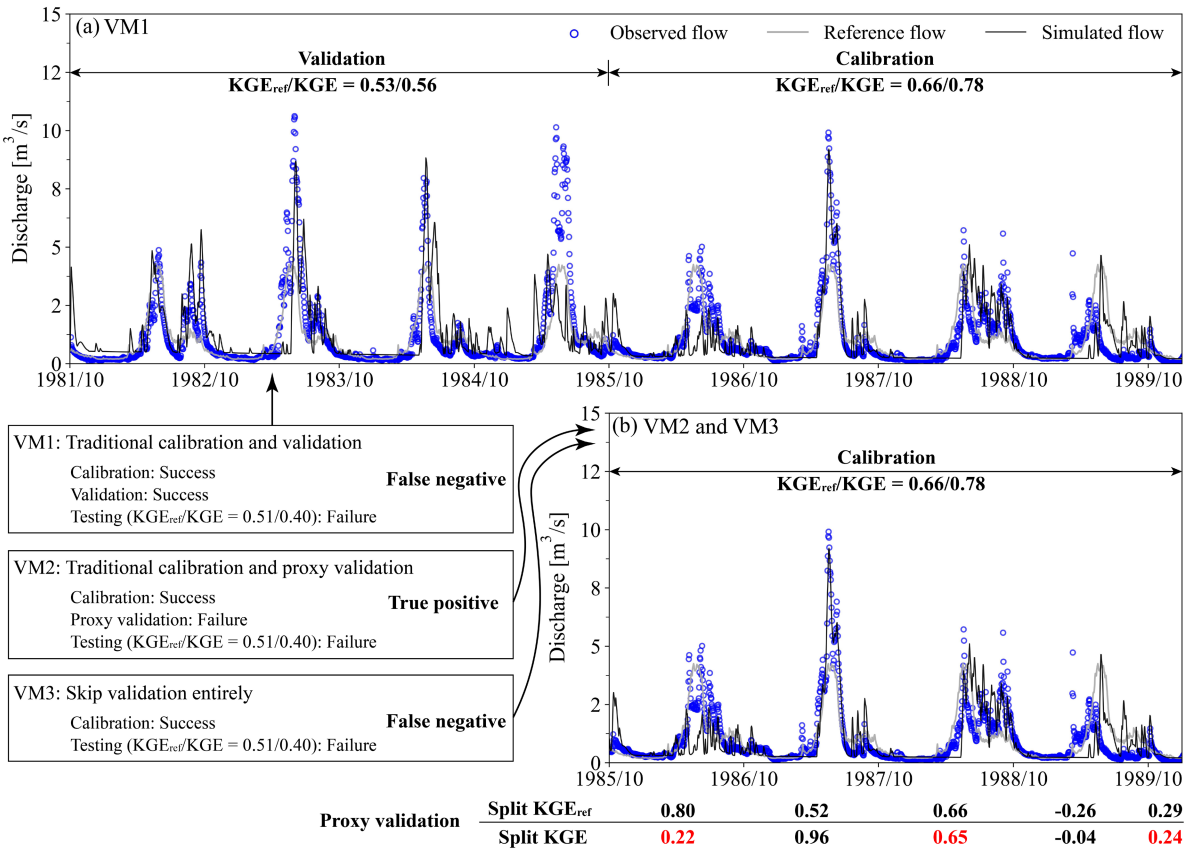

Figure 5-2. Example hydrographs at CAMELS gauge 08377900 (RIO MORA NEAR TERRERO, 139 km^2). The hydrographs contain the observed flow, the reference flow and the HMETS-simulated flow in (a) calibration and validation periods (for validation method 1 (VM1)) and (b) calibration period and proxy validation (for VM2 and VM3). The HMETS model building in this example is based on the SST decision CSP-50%C$_{1990}$. The KGE, reference KGE (KGE$_{ref}$), Split KGE, and Split Reference KGE (Split KGE$_{ref}$) metrics are highlighted in the two panels. Note that flows in the spin-up period and testing period are not displayed, and the testing period in this example (testing HMETS model in the first 5 years of the testing period) is identified as model failure. The model building and testing states classified by the three validation methods are summarized in the lower-left boxes. False negative stands for the model built as a success is actually inadequate in testing period. True positive stands for the inadequate model is correctly predicted in testing period.

5.3.2 Ability to correctly classify model failures in the model building and testing

The classification accuracy and fractions of False Negative (FN) and False Positive (FP) derived from each gauge are reported in this section. A validation method corresponds to 20 trials results per SST decision per catchment. To compare how these metrics vary across the large-sample catchments, we aggregate 20 trials results of a single gauge as one data point in the following figures.

Since the result patterns are similar between different calibration thresholds and different testing periods, we only report results based on reference KGE-based thresholds with zero additional performance expectation in the first five years of testing period for brevity. Other results are considered in the assessment. Figure 5-3 displays the empirical cumulative distribution functions (ECDFs) of the three validation methods (VMs).

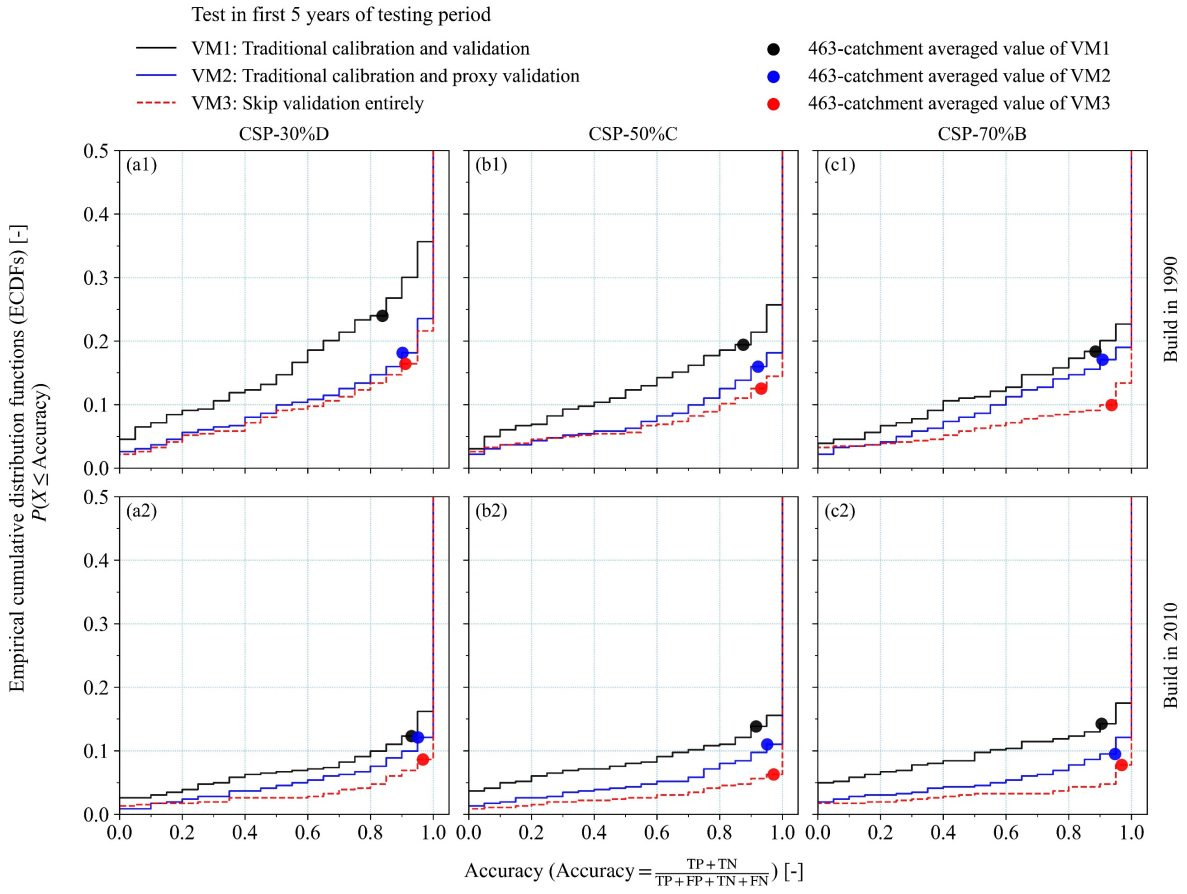

Figure 5-3. Empirical cumulative distribution functions (ECDFs) of classification accuracy for the three different validation methods (VMs) in the first five years of testing period. Panels in each column stand for different continuous calibration sub-periods (CSPs), while the top and bottom panels are for build year 1990 and 2010, respectively. The thresholds for classifying model failures in building and testing are the reference KGE added with a variable value Δ (0 in this case).

It can be seen no matter what SST decisions or validation methods are applied in model building and testing, the majority of gauges have 1.0 accuracy values, indicating SST decisions work correctly in testing period. This is more evident in 2010 (bottom panels in Figure 5-3). Comparing the three validation methods, it can be seen VM3 (skipping validation entirely; dashed line in Figure 5-3) is overall the best one with respect to less gauges with small accuracy values. The VM1 (traditional

calibration and validation; black line) appears to be the worst, especially in 1990. The VM2, which applies the proxy validation to replace traditional validation, achieves improvement compared to the traditional validation. Figure 5-3 also presents a tendency that when model is built in 1990 with only 10 years of data available for calibration and validation, the SST decisions with longer calibration data may lead to a better accuracy ECDF over the 463 catchments.

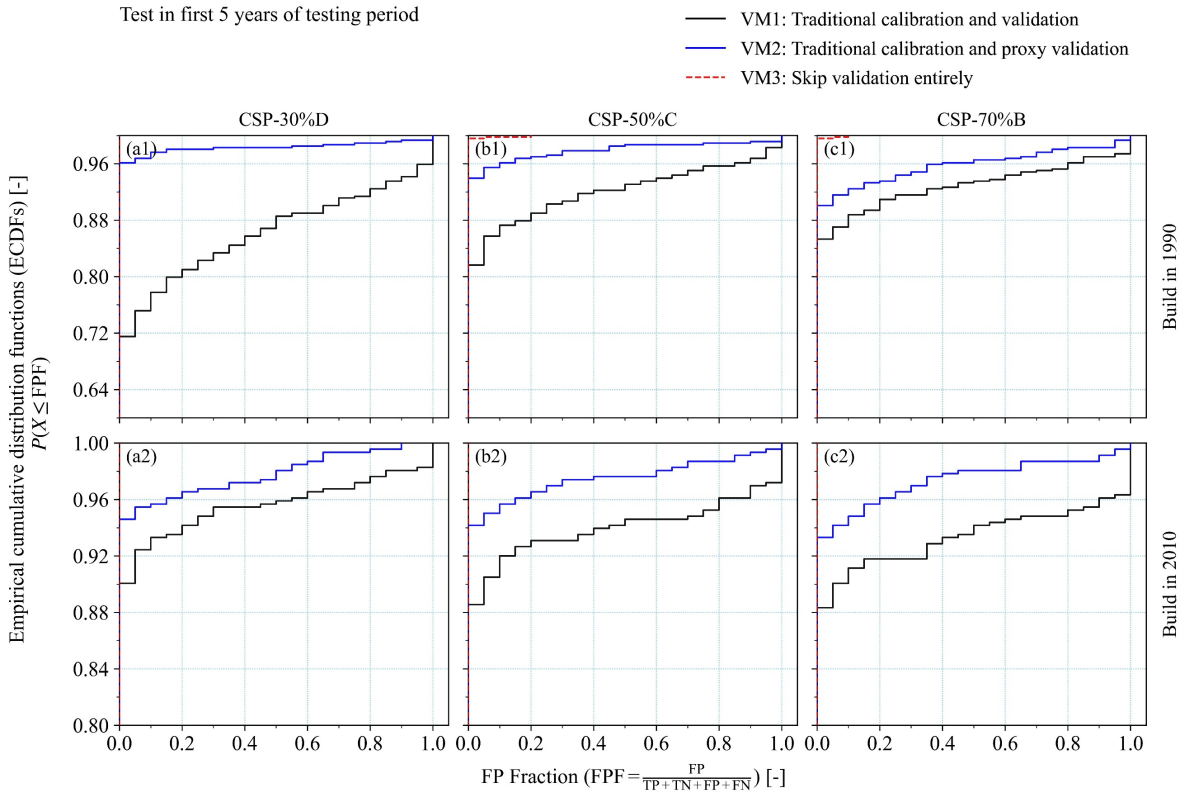

Figure 5-4. Empirical cumulative distribution functions (ECDFs) of fractions of false positive (FPF) for the three different validation methods (VMs) in the first five years of testing period. Panels in each column stand for different continuous calibration sub-periods (CSPs), while the top and bottom panels are for build year 1990 and 2010, respectively. The thresholds for classifying model failures in building and testing are the reference KGE added with a variable value Δ (0 in this case). Note that the dashed lines for VM3 almost coincide with the *y*-axis in each panel.

The accuracy metric provides the total fraction when an SST decision with a validation method is working correctly. We further assess the two classes when such an SST decision is working incorrectly, i.e., the fractions of FP and FN shown in Figure 5-4 and Figure 5-5, respectively.

False positive instances is classified when model building is failure and testing is success. Figure 5-4 shows the VM1 in all panels is the worst, and gauges with FP instances can be near 0.3 in CSP-30%D_{1990}. The VM3 is overall the best, as its FP for all gauges approaches zero (i.e., ECDFs

coincide with *y*-axis). The VM2 shows a notable improvement to the VM1. The opposite pattern is observed in the fraction of FN in Figure 5-5, which displays the VM3 is usually the worst one while the VM1 can be the best, especially in 1990. The VM2 is generally in between the other two methods, showing its balance between the FP and FN. Considering only the false negative metric in model testing (rate of not identifying inadequately calibrated models) the VM2 improved upon VM3 by using Split KGE and Split Reference KGE. Considering only the false positive metric in model testing (rate of not identifying adequately calibrated models), the VM2 improved upon the traditional approach VM1.

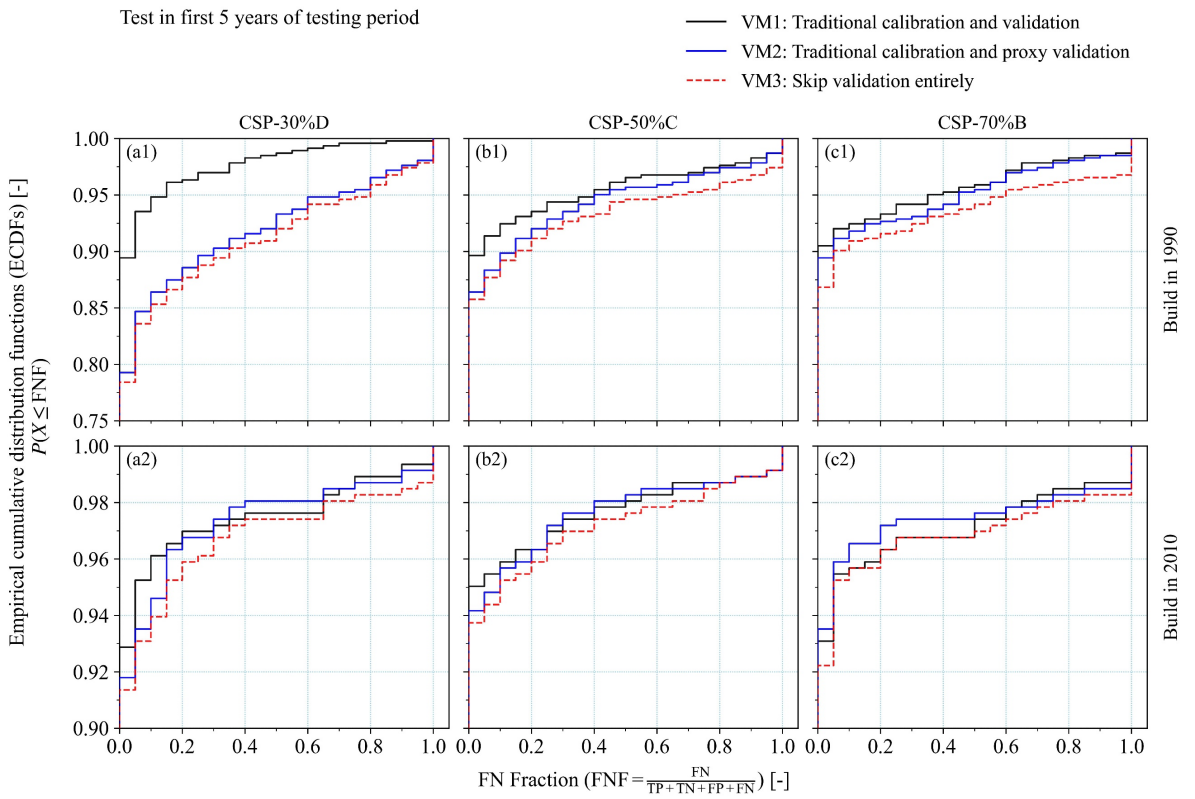

Figure 5-5. Empirical cumulative distribution functions (ECDFs) of fractions of false negative (FNF) for the three different validation methods (VMs) in the first five years of testing period. Panels in each column stand for different continuous calibration sub-periods (CSPs), while the top and bottom panels are for build year 1990 and 2010, respectively. The thresholds for classifying model failures in building and testing are the reference KGE added with a variable value Δ (0 in this case).

In practice, the FP instances may require modelers to put more efforts in rebuilding the model until it yields acceptable results in calibration and validation to ensure there is at least one model available for future period applications, which can be much computationally expensive. However, FN instances may lead to model failure in the future period applications (e.g., wrong prediction) and thus, failing to support specific decision-making. Therefore, it is necessary to weigh the consequences of

both FP and FN for specific applications, and the above results of the three validation methods provide distinctive advantages in reducing one of the costs or balancing both of them.

5.3.3 Multi-objective assessment of validation methods

The more robust multi-objective analysis between median KGE and accuracy at each of the 463 gauges are measured by the percent distance (PD) metric in Figure 5-6. It is seen that the VM3 consistently has the largest PD values in all SST decisions, showing that skipping validation entirely can always be the optimal solution with respect to both median KGE and accuracy compared to other two validation methods. The VM2, though still in between the other two methods, shows about 10% improvement in PD compared to VM1 in 1990 for both median KGE and accuracy. The difference between the VM2 and VM3 are minor compared to difference between the VM1 and VM3.

Overall, the result pattern of VM1 and VM3 is consistent with the findings reported in Chapter 3 and Chapter 4 that calibrating models to all data is superior to other possible splits we assessed. The evidence also supports using the proxy validation, i.e., Split KGE and Split Reference KGE in calibration, to identify unacceptable model calibration results. Even though the PD in accuracy and median KGE of VM2 are slightly worse than the VM3, the VM2 can still be valuable in practical model building, since the costs of FP and FN may be critical factors that need more attention. The VM2 can better balance the two costs than the VM1 and VM3, which shows some promise to apply the proxy validation in practical model building, especially when all data is used in model calibration.

This study adopted three validation methods to explore possible approaches to identifying unacceptable model calibration results that are not detected by the overall KGE metric. We tested these validation methods in six recent CSPs from short to long length; however, it should be noted that such exploration can still be varied, as such more alternative validation methods can be added into this experiment. For example, one of the most important implications from this study is the VM2 improves both VM1 and VM3 by using Split KGE and Split Reference KGE. This may be further extended to applying the Split KGE in model calibration to directly consider the disproportional influence of different years on KGE metric, which is similar to what was done by Fowler, Peel, et al. (2018). However, as discussed in Section 5.2.4, directly averaging Split KGE of different years may be problematic due to potential very negative values in some years. As such, new ways to aggregating Split KGE values as a practical objective function in optimization are needed. Another example for validating models is the K-fold cross-validation, which ensures every fold can be adopted in both calibration and validation, thus making use of the information of the full dataset in model building (Kohavi, 1995). As such, model can be more deeply validated. However, this may eventually require

calibrating each of the K-fold splits instead of post-processing what we have done in previous chapter, and the computational cost will be extremely significant in practice.

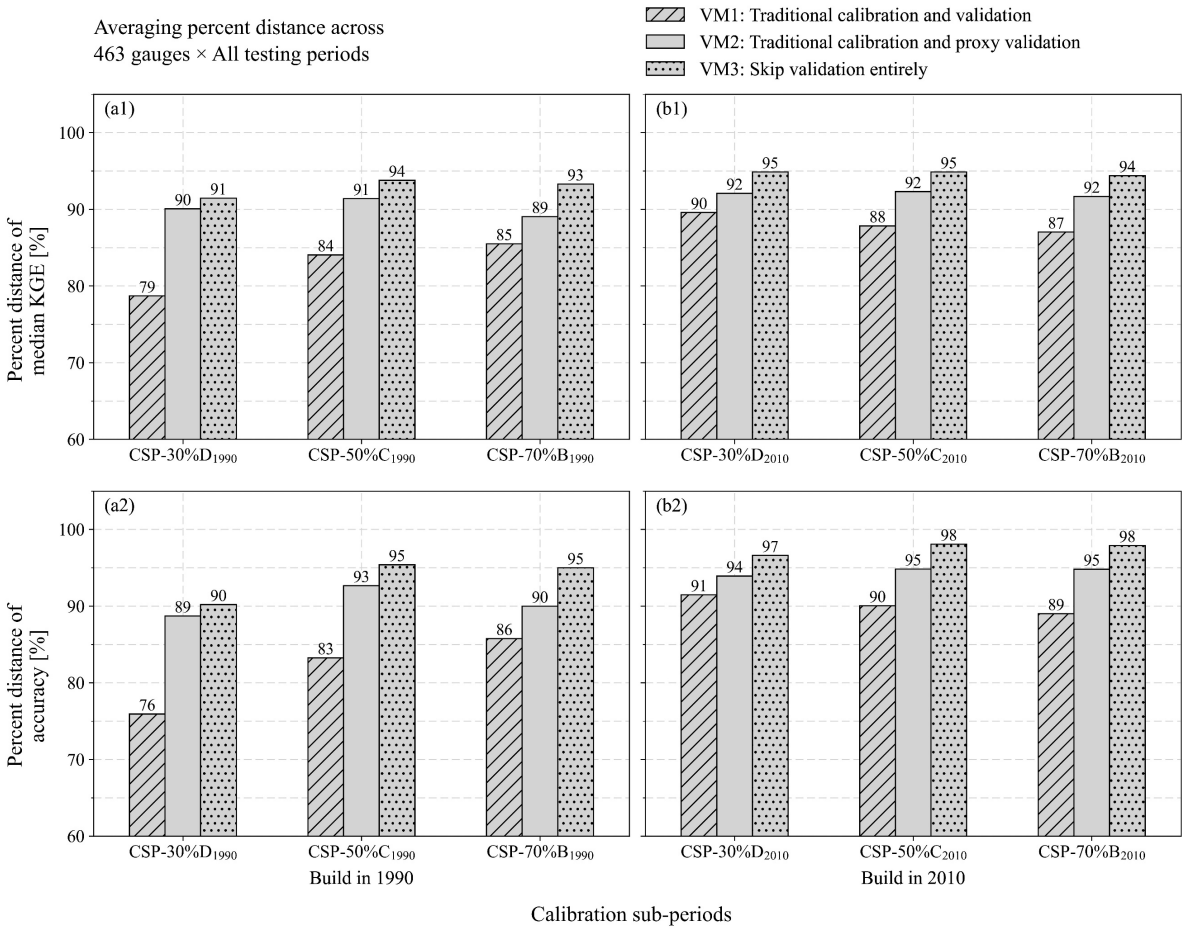

Figure 5-6. Percent distance (PD) of median KGE (top panels) and accuracy (bottom panels) for the three validation methods (VMs) based on the multi-objective decision-making problem to simultaneously optimize median KGE and accuracy in model testing periods. The percent distance of each split is calculated at single gauges and averaged across all 463 gauges and all testing periods in build year 1990 (left column) and 2010 (right column).

5.4 Conclusions

In this study, we post-processed some of the model building and testing experiment results from Chapter 4 to further explore alternative model validation methods that could enhance model robustness. We tested six continuous calibration sub-periods (CSPs) that utilize recent data for calibration and older data for validation. These recent CSPs-based model building and testing results are post-processed in three different ways corresponding to three validation methods (VMs). Our validation methods include the traditional calibration and validation approach (VM1), the traditional

calibration and a proxy validation that use Split KGE and Split Reference KGE in calibration to identify unacceptable models (VM2), and skipping validation entirely (VM3). The experiments were conducted in 463 catchments across the US and were assessed in multiple aspects such as the accuracy of each CSP with a validation to correctly predict model failures in model building and testing and the costs of failed to do so. The accuracy and median KGE in the testing period of each validation method are also framed as a multi-objective problem to be optimized simultaneously.

Our validation experiments tested on a large-sample of catchments generally indicated using Split KGE and Split Reference KGE can be more effective to detect hydrographs that may be overfitted to specific years such as high flow year. The results further showed that surprisingly, VM3 showed overall the best performance in the model testing period with respect to the accuracy and KGE implying that even completely disregarding available data for validation (e.g., not using in the model build process) is a better approach than using it to try and validate the model. This is a different finding relative to previous chapters where not validating the model meant instead using all available data for calibration. However, looking beyond accuracy and KGE, considering only the false negative metric in model testing (rate of not identifying inadequately calibrated models) the VM2 improved upon VM3 by using Split KGE and Split Reference KGE. Considering only the false positive metric in model testing (rate of not identifying adequately calibrated models), the VM2 improved upon the traditional approach VM1. Such alternative validation method provides a good example for model calibration based on all data that a proxy validation can still be possible to further improve model robustness. We believe the exploration shown in this study is promising to enhance model robustness when all data are used in calibration. We suggest exploration on more possible alternative validation methods be added into the experiment in the future.

Chapter 6
Conclusions

This thesis achieves the four objectives summarized in Chapter 1. Findings in this thesis are based on massive modeling experiments totaling over 1.3 million model calibration experiments (each solved by an optimization algorithm) and 1.9 million model testing assessments. Major findings from Chapter 3 to Chapter 5 are summarized in Section 6.1. Limitations and future work are discussed in Section 6.2.

6.1 Major findings

Chapter 3 established a novel and comprehensive split-sample test (SST) experimental assessment and applied to two conceptual hydrological models in 463 catchments across the United States. We evaluated 50 different continuous calibration sub-periods (CSPs) for model calibration (varying data period length and recency) across five different model build year scenarios to ensure results are robust across three 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. Key findings are that:

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. This is exactly the opposite approach to what is typically done in hydrological modeling studies.
2. Calibrating a model to the full available data period and skipping temporal model validation entirely is the most robust choice.

Chapter 4 employed the evaluation framework proposed in Chapter 3 to further assess 44 representative continuous and discontinuous split-sample test (SST) decisions using a conceptual hydrological model in 463 catchments across the United States, and the extensive experiments results are thoroughly assessed in similar ways we did in Chapter 3 except that all model optimization trials were all considered for even more robust conclusion drawn. Key findings are that:

1. The empirical results considering new discontinuous SST decisions in model building and testing assessment strongly support that the SST recommendations made in Chapter 3 (e.g., calibrating models to all available data is the most robust choice) is still valid even when representative discontinuous SST decisions are evaluated under the same framework.

2. Strong evidence shows the superiority in calibrating to all data, while those discontinuous SST decisions have no clear advantage over the full-period CSP.
3. We recommend that hydrological modelers rebuild models after their validation experiments, but prior to operational use of the model, by calibrating models to all available data.

Chapter 5 post-processed the model building and testing experiment results from Chapter 4 to explore alternative model validation methods that could enhance model robustness. Three validation methods were tested with six continuous SST decisions in 463 catchments across the United States. The three validation methods consisted of the traditional calibration and validation approach (VM1), the traditional calibration and a proxy validation that use Split KGE and Split Reference KGE in calibration to identify unacceptable models (VM2), and skipping validation entirely (VM3). Key findings are that:

1. Compared to VM1 (traditional approach), using Split KGE and Split Reference KGE can more effectively detect hydrographs that may be overfitted to specific years such as high flow years.
2. The VM3 showed overall the best performance in model testing period with respect to the accuracy and KGE. However, considering only the false negative metric in model testing (rate of not identifying inadequately calibrated models) the VM2 improved upon VM3 by using Split KGE and Split Reference KGE.
3. The exploration shown in this study has some promise to enhance model robustness when all data are used in calibration and a proxy validation type of approach like the Split KGE is utilized. We suggest exploration on other alternative proxy validation methods be added into the experiment in the future.

6.2 Limitations and future work

Several limitations are listed below to motivate future work.

1. The scope of this thesis is addressing the problem of single-site and temporal model calibration and validation. Our recommendations may turn out to be applicable for multi-site model calibration and spatial validation problems (e.g., see Mai, Shen, et al., 2022) but further exploration is needed to determine this.
2. Our recommendations do not apply to climate change impact assessment studies focused on carefully assessing and ensuring parameter transferability under contrasting climates. An assumption in this thesis is that we do not know how testing (future) conditions will be in our experiments. However, if the future conditions are known with high certainty, such as provided by the global climate model (GCM) or regional climate model (RCM), plus the intention is to assess

how hydrological changes will manifest in the future period, it is appropriate to build models over a historical period with similar conditions to the future, which can be referred to the DSST studies (e.g., see Bai et al., 2021; Coron et al., 2012; Fowler et al., 2016; and Motavita et al., 2019).

3. We only used one model calibration objective function when assessing different data splitting methods: KGE of daily discharge. The integrated KGE metric, having the same constitutive components as 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. Recent work in Clark et al. (2021) could be used in future work to account for the uncertainty in KGE in our experimental design. As we showed in Chapter 5, the Split KGE is useful to identify unacceptable hydrographs and thus, it may be possible to apply Split KGE as the calibration objective function in the future.
4. In all experiments in this thesis, it was assumed the model development process is intended to generate a model that is calibrated deterministically, thus using only a single SST. However, doing a single split for model building naturally skips a particular sub-period in calibration or validation. As results imply in Chapter 4, validating models to some different data can be valuable to identify bad models. Future work could focus on making use of all data for both calibration and validation during model building, e.g., using the K-fold cross-validation method.
5. Here, we only applied two conceptual lumped hydrological models in the experiments. Testing other models with different complexities may be helpful to further generalize our conclusions. For example, distributed models can be appealing, as they can be better representation of physical processes and can make use of information from discretized watershed. To achieve this goal, the CAMELS dataset with only lumped forcing data at the basin scale may be undesirable, and hence other large-sample catchment datasets are needed, such as the dataset in the Great Lakes Region (Mai, Shen, et al., 2022) which contains long-term gridded meteorological data and high-resolution geo-spatial data required for distributed modeling. Studies on watershed discretization can also be beneficial for this purpose to provide basic GIS layers for distributed modeling (e.g., see Han et al., 2023).
6. There are numerous future studies that could mine the existing suite of experimental results in this thesis. Examples include replacing accuracy metric with alternative confusion matrix metrics focused on false negatives, mining how findings change spatially and fully investigating the early findings in Chapter 5 suggesting there are benefits associated with not using older data at all in model building (either in calibration or validation).

References

Abbott, M. B., & Refsgaard, J. C. (2012). *Distributed hydrological modelling* (Vol. 22). Springer Science & Business Media.

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'Connell, P. E., & Rasmussen, J. (1986). An introduction to the European Hydrological System—Systeme Hydrologique Europeen,"SHE", 1: History and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology*, *87*(1–2), 45–59.

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, Nans, 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

Addor, Nans, Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., & Mendoza, P. A. (2020). Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal*, *65*(5), 712–725. https://doi.org/10.1080/02626667.2019.1683182

Ahmed, M. I., Stadnyk, T., Pietroniro, A., Awoye, H., Bajracharya, A., Mai, J., et al. (2023). Learning from hydrological models' challenges: A case study from the Nelson basin model intercomparison project. *Journal of Hydrology*, *623*(June). https://doi.org/10.1016/j.jhydrol.2023.129820

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.

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., et al. (2018). The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies–Chile dataset. *Hydrology and Earth System Sciences*, *22*(11), 5817–5846.

Anctil, 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

Andréassian, V., Perrin, C., Berthet, L., Moine, N. Le, Lerat, J., Loumagne, C., et al. (2009). HESS Opinions" Crash tests for a standardized evaluation of hydrological models". *Hydrology and Earth System Sciences*, *13*(10), 1757–1764.

Arsenault, R., Essou, G. R. C. C., & Brissette, F. P. (2017). Improving hydrological model simulations with combined multi-input and multimodel averaging frameworks. *Journal of Hydrologic Engineering*, *22*(4), 1–11. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001489

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

Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., et al. (2020). A

comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds. *Scientific Data*, *7*(1), 243.

Ayzel, G., & Heistermann, M. (2021). The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: A case study for six basins from the CAMELS dataset. Computers & Geosciences, 149, 104708.

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*. *British Columbia, Canada*. Retrieved from http://epe.lac-bac.gc.ca/100/200/300/forrex/forrex_series/FS25.pdf

Bedient, P. B., Huber, W. C., & Vieux, B. E. (2008). Hydrology and floodplain analysis.

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 and 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

Beven, K, Calver, A., & Morris, E. M. (1987). The Institute of Hydrology distributed model.

Beven, Keith. (1989). Changing ideas in hydrology—the case of physically-based models. *Journal of Hydrology*, *105*(1–2), 157–172.

Beven, Keith. (1997). TOPMODEL: a critique. *Hydrological Processes*, *11*(9), 1069–1085.

Beven, Keith. (2012). *Rainfall-Runoff Modelling*. *Rainfall-Runoff Modelling*. https://doi.org/10.1002/9781119951001

Bicknell, B. R., Imhoff, J. C., Kittle Jr, J. L., Donigian Jr, A. S., & Johanson, R. C. (1997). Hydrological simulation program—FORTRAN user's manual for version 11. *Environmental Protection Agency Report No. EPA/600/R-97/080. US Environmental Protection Agency, Athens, Ga*.

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.

Boughton, W. (2004). The Australian water balance model. *Environmental Modelling & Software*, *19*(10), 943–956.

Bowles, D. S., & O'Connell, P. E. (2012). *Recent advances in the modeling of hydrologic systems* (Vol. 345). Springer Science & Business Media.

Brunner, P., & Simmons, C. T. (2012). HydroGeoSphere: a fully integrated, physically based hydrological model.

Groundwater, *50*(2), 170–176.

Budyko, M. I., Miller, D. H., & Miller, D. H. (1974). *Climate and life* (Vol. 508). Academic press New York.

Burn, D. H., & Whitfield, P. H. (2018). Changes in flood events inferred from centennial length streamflow data records. *Advances in Water Resources*, *121*, 333–349.

Carlson, T. N., & Arthur, S. T. (2000). The impact of land use—land cover changes due to urbanization on surface microclimate and hydrology: a satellite perspective. *Global and Planetary Change*, *25*(1–2), 49–65.

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., & Siqueira, V. A. (2020). CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil. *Earth System Science Data*, *12*(3), 2075–2096.

Chen, Jie, Brissette, F. P., Chaumont, D., & Braun, M. (2013). Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America. *Water Resources Research*, *49*(7), 4187–4205. https://doi.org/10.1002/wrcr.20331

Chen, Junyi, Zheng, F., May, R., Guo, D., Gupta, H., & Maier, H. R. (2022). Improved data splitting methods for data-driven hydrological model development based on a large number of catchment samples. *Journal of Hydrology*, *613*(PA), 128340. https://doi.org/10.1016/j.jhydrol.2022.128340

Chen, Z., Zhu, R., Yin, Z., Feng, Q., Yang, L., Wang, L., et al. (2022). Hydrological response to future climate change in a mountainous watershed in the Northeast of Tibetan Plateau. *Journal of Hydrology: Regional Studies*, *44*, 101256.

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*, e2020WR029229.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V, et al. (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, *44*(12).

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, *51*(4), 2498–2514.

Clark, M. P., Wilby, R. L., Gutmann, E. D., Vano, J. A., Gangopadhyay, S., Wood, A. W., et al. (2016). Characterizing Uncertainty of the Hydrologic Impacts of Climate Change. *Current Climate Change Reports*, *2*(2), 55–64. https://doi.org/10.1007/s40641-016-0034-x

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

Condon, L. E., & Maxwell, R. M. (2015). Evaluating the relationship between topography and groundwater using outputs from a continental-scale integrated hydrology model. *Water Resources Research*, *51*(8), 6602–6621.

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). *Journal of Geophysical Research: Atmospheres*, *108*(D22).

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., et al. (2020). CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. *Earth System Science Data*, *12*(4), 2459–2483.

Craig, J. R. (2023). Raven user's and developer's manual (Version 3.7). 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.

Crawford, N. H., & Linsley, R. K. (1966). Digital Simulation in Hydrology'Stanford Watershed Model 4.

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

Dakhlaoui, H., Ruelland, D., Tramblay, 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

Dakhlaoui, Hamouda, Ruelland, D., & Tramblay, 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

Dembélé, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., & Schaefli, 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

Dooge, J. C. I. (1959). A general theory of the unit hydrograph. *Journal of Geophysical Research*, *64*(2), 241–256.

Downer, C. W., & Ogden, F. L. (2006). *Gridded Surface Subsurface Hydrologic Analysis (GSSHA) User's Manual; Version 1.43 for Watershed Modeling System 6.1*. ENGINEER RESEARCH AND DEVELOPMENT CENTER VICKSBURG MS COASTAL AND HYDRAULICS LAB.

Droogers, P., & Immerzeel, W. (2008). Managing the real water consumer: Evapotranspiration. *World Bank Documents*.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., et al. (2006). Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *Journal of Hydrology*, *320*(1–2), 3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031

Duan, Qingyun, 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.

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

Falkenmark, M., & Chapman, T. (1989). Comparative hydrology: An ecological approach to land and water resources. *(No Title)*.

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., 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., 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. 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.

Fowler, K. J. A., Acharya, S. C., Addor, N., Chou, C., & Peel, M. C. (2021). CAMELS-AUS: hydrometeorological time series and landscape attributes for 222 catchments in Australia. *Earth System Science Data*, *13*(8), 3847–3867.

Freeze, R. A., & Harlan, R. L. (1969). Blueprint for a physically-based, digitally-simulated hydrologic response model. *Journal of Hydrology*, *9*(3), 237–258.

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., Ricard, S., Lachance-Cloutier, S., Anctil, F., & Turcotte, R. (2015). Comparing global and local calibration schemes from a differential split-sample test perspective. *Canadian Journal of Earth Sciences*, *52*(11), 990–999. https://doi.org/10.1139/cjes-2015-0015

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

Gao, H., Tang, Q., Shi, X., Zhu, C., Bohn, T., & Su, F. (2009). Water Budget Record from Variable Infiltration Capacity (VIC) Model Algorithm Theoretical Basis Document. *Rapport - Version 1.2*, (Vic), 57.

Garrick, M., Cunnane, C., & Nash, J. E. (1978). A criterion of efficiency for rainfall-runoff models. *Journal of*

Hydrology, *36*(3–4), 375–381.

Gauch, M., Kratzert, F., Gilon, O., Gupta, H., Mai, J., Nearing, G., et al. (2022). In Defense of Metrics: Metrics Sufficiently Encode Typical Human Preferences Regarding Hydrological Model Performance. https://doi.org/https://doi.org/10.31223/X52938

Gharari, S., Hrachowitz, M., Fenicia, F., & Savenije, H. H. G. (2013). An approach to identify time consistent model parameters: Sub-period calibration. *Hydrology and Earth System Sciences*, *17*(1), 149–161. https://doi.org/10.5194/hess-17-149-2013

Grayson, R. B. (1996). Distributed parameter hydrologic modeling using vector elevation data: THALES and TAPES-C. *Computer Models of Watershed Hydrology*, 669–696.

Gudmundsson, L., Leonard, M., Do, H. X., Westra, S., & Seneviratne, S. I. (2019). Observed Trends in Global Indicators of Mean and Extreme Streamflow. *Geophysical Research Letters*, *46*(2), 756–766. https://doi.org/10.1029/2018GL079725

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.

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.

Han, M., Shen, H., Tolson, B. A., Craig, J. R., Mai, J., Lin, S. G. M., et al. (2023). BasinMaker 3.0: A GIS toolbox for distributed watershed delineation of complex lake-river routing networks. *Environmental Modelling and Software*, *164*(June 2022), 105688. https://doi.org/10.1016/j.envsoft.2023.105688

Hartmann, G., & Bárdossy, A. (2005). Investigation of the transferability of hydrological models and a method to improve model calibration. *Advances in Geosciences*, *5*, 83–87. https://doi.org/10.5194/adgeo-5-83-2005

Helsel, D. R., Hirsch, R. M., Ryberg, K. R., Archfield, S. A., & Gilroy, E. J. (2020). *Statistical methods in water resources*. *Techniques and Methods*. Reston, VA. https://doi.org/10.3133/tm4A3

Horton, R. E. (1933). The role of infiltration in the hydrologic cycle. *Eos, Transactions American Geophysical Union*, *14*(1), 446–460.

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.

Hunt, R. J., Feinstein, D. T., Pint, C. D., & Anderson, M. P. (2006). The importance of diverse data types to calibrate a watershed model of the Trout Lake Basin, Northern Wisconsin, USA. Journal of Hydrology, 321(1–4), 286–296. https://doi.org/10.1016/j.jhydrol.2005.08.005

Hunt, R. J., Walker, J. F., Selbig, W. R., Westenbroek, S. M., & Regan, R. S. (2013). Simulation of climate-change effects on streamflow, lake water budgets, and stream temperature using GSFLOW and SNTEMP, Trout Lake Watershed, Wisconsin.

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.

Kami, B., & Jakubczyk, M. (2018). A framework for sensitivity analysis of decision trees, 135–159. https://doi.org/10.1007/s10100-017-0479-6

Kennard, R. W., & Stone, L. A. (1969). Computer Aided Design of Experiments. *Technometrics*, *11*(1), 137–148. https://doi.org/10.1080/00401706.1969.10490666

Kim, U., & Kaluarachchi, J. J. (2009). Hydrologic model calibration using discontinuous data: An example from the upper Blue Nile River Basin of Ethiopia. *Hydrological Processes*, *23*(26), 3705–3717. https://doi.org/10.1002/hyp.7465

KlemeŠ, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24. https://doi.org/10.1080/02626668609491024

Klingler, C., Schulz, K., & Herrnegger, M. (2021). LamaH-CE: LArge-SaMple DAta for hydrology and environmental sciences for central Europe. *Earth System Science Data*, *13*(9), 4529–4565.

Knoben, W. J.M., Freer, J. E., Fowler, K. J. A., Peel, M. C., & Woods, R. A. (2019). Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT) v1.2: An open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. *Geoscientific Model Development*, *12*(6), 2463–2480. https://doi.org/10.5194/gmd-12-2463-2019

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

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, Wouter J.M., Woods, R. A., & Freer, J. E. (2018). A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data. *Water Resources Research*, *54*(7), 5088–5109. https://doi.org/10.1029/2018WR022913

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.

International Joint Conference of Artificial Intelligence, (June).

Kovács, G. (1984). Proposal to construct a coordinating matrix for comparative hydrology. *Hydrological Sciences Journal*, *29*(4), 435–443.

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

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., et al. (2023). Caravan - A global community dataset for large-sample hydrology. *Scientific Data*, *10*(1), 1–11. https://doi.org/10.1038/s41597-023-01975-w

Kuentz, A., Arheimer, B., Hundecha, Y., & Wagener, T. (2017). Understanding hydrologic variability across Europe through catchment classification. *Hydrology and Earth System Sciences*, *21*(6), 2863–2879. https://doi.org/10.5194/hess-21-2863-2017

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

Leavesley, G. H., & Stannard, L. G. (1995). The precipitation-runoff modeling system-PRMS. *Computer Models of Watershed Hydrology.*, 281–310.

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

Li, Chuan-zhe Zhe, Wang, H., Liu, J., Yan, D. H., Yu, F. L., & Zhang, L. (2010). Effect of calibration data series length on performance and optimal parameters of hydrological model. *Water Science and Engineering*, *3*(4), 378–393. https://doi.org/10.3882/j.issn.1674-2370.2010.04.002

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.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, *201*(1–4), 272–288.

Lund, J. R. (1991). Random variables versus uncertain values: stochastic modeling and design. *Journal of Water Resources Planning and Management*, *117*(2), 179–194.

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.

Mai, Juliane. (2023). Ten strategies towards successful calibration of environmental models. *Journal of Hydrology*, *620*(PA), 129414. https://doi.org/10.1016/j.jhydrol.2023.129414

Mai, Juliane, 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.

Mai, Juliane, Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., et al. (2022). The Great Lakes Runoff Intercomparison Project Phase 4: The Great Lakes (GRIP-GL). *Hydrology and Earth System Sciences*, *26*(13), 3537–3572. https://doi.org/10.5194/hess-26-3537-2022

Mai, Juliane, 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

Maier, H. R., Zheng, F., Gupta, H., Chen, J., Mai, J., Savic, D., et al. (2023). On how data are used in model development: The elephant in the room. *Environmental Modelling and Software*, *1*(3394), 7–20. https://doi.org/10.1016/j.envsoft.2023.105779

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60.

Markstrom, S. L., Niswonger, R. G., Regan, R. S., Prudic, D. E., & Barlow, P. M. (2008). GSFLOW-Coupled Ground-water and Surface-water FLOW model based on the integration of the Precipitation-Runoff Modeling System (PRMS) and the Modular Ground-Water Flow Model (MODFLOW-2005). *US Geological Survey Techniques and Methods*, *6*, 240.

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*(January), 124698. https://doi.org/10.1016/j.jhydrol.2020.124698

Matott, L. S. (2017). OSTRICH-An Optimization Software Toolkit for Research Involving Computational Heuristics Doacumentation and User's Guide. Buffalo, NY: State University of New York. Retrieved from www.eng.buffalo.edu/~lsmatott/Ostrich/OstrichMain.html

May, R. J., Maier, H. R., & Dandy, G. C. (2010). Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Networks*, *23*(2), 283–294. https://doi.org/10.1016/j.neunet.2009.11.009

Mei, Y., Mai, J., Do, H. X., Gronewold, A., Reeves, H., Eberts, S., et al. (2023). Can Hydrological Models Benefit From Using Global Soil Moisture, Evapotranspiration, and Runoff Products as Calibration Targets? Water Resources Research, 59(2). https://doi.org/10.1029/2022WR032064

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.

Melsen, Lieke A. (2022). It Takes a Village to Run a Model—The Social Practices of Hydrological Modeling. *Water Resources Research*, *58*(2). https://doi.org/10.1029/2021WR030600

Melsen, Lieke A., Addor, N., Mizukami, N., Newman, A. J., Torfs, P. J. J. F., Clark, M. P., et al. (2018). Mapping

(dis)agreement in hydrologic projections. *Hydrology and Earth System Sciences*, *22*(3), 1775–1791. https://doi.org/10.5194/hess-22-1775-2018

Melsen, Lieke 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

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

Mishra, S. K., & Singh, V. P. (2013). *Soil conservation service curve number (SCS-CN) methodology* (Vol. 42). Springer Science & Business Media.

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

Moriasi, Daniel 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.

Motavita, D. F., Chow, R., Guthke, A., & Nowak, W. (2019). The comprehensive differential split-sample test: A stress-test for hydrological model robustness under climate variability. *Journal of Hydrology*, *573*(March), 501–515. https://doi.org/10.1016/j.jhydrol.2019.03.054

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 Eamonn, & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290.

Nash, James Edward, & HRS. (1958). Determining run-off from rainfall. *Proceedings of the Institution of Civil Engineers*, *10*(2), 163–184.

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*. 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 hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223.

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

Nicolle, P., Andreássian, V., Royer-Gaspard, P., Perrin, C., Thirel, G., Coron, L., et al. (2021). Technical Note – RAT : a Robustness Assessment Test for calibrated and uncalibrated hydrological models Key Words Key

Points 1 Introduction. *Hydrology and Earth System Sciences*, *25*(March), 1–22. https://doi.org/10.5194/hess-25-5013-2021

Nohara, D., Kitoh, A., Hosaka, M., & Oki, T. (2006). Impact of climate change on river discharge projected by multimodel ensemble. *Journal of Hydrometeorology*, *7*(5), 1076–1089. https://doi.org/10.1175/JHM531.1

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

Perrin, C., Oudin, L., Andreassian, 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

Pomeroy, J. W., Gray, D. M., Brown, T., Hedstrom, N. R., Quinton, W. L., Granger, R. J., & Carey, S. K. (2007). The cold regions hydrological model: a platform for basing process representation and model structure on physical evidence. *Hydrological Processes: An International Journal*, *21*(19), 2650–2667.

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

Quick, M. C., & Pipes, A. (1977). UBC WATERSHED MODEL/Le modèle du bassin versant UCB. *Hydrological Sciences Journal*, *22*(1), 153–161.

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

Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. Retrieved from http://arxiv.org/abs/1811.12808

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

Razavi, S., & Tolson, B. A. (2013). An efficient framework for hydrologic model calibration on long data periods. *Water Resources Research*, *49*(12), 8418–8431. https://doi.org/10.1002/2012WR013442

Refsgaard, J. C. (1997). Parameterisation, calibration and validation of distributed hydrological models. Journal

of Hydrology, 198(1–4), 69–97.

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

Reitermanov, Z. (2010). Data Splitting, 31–36.

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

Santos, L., Thirel, G., & Perrin, C. (2018). Technical note: Pitfalls in using log-transformed flows within the KGE criterion. *Hydrology and Earth System Sciences*, *22*(8), 4583–4591. https://doi.org/10.5194/hess-22-4583-2018

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.

Schoups, G., Van De Giesen, N. C., & Savenije, H. H. G. (2008). Model complexity control for hydrologic prediction. *Water Resources Research*, *44*(1), 1–14. https://doi.org/10.1029/2008WR006836

SCS (1956). Hydrology, National Engineering Handbook, Supplement A, Section 4, Chapter 10, Soil Conservation Service, USDA, Washington.

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.

Seiller, G., Anctil, F., & Perrin, C. (2012). Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. *Hydrology and Earth System Sciences*, *16*(4), 1171–1189. https://doi.org/10.5194/hess-16-1171-2012

Shafii, M., & Tolson, B. A. (2015). Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resources Research*, *51*(5), 3796–3814.

Sharma, G. (2017). Pros and cons of different sampling techniques. International journal of applied research. *International Journal of Applied Research*, *3*(7), 749–752. Retrieved from www.allresearchjournal.com

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, H., Tolson, B. A., & Mai, J. (2022a). Time to Update the Split-Sample Approach in Hydrological Model Calibration. *Water Resources Research*, *58*(3), 1–26. https://doi.org/10.1029/2021WR031523

Shen, H., Tolson, B. A., & Mai, J. (2022b). Time to Update the Split-Sample Approach in Hydrological Model Calibration v1.1. https://doi.org/10.5281/ZENODO.6578924

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.

Sherman, L. K. (1932). Streamflow from rainfall by the unit-graph method. *Eng. News Record*, *108*, 501–505.

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. (2018). Hydrologic modeling: progress and future directions. *Geoscience Letters*, *5*(1). https://doi.org/10.1186/s40562-018-0113-z

Singh, V. P., & Chow, V. T. (2016). *Handbook of applied hydrology* (Second edi). 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(2002)7:4(270)

Sivapalan, M., & Blöschl, G. (2015). Time scale interactions and the coevolution of humans and water. *Water Resources Research*, *51*(9), 6988–7022.

Skøien, J. O., Blöschl, G., & Western, A. W. (2003). Characteristic space scales and timescales in hydrology. *Water Resources Research*, *39*(10). https://doi.org/10.1029/2002WR001736

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

Smith, M. B., Koren, V., Reed, S., Zhang, Z., Zhang, Y., Moreda, F., et al. (2012). The distributed model intercomparison project - Phase 2: Motivation and design of the Oklahoma experiments. *Journal of Hydrology*, *418–419*, 3–16. https://doi.org/10.1016/j.jhydrol.2011.08.055

Snee, R. D. (1977). Validation of regression models: methods and examples. *Technometrics*, *19*(4), 415–428.

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*. https://doi.org/10.1029/WR019i001p00251

Sorooshian, S., Duan, Q., & Gupta, V. K. (1993). Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting Model. *Water Resources Research*, *29*(4), 1185–1194. https://doi.org/10.1029/92WR02617

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).

https://doi.org/10.1029/2019WR027009

Stephenson, N. L. (1990). Climatic control of vegetation distribution: the role of the water balance. *The American Naturalist*, *135*(5), 649–670.

Sugawara, M. (1974). Tank model and its application to Bird Creek, Wollombi Brook, Bikin River, Kitsu River, Sanaga River and Nam Mune. *Research Notes of the National Research Center for Disaster Prevention*, *11*, 1–64.

Sungmin, O., Dutra, E., & Orth, R. (2020). Robustness of process-based versus data-driven modeling in changing climatic conditions. *Journal of Hydrometeorology*, *21*(9), 1929–1944. https://doi.org/10.1175/JHM-D-20-0072.1

Taheri, M., Ranjram, M., & Craig, J. R. (2023). An Upscaled Model of Fill-And-Spill Hydrological Response. *Water Resources Research*, *59*(5), 1–21. https://doi.org/10.1029/2022WR033494

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

Thompson, S. E., Harman, C. J., Konings, A. G., Sivapalan, M., Neal, A., & Troch, P. A. (2011). Comparative hydrology across AmeriFlux sites: The variable roles of climate, vegetation, and groundwater. *Water Resources Research*, *47*(7), 1–17. https://doi.org/10.1029/2010WR009797

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

Towler, E., Foks, S. S., Dugger, A. L., Dickinson, J. E., Essaid, H. I., Gochis, D., et al. (2023). Benchmarking high-resolution hydrologic model performance of long-term retrospective streamflow simulations in the contiguous United States. *Hydrology and Earth System Sciences*, *27*(9), 1809–1825. https://doi.org/10.5194/hess-27-1809-2023

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 1 - Comparison of six snow accounting routines on 380 catchments. *Journal of Hydrology*, *517*, 1166–1175. https://doi.org/10.1016/j.jhydrol.2014.04.059

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

Vereecken, H., Huisman, J.-A., Hendricks Franssen, H.-J., Brüggemann, N., Bogena, H. R., Kollet, S., et al. (2015). Soil hydrology: Recent methodological advances, challenges, and perspectives. *Water Resources Research*, *51*(4), 2616–2633.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0:

fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272.

Vrugt, J. A., & de Oliveira, D. Y. (2022). Confidence intervals of the Kling-Gupta efficiency. *Journal of Hydrology*, *612*(PA), 127968. https://doi.org/10.1016/j.jhydrol.2022.127968

Vrugt, J. A., & Sadegh, M. (2013). Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. *Water Resources Research*, *49*(7), 4335–4345. https://doi.org/10.1002/wrcr.20354

Wasko, C., Guo, D., Ho, M., Nathan, R., & Vogel, E. (2023). Diverging projections for flood and rainfall frequency curves. *Journal of Hydrology*, *620*(PA), 129403. https://doi.org/10.1016/j.jhydrol.2023.129403

Wheater, H. S., McIntyre, N., & Wagener, T. (2008). *Calibration, uncertainty and regional analysis of conceptual rainfall-runoff models*. Cambridge University Press, Cambridge, England.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics* (pp. 196–202). Springer.

Wu, W., May, R. J., Maier, H. R., & Dandy, G. C. (2013). A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks. *Water Resources Research*, *49*(11), 7598–7614. https://doi.org/10.1002/2012WR012713

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, W., Chen, H., Xu, C. Y., Huo, R., Chen, J., & Guo, S. (2020). Temporal and spatial transferabilities of hydrological models under different climates and underlying surface conditions. *Journal of Hydrology*, *591*(February), 125276. https://doi.org/10.1016/j.jhydrol.2020.125276

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.

Zhang, G. P., & Berardi, V. L. (2001). Time series forecasting with neural network ensembles: An application for exchange rate prediction. *Journal of the Operational Research Society*, *52*(6), 652–664. https://doi.org/10.1057/palgrave.jors.2601133

Zhao, R. (1992). The Xinanjiang model applied in China. *Journal of Hydrology*, *135*(1–4), 371–381.

Zheng, F., Chen, J., Maier, H. R., & Gupta, H. (2022). Achieving Robust and Transferable Performance for Conservation-Based Models of Dynamical Physical Systems. *Water Resources Research*, *58*(5), 1–18. https://doi.org/10.1029/2021WR031818

Zheng, F., Chen, J., Ma, Y., Chen, Q., Maier, H. R., & Gupta, H. (2023). A Robust Strategy to Account for Data Sampling Variability in the Development of Hydrological Models Water Resources Research. https://doi.org/10.1029/2022WR033703

Appendices

A-1 Spatial location of the 463 CAMELS catchments used in this study

Figure A1-1 displays the spatial locations of the 671 CAMELS catchments, where the 463 catchments selected in this study and other filtered catchments are distinguished by colors.

We filtered gauges in the CAMELS dataset by three strict criteria: (1) Catchment area discrepancies (calculated from the CAMELS derived drainage areas and the USGS reported ones) are smaller than 10%; (2) Water balance errors derived from Budyko curve (Budyko et al., 1974) are reasonably limited as reported in Knoben et al. (2020); and (3) Consecutive missing data periods in a streamflow record must be less than four months in every year from 1980–2014 *and* all missing data for the 1980–2014 is less than six months in total.

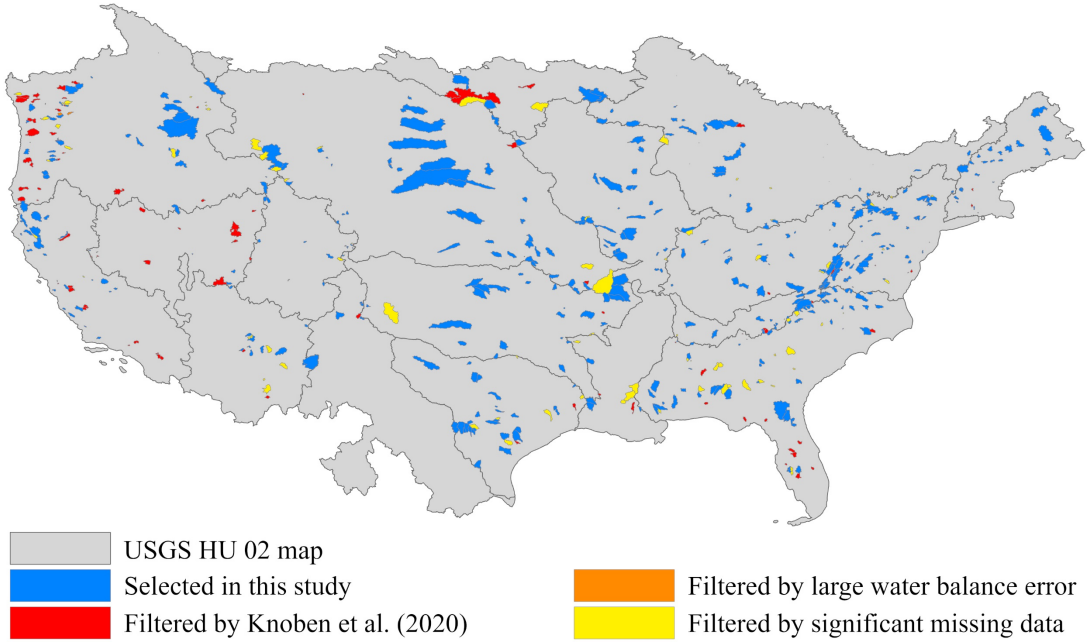

Figure A1-1. Spatial locations of catchments provided in the CAMELS dataset (671 in total). 463 catchments denoted as blue are selected in this study after removing 2 catchments with large water balance error (orange), 112 catchments filtered out by Knoben et al. (2020) (red), and 94 catchments with significant missing data in between the daily streamflow series. Note that the base layer depicts the hydrological unit (HU) 02 level map in the Watershed Boundary Dataset (WBD) provided by the USGS (available at https://www.usgs.gov/national-hydrography/watershed-boundary-dataset)

A-2 Repositories of model, data and results

The Raven formatted Daymet forcing and the USGS gauge streamflow data for the 463 CAMELS catchments used in this thesis are available on Zenodo by Shen et al. (2022b) (https://doi.org/10.5281/zenodo.6578924).

The reference KGE and KGE derived in Chapter 3 for each catchment-model combination in calibration, validation, and testing periods used for modeling results analysis are also available on Zenodo by Shen et al. (2022b) (https://doi.org/10.5281/zenodo.6578924).

The Raven Hydrologic Modeling Framework v3.0.4 and v3.6 used in this thesis and more recent versions, as well as the Raven templates for emulating GR4J and HMETS models are available at http://raven.uwaterloo.ca/Downloads.html.

The DDS algorithm and Ostrich software v17.12.19 used in this thesis are available at http://www.civil.uwaterloo.ca/envmodeling/Ostrich.html.

A-3 Details of the GR4J model parameters

The GR4J model emulated by Raven in this thesis contains six parameters in model calibration. Table A3-1 lists all the six parameters with their definitions and ranges in calibration. The parameters x_1 to x_4 are originally included in the GR4J model structure in Perrin et al. (2003), while parameters x_5 and x_6 are for the CemaNeige degree-day snow model (Valéry, 2010).

Table A3-1. The GR4J model parameters for calibration.

Parameter	Brief description	Unit	Lower bound	Upper bound
x_1	Maximum capacity of the production store	mm	1	2500
x_2	Groundwater exchange coefficient	mm	-15	10
x_3	One day ahead maximum capacity of the routing store	mm	10	700
x_4	Time base of unit hydrograph	days	0	7
x_5	Annual average snow depth	mm	1	30
x_6	Cold content factor	mm/d	0	1

A-4 Details of the HMETS model parameters

The HMETS model contains 21 parameters (Martel et al., 2017), and this model is also emulated by Raven in this study. Table A4-1 lists all the 21 parameters with their definitions and ranges in calibration. These parameter ranges are adapted from Mai et al. (2020).

Table A4-1. The HMETS model parameters for calibration.

Parameter	Brief description	Unit	Lower bound	Upper bound
x_1	Shape parameter for the gamma distribution used on the surface unit hydrograph	-	0.3	20.0
x_2	Rate parameter for the gamma distribution used on the surface unit hydrograph	1/d	0.01	5.0
x_3	Shape parameter for the gamma distribution used on the delayed unit hydrograph	-	0.5	13.0
x_4	Rate parameter for the gamma distribution used on the delayed unit hydrograph	1/d	0.15	1.5
x_5	Minimum fraction for the snowpack water retention capacity	mm/d/C	0.0	20.0
x_6	Maximum fraction of the snowpack water retention capacity	mm/d/C	0.0	20.0
x_7	Base melting temperature	C	-2.0	3.0
x_8	Degree day increase rate with cumulative melt	1/mm	0.01	0.2
x_9	Minimum water saturation fraction of snow	-	0.0	0.1
x_{10}	Maximum water saturation fraction of snow	-	0.01	0.3
x_{11}	Parameter for the calculation of water retention capacity	1/mm	0.005	0.1
x_{12}	Degree day reference (freezing) temperature	C	-5.0	2.0
x_{13}	Maximum refreeze factor used in degree day models	mm/d/C	0.0	5.0
x_{14}	Empirical exponent for the freezing equation	-	0.0	1.0
x_{15}	Fraction of the potential evapotranspiration	-	0.0	3.0
x_{16}	Fraction of the water for surface and delayed runoff	-	0.0	1.0
x_{17}	Fraction of the water for groundwater recharge	-	0.00001	0.02
x_{18}	Fraction of the water for hypodermic flow	-	0.0	0.1
x_{19}	Fraction of the water for groundwater flow	-	0.00001	0.01
x_{20}	Maximum level of the vadose zone	m	0.0	0.5
x_{21}	Maximum level of the phreatic zone	m	0.0	2.0

A-5 Net percentage of catchments that a split X significantly outperforms a split Y based on the Wilcoxon rank-sum test

Figure A5-1 displays the net percentage of catchments that split X significantly outperforms split Y during the first 5 years of testing period. The description of this assessment was presented in Section 4.2.4.2 and the example results were reported in Section 4.3.2.

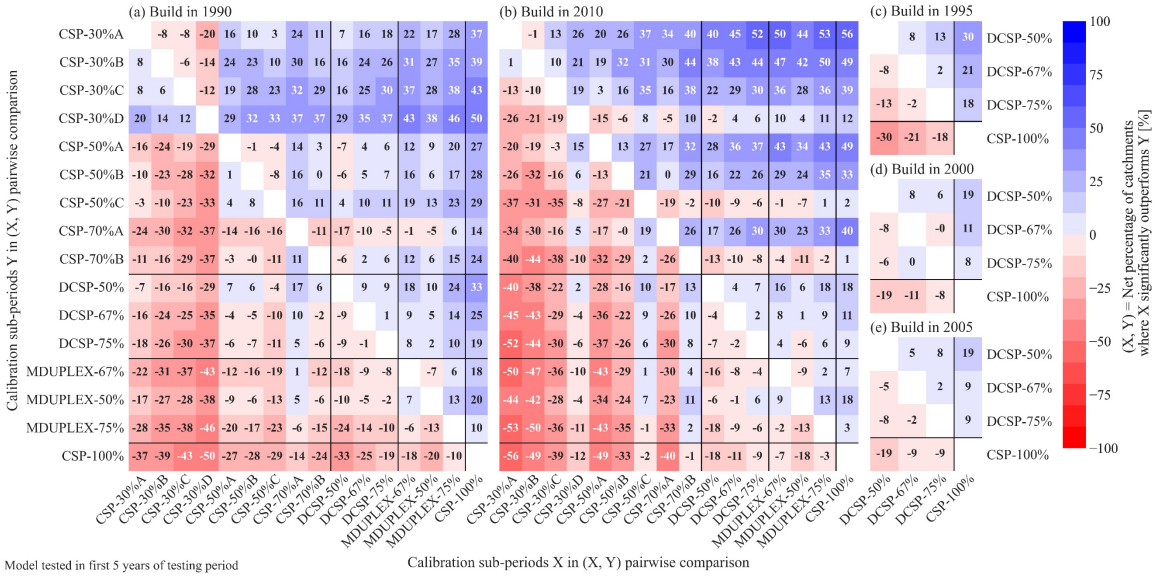

Figure A5-1. Overview of pairwise comparisons of two splits (X, Y) based on the Wilcoxon rank-sum test during the first five years of the model testing period. Results are reported as net percentage of catchments that X significantly outperforms Y in the total count of catchment 463. Note that (a) was reported as an example in Figure 4-3.

A-6 Tradeoff between median KGE and accuracy for all catchments

Tradeoff analysis on median KGE versus classification accuracy using all results of the 463 catchments are displayed in Figure A6-1 and it is seen that the full-period CSPs are identified as non-dominated solutions in 7 out of 8 instances.

Non-dominated solutions in all 23 tradeoff analyses (5 instances in each model build years of 1990, 1995, 2000 and 2005, and 3 instances in 2010 where repeated testing periods are counted once) are summarized in Figure A6-2. In addition, Figure A6-2 displays another informative metric in this multi-objective assessment, i.e., the frequency a calibration split dominating other splits, which denotes a split being better than another split with respect to both axes. In total, each calibration split has 75 pairwise comparisons in 1990 (15 pairwise comparisons × 5 testing periods), 12 pairwise comparisons in 1995, 2000 and 2005 (3 pairwise comparisons × 5 testing periods), and 45 pairwise comparisons in 2010 (15 pairwise comparisons × 3 testing periods excluding repeated testing periods).

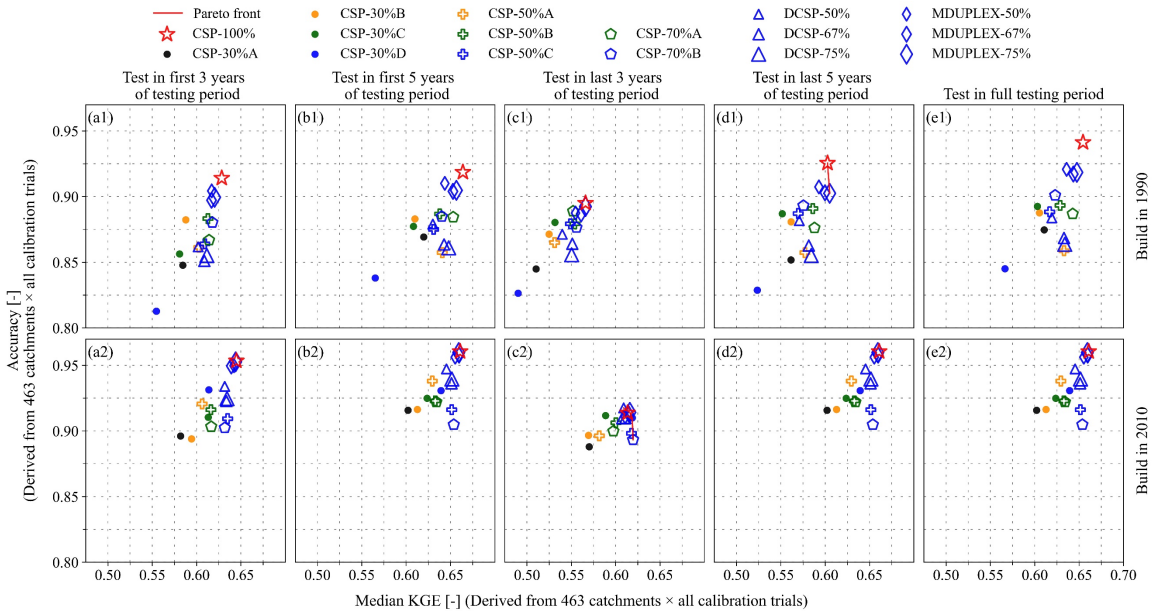

Figure A6-1. The Pareto solutions in the two-dimensional space with respect to median KGE and classification accuracy (both derived from all calibration trials (20 or 21 trials) × 463 catchments) of the 16 calibration splits in model build year (a1, b1, c1, d1, e1) 1990 and (a2, b2, c2, d2, e2) 2010 during all five different model testing periods. Note that the (b2) second, (d2) fourth, and (e2) fifth testing periods are the same when model is built in 2010. Each Pareto solution in these panels is derived from 9,260 trials (9,723 trials for DCSP-67%) of each split. Model failures are handled in this analysis when calculating median KGE that failed models use reference flow for testing period prediction

instead. There are 16 markers on each panel and markers are distinguishable with respect to their types, colors and sizes shown in the legend. Note that blue markers are the recent splits containing most-recent data, and black markers are the oldest splits. The solutions lying in the upper-right of each panel with larger values in both axes are dominating solutions in the lower-left of the corresponding panel. The red solid line is the Pareto front indicating the set of all non-dominated solutions. There is no Pareto front drawn in panels except (d1) and (c2) due to the sole non-dominated solution in each of the plots.

Figure A6-2 shows full-period CSP in all five build years consistently has the largest frequency being non-dominated solution (the main *y*-axis and bars in Figure A6-2), and the frequency is 1.0 in 4 out of 5 build years. MDUPLEX splits rank at the top quartiles with respect to the frequency of being non-dominated solutions but are no better than CSP-100% in both 1990 and 2010. DCSP splits are less competitive than CSP-100% and MDUPLEX splits in 2010 and can even be worse than some continuous CSPs in 1990.

Moreover, Figure A6-2 displays full-period CSP in all five build years is most frequently dominating other splits (the secondary *y*-axis and markers in Figure A6-2) with the frequency value ranging from 0.87 (build in 2010 in Figure A6-2e) to 1.0 (build in 1995 and 2000 in Figure A6-2b and Figure A6-2c, respectively).

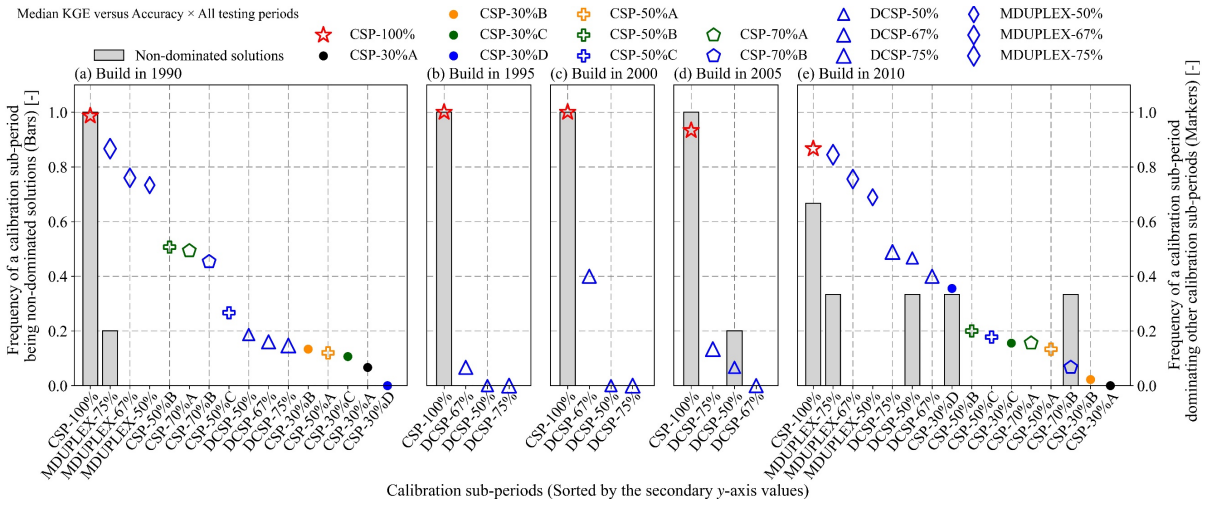

Figure A6-2. Overview of the tradeoffs between median KGE and classification accuracy considering all optimization trials (20 or 21 trials) × 463 catchments. Examples of the Pareto solutions in model build year 1990 and 2010 are displayed in Figure 6. Model failures are handled in this analysis when calculating median KGE that failed models use reference flow for testing period prediction instead. The five panels are for results when model is built in different years. Bars (the main *y*-axis) indicate the frequency of each split being non-dominated solutions from 0 to 1.0. The total tradeoff analyses

are 5, 5, 5, 5, and 3 in 1990, 1995, 2000, 2005, and 2010, respectively. Markers (the secondary *y*-axis) indicate the relative frequency of each split dominating other splits. The total count of pairwise comparisons is 75, 15, 15, and 45 in 1990, 1995, 2000, 2005, and 2010, respectively. The *x*-axis is sorted in descending order by the values with respect to the secondary *y*-axis. Markers in each of the panels are also distinguishable with respect to their types, colors and sizes shown in the legend.