Author:
Kiraz, Melike

Title:
Addressing the challenges of catchment characterisation, model selection and evaluation in large-sample hydrology
application to Great Britain

Addressing the Challenges of Catchment Characterisation, Model Selection and Evaluation in Large-Sample Hydrology: Application to Great Britain

By

Melike Kiraz

Department of Civil Engineering

UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in the Faculty of Engineering.

JUNE 2023

Word count: 25099

Abstract

Rainfall-runoff models play a vital role in understanding hydrologic processes, estimating streamflow, and predicting flood and drought risks across various scales. However, hydrological modelling still faces significant uncertainties and challenges. Difficulties arise in identifying and characterizing hydrological processes (e.g. subsurface losses), selecting and evaluating model structures, and dealing with uncertainties in observational data and model structures. These problems become even more complex in large-sample hydrology due to the heterogeneity of catchments, the abundance of catchment types, and the variability in data quality and human influence.

In this thesis, we address three challenges in rainfall-runoff modelling across a large sample of Great Britain catchments. Firstly, we assess the role of catchment location in understanding water balance issues in highly permeable catchments when available catchment descriptors are insufficient. We find that catchment location relative to the coast and within a wider river basin shed light on water balance issues in highly permeable catchments. Secondly, we explore the importance of prior model selection through a comparison of two modular modelling frameworks. By selecting model structures consistent with expected hydrologic variability, we demonstrate the possibility of observing meaningful performance differences between model structures in specific catchments. Lastly, we develop a signature-based hydrologic efficiency metric that proves comparable to traditional statistical evaluation metrics. This metric shows promise for model evaluation in ungauged catchments if its signatures can be well-regionalized.

All three contributions pave the way for follow-up research on new location-based catchment descriptors, hydrologically tailored efficiency metrics in gauged and ungauged basins, and identifying appropriate components for modular modelling system. We end by defining two specific ideas for future research. Firstly, quantifying hydrologic ecosystem services through new signatures to assess benefits and understand spatial-temporal variations. Secondly, coupling national-scale groundwater modelling across Great Britain with catchment-scale modelling to estimate inter-catchment groundwater flow between neighbouring catchments.

Dedication and Acknowledgements

I dedicate this work to my loving family, whose unwavering support and encouragement have been the cornerstone of my journey. Their belief in my abilities has been my constant source of inspiration, and I am forever grateful for their love, understanding, and sacrifices.

I would like to express my deepest gratitude to my advisors, Thorsten Wagener, Gemma Coxon and Shams Rahman, and to my annual progress reviewer, Francesca Pianosi, for their invaluable guidance, expertise, and unwavering support throughout the entire research process. Their insightful feedback, encouragement, and commitment to excellence have played a crucial role in shaping this thesis.

I am also grateful to the members of my thesis committee, Miguel Rico-Ramirez and Michaela Bray, for their valuable input, constructive criticism, and valuable suggestions, which have significantly enhanced the quality of this research.

I am thankful to my special friends Elisa, Giulia, Lina, Maria, Valentina and other colleagues in Water Group and other close friends in Bristol for their constant encouragement, insightful discussions, and moral support. Their friendship and camaraderie have made this academic journey enjoyable and memorable.

I would like to acknowledge that this work was funded by Ministry of National Education, the Republic of Türkiye.

Lastly, this work is the result of collective efforts, guidance, and support from numerous individuals and institutions. While the responsibility for any errors or shortcomings rests solely with me, their contributions have undeniably shaped and enriched this thesis.

Author’s Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: .....

Table of Contents

List of Figures

List of Tables

List of Acronyms

2PAR: Two Conceptual Reservoirs in Parallel

AET: Actual Evapotranspiration

AI: Aridity Index

BFI: Baseflow Index

BFI-HOST: Baseflow index derived from the 29-class HOST classification

CAMELS: Catchment Attributes and MEteorology for Large-sample Studies

CDF: Cumulative distribution function

CRES: Conceptual Reservoir

dE: Delta Elevation

dGW: Delta Groundwater

dRR: Delta Runoff Ratio

ER: Effective Rainfall

FDC: Flow Duration Curve

FUSE: Framework for Understanding Structural Errors

GB: Great Britain

gw: Groundwater

HOST: Hydrology Of Soil Types

LEAK: Leaky Aquifer Model Structure

NP: Non-parametric version of Kling Gupta Efficiency

P: Precipitation

PAF: Permeable Area Fraction

PDM: Probability distributed model

PEN: Model based on Penman Drying Curve

PET: Potential Evapotranspiration

PC: Pearson Correlation

PRMS: Precipitation-Runoff Modelling System

Q: Streamflow

RR: Runoff Ratio

RRMT: Rainfall-Runoff Modelling Toolbox

SHE: Signature-based Hydrologic Efficiecny

SI: Strahler Index

SRC: Spearman Rank Correlation

SSI: Strahler Sequence Index

UK: United Kingdom

VIC: Variable Infiltration Capacity

VR: Variance ratio

Chapter 1 Introduction

1.1 The importance and challenges of rainfall-runoff modelling

Hydrology analyses the storage and movement of water in complex environmental systems ranging from centimetres to the whole Earth system (Thompson, 2017). Over the past century, terrestrial water fluxes have been rapidly changing in many parts of the world. Climate change is changing the frequency, severity and duration of hydrological extremes such as floods and droughts (IPCC, 2001; Lehner et al., 2006; Modarres et al., 2016; Brunner et al., 2021; Dutta and Maity, 2021; Lane and Kay, 2021; Wang et al., 2022; Gebrechorkos et al., 2023). Moreover, human activities such as land use/land cover changes, deforestation, urbanisation, dams and water abstractions are having significant impact streamflows in most parts of the world (Li et al., 2007; Vogel, 2011; Dey and Mishra, 2017; Singh and Basu, 2022; Van Loon et al., 2022; Malede et al., 2023). In order to develop sustainable water resources management strategies and to provide long-term water security for people and the environment, hydrologists are interested in understanding and modelling the hydrologic cycle across spatial and temporal scales (Wagener et al., 2010; Peel and McMahon, 2020, Yang et al., 2021). Rainfall runoff models, i.e. tools representing how a catchment responds to rainfall under different conditions (Dawdy and O'Donnell, 1965; Beven, 2001), have become increasingly important for a wide range of issues, including for understanding hydrologic processes (Arnold et al., 1998; Watson et al., 2019), for estimating streamflow and other hydrologic variables (Perrin et al., 2003; Young, 2006), and for predicting flood and drought risks (Guo et al., 2020).

The rainfall-runoff modelling process consists of several key steps any modeller has to address (e.g. Beven, 2001). These are; 1) specifying the hydrologic processes of the underlying catchment (i.e. the perceptual model), 2) turning these processes into equations (i.e. the conceptual model), 3) developing or selecting one or more suitable model structures that integrate these equations (i.e. the procedural model) , 4) defining an

objective function to assess how well combinations of a specific model structure and a set of parameters perform, 5) identifying suitable parameter sets to represent the catchment under study (i.e. model calibration) and 6) ensuring accuracy and applicability of the model (i.e. model validation).

All of these steps in the modelling process still contain unsolved problems and uncertainties originating from imperfect observational data, parameter or model structures (Blöschl et al., 2019; Pan et al. 2019; Knoben et al., 2020; Klotz et al., 2022); model calibration and evaluation problems under changing conditions (Bathurst et al., 2004; Duan et al., 2006; Vaze et al., 2010; Thirel et al., 2015; Saft et al., 2016; Fowler et al., 2018; Yang et al., 2022), equifinality (Beven, 2006; Ebel and Loague, 2006; Lee et al., 2012; Kelleher et al., 2017; Khatami et al., 2019; Wu et al., 2022) of rainfall-runoff models in gauged and ungauged locations (Beven, 2019; Blair et al., 2019). Moreover, catchment behaviour is poorly captured by existing physical and climatic descriptors in many catchments due to unaccounted for issues such as subsurface losses and anthropogenic activities (Le Moine et al., 2007, Schaller and Fan, 2009; Munoz et al., 2016; Kuentz et al., 2017; Bouaziz et al., 2018; Fan 2019; Liu et al., 2020; Luijendijk et al., 2020). It is also challenging to select an adequate model structure and parameters (Uhlenbrook et al., 1999; Beven, 2000; Bai et al., 2009; Coxon et al., 2014; Paul et al., 2021; David et al., 2022), including the problem of evaluating model performance in gauged and ungauged catchments (Gupta et al., 2009; Hrachowitz et al., 2013; Knoben et al. 2019, Clark et al., 2021).

1.2 The challenges of rainfall-runoff modelling in large-sample hydrology

Regardless of whether we study one or many catchments, these modelling problems persist. Nevertheless, large-sample hydrology entails further challenges such as dealing with many types of catchments or the need to select multiple model structures. Large-sample hydrology focuses on the evaluation of hydrologic systems using large number of

catchments (Addor et al., 2020). It originates from comparative hydrology which is about learning from the differences and similarities between places and about transferring hydrologic knowledge across regions (Falkenmark and Chapman, 1989). Gupta et al. (2014) highlight four fundamental benefits of using large-sample datasets: 1) *improved understanding*, providing wider range of applicability, higher extrapolation capabilities and better identification of limitations by rigorously testing and comparing rainfall runoff model structures and hypotheses, 2) *robustness of generalizations*, enabling to diminish the effects of severe data errors and assisting in identifying and addressing outliers via statistical analyses with large number of data, 3) *classification, regionalization and model transfer*, by providing a heterogeneous domain including diverse climatic, hydrologic and physical characteristics, 4) *estimation of uncertainty*, providing a better comprehension of uncertainty by enabling a statistical regionalization of uncertainty estimates. There is an increasing number of large-sample datasets available around the world such as the US MOPEX (Hogue et al., 2004) and CAMELS (Newman et al., 2015; Addor et al., 2017) for the USA, CAMELS-CL (Alvarez-Garreton et al., 2018) for Chile, CAMELS-GB (Coxon et al., 2020b) for UK, CAMELS-BR (Chagas et al., 2020) for Brazil, LamaH-CE (Klingler et al., 2021) for central Europe and CAMELS-AUS (Fowler et al., 2021) for Australia. Such datasets have been used for different purposes such as catchment classification (e.g. Sawicz et al., 2011; Jehn et al., 2020; Brunner et al., 2020), investigating extreme events (e.g. Berghuijs et al., 2017; Stein et al., 2021; Vega-Briones et al., 2023), quantifying uncertainties in hydrologic data and models (e.g. Coxon et al., 2015; Knoben et al., 2020; Yan et al., 2023), and for hydrologic model evaluation and benchmarking (Rakovec et al., 2019; Lane et al., 2019; Lees et al., 2021).

This thesis addresses three core challenges of rainfall runoff modelling in large-sample hydrology. These are challenges of accounting for subsurface groundwater losses, establishing distinct relationships between model structure and catchment types, and offering a hydrologically diagnostic model evaluation that can be utilized in both gauged

and ungauged catchments. They are selected to be addressed in this thesis due to their enduring statues as longstanding issues in rainfall-runoff modelling. Potential groundwater flow pathways (such as local and regional flows) have been subject to investigation for over five decades (e.g. Toth, 1963). However, comprehending these pathways and their impacts on catchment water balance remains uncertain, particularly in the context of large-sample hydrology. Similarly, modular modelling frameworks have been developed for more than two decades (e.g. Leavesley et al., 1996) to provide flexibility in model selection for diverse catchments but there is still lack of a robust strategy to select right model structures for certain catchment types. Lastly, performance evaluation metrics have been evolved in a long time to be more diagnostic (e.g. Nash and Sutcliffe, 1970), yet they are still lacking valuable hydrologic information and cannot be calculated in ungauged catchments. Tackling these challenges is a pivotal undertaking that will contribute to an enhanced understanding of hydrologic processes and improve the selection and evaluation of model structures. They are discussed in detail in the following sections.

1.2.1 Can we identify and characterise leaky catchments based on their location?

Accounting for subsurface groundwater losses is still an unsolved problem in runoff-rainfall modelling (Le Moine et al., 2007, Liu et al., 2021). Many modellers present topographic catchments as self-contained hydrologic systems by commonly assuming that the net groundwater outflow is negligible (Fan, 2019). However, the water balances of many catchments are not closed but influenced by subsurface losses (Figure 1-1) (Schaller and Fan, 2009; Genereux and Jordan, 2005). Investigating these water balance issues at the catchment scale is necessary to provide more robust understanding of subsurface losses and to guide model development and selection processes in rainfall-runoff modelling (Oldham et al., 2023). However, understanding these losses is not an easy task, especially when trying to do with currently available catchment descriptors (Bouaziz et al., 2018). Descriptors include catchment physical properties (e.g. soil type distribution, land cover, geology), climatic boundary conditions (e.g. aridity index), catchment topography (e.g.

average topographic slope, elevation) and water management (e.g. abstractions, reservoir capacity). Due to complex and hard to observe groundwater gains and losses, the exact reasons of water balance discrepancies between neighboring catchments are mostly unknown (Genereuz et al. ,2005; Munoz et al., 2016). Since locational aspects of catchments are not generally included (or limited) in large-sample hydrology datasets (Addor et al., 2018), their contribution to understand these subsurface losses are not investigated yet. Some recent papers did suggest that the location of catchments contains at least some information that could provide some insight into catchment losses. For example, Liu et al. (2020) investigated several thousand catchments around the globe. Their results suggest that factors such as location to coast or within a wider river basin could be informative in this regard. We will pursue this issue in chapter 2.

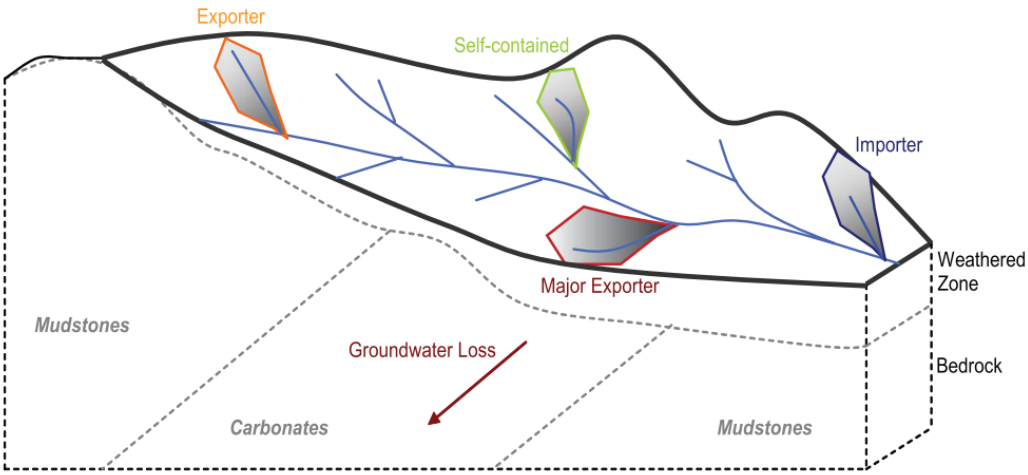
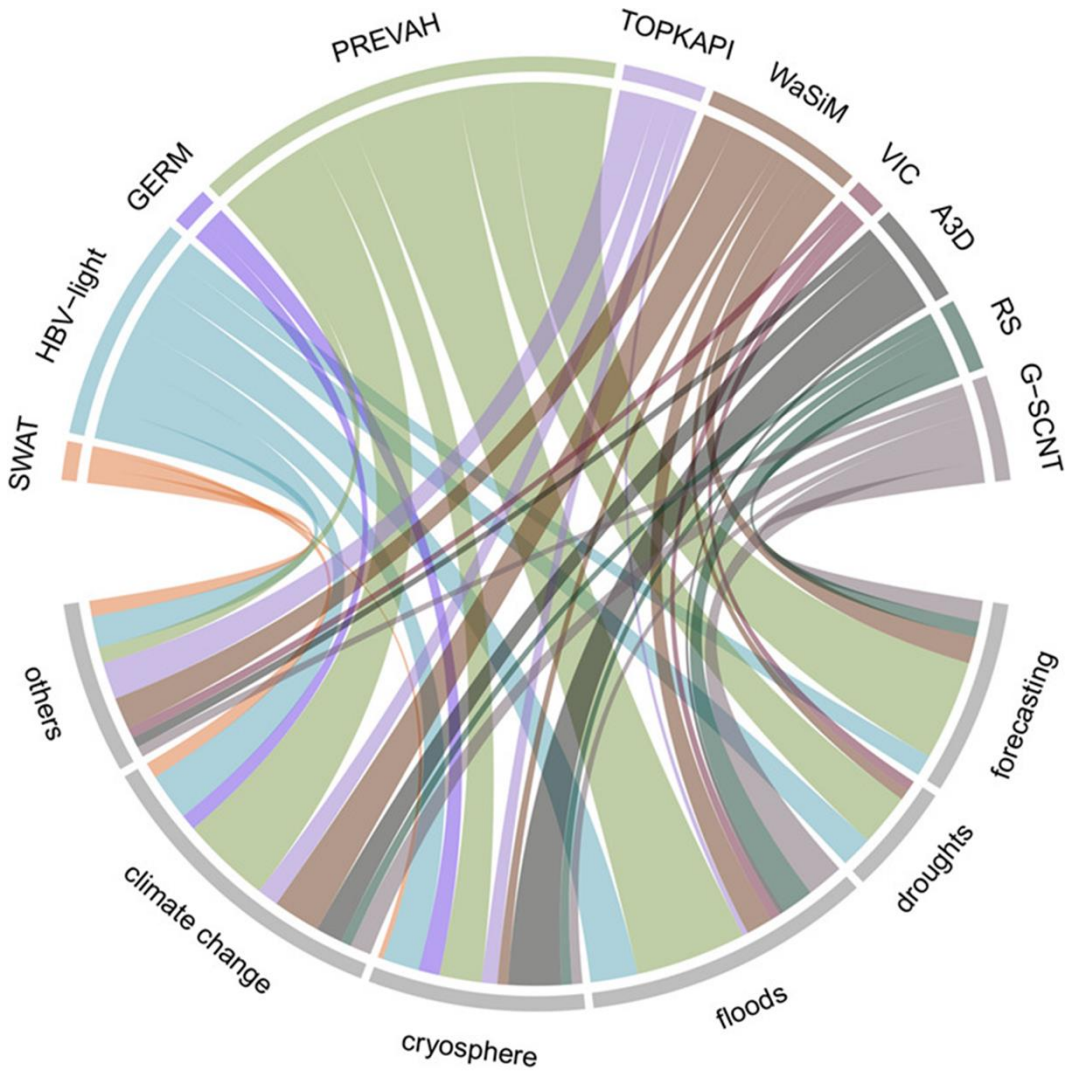


Figure 1-1. A hypothetical situation where the carbonate unit diverts water away from the river basin under study. In this case, the surface drainage does not coincide with the subsurface drainage, with regard to flow boundaries as well as flow directions. Lower reach basins can be exporters, complicating the "elevation dependence," and the large basin as a whole is not self-contained, complicating the "scale dependence.". Taken from (Schaller and Fan, 2009).

1.2.2 How do we select model structures to include in modular modelling frameworks applied in a large-sample hydrology?

Multi-model studies have become common in large sample hydrology to represent diverse catchment characteristics and hydrologic processes across heterogeneous domains (Rakovec et al., 2019). Model selection is an important step to represent the dominant

hydrologic processes of diverse catchments across a study domain while aligning with the purpose of the study. Selection criteria depend on the modelling purpose such as understanding specific hydrologic processes, determining the frequencies of runoff events, or predicting runoff yield for management purposes (Vaze et al., 2012), though one generally would like to represent the relevant processes of the catchment at hand. However, model selection has been challenging due to an abundance of models in the hydrological community (Clark et al., 2011). Having a diverse range of model applications, each with unique requirements and demands in hydrology and the lack of consensus on a standardized set of concepts for process representations further contribute to the creation of more models and/or modular modelling frameworks (Figure 1-2) (Horton et al., 2022; Weiler and Beven, 2015). Moreover, models should possess two characteristics, adequacy and parsimony, i.e. they should not be over-complicated to be interpretable and should be effective enough for the task at hand (Horton et al., 2022; Höge et al., 2022). At the same time, other reasons such practicality, convenience and experience with existing model set ups might also influence model selection (Addor and Melsen, 2019).



Model		Application
SWAT	→	climate change, cryosphere, others
HBV-light	→	climate change, cryosphere, forecasting, droughts, floods, others
GERM	→	climate change, cryosphere
PREVAH	→	climate change, cryosphere, droughts, forecasting, floods, others
TOPKAPI	→	climate change, cryosphere, floods, others
WaSiM	→	climate change, cryosphere, floods, droughts, forecasting, others
VIC	→	floods, droughts, others
A3D	→	climate change, cryosphere, floods, others
RS	→	climate change, cryosphere, forecasting, floods
G-SCNT	→	climate change, cryosphere, forecasting, floods, others

Figure 1-2. Visualization of how model diversity did arise based on the example of Switzerland. Hydrological models applied to different contexts in Switzerland. The importance of the link is proportional to the number of scientific articles. The importance of some models can be inflated by the fact that an article can address multiple contexts, such as floods and climate change. Models with too few use cases (less than three) are not included for the sake of clarity. Different colours represent different models. The models given in the visualisation are SWAT (Soil and Water Assessment Tool), HBV-light (Hydrologiska Byråns Vattenbalansavdelning—light), GERM (Glacier Evolution Runoff Model), PREVAH (Precipitation–Runoff–Evapotranspiration HRU Model), TOPKAPI (TOPographic Kinematic APproximation and Integration), WaSiM (Water Flow and Balance Simulation Model), VIC (Variable Infiltration Capacity model), A3D (Alpine3D), RS (Routing System) and G-SCNT (GSM-SOCONT, Glacier and SnowMelt SOil CONTribution model). The visualization is taken from Horton et al., 2022. Below the visualization, applications of each model are listed.

Modular modelling frameworks attempt to consider the possible need for representing different catchment with different model structures from the beginning. The challenge then often becomes even more complicated on how to select suitable models from a large number of possible options. Some studies struggle to clearly discern model performances of the included model structures in different catchment types (e.g., Lane et al., 2019; Knoben et al., 2020). This challenge arises from the selection of multiple model structures that share similar process representations or complexities, or that can recreate very similar catchment behaviours. To overcome these issues, modellers or modular modelling framework developers need a robust strategy for (priori) selection of model(s) to include in their multi-model studies or modular modelling frameworks. Little focus has so far been placed on the process to identify and include model structures a priori, i.e. before running all of them to distinguish them based on performance differences alone. We take a different look at this problem in chapter 3.

1.2.3 How can model performance be evaluated in a hydrologically meaningful way, with and without streamflow observations?

Application of rainfall-runoff models to obtain reliable simulations requires some degree of parameter estimation (i.e. calibration) (Mizukami et al., 2019). Statistical performance metrics are commonly used for parameter estimation to evaluate model performance based on the comparison of simulated and observed streamflow values (Triana et al., 2019). A number of statistical performance metrics have been developed and used over the years

such as the Root Mean Squared Error (RSME) (Gershenfeld, 1999), Nash-Sutcliffe efficiency metric (NSE) (Nash and Sutcliffe, 1970) and Kling-Gupta efficiency (Gupta et al., 2009). Evolution of these metrics over time results from the need of more diagnostic methods to be used to evaluate and correct (or improve) models (Gupta et al., 2008), which has led to the use of hydrological signatures in addition to statistical performance metrics.

Currently available performance metrics are still criticized for their lack of useful hydrological information regarding the behaviour of a model in the catchment of interest, i.e. they mainly provide some statistical summary metric with little hydrologic information (Schaefli and Gupta, 2007). The widespread utilization of hydrologic signatures does provide an interesting way forward though, given that they can be used to quantify hydrologically relevant information while also reproducing the information contained in some statistical metrics, such as the bias. Moreover, current metrics are only applicable to gauged catchments, i.e. they require historical time series of observed streamflow. In large-sample hydrology context, hydrologically meaningful diagnostic methods are necessary to evaluate the performances of models in terms of hydrological suitability across diverse catchments for both gauged and ungauged cases. We investigate this possibility in chapter 4.

1.3 Introduction to the main research questions

In this thesis, these three open challenges of rainfall-runoff modelling will be addressed across a large sample of catchments in Great Britain (GB). We choose GB as our key study area due to the availability of large-sample catchment hydrology datasets (Coxon et al., 2020b and due to presence of key knowledge gaps in our perceptual model of GB hydrology (Wagener et al., 2021). Water balance issues across GB catchments need to be addressed by accounting for groundwater exchange in local and regional scale. Investigating these water balance issues using different locational aspects can provide better insight into potential future hydrologic changes in specific regions such as coastal regions (Fan, 2019; Liu et al., 2020) where significant challenges (e.g. flooding, erosion)

already exist in GB (de la Vega-Leinert et al., 2008). Moreover, significant challenges remain in our ability to make meaningful connections between model structures and catchment characteristics, and thus to develop a coherent strategy for model selection to develop tailored multi-model ensembles in GB (e.g. Lee et al., 2005; Coxon et al., 2014; Lane et al., 2019). These gaps need to be addressed through improved understanding of catchment functions and their model-based representations in a nationally consistent framework. Hydrological signatures have an important role in understanding large-sample hydrology (Addor et al., 2020; McMillan, 2021). The catchment functions can be captured using hydrologic signatures describing hydrologic behavior and dominant processes of catchments (Sivapalan, 2005; Wagener et al., 2008; McMillan et al., 2017). Hydrologic signatures have been used in literature with different purposes (McMillan, 2021) such as understanding space-time variability of hydrologic processes (e.g. Troch, 2009; McMillan, 2020, McMillan et al., 2022b), catchment classification (e.g. Sawicz et al., 2011; Kuentz et al., 2017; Johnson et al., 2022), defining hydrologic similarity between catchments (e.g. Wagener, 2007; Toth, 2013; Neri et al., 2022), predictions in ungauged basins (Blöschl et al., 2013; Guo et al. 2021; Pool et al., 2021; Dal Molin et al., 2023) and to assess hydrologic model performance (Yilmaz et al. 2008, Euser et al. 2013; Shafii et al. 2017; Sahraei et al., 2020; Saavedra et al., 2022) with the aim of focusing model calibration on relevant hydrograph aspects or major catchment functions (Pool et al., 2018; Todorović et al., 2022). In large sample hydrology, hydrological signatures can help to define and quantify the observed variability in hydrologic behaviours and their control mechanisms. Therefore, it is expected that implementing a consistent approach based on suitable hydrologic signatures will make hydrological analyses, modelling and water-management applications more meaningful (Wagener et al., 2008).

The aim of this thesis is to address three challenges of rainfall-runoff modelling outlined across a large-sample of GB catchments using a consistent hydrologic signature-based

approach. We use the following research questions to guide our efforts for the main aim of my thesis:

1. Does catchment location relative to its surrounding area explain catchment subsurface losses?
2. Does a priori model selection in multi-model studies enhance our ability to meaningfully explain differences in model performance in different GB catchments?
3. Can we define a signature-based hydrologic efficiency metric, which is comparable to commonly used statistical evaluation metrics in model evaluation and can be estimated in ungauged catchments?

The research presented in this thesis addressing research questions 2 and 3 above is in review in peer reviewed journals:

1. Kiraz, M., Coxon, G. and Wagener, T. (2023). A priori selection of hydrological model structures in modular modelling frameworks: Application to Great Britain. Hydrological Sciences Journal. Just-Accepted,
2. Kiraz, M., Coxon, G. and Wagener, T. (2023). A signature-based hydrologic efficiency metric for model calibration and evaluation in gauged and ungauged catchments. Water Resources Research. Under review,

1.4 Thesis structure

This thesis consists of three research chapters, a conclusion chapter and four appendices. Our three research questions are addressed in each research chapter individually as shown below:

- In Chapter 2, we quantify the water balance issues of catchments with unaccounted losses or gains of water using a hydrologic signature and relate these issues to possible subsurface losses using several locational aspects of catchments,

- In Chapter 3, we test the importance of selecting model structures that are consistent with the hydrologic variability across the GB domain on understanding of what kind of model structures should be used or can work better for certain catchment types.
- In Chapter 4, we link hydrologic signatures to the components of statistical evaluation metrics to develop a hydrologic signature-based efficiency metric, to eventually calibrate and evaluate the rainfall-runoff models in hydrologically meaningful way for both gauged and ungauged catchments,

The Appendices A-C contain supplemental material for Chapters 2-4, respectively. In Appendix D, the Curriculum Vitae (CV) of the author, providing a comprehensive overview of their professional qualifications, experiences, and accomplishments, is given.

Chapter 2 Location, Location, Location – Considering Relative Catchment Location to Understand Surface Losses

This chapter has been prepared for submission to a relevant journal and has undergone slight modifications to align with the general layout of this thesis. The study was conceptualized by Melike Kiraz, Gemma Coxon, and Thorsten Wagener. Melike Kiraz conducted the data processing, with assistance from Gemma Coxon and Mostaquimur Rahman. The analyses and creation of figures were performed by Melike Kiraz, under the guidance of Gemma Coxon, Mostaquimur Rahman, and Thorsten Wagener. The manuscript was primarily written by Melike Kiraz, with input and comments from all co-authors.

Citation: Kiraz, M., Coxon, G., Rahman, M. and Wagener, T. (2023). Location, location, location – Considering relative catchment location to understand subsurface losses. (In preparation).

2.1 Introduction

A wide range of catchment descriptors have been developed and utilized to characterize hydrologically relevant catchment characteristics for large sample hydrology (e.g. CEH, 1999; Addor et al., 2017). Descriptors include catchment physical properties (e.g. soil type distribution, land cover), climatic boundary conditions (e.g. aridity index), catchment topography (e.g. average topographic slope) and water management (e.g. reservoir capacity) etc. These descriptors have been deployed for the purpose of regionalization of hydrologic signatures or of hydrologic model parameters (e.g. Merz et al., 2004; Young, 2006; Yadav et al., 2007; Westerberg et al., 2016; Prieto et al., 2019; Beck et al., 2020; Pool and Seibert, 2021), for catchment classification (e.g. Moliere et al., 2009; Sawicz et

al., 2011; Fang et al., 2017; Kuentz et al., 2017; Tumiran & Sivakumar, 2021), and for comparative hydrology studies (Wagener et al., 2007; Gupta et al., 2014; Addor et al., 2019; McMillan et al., 2022b).

However, various studies have pointed to the problem that available catchment descriptors – often those related to subsurface properties – are frequently insufficient to describe hydrological differences between catchments (e.g. Genereux et al., 2005; Almeida et al., 2016; Frisbee et al., 2016; Munoz et al., 2016; Addor et al., 2018). In particular, they are often insufficient to explain the widely discussed problem that the water balance of many catchments is not closed but influenced by sub-surface losses (e.g. Schaller & Fan, 2009; Munoz et al., 2016; Bouaziz et al., 2018; Liu et al., 2020; Luijendijk et al., 2020). The exact reasons for water balance gains/losses between similar neighboring catchments based on available rainfall, topography, land cover, soil and geology data are often unknown. This is due to complex and hard-to-observe groundwater interactions (Genereux et al. ,2005; Munoz et al., 2016). While some local or regional-scale studies have used more detailed descriptors such as major ion concentrations (Genereux and Jordan, 2006) and geomorphic metrics (Frisbee et al, 2016) to provide more insight into local and regional flow systems, their availability and accessibility in large-scale datasets is very limited (Addor et al., 2018).

Across large domains, Liu et al. (2020) and Schwamback et al. (2022) find that aridity is a key indicator with drier catchments tending to lose water when assessing 2760 catchments around world and for 733 Brazilian catchments, respectively. In addition to climate, both studies also find that catchment area and slope (or elevation) are some other factors to lose or gain water for catchments. The study by Liu et al. (2020) also shows that there is some information in the location of catchments, e.g. in relation to the coast. However, Liu et al. (2020) do not go into much depth regarding this issue (given the large geographical scale of their study), and few other large-sample studies have investigated the issue. Typically, locational aspects are not included in large sample catchment hydrology datasets, thus have rarely been assessed for their potential as informative catchment descriptors. In this study,

we propose that locational information is more informative than previously thought and we test this hypothesis by investigating water balance differences between catchments using different locational aspects in combination with basic geological and topographic information.

We conduct our study by comparing observed and expected (based on climate only) long-term water balances for 660 catchments across Great Britain. If we assume that the dominant control on water balance is climate, then catchment water balances should vary smoothly in space (similar to climate). If they do not, then something else must exert an additional control to cause this deviation. We study the differences between observed and expected catchment water balances to understand the role of catchment location on water balance issues, i.e. location to coast, location within river drainage basin and location to a relevant neighbor.

2.2 Data

We analyze 660 catchments spread across Great Britain. Great Britain – consisting of England, Wales and Scotland – is characterized by a temperate climate, moderate topographic variability and significant geological heterogeneity. Precipitation decreases from northern west to southern east, with mean annual precipitation values ranging from 3500 to about 550 mm/year (Coxon et al., 2020b). Conversely, potential evaporatranspiration (PET) increases from northern west (minimum of about 350 mm/year) to southern east (maximum of about 550 mm/day). The ratio of PET to P is generally below 1, which means that we are in an energy limited domain. Most of England is dominated by lowland terrain, whereas Wales and Scotland are dominated by more mountainous regions with the highest catchment mean elevations of 527m and 682m respectively (Coxon et al., 2020b). Great Britain has a diverse geology including aquifers consisting of Chalk, Magnesian, Jurassic, Devonian/Carbonifero limestone and Permo-Triassic sandstone. Chalk is the principal aquifer of Great Britain, and it accounts for more than 50% of the groundwater abstractions in the country due to its high productivity, while Permo-Triassic

sand stones provide approximately 25% of the groundwater abstractions in England and Wales (Allen et al., 1997). Streamflow in the south-east and the midlands of England are further influenced by human modifications such as abstractions, effluent discharges (i.e. effluent returns), urbanisation and/or reservoirs. Very little to no snow is observed in most (>90%) catchments, i.e. no more than 5% of all precipitation falls as snow resulting in snow fractions of no more than 0.05 (Coxon et al., 2020). Only twelve catchments in Scotland have higher snow fractions than 0.1 up to 0.17 (Coxon et al., 2020).

For each catchment, we calculate its long-term (decadal) water balance using daily rainfall, potential evapotranspiration and streamflow time series for a ten-year period (October 1, 1999 – September 30, 2009) compiled from the CAMELS-GB dataset (Coxon et al., 2020a; 2020b). CAMELS-GB is a large sample, open-source, hydro-meteorological dataset for Great Britain. It includes hydro-meteorological time series (consisting of rainfall, streamflow, potential evapotranspiration, temperature, radiation and humidity for 1970-2015 years), catchment attributes (including topography, climate, hydrology, land cover, soils, hydrogeology and human influences) (see Table A1 in APPENDIX A) and catchment boundaries for 671 catchments across Great Britain (Coxon et al., 2020b). Considering climatic variability (i.e. wet and dry periods), ten years of data is assumed to be sufficient to capture long-term climatic and hydrologic characteristics of our catchments for the purpose of this study. About 96% of the 671 catchments have >90% complete streamflow data in this 10-year period (i.e. 1999-2009). While CAMELS-GB dataset includes 671 catchments, we remove a small number of catchments (11 of 671) that have no available data for the time period of our study, where suspected flows were observed due to instrumentation problems or have unrepresentative runoff due to heavy urbanization. These catchments were removed based on the information acquired from the National River Flow Archive (NRFA) website (https://nrfa.ceh.ac.uk/data/search), which reported suspected flow issues due to instrumentation problems and unrepresentative runoff due to heavy urbanisation. It is worth noting that the NRFA website does not provide a timeline for when

these factors might have influenced catchment runoff. As a result, it is presumed that the affected catchments experienced the impacts of these factors during the 10-year study period.

We use a variety of data and catchment descriptors to test their ability to explain water balance differences between catchments. Firstly, in order to determine if water balance differences can be understood based on readily available catchment descriptors, we use the catchment attributes (e.g. topographic, climatic, hydrologic, land cover, soil etc.) supplied by CAMELS-GB. In addition to the hydrogeological descriptors available in CAMELS-GB, we derive permeability information from the 1:50,000-scale digital geological map of Great Britain prepared by the British Geological Survey (BGS). From this digital geological map, we use only basic bedrock permeability information which classifies subsurface properties as very low, low, moderate, high and very high. We assume that the areas with ‘very high’ class in bedrock permeability have highly permeable geology and the remaining areas do not. This ‘very high’ class covers the regions where Chalk, Jurassic, Magnesian aquifers and some parts of Devonian/Carboniferous aquifers are located. It is accepted as the permeability information for our study. To calculate catchment permeable area fractions, we utilize an intersection process within ArcMap, overlaying the 'very high' bedrock permeability class map with the map of studied catchments. The permeable fraction of a catchment is determined by the ratio of the intersected area, where the 'very high' class bedrock permeability map aligns with the catchment's topographic area, to the overall catchment area.

Secondly, we define locational aspects of the catchments including location to the coast and location within the river basin. Most CAMELS-GB catchments do not share a border with the coastline because river gauges will typically be located some distance from the coast. Therefore, we assume here that catchment boundaries that intersect a 10 km buffer to the coastline are catchments with a subsurface connection to the coast. A 10 km buffer is chosen as a suitable distance after manually testing multiple distances (5, 10, 15, 20 km

etc.). Certain coastal catchments were excluded when a shorter distance (e.g. 5 km) was applied, or inner catchments (i.e. non-coastal catchments) were inadvertently selected when a longer distance (e.g. 15, 20 km) was used. However, these issues did not arise when testing a 10 km buffer zone, which is why it was chosen as the appropriate distance. In our study, we defined coastal catchments as those with their gauging station situated within the designated buffer zone and with no intervening catchments located between them and the coast. Any catchments with others situated between them and the coast were not categorized as coastal catchments. Location within a river basin is described in a new simple index described in Section 2.3.2.

Finally, in order to investigate the relationship between mean topographic elevation and groundwater levels of catchments, we use daily groundwater level data of 878 wells in GB from National Groundwater Level Archive of British Geological Survey (BGS, 2022). 878 wells are selected to include in our study because they have daily groundwater level data for every season and multiple years (i.e. ranging from 2 to 10 years). The groundwater levels are quite variable between seasons and also some variance between years due to dry and wet years. Hence, we calculate the average groundwater levels of wells using daily groundwater level time series for a ten-year period (October 1, 1999 – September 30, 2009) and its map is given in Figure A9. The average groundwater levels of a catchment are quantified by taking the mean of average groundwater levels of wells located in the catchment.

2.3 Methods

2.3.1 Expected water balance based on climate alone

The relationship between runoff ratio (RR) and aridity index (AI) is widely used as a reference for the long-term catchment water balance (Budyko, 1961). RR is the ratio of long-term average streamflow (Q) to long-term average precipitation (P), indicating how much precipitation is released from the catchment as streamflow, rather than as evapotranspiration (assuming no change in storage and no regional groundwater flux). AI

is the ratio of long-term average evapotranspiration (PET) to long-term average precipitation (P) and it represents the relative availability of moisture and energy in a catchment. If a catchment has an AI value less than 1, then the available energy is limiting the amount of actual evapotranspiration, and if AI is larger than 1, available water is limiting the amount of actual evapotranspiration.

We calculate the expected water balance (expected RR) for all catchments, under the assumption that the water balance is only controlled by climate, using the Turc-Mezentsev curve, which is based on the widely studied Budyko framework (Budyko, 1961). It provides reference conditions for energy and water limits on the catchment water balance. Catchments with water balances unimpacted by other natural or human controls beyond climate are expected to plot close to the Budyko curve located in between water (AET = P) and energy limits (AET = PET). Similarly, Turc and Mezentsev link the long-term average evaporation to long term average precipitation (Turc, 1955; Mezentsev, 1955). The formula developed by Turc is;

$$\frac{\mathrm{AET}}{\mathrm{P}} = \frac{1}{[1+\left(\frac{\mathrm{P}}{PET}\right)^2]^{1/2}} \qquad (2\text{-}1)$$

Given that actual evapotranspiration is not measured at the catchment scale, we adjust the formula using 1- (Q/P) term instead of AET/P. Our formula for the Turc-Mezentsev Curve is therefore;

$$1 - \frac{Q}{P} = \frac{1}{[1+\left(\frac{P}{PET}\right)^2]^{1/2}} \quad \text{where } RR = \frac{Q}{P} \qquad (2\text{-}2)$$

The reason of using the Turc-Mezentsev Curve rather than the Budyko Curve is that it provides a more straightforward and simple formula to estimate RR based solely on P and PET. It is worth noting that the Turc-Mezentsev Curve assumes a linear relationship between streamflow and potential evapotranspiration. This means that it assumes a constant proportion of potential evapotranspiration contributes to streamflow across different catchments. Since it is primarily based on climate only, it disregards potential spatial

variability in topography, geology, land use, soil properties and other factors that can influence streamflow. Using the Turc-Mezentsev curve as a benchmark for estimating RR is easier and quicker due to its simple formula and minimal data requirement (i.e. only P and PET data). Its reliability for estimating RR varies depending the specific analysis objectives. It can provide useful estimates for catchments with relative uniform attributes and under certain simplifying assumptions. However, its accuracy might decrease when applied to catchments with diverse characteristics or complex hydrological behaviours. The choice between using this curve or more complex methods depends on the trade-off between the simplicity and accuracy. In this study, we use it as initial estimation tool, providing a preliminary approximation of RR values across GB catchments without the need for extensive hydrological modelling.

We assess the sensitivity of our results to this choice by calculating delta RR (dRR) as the difference between the observed RR of catchments and the estimated RR values calculated using the Turc-Mezentsev Curve (Figure 2-2). We also test a simple linear regression fit to our RR vs AI data as a baseline, which is shown and analysed in APPENDIX A (Figure A1).

2.3.2 A simple index of catchment location within the river basin

One of the locational aspects that we consider in our study is the location within the wider river basin in order to investigate water balance issues of catchments. While the Strahler Index indicates order of a stream in a river network (Horton, 1945; Strahler, 1952 and Strahler, 1957), it does not define the catchment location within the wider river basin. A first order stream can for example either occur in the headwaters of a basin or towards the outlet.

In our study, we introduce a new index to define catchment location within each river basin. The Strahler Sequence Index (SSI) is calculated as the difference between the Strahler index (SI) at the outlet of the catchment under study and the subsequent receiving river (i.e.

SSI = $SI_{receiving} - SI_{source}$). If a first or second-order catchment (i.e. having SI value 1 or 2) has a low SSI value (i.e. 1 or 2), it means that it is in the upper parts of the river basin. If it has a high SSI value (i.e. 3 or 4), it is in the lower parts of the river basin (Figure 2-1). It is important to highlight that SSI index values do not directly increase from upper parts to lower parts in a river network for all streams. However, it is a good proxy to define the location of lower order catchments within the wider river basin.

In order to calculate SSI values based on our formulation, we first calculate SI values of the river network. We use the UK NEXTMap 50 m gridded digital elevation model (Intermap Technologies, 2009) to derive GB river network and quantify SI values of streams. The percentage of total stream length is highest for Strahler index 1 (58%) and gradually decreases as the Strahler index increases (maximum of Strahler index of 8). After calculating SI values of 660 CAMELS-GB catchments, we calculate their SSI values based on our formulation.

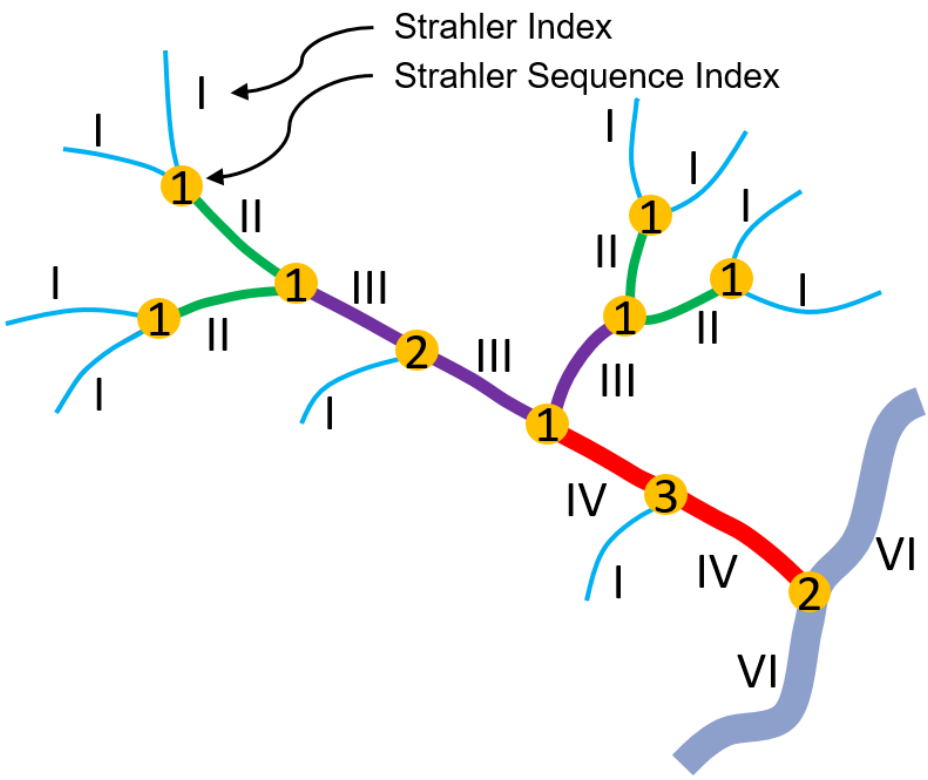


Figure 2-1. Visualization of Strahler Index (SI) and Strahler Sequence Index (SSI). SI is the order of streams in a river network. SSI is the order difference between a stream and the next stream which it drains into.

2.4 Results

Our first step is to evaluate water balance deviations from a climate only expected value across Great Britain (GB) - subsequently we call these water balance errors, dRR. We find that water balance errors vary significantly across GB (Figure 2-2). When considering highly permeable regions of GB (Figure 2-2c), catchments located in highly permeable regions appear to have the largest water losses (i.e. negative dRR values). Moreover, Figure 2-2b shows that AI values of the CAMELS-GB catchments are all lower than 1 (i.e. they are energy limited). The largest water losses are mostly observed in the catchments with high AI values (i.e. AI>0.5). There are also five catchments that have RR values higher than 1, which implies that these catchments are gaining water beyond precipitation (or are affected by unknown for anthropogenic activities). dRR values calculated based on linear regression using AI as the predictor (Figure A1) indicates quite similar variability across GB.

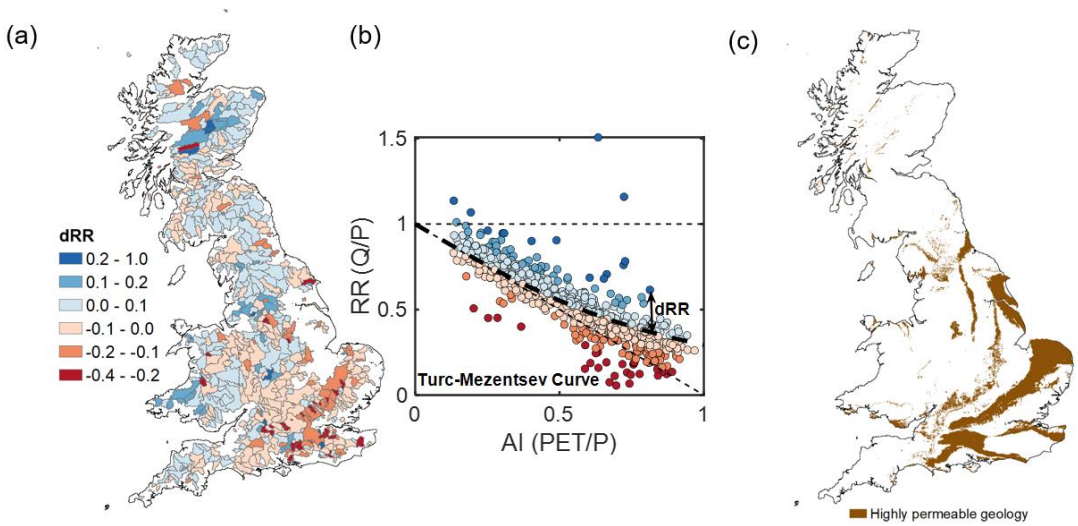


Figure 2-2. (a) Map of dRR values calculated according to Turc-Mezentsev, (b) Scatter plot of RR vs. AI values for 660 CAMELS-GB catchments. The thick black dashed curve is the Turc-Mezentsev Curve and dRR values for each catchment are calculated as the vertical difference between the observed RR and their corresponding points on the Turc-Mezentsev Curve. The thin dashed lines reflect energy and water limits. (c) Map of highly permeable geology of GB.

By assessing the relationship between dRR values and permeable area fraction (i.e. PAF), which is the fraction of catchment area that is underlained by highly permeable geology, we find that the median dRR values change from positive to negative when PAF values of catchment groups change from lower to higher (Figure 2-3a). We also find that while the 80th percentile of the catchment group with the highest PAF values have dRR values less than zero (down to -0.4), these percentiles decrease for other catchment groups with lower PAF values (Figure 2-3b). However, both Pearson linear and Spearman rank correlation values between dRR and PAF values of catchments are only -0.29 and -0.23, respectively. These findings imply that permeable area fraction is a factor affecting water balance issues of catchments, but it is not the only factor contributing to water balance issues. We also test the correlations between dRR and an extensive number of catchment attributes from the CAMELS-GB dataset (see Table A1) and find that catchment attributes do not show a strong correlation with dRR apart from some hydrological attributes (e.g. runoff, Q_{mean} etc.) (*Figure 2-4* and *Figure 2-5*). This indicates that currently available CAMELS-GB catchment attributes are not enough to understand water balance issues.

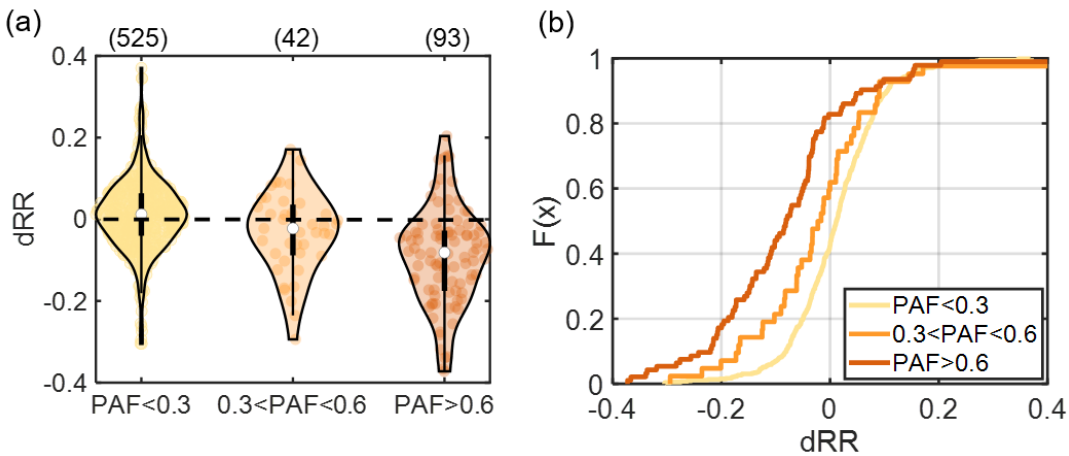


Figure 2-3. (a)Violin plot (b) cdf plot of dRR values for the groups of catchments with different permeable area fraction (PAF) ranges for 660 CAMEL-GB catchments. dRR values of catchments in each group are also shown as circles on the violin plot. The central white circles of the violin plot represent the median dRR value of each group. The bottom and top edges of the black-filled rectangular box indicate 25th and 75th percentiles, respectively. The bottom and top edges of the vertical black line demonstrate the lowest and highest data point in the dataset excluding any outliers, respectively. The dashed horizontal black line in the violin plot indicates where dRR equals to zero. The numbers

given in parentheses above violin plot are the number of catchments in each group. dRR values are calculated according to Turc-Mezentsev Curve.

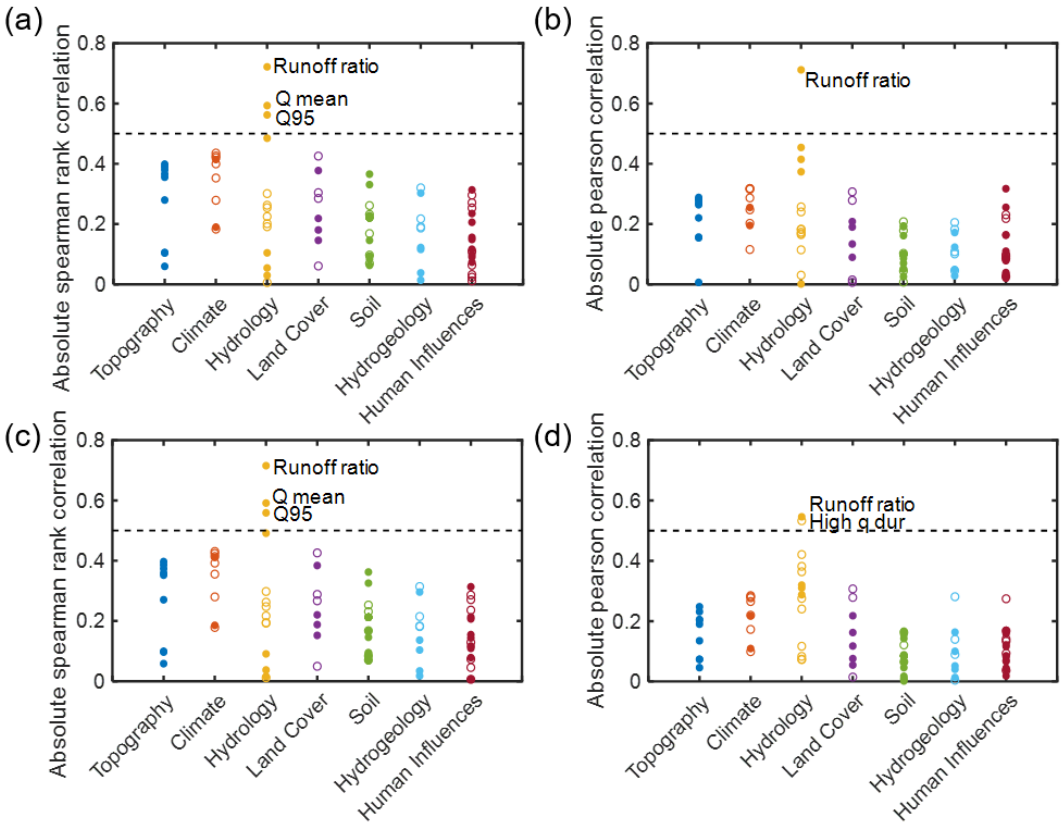


*Figure 2-4. (a) Absolute spearman rank and (b) pearson correlation values between dRR and CAMEL-GB attributes. (c) Absolute spearman rank and (d) pearson correlation values between WBE (i.e. dRR/RRobserved*100) and CAMEL-GB attributes. The filled and empty circles are positive and negative values, respectively. Descriptions and units of used catchment attributes are listed in Table A1.*

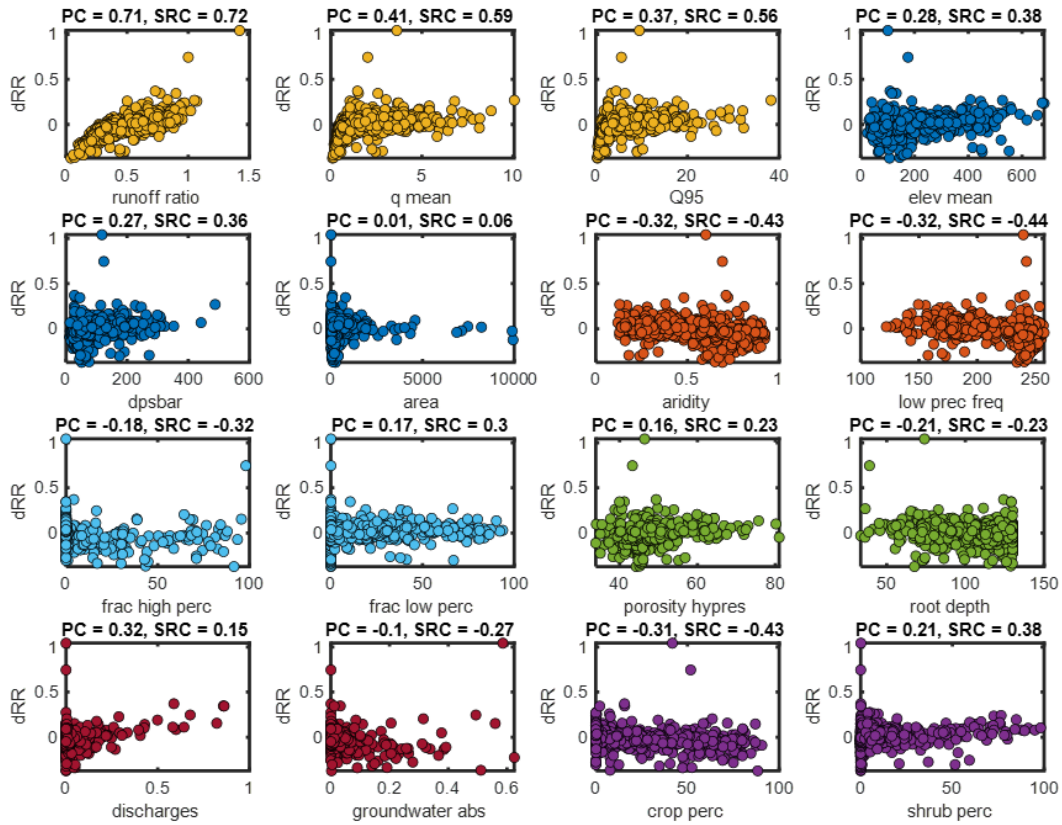


Figure 2-5. Scatter plots of dRR vs. some of CAMEL-GB attributes. PC and SRC represents Pearson Correlation and Spearman Rank Correlation, respectively. Different colors represent different attributes types (e.g. topography, land cover, soil etc.) as shown in Figure 2-4. *Descriptions and units of used catchment attributes are listed in Table A1.*

In our study, we aim to investigate whether the addition of location information can help to explain these water balance issues. The first locational aspect that we investigate is location to coast by analyzing the water balance of coastal catchments, defined as those whose catchment boundaries intersect with a 10km coastal buffer we defined. Figure 2-6a indicates the coastal catchments with and without highly permeable geology in the area between catchment and coast. Figure 2-6b shows that approximately 80th percentile of those coastal catchments with a highly permeable geology connection between catchment and coast have dRR values less than zero whereas only 40th percentile of other coastal catchment group have dRR values less than zero. This implies that coastal catchments with highly permeable geology extended all the way to coastline more likely lose water than coastal catchments without such geology (also see Figure A2, A3 and A4 in APPENDIX A). When we compare the groundwater levels of wells located in some coastal catchments

with and without highly permeable geology connection between catchment and coast (Figure A5), we observe that wells of catchments that have highly permeable geology connection with coast have lower groundwater levels and these catchments have lower dRR values than other coastal catchments compared. The result suggests that information about location to coast and geology is helpful in defining the water balance of coastal catchments.

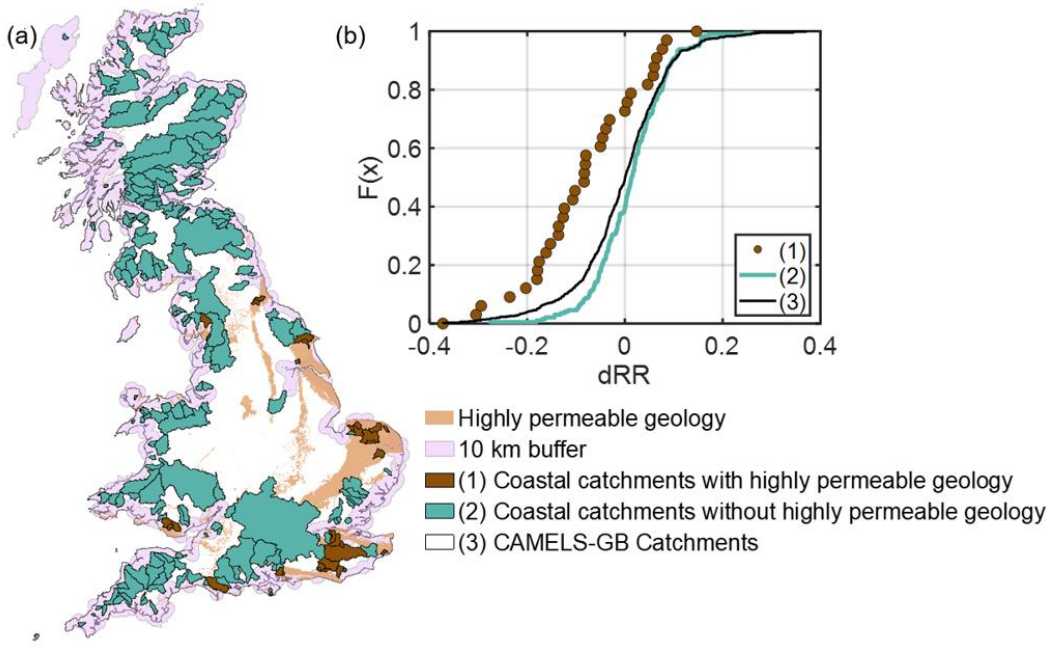


Figure 2-6. (a) Map and (b) cdf plot of coastal catchment groups. Map indicates the catchment groups analyzed in the hypothesis. dRR values are calculated according to Turc-Mezentsev Curve.

The second locational aspect we consider is the catchment's location within the wider river basin. In this analysis, we use a proxy to define the location of a catchment within wider river basin called the Strahler Sequence Index (SSI) introduced in section 2.2.2. dRR vs. SI values of catchments with highly permeable geology (Figure 2-7a) indicate that catchments with low SI values (i.e. I or II order streams) seem to be mostly losing water while catchments with SI greater than two have a range of dRR values centered around zero. For catchments without highly permeable geology, the dRR vs SI values show that they seem to be losing or gaining water regardless of their SI values (Figure 2-7b). If we further group these catchments based on their SI values and check their dRR values with respect to their SSI values (Figure 2-7c and Figure 2-7d), we observe that when smaller

catchments with highly permeable geology (i.e. SI=I or SI=II) are located in the upper regions of wider river basin (i.e. low SSI values), they lose more water than the ones located in the lower regions of wider river basin (i.e. high SSI values). We do not observe any relationships in dRR values of catchments without highly permeable geology based on their location within wider river basin. Consequently, understanding water balance issues of small catchments with highly permeable geology might be aided by information regarding their location within wider river basin (also see Figure A6, A7 and A8 in APPENDIX A).

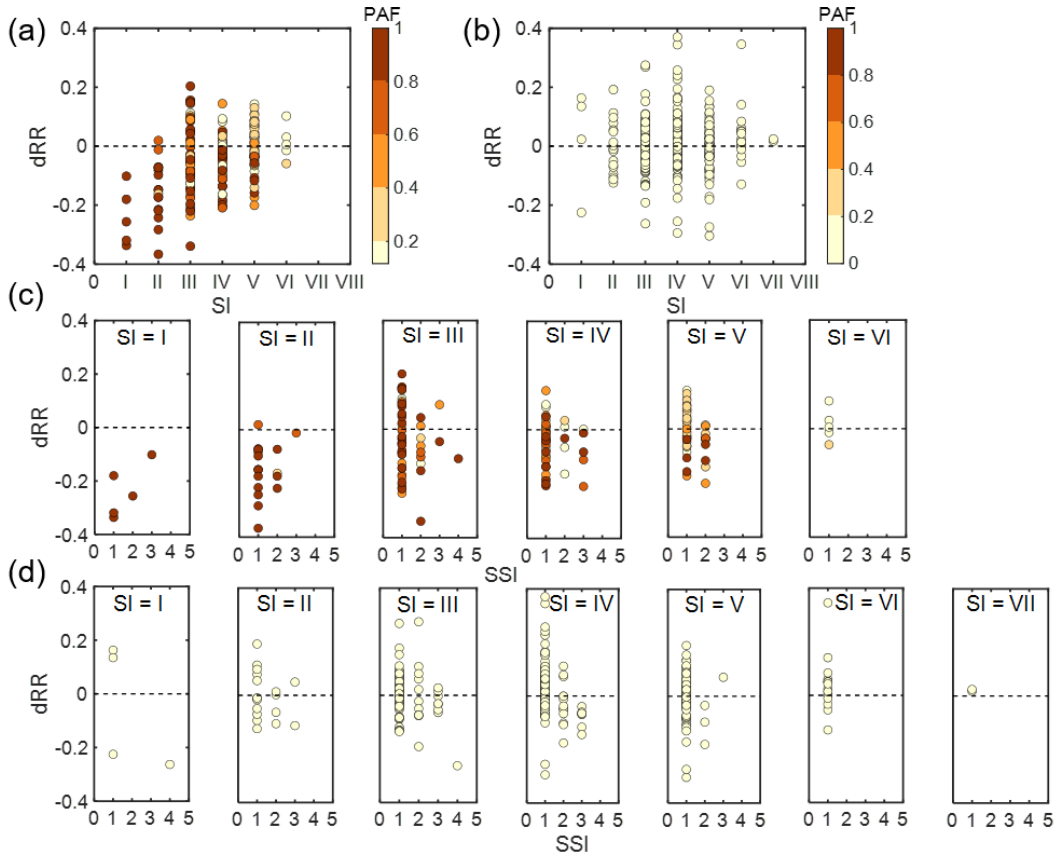


Figure 2-7. (a) Scatter plot of dRR vs. SI of 174 CAMELS-GB catchments with highly permeable geology (PAF (permeable area fraction)>0.1), (b) Scatter plot of dRR vs. SI (Strahler Index) of 320 CAMELS-GB catchments without highly permeable geology (PAF<0.1), (c) Scatter plots of dRR vs. SSI (Strahler Sequence Index) of catchment in (a) subgrouped based on their SI values and (d) Scatter plots of dRR vs. SSI of catchment in (b) subgrouped based on their SI values. dRR values are calculated according to Turc-Mezentsev Curve.

More locational aspects (e.g. location to a relevant neighbor catchment) could be useful to explain these water balance issues. In order to investigate this further by considering location to a relevant neighbor catchment, groundwater level data can be informative

regarding the potential direction of groundwater exchange between catchments. However, this data is not easily accessible for all catchments. Therefore, we investigate if topographic elevation which is a widely available and accessible proxy could be used instead of groundwater level. When we check the relationship between average groundwater levels and topographic elevation of 887 BGS wells, we observe a high correlation with both Spearman rank (SRC) and Pearson coefficient (PC) of 0.9 (Figure 2-8a). If we check this relationship in catchment scale, we still observe relatively high correlation values (i.e. between 0.7 and 0.9) between average groundwater levels and mean elevation or 10th percentile elevation of catchments (Figure 2-8b and Figure 2-8c). The information regarding the computation of average groundwater levels for wells and catchments can be found in Section 2.2. In Figure 2-8, catch. mean elev. and catch. elev. 10 represents mean elevation and 10th percentile elevation of catchments, respectively. These elevation values are obtained from CAMELS-GB dataset. In addition, the correlation between groundwater level difference and elevation difference of neighbouring catchment pairs is also high (i.e. between 0.8 and 0.9) especially when elevation difference is calculated using the wells' elevation located in these neighbor catchments (Figure 2-8d). If we use mean or 10th percentile elevation of these neighbouring catchments rather than well elevation to calculate elevation difference, we observe lower correlation values (Figure 2-8e and Figure 2-8f). Overall, these plots suggest that groundwater levels correlate with topographic elevation.

By using topographic elevation instead of groundwater level data, one speculation could be that there might be groundwater transfer between neighboring catchments from topographically higher catchment to lower catchment if they are hydrogeologically connected. This might be a factor contributing to water balance issues of catchments. In order to test this speculation, we attempt to formulate a proxy including both permeability (e.g. permeable area fraction, permeable shared boundary between neighbouring catchments) and topographic elevation (e.g. mean elevation, 10th percentile elevation)

attributes of catchments, but this does not help us to relate the water balance issues of catchments with a relevant neighbouring relationship as a locational aspect (results not shown). The reason might be that permeability and topographic attributes used are not specific enough or our approach to investigate this locational aspect does not work. These possible reasons are discussed further in the following section.

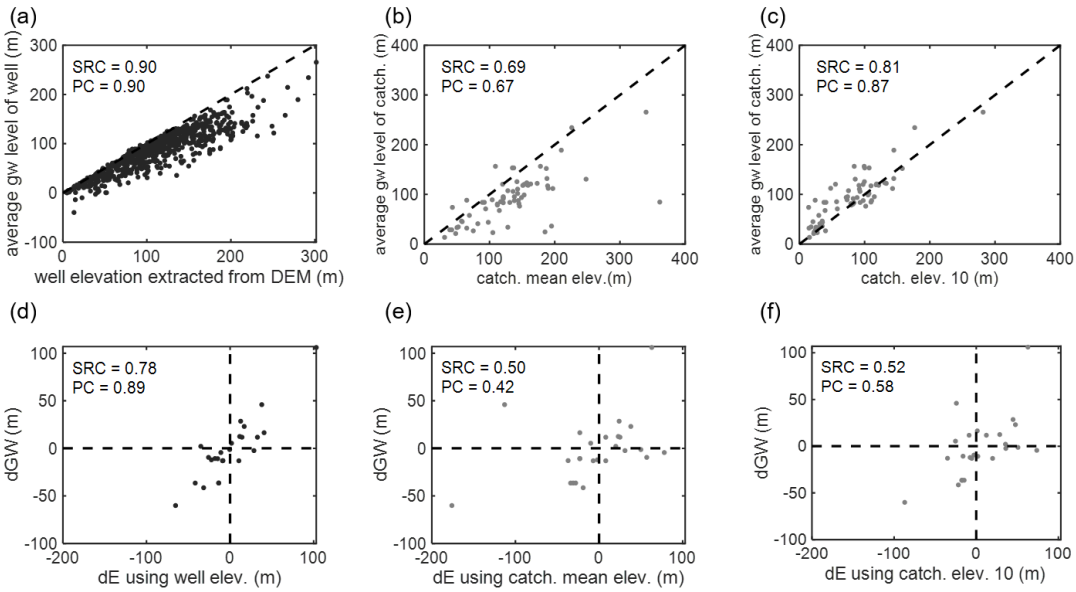


Figure 2-8. Scatter plots of (a) average gw level vs. elevation of 878 BGS wells, (b) average gw level of catchments vs. mean elevation and (c) gw level vs. 10th percentile elevation of 66 catchments, (d) dGW vs. dE (calculated based on well elevation), (e) dGW vs. dE (calculated based on catchment mean elevation) and (f) dGW vs. dE (calculated based on catchment 10th percentile elevation) of 25 neighboring catchment pairs. Where neighboring catchments are concerned, "dGW" denotes the difference between their average groundwater levels, while "dE" stands for the difference between their elevations. The computation of "dE" involves utilizing the average elevation of wells situated in neighboring catchments, the mean and 10th percentile elevations of neighbouring catchments in parts (d), (e), and (f), respectively. SRC and PC are Spearman Rank and Pearson correlation values for each scatter plot, respectively.

2.5 Discussion

Our study indicates that catchment location is informative to understand water balance losses in catchments especially when available catchment descriptors are not sufficient. We find that highly permeable coastal catchments are losing more water than others by combining information regarding location to coast with geology. Other studies conducted in the UK (e.g. Gale and Rutter, 2006; Allen and Crane, 2019) report that some regions (e.g. Flamborough, Arish Mell and West Lulworth) have groundwater flows in seaward

direction and seepages direct to ocean in chalk dominated basins (e.g. Yorkshire and Wessex basins). This effect of geology on groundwater flux of coastal catchments to ocean is also observed in the large-sample study by Liu et al. (2020) discussed earlier, and in the global coastal groundwater flow modeling study by Luijendijk et al. (2020). The authors investigate the response of marine groundwater discharge to variation in groundwater recharge, size of contributing area, subsurface permeability and topographic gradient. They find that globally groundwater flux of coastal catchments to the ocean is mainly controlled by aquifer permeability and topographic gradient, not by catchment length or groundwater recharge magnitudes. Looking closer at their results for Great Britain, we find that geology is the only controlling factor for coastal groundwater discharge in our domain (likely because coastal topography in relevant locations is rather low). These results are in line with our findings given that we also found no further improvement when including the topographic difference between catchment and coastline (results not shown).

Our study introduces an index to define the location of catchments within a wider river basin – the Strahler Sequence Index (SSI). Our findings show that highly permeable headwater catchments (i.e. those that have low SI values) located in the upper regions of wider river basins lose more water than the ones in the lower regions. This result is consistent with that of Buoaziz et al. (2018) who showes that water balance losses in the Meuse basin are mainly found in small headwater catchments and water loss positively correlates with increasing permeability (i.e. the percentage of highly fissured aquifers). Their modelling results indicate that while the upstream catchment of the Semois catchment (i.e. Sainte-Marie) loses 17-20% of observed discharge annually, these water balance losses decrease further downstream of the catchment. However, it is not consistent across studies to find such a result. For example, Schaller and Fan (2009) expect upper catchments to lose water and lower catchments to gain water. However, the ratio of river streamflow to the difference between precipitation and evapotranspiration (i.e. Q/P-ET) distribution in the

Cedar River Basin indicates that upper catchments can gain water and lower catchments can lose water depending on their underlying geologic structures.

One limitation of our study is the limited number of headwater catchments (i.e. only 14 and 49 of the CAMELS-GB catchments have SI of 1 and 2, respectively) to test the relation of location within drainage basin with water balance losses. This is despite the fact that the largest portion of total stream length of GB river network are headwaters (i.e. SI=1) (Figure 2-9). Given this limitation, more data are required to test this further across Great Britain and globally.

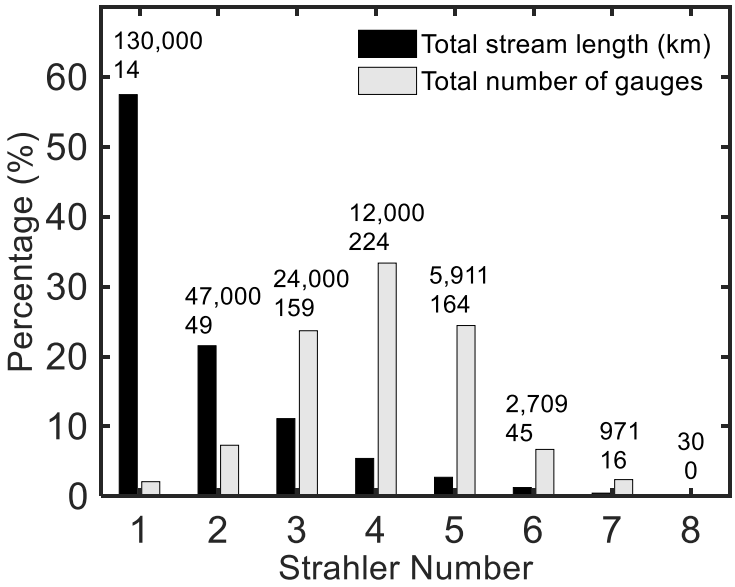


Figure 2-9. Percentages of total stream lengths and total number of CAMELS-GB gauges in GB river network according to their Strahler Number. Upper and lower values on the top of the bars represent total stream lengths (km) of streams (shown as black bars) and total number of CAMELS-GB gauges (show as gray bars), respectively.

High correlations observed between topographic elevation and groundwater levels of catchments implies that topographic relief correlates with the local groundwater flow system between neighbouring catchments. Using available groundwater levels and their topographic elevation, we are able to indicate that they are strongly correlated in Great Britain. Moreover, previous studies have suggested that topographic variability can be an indicator of groundwater connectivity. Theoretically, Toth (1963) suggest that topographic relief increases the importance of local flow systems and stimulates regional groundwater

flows. Condon and Maxwell (2015) conduct a modeling study to evaluate the relationship between topography and groundwater behavior in the US. They find that groundwater fluxes are mostly driven by topographic gradients. According to Munoz et al. (2016), three neighbouring headwater catchments with the same climate and land cover have dissimilar water balances due to interbasin water exchange, which is consistent with elevation differences based on topographic information provided.

Nonetheless, our study demonstrates that more specific catchment attributes or different approaches might be necessary to define groundwater transfer between neighboring catchments and link this with water balance issues of catchments. Even though groundwater transfer between some neighbouring catchments has been postulated in Chalky regions of Great Britain (e.g. between Pang and Thames catchments or Frome and Piddle catchments), there is a complex pattern of groundwater movement in different regions due to the various fractures, geological faults and dissolution features (Bradford, 2002; Griffiths et al., 2006; Allen and Crane, 2019). Due to this heterogeneity, it is challenging to introduce a single simple index to define this hydrogeological connectivity. More specific attributes are likely necessary to define this connectivity between neighbouring catchments beyond simple permeability. Moreover, considering only catchment-scale attributes as a top-down approach was sufficient for the earlier two locational aspects but it is not for the location to a relevant neighbor. Investigating this locational aspect by starting from more local scale (e.g. shared boundary regions of neighbouring catchments) to catchment-scale as a bottom-up approach might be a way to tackle this problem. However, it points to an interesting aspect that there is no available hydrogeological connectivity or topographic elevation descriptors for only dealing with a region of catchment and being still relevant for the whole catchment. Eventually, we could not justify groundwater transfer between neighbouring catchments and its relationship with the water balance issues of catchments in our study.

Lastly, it is important to highlight that the proxy we use to quantify water imbalances of catchments (i.e. dRR) does not differentiate the effect of water management activities (i.e. human disturbances) from natural catchment features. Moreover, dRR values of human impacted catchments have the largest deviations compared to benchmark catchments (Figure 2-10a). The net effects of water management activities on runoff are still unknown and poorly quantified to relate with water balance issues (Figure 2-10b) – even though the catchments included should largely be free of major management activities. This lack of information on human activities has also been also emphasized in previous studies (e.g. Schwamback et al., 2022).

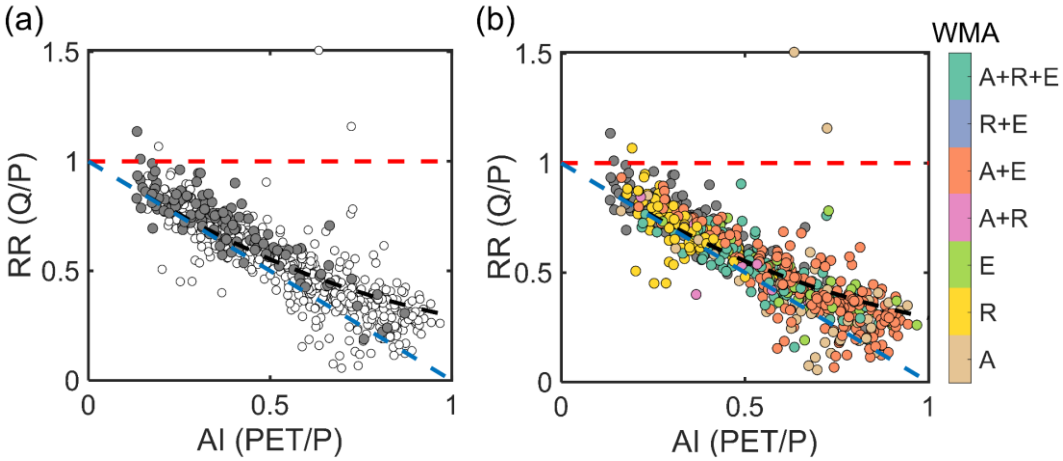


Figure 2-10. (a) Natural catchments (i.e. gray filled circles) (b) Human impacted catchments colored based on their water management activities (i.e. WMA). In both (a) and (b), red horizontal and blue diagonal dashed line indicates the water limit (i.e. Q=P) and the energy limit (i.e. Q = PET), respectively. The black dashed curve is the Turc-Mezentsev Curve. A is abstraction, R is reservoirs and E is effluent returns. Natural catchments have no abstractions and discharges or the variation due to them is so limited that the gauged flow is considered to be within 10% of the natural flow at, or in excess of, the Q95 flow. Abstraction means that natural runoff is reduced by the quantity abstracted from a reservoir or by a river intake for different purposes (e.g. public water supply, industry and/or agriculture) and/ or reduced or augmented by groundwater abstraction or recharge. Effluent returns are outflows from sewage treatment works will augment the river flow if the effluent originates from outside the catchment (Marsh and Hannaford, 2008). Water management activities shown here are based solely on the information in Factors Affecting Runoff section which is provided at NRFA website (https://nrfa.ceh.ac.uk/data/search) for each gauge. This information is subject to ongoing evaluation and is routinely refreshed on the website. Nevertheless, specifics concerning the timing and duration of these activities, as well as the impacts stemming from factors like population growth, climate variations, and alterations in land cover/use, remain undisclosed. In the scope of this investigation, it is posited that the runoff within the examined catchments is influenced by these water management activities during the designated study period.

2.6 Summary

The analysis of large samples of catchments has become a standard tool in hydrologic analysis. However, available catchment descriptors are regularly shown to be insufficient in defining all relevant aspects of the hydrological behaviour of catchments. In our study, we test the hypothesis that the location of a catchment relative to its surrounding area can add further information. Specifically, we focus on three different locational aspects to determine their value in defining subsurface leakage of GB catchments. We find that location relative to the coast and within wider river basin explains significant parts of the water balance issues of highly permeable catchments. We further find that there is a strong relationship between topography and groundwater levels. We have not found a way to use this information in a helpful manner though, which might partially be related to the need to add more detail, which we did not do here. The strong bias in streamgage location is also problematic in GB, where we also lack information in first and second order streams, even though they make up most of the stream length. Also, in how far our results are specific to the GB setting with its Chalk geology and other characteristics remains open to future studies.

Overall, we believe that our results show that location specific information should be considered in large sample hydrology. It is likely that other informative indices can be defined which are full or partially based on catchment location. We believe that this consideration has been largely overlooked so far and deserves further investigation. Given the limitations of our dataset, we could not test some location specific expectations for catchment water balances. For example, climatic change lines or strong topographic variability, e.g. mountain fronts, cannot be found in GB but have been suggested to play a role (Fan, 2019). We are further limited by the problem that GB has limited topographic variability compared with other domains. Hence, we hope that others will perform similar analyses across more topographically and climatically diverse domains.

Chapter 3 A Priori Selection of Hydrological Model Structures in Modular Modelling Frameworks

This chapter has been submitted to Hydrological Sciences Journal and has undergone slight modifications to align with the general layout of this thesis. The study was conceptualized by Melike Kiraz, Gemma Coxon, and Thorsten Wagener. Melike Kiraz conducted the data processing, model simulations and creation of figures under the guidance of Gemma Coxon and Thorsten Wagener. The manuscript was primarily written by Melike Kiraz, with input and comments from all co-authors.

Citation: Kiraz, M., Coxon, G. and Wagener, T. (2023). A priori selection of hydrological model structures in modular modelling frameworks: Application to Great Britain. Hydrological Sciences Journal. (Under review).

3.1 Introduction

Modular rainfall-runoff modelling frameworks have been widely used to provide a more flexible approach to modelling diverse catchments. These frameworks consist of model structures or model structural elements that can be combined in different ways to represent different dominant hydrological processes (often at the catchment scale). Some of the more widely used frameworks include the Modular Modeling System (MMS) (Leavesley et al., 1996), the Rainfall-Runoff Modelling Toolbox (RRMT) (Wagener et al., 2001a), the Framework for Understanding Structural Errors (FUSE) (Clark et al., 2008), Catchment Modelling Framework (CMF) (Kraft et al. 2011), SUPERFLEX (Fenicia et al., 2011; Kavetski and Fenicia, 2011), the Structure for Unifying Multiple Modeling Alternatives (SUMMA) (Clark et al., 2015a;b), the Eco-hydrological Simulation Environment ECHSE (Kneis, 2015), Dynamic fluxEs and ConnectIvity for Predictions of HydRology framework DECIPHeR (Coxon et al., 2019),the Nonstationary Rainfall-Runoff Toolbox (NRRT)

(Sadegh et al., 2019), the Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT) (Knoben et al., 2019), and RAVEN (Craig et al., 2020) among others. These frameworks vary in their spatial resolution, in the model structures and structural elements included, in the granularity of the components that make up the model structures and in other ways such as the optimization or uncertainty quantification tools included.

An initial step in any modular modelling exercise is the selection of the model structures (or model structural components) to be considered given that they should be both potential representations of the system(s) under study and appropriate for the modelling objective(s). We might assume that a framework is so flexible that it can reflect any system, but even then, some pre-selection might be helpful to avoid testing model structures that can already be considered a priori unsuitable– e.g. those based on process perceptions that are not present in the study domain to avoid getting the right result for the wrong reasons (e.g. Grayson et al., 1992; Kirchner, 2006). Experience, embedded in a perceptual model(s) of the underlying system(s), is one way to identify the differences between systems such as catchments (Seibert and McDonnell, 2002; Beven and Chappell, 2021; Wagener et al., 2021; Fenicia and McDonnell, 2022). One recurring problem in this context is that various multi-model studies found that there is no specific model that performs better than all others for a specific catchment, and that we might only find some basic trend that models with more parameters have more flexibility to fit rainfall-runoff relationships (e.g. Perrin et al., 2001; Kollat et al., 2012; Van Esse et al., 2013; Orth et al., 2015), but even that conclusion is not always clear and confounded with the adequacy of a certain process representation (Knoben et al., 2020). As a consequence, there is often no clear relationship between catchment type and well-performing model structures (e.g. Nicolle et al., 2014; Ley et al., 2016; Knoben et al., 2020). One aspect that has so far been studied less extensively, is that the actual choice of model structures included in such studies might be increasing the problem of equifinality in model performances, and thus that the problem can be reduced through better a priori selection.

Two of the above-mentioned modular modelling frameworks have so far been used in multiple studies in the UK - the RRMT (Wagener et al., 2001a; Lee et al., 2005) and FUSE (Coxon et al., 2014; Lane et al., 2019) frameworks. The former framework was specifically developed for the UK, a region with mostly small catchments located in a temperate climate, while the latter includes a wider range of model components based on globally used models (more details in the methods section). Using RRMT, Lee et al. (2005) tested 12 model structures on 28 UK catchments. The authors could not find evidence for a relationship between catchment type and model structure, though they identified a subset of models that performed better than the rest. Lee et al. (2005) characterized catchment types using only descriptors that are available for both gauged and ungauged UK catchment– area, regionalized baseflow index and average rainfall – which have limited value in characterizing hydrologic differences (Addor et al., 2018). Using the FUSE framework, Coxon et al. (2014) evaluated performances of 78 model structures across a different set of 24 catchments in England and Wales. They found that statistical model performance increased with catchment wetness and that only certain model structures provided good model performance in baseflow dominated catchments. Moreover, they highlighted the possibility of identifying more informative signatures for better model identification (in gauged catchments). Lane et al. (2019) followed up on Coxon et al.'s study by selecting only four model structures from the FUSE framework but by implementing them in over 1000 GB catchments. The performance patterns of all four models across GB were very similar. They performed better (worse) in wetter (drier) catchments and were particularly poor in catchments with groundwater leakage to neighboring catchments, which none of the model structures accounted for. Even though some performance differences exist between these model structures in different types of catchments, the authors were not able to explain them through model structure/complexity differences.

In this study, we aim to test how much distinguishing between model structures in multi-model studies is a function of which model structures are selected beforehand. In other words, we would like to investigate the importance of a priori model selection step (i.e. deciding which model structures will be included in a multi-model study) on our ability to observe relevant performance differences between model structures. To do so, we compare two different modular rainfall-runoff modelling frameworks (and the model structures they include) on 998 GB catchments in a Monte Carlo framework. For this comparison, we select six model structures from the Rainfall-Runoff Modelling Toolbox (Wagener et al., 2001a) and compare these to simulation results from the four FUSE model structures reported in Lane et al. (2019). We then attempt to explain the resulting differences between model structures using hydrologic signatures – as more informative descriptors than catchment properties. Lane et al. (2019) used model structures of similar complexity and found no distinguishable differences between their process representations that can easily be linked to our perceptions of different dominant hydrologic processes across the UK. We hypothesize that a different a priori model selection will improve our ability to distinguish the performance of model structures across catchment types. So, we are trying to overcome two problems we were left with after the study by Lane et al. (2019): (1) The lack of well-performing model structures in leaky catchments. (2) The lack of distinctive and hydrologically relevant performance differences between model structures and catchment types.

3.2 Data, Modular Modelling Framework and Methods

3.2.1 Data and Hydrologic Signatures

In this study, we analyse the same catchments across Great Britain as selected by Lane et al. (2019) to ensure comparability of our results. These catchments were selected from the National River Flow Archive (Centre for Ecology and Hydrology, 2016) based on the quality and availability of flow time series. They represent a diverse range of catchment characteristics in terms of topography, geology and climate, thus capturing much of the

variability found across Great Britain (GB). Details about the general characteristics (e.g. climatic, topographic, geologic) of Great Britain are provided in Section 2.2. of Chapter 2.

We use daily rainfall, streamflow, and potential evapotranspiration time series for twenty-one years (January 1, 1988 – December 31, 2008) to cover the same time period used by Lane et al. (2019). Daily rainfall and potential evaporation data are derived from the Centre for Ecology and Hydrology Gridded Estimates of Areal Rainfall (CEH-GEAR) (Tanguy et al., 2021) and the Climate Hydrology and Ecology Research Support System Potential Evapotranspiration (CHESS-PE) (Robinson et al., 2015a), respectively. Daily potential evapotranspiration (mm/day) was calculated using the Penman –Monteith equation (Monteith, 1965) for a well-watered grass surface (Allen et al., 1998) with meteorological data from the Climate Hydrology and Ecology research Support System dataset (CHESS-met) (Robinson et al., 2015a; b). Daily observed streamflow data from the National River Flow Archive (NRFA) are used to evaluate model performances (Centre for Ecology and Hydrology, 2020).

To compare model structure performances across different catchment types, we organise the catchments based on four hydrological signatures which have been found helpful for distinguishing UK catchments in the past (Coxon et al., 2014; McMillan et al., 2022b): baseflow index (BFI), runoff ratio (RR), the deficit in water balance (dRR) and slope of flow duration curve (Slope of FDC). We choose these hydrological signatures because they differ in the information they provide about the runoff processes of catchments. Runoff ratio represents the proportion of precipitation becoming streamflow, while the baseflow index represents the proportion of streamflow sourced from groundwater. The deficit in the water balance indicates if catchments produce more or less runoff than expected based on climate only, while the slope of the flow duration curve indicates whether catchments have more or less flashy (i.e. variable) flow regimes (e.g. Yadav et al., 2007). We further use streamflow-derived baseflow index (BFI) values from the UK Hydrometric Register (Marsh and Hannaford, 2008).

For each catchment, we calculate its long-term water balance (i.e. runoff ratio, RR = Q/P), aridity index (AI = PET/P) and slope of flow duration curve (33-66 percentile; Sawicz et al., 2011) values using daily precipitation, P, potential evapotranspiration, PET, and streamflow, Q, data. We also calculate the expected water balance (expected RR) based only on climate using the Turc-Mezentsev curve. The formula of Turc-Mezentsev Curve which is used to estimate expected RR is provided in Section 2.3.1 of Chapter 2. Delta runoff ratio (dRR) is calculated by taking the difference between observed runoff ratio and the expected runoff ratio derived from the Turc-Mezentsev Curve.

3.2.2 Modular Modelling Framework

In this study, we selected a minimal set of model structures that covered the range of dominant hydrological processes across our study domain of GB. We wanted to ensure that the model structures had different levels of complexity (i.e. the number of parameters) to represent these dominant hydrological processes and that model structural choices (such as the type of flow routing module) could be evaluated in isolation to demonstrate the impact of different model structural modules on model performance.

To achieve these goals, we first selected model structures from the Rainfall-Runoff Modelling Toolbox (RRMT) (Figure 3-1a) (Wagener et al., 2001a). RRMT is a flexible modelling framework that allows the user to develop model structures with different complexity levels by combining soil moisture accounting and flow routing modules that are of a low and medium complexity (Wagener et al., 2004). We selected six model structures from the toolbox consisting of different combinations of two soil moisture accounting modules and three flow routing modules. The soil moisture accounting modules (PEN and PDM) are based on long standing experience with UK catchments (see discussions in Wagener et al., 2004; Lee et al., 2005; Moore, 2007) and capture key runoff generation processes across GB. The flow routing models (CRES, 2PAR and LEAK) capture different levels of complexity from one to two linear reservoirs and a leaky routing component to reflect different flow pathways because of different soils and regional

aquifers that occur across GB (Moore, 2007). The model structural modules are explained in detail below and the parameters of each module are listed in Table B2.1. (APPENDIX B). Full names of used soil moisture accounting and routing modules are given in List of Acronyms.

PEN is a parsimonious two-store structure based on an empirical drying curve concept developed from observed drying patterns in UK soils by Penman (1949). The upper and lower store represent the root zone and an infinite soil reservoir, respectively. Analysing UK soils, Penman (1949) found that actual evapotranspiration occurs close to the potential rate whenever water is available in the root zone reservoir. The actual rate decreases to a very small percentage of the potential rate (8%) when the upper store is depleted. Effective rainfall – the part of the rainfall that contributes to runoff – is created in two ways. Either as rainfall bypass to represent processes such as rapid groundwater recharge or rainfall falling close to a river, or as saturation-excess runoff which is produced when both stores are full. The model parameters define the size of the root zone storage, S_{max1}, and the fraction of bypass flow, φ.

PDM is the probability-distributed soil moisture accounting component, which represents the variability in soil moisture storage across a typical humid catchment using a distribution of storage depths (Moore, 2007). Effective rainfall is produced as overflow from the stores which are described as Pareto distribution based on two parameters, the maximum storage capacity, C_{max}, and parameter, b, describing the shape of the distribution. While PEN module primarily focuses on water loss due to the evapotranspiration and features a fixed storage capacity across the entire catchment, PDM module represents more flexibility through its distribution function for soil moisture storages with different capacities. This accounts for heterogeneity in the catchment.

There are three different routing components. Firstly, CRES is a single linear reservoir defined only by a time constant. 2PAR is a combination of two linear reservoirs in parallel for routing, one representing fast flow and the other representing slow flow. The effective

rainfall (ER) is distributed with respect to parameter a describing the fraction of flow through the fast reservoir, while both reservoirs are defined by a time constant. And thirdly, LEAK is a leaky aquifer routing component, which allows the model to consider the situation when the water balance of a catchment is not closed. The flow from the bottom outlet represents leakage from the catchment, while the middle and upper outlets contribute to routing the effective rainfall. While CRES modules has simpler structure that renders it well-suited for catchments characterized by uncomplicated hydrologic processes, 2PAR and LEAK modules which include multiple reservoirs, are better equipped to capture more complex hydrologic processes such as subsurface flows and leakage of water within catchments.

The four model structures provided by the FUSE modelling framework (Clark et al., 2008) and used by Lane et al. (2019) are shown in Figure 3-1b. These model structures are based on four hydrological models which are TOPMODEL (Beven and Kirkby, 1979), the Variable Infiltration Capacity (ARNO/VIC) (Liang et al., 1994; Todini, 1996), the Precipitation-Runoff Modelling System (PRMS) (Leavesley et al., 1983) and SACRAMENTO (Burnash et al., 1973). The details of model parameters are listed in Table B2 (APPENDIX B). The modelling decisions are described by Lane et al. (2019) (See Table 3 in Lane et al.'s study). Even though these models have similar complexity, their structures are different in terms of the structures of upper and lower soil layers and the parametrizations of water balance components such as evaporation, surface runoff, percolation, interflow and baseflow. Since only a small proportion of the catchments (1%) have a snow fraction higher than 0.1 and likely to be snow impacted, no snow modules are used in any model structures selected from both RRMT and FUSE frameworks.

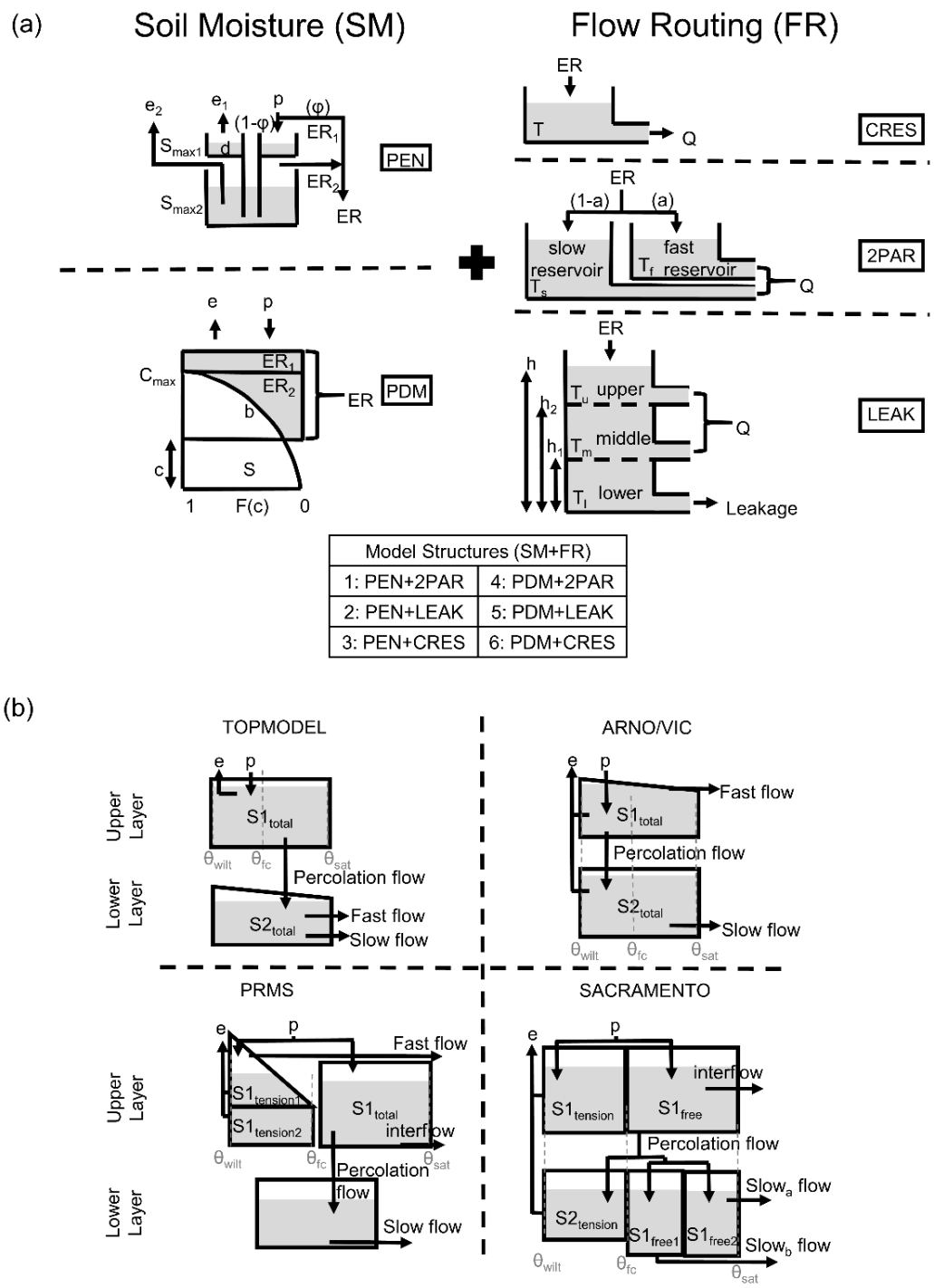


Figure 3-1. Structures of models used in the study. (a) six model structures consisting of different combinations of two soil moisture modules and three flow routing modules provided by Rainfall – Runoff Modelling Toolbox (RRMT). PEN/PDM + 2PAR have 5 parameters, PEN/PDM + LEAK have 7 parameters and PEN/PDM + CRES have 3 parameters. (b) four models provided by The FUSE modelling framework. Schematic illustrations of their structures are taken from Lane et al. (2019). TOPMODEL and

ARNO/VIC have 10 parameters, PRMS has 11 parameters and SACRAMENTO has 12 parameters. In the diagram, p, e, ER, S and Q represent precipitation, evaporation, effective rainfall, storage and outflow respectively. In PEN module, S_{max1}, S_{max2}, d, φ represent size of the upper store (i.e. root constant), size of the lower store, initial deficit in upper store and bypass value, respectively. In PDM module, C_{max}, b and c represent maximum storage capacity, degree of spatial variability and initial critical capacity, respectively. In CRES module, T represent the residence time of reservoir. In 2PAR module, a, T_s and T_f represent the fraction of effective rainfall going through fast reservoir, the residence times of reservoirs for slow flow and fast flow, respectively. In LEAK module, T_u, T_m, T_l, h_1 and h_2 represent the residence times of upper, middle, lower parts, lower threshold and upper threshold, respectively.

Model equations in the FUSE framework are solved by an implicit version of Newton-Raphson method (See Appendix A in Clark et al.'s study (2008)). Equations in the soil moisture accounting and routing modules of RRMT framework are the first order equations which are solved in MATLAB programming environment (Wagener et al., 2001a). However, our focus is not the analysis of the relative performance between the two frameworks, but rather the differences between model structures within each framework. Since each framework utilizes a consistent numerical implementation across its model structures, the distinctions among model structures within each framework are not linked to their numerical implementation.

3.2.3 Methods

3.2.3.1 Model Set-up

To enable comparison with the results from Lane et al. (2019) we replicated the modelling setup the authors employed. Consequently, 10,000 parameter values for the six model structures in this study are independently and randomly sampled from uniform distributions. The model parameter ranges used for RRMT and FUSE frameworks are given in Tables B2.1 and B2.2. These parameter ranges are suggested as feasible for RRMT (Wagener et al., 2001b) and FUSE model structures (Clark et al., 2008; Coxon et al., 2014). The first five years of 21-year period (1988-2008) are used as a warm-up period. A shorter warm-up period (e.g. 5% of the study period that is used in Chapter 4 – See Section 4.3.3.) can be enough to allow the model to reach a steady state by accounting for initial conditions and ensuring that transient effects from the starting point do not significantly affect the

simulation results. In this study, the utilization of this warm-up period (approximately 24%) was solely driven by the intention to mirror the modelling configuration employed by Lane et al. (2019) in their research, thereby facilitating a direct comparison with their findings.

3.2.3.2 Model performance evaluation

In this study, it is crucial to be able to compare the performance of multiple model structures across many catchments and to make results comparable with Lane et al. (2019). Considering this, we use the Nash Sutcliffe (NSE) - which was used by Lane et al. (2019) - and Kling Gupta (KGE) Efficiency metrics because they are normalized and unit-free metrics enabling the comparison of model performances across catchments. Both metrics are calculated for the time period of 1993-2008. We only have the best runs based on NSE for the Lane et al. (2019) study, which is why we calculate the KGE values for those and not identify the best KGE run separately.

Nash-Sutcliffe efficiency (NSE) metric is calculated as (Nash and Sutcliffe, 1970);

$$NSE = 1 - \frac{\sum_{t=1}^{n}(x_{s,t} - x_{o,t})^2}{\sum_{t=1}^{n}(x_{o,t} - \mu_o)^2} \qquad (3\text{-}1)$$

where $x_{s,t}$ is the simulated value at time-step t, $x_{o,t}$ is the observed value at time-step t, n is the total number of time-steps and μ_O is the mean of observed values. NSE ranges from -∞ to 1, with a value of 1 indicating a perfect correspondence between simulations and observations. NSE=0 indicates that simulations have the same predictive skill as the mean of the observations, while NSE<0 indicates that simulations are a worse predictor (Schaefli and Gupta, 2007).

Kling Gupta efficiency (KGE) metric is calculated as (Gupta et al., 2009);

$$KGE = 1 - \sqrt{(\alpha - 1)^2 + (\beta - 1)^2 + (r - 1)^2} \qquad (3\text{-}2)$$

with $\alpha = \sigma_S / \sigma_O$ and $\beta = \mu_S/\mu_O$ where σ_o and σ_S are the standard deviations of observed and simulated values, μ_o and μ_S are the mean of observed and simulated values

and r is the linear correlation coefficient between observed and simulated values, respectively. Like NSE, KGE metric also ranges from -∞ to 1. KGE=1 also means that simulations are perfectly in agreement with observations. Knoben et al. (2019) found that when KGE is approximately -0.41, simulations have the same predictive skill as the mean of the observations.

To establish whether the performance of specific model structures (measured using KGE or NSE) varies with the magnitude of a specific hydrologic signature (section 2.1), it is difficult to simply create scatter plots of one against the other as there is a lot of noise that makes it difficult to see trends. We therefore smooth the data to lower the effects of variability across catchments so that the separations between increasing or decreasing trends in the relative performance differences of model structures can be observed more clearly (e.g. Burn and Elnur, 2002). Without smoothing, it is difficult to observe the increasing and decreasing trends on the scatter plots (see Figure B1.1. in APPENDIX B). We use a nonparametric local weighted regression (LOWESS) approach which includes a bi-square weight function to minimize the effect of the outliers in the smoothed values (Cleveland, 1979; Coxon et al., 2015). In the LOWESS (locally weighted scatterplot smoothing) method, the highest NSE or KGE value for each catchment is sequentially selected as the central point (x) among a set of 2k+1 data points (k is called as span which is the half of window size selected for smoothing). The catchments are sorted based on chosen attributes. Within these 2k+1 points, k data points are before the central NSE or KGE point, and k are after. A smoothed NSE or KGE value and its variability are then determined by fitting a weighted linear regression to this set of data points. This process is repeated for all data points to obtain a final LOWESS fit. More details regarding the LOWESS smoothing process are given in APPENDIX B (Section B3 and Figure B1.1). We find that a smoothing window size of 40 catchments reflects the performance changes across catchments without overly smoothing the results. We then calculate the performance

difference (i.e. NSE or KGE difference) between each model structure and the best model structure (i.e. the model structure having the highest smoothed NSE or KGE value).

3.3 Results

3.3.1 RRMT and FUSE model performance across Great Britain

First we compare the performance of model structures from both frameworks across Great Britain. While Lane et al. (2019) included 1013 catchments in their analysis, we remove 15 catchments because they have unrealistic runoff ratio values (i.e. RR>1) or because all model structures fail to work (i.e. NSE<0). We assume that these problems are caused by unknown and thus unaccounted for anthropogenic impacts. Figure 3-2a and Figure 3-2b show the best NSE performance from all the model structures from RRMT and FUSE frameworks for 998 GB catchments, respectively. Both frameworks simulate 95% of the studied catchments with NSE values higher than 0.5 as shown in Figure 3-2c. The more complex FUSE models perform slightly better in catchments where both frameworks achieve high NSE values.

The spatial patterns of model performance in Figure 3-2a and Figure 3-2b are largely similar. However, there are 40 catchments located in south-eastern GB where we find larger performance differences (i.e. >±0.2 NSE) between the frameworks (see Figure B4.1a and c in APPENDIX B). In 28 of them, highest NSE values are obtained by the RRMT framework, and they are significantly higher than ones by the FUSE framework (i.e. NSE difference>0.2). More than 80% of these catchments have highly permeable geology covering more than 60% of their respective catchment areas. Among them, there are 6 catchments where the FUSE models perform particularly poorly (i.e. NSE<0) but RRMT is able to simulate their streamflow with NSE>0.7. In this Chalky region, the catchments are mostly baseflow- dominated and some of them are losing water through regional groundwater flows. The inclusion of a LEAK routing component in the RRMT framework enables better performances under those conditions.

Lower NSE values are also seen in some catchments of northern east and central Scotland and north Wales, likely due to snow or reservoirs. There are three catchments in north-east Scotland (i.e. snow fractions>0.1) for which model performances show NSE values less than 0.5. We did not focus on this any further given that these are just three out of almost 1,000 catchments. To investigate the impact of reservoirs in more detail, we investigated the relationship between two reservoir related descriptors (contributing area upstream of the reservoir and normalized upstream capacity; Salwey et al., 2023) and highest NSE scores obtained by RRMT and FUSE model structures for 252 catchments (Section B5 in APPENDIX B). We found that there is a small decline in model performance the closer a reservoir exists to the catchment outlet, and to a lesser degree the larger it is (Figure B5.1 in APPENDIX B). However, the variability in performance change is very large and it would take consideration of additional aspects such as reservoir management to add reservoirs to the models used here (e.g. Payan et al., 2008), which is beyond the main aims of this study. We also calculate KGE values for the best NSE model runs for comparison (Figure 3-2c, Figure 3-2d and Figure B4.2 in APPENDIX B). Overall, they indicate similar patterns in comparison with NSE values across GB. However, Figure 3-2d shows that RRMT has a larger number of catchments with KGE values>0.4 and both frameworks have quite similar distributions after KGE>0.8, whereas FUSE has a larger number of catchments with NSE values >0.6 (Figure 3-2c). Performance differences between RRMT and FUSE frameworks are not fully consistent when comparing NSE and KGE values which are calculated based on the simulation producing best NSE in each catchment, due to the difference between the formulations of NSE and KGE. While bias, variance and correlation components of streamflow are equally weighted in KGE formulation, they are weighted differently in the NSE formulation (i.e. variance term is more dominant in NSE formulation than other terms) (Gupta et al., 2009). Moreover, the relationship between NSE and KGE will be different for different catchments, mostly depending on the coefficient of variation of the observed streamflow (Knoben et al., 2019; Lamontagne et al., 2020). Having more complex model structures (i.e. having higher number of parameters) might

provide FUSE with more ability to capture the variance of streamflow than RRMT but RRMT seems to be performing as well as FUSE considering all three components of streamflow equally.

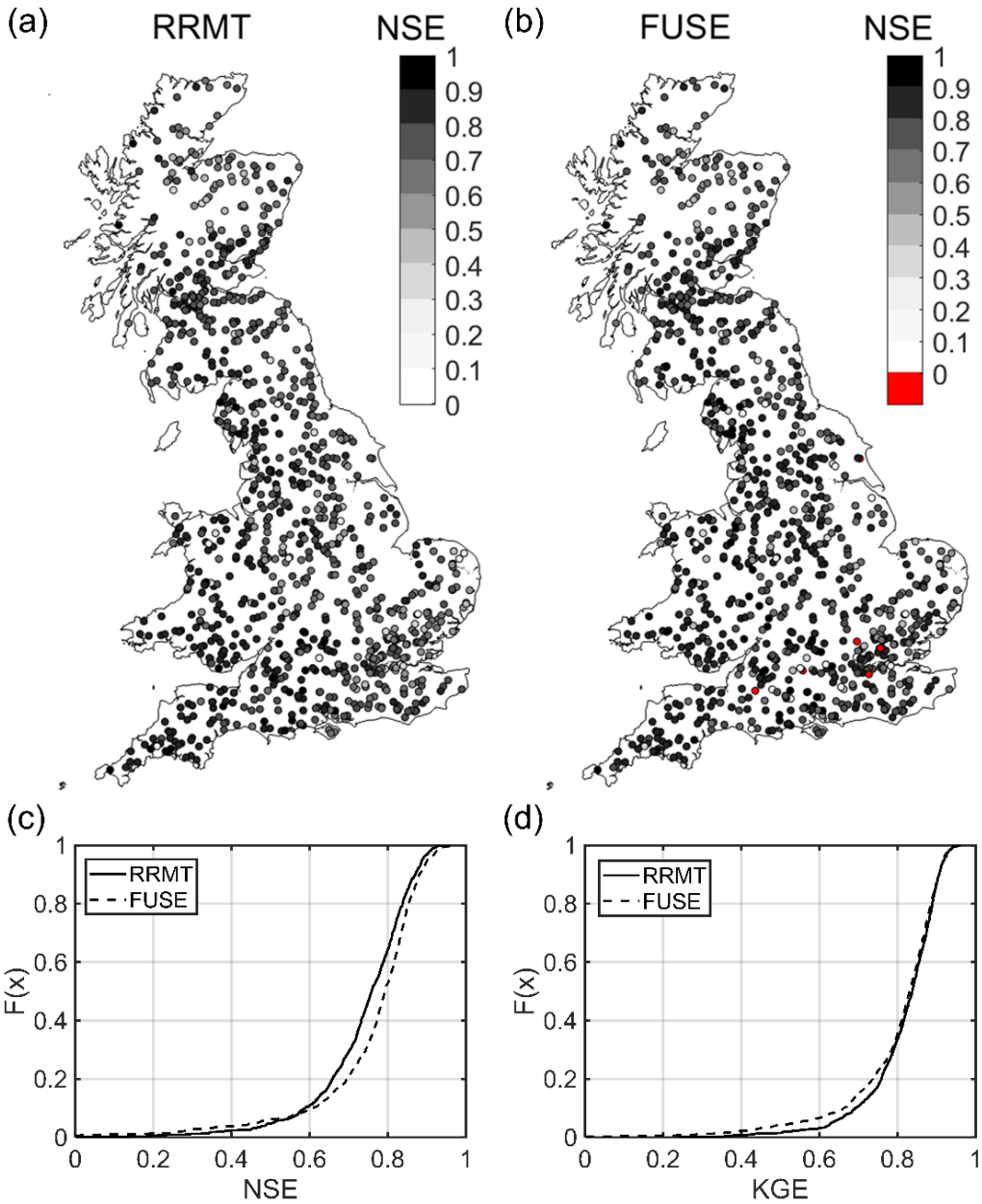


Figure 3-2. NSE values of best simulations performed by any of model structures which are selected from a) RRMT and b) FUSE frameworks and Cumulative Distribution Function (CDF) plots of (c) NSE and (d) KGE values of these frameworks.

3.3.2 Linking model structure performance with hydrologic signatures and catchment characteristics

Figure 3-3 and Figure 3-5 show the differences in NSE values of the six RRMT model structures (PEN+2PAR, PEN+LEAK, PEN+CRES, PDM+2PAR, PDM+LEAK, PDM+CRES) and the four FUSE model structures (TOPMODEL, ARNO/VIC, PRMS, SACRAMENTO) in relation to the best performing model structure in each framework. We plot these results against four hydrologic signatures: (a) BFI, (b) dRR, (c) RR and (d) slope of FDC, which have in the past been shown to be informative for UK settings (Yadav et al., 2007). We visualize the results in two different ways. The left columns of Figure 3-3 and Figure 3-5 show scatter plots of model structure performances in percent difference compared to the best performing model (in each framework) against hydrologic signatures after the smoothing process described in section 2.3.2 has been applied. The left column also shows a threshold of 10% to visualize (as dashed horizontal line) which model structures are similar in their performance (i.e. having NSE difference less than 10% with respect to the best model structure which has the highest NSE value). Choosing 10% is a subjective decision (for visualisation purposes only) but clearer separations are observed between performances of model structures using this value in comparison with other thresholds we tried (i.e. 5%, 8%, 15%) as shown in Figure B6.1 (APPENDIX B). The panel bar plots in the right column of Figure 3-3 and Figure 3-5 indicate where model structures show NSE differences less than 10% compared to the best performing structure for selected attributes to better show which model structures stop performing well as a function of different signature values.

Figure 3-3 shows that there are clear separations between the performances of the six RRMT model structures. We find that the model structures containing a parallel flow routing module to represent fast and slow flows (i.e. PEN/PDM + 2PAR) and the model structures with the leaky flow routing module (i.e. PEN/PDM + LEAK) outperform other models in catchments with a high baseflow contribution (BFI> 0.7) (Figure 3-3a). Both

structures allow for slower responses and hence better baseflow representation. For small BFI values, all model structures have similar performances (i.e. NSE difference < 10%) which suggests that model structures with a single routing reservoir (i.e. PEN/PDM+CRES) are sufficient. Figure 3-3b shows that PEN/PDM + LEAK outperform other models in catchments with significantly negative delta-RR values (dRR < -0.2, indicating subsurface losses or large abstractions). Model structures with the leaky flow routing module perform best in catchments that lose water. Interestingly, only PDM + 2PAR outperforms other models in catchments that have high water gains (i.e. dRR > 0.2), suggesting that the flexibility of this model in runoff generation and routing is sufficient to capture this situation.

To explore these interactions in more detail, Figure 3-4 shows the relationship between BFI, dRR and model performance of PEN/PDM+2PAR and PEN/PDM+LEAK. We find that the majority (~66%) of catchments with high BFI values (i.e. BFI >0.7) have higher NSE values when using PEN/PDM+2PAR. The ones where PEN/PDM+LEAK outperforms PEN/PDM+2PAR have very negative dRR values (i.e. dRR < -0.2). This implies that there are some catchments that have both high BFI and very negative dRR values and that PEN/PDM+LEAK outperforms the other RRMT models in these catchments.

Similarly, all model structures have NSE difference < 10% for catchments with high runoff ratio values (RR > 0.6) except for the model structures with the PEN module that have NSE difference > 10% for catchments with RR>0.9 (Figure 3-3c). These catchments with RR>0.9 tend to gain significant amounts of water (i.e. dRR>0.4) due to water management activities (e.g. effluent returns, groundwater augmentation etc.) and this makes these catchments artificially wet. Simpler model structures based on the Penman Drying Curve (i.e. PEN+LEAK/CRES) fail here. The possible reason for this is that PDM module represents more flexibility through its distribution function even though it is likely for the wrong reasons. If these artificially wet catchments are ignored, all model structures perform

well in wet catchments. Therefore, the simplest model structures with a single conceptual reservoir for flow routing (i.e. PEN/PDM+CRES) is already suitable under those conditions. On the other end of the RR range, it is interesting that only PDM + LEAK shows sufficient flexibility, i.e. that both soil moisture accounting and routing have to be rather flexible.

Lastly, Figure 3-3d shows that all model structures except the simplest ones with a single flow routing reservoir (i.e. PEN/PDM+CRES) perform well in catchments showing very high streamflow variability. Larger streamflow variability correlates with larger slope of FDC (i.e. Slope of FDC>4). On the other end of this signature, only the PDM model with 2PAR and (to a lesser extent) LEAK seems to be able to capture the lack of streamflow variability (i.e. low slope of FDC values). It is interesting that this is not just a question of the routing function, but again requires a flexible runoff production function (i.e. PDM).

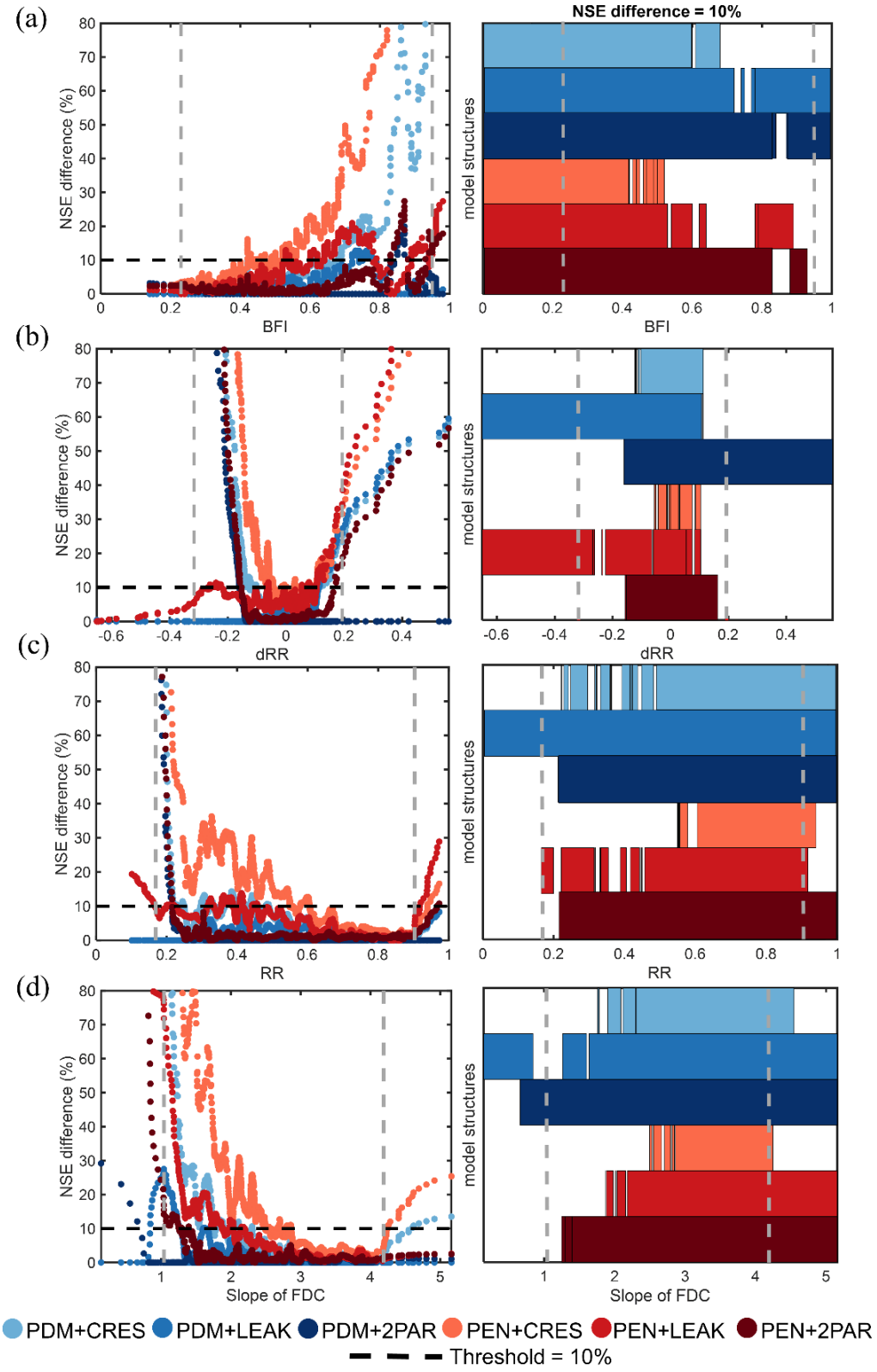


Figure 3-3. NSE difference (%) values and bar plots of six model structures (PEN+2PAR, PEN+LEAK, PEN+CRES, PDM+2PAR, PDM+LEAK, PDM+CRES) plotted against their BFI (a), dRR (b), RR (c) and slope of FDC (d) attributes. NSE difference values are calculated by taking the difference between maximum NSE value obtained by any model structure and NSE values of remaining model structures and divided by maximum NSE value and multiply by 100 for every catchment. NSE values of model structures are

obtained by moving means with 40 point - window size. Through visual inspection, 10% is selected as the most helpful threshold to show which model structure is performing differently in relation to a specific attribute. The range between two grey dashed vertical lines indicates the ranges where the smoothing is based on 20 left and right of the average calculated. Outside these ranges, points become increasingly biased by the points at the minimum and maximum signature values.

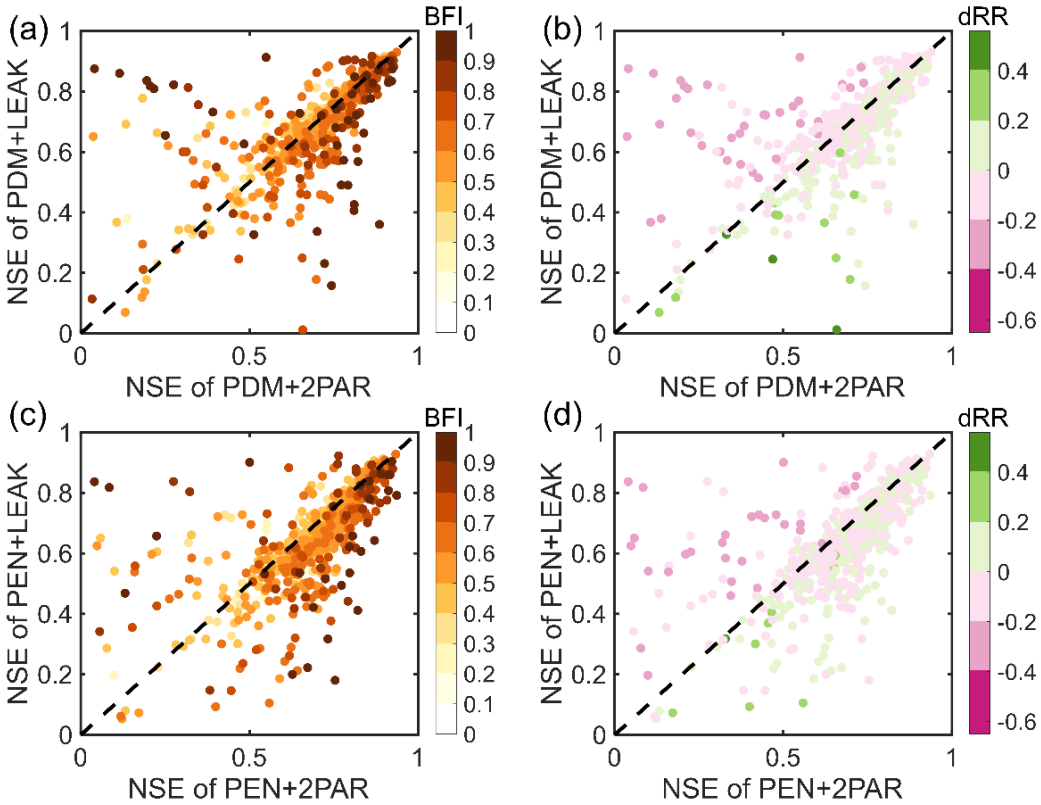


Figure 3-4. Scatter plots of NSE values of PEN+2PAR vs. PEN+LEAK color coded by BFI (a) and dRR (b) and PDM+2PAR vs. PDM+LEAK color coded by BFI (c) and dRR (d).

Figure 3-5 indicates that there are also some separations between the four FUSE model structures (TOPMODEL, ARNO/VIC, PRMS, SACRAMENTO) though not clearly related to hydrologic process differences. We find that ARNO/VIC performs well across all BFI values, and outperforms the rest in catchments with high BFI values (BFI > 0.7). It is difficult to explain why this model structure outperformed the other model structures in baseflow dominated catchments because all four models have a slow flow component as shown in Figure 3-1b. On the lower end of BFI values, all FUSE model structures are within 10% NSE difference range and there is therefore no significant difference (Figure 3-5a).

Interestingly, the performance of the ARNO/VIC model is quite robust across a wide range of catchment behaviours. It performs better or is sufficiently close (within 10%) to the best

performing model across the whole range of RR and slope of the FDC values (Figure 3-5c, Figure 3-5d). Also the TOPMODEL implementation works across all RR values, while the other two models work across a slightly narrower range of RR values only. We know that TOPMODEL can produce simulations with less bias from Lane et al.'s study (2019), but the reason for its performance advantage is unclear. It seems that there is no specific model structure, except ARNO/VIC, that outperforms the others in catchments with very high RR values. This is again due to the artificially wet catchments we discussed above (Figure 3-3c). Without those catchments, all FUSE model structures are within 10% NSE difference range. The SACRAMENTO and PRMS models struggle if the water balance deviates more than about 20% from the one we expect using climate only (i.e., dRR values of +/-0.2) (Figure 3-5b). The models are thus quite sensitive to water balance problems. ARNO/VIC and TOPMODEL are robust in this regard, though they do so for negative and positive dRR values respectively. And finally, ARNO/VIC, TOPMODEL, SACRAMENTO and PRMS work increasingly poorly in this order when it comes to fitting flow variability as expressed through the Slope of the FDC (Figure 3-5d). All model structures except PRMS perform well in catchments with high slope of FDC values (Figure 3-5d).

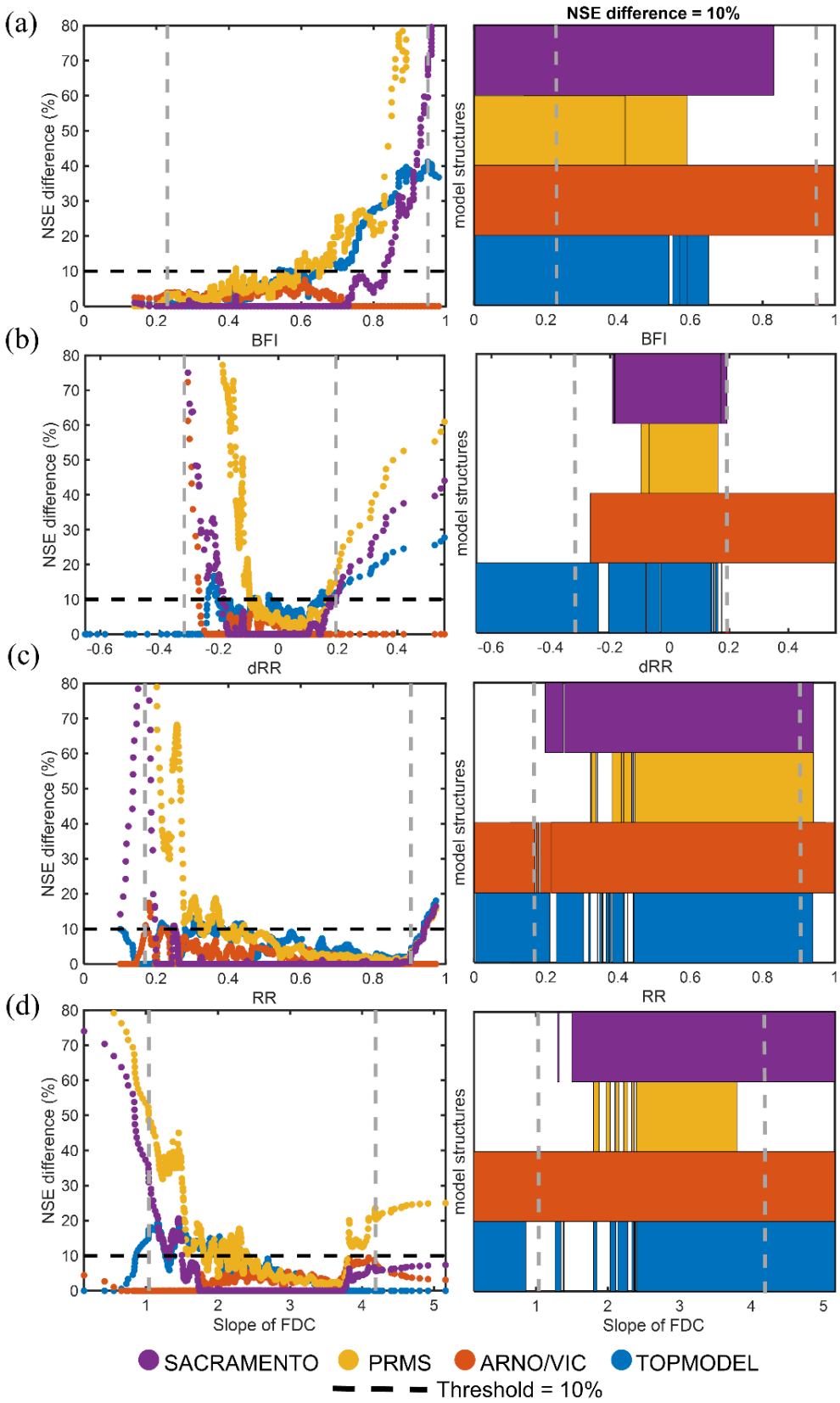


Figure 3-5. NSE difference (%) values and bar plots of four model structures (TOPMODEL, ARNO/VIC, PRMS, SACRAMENTO) plotted against their BFI (a), dRR (b), RR (c) and slope of FDC (d) attributes. NSE difference values are calculated by taking the difference between maximum NSE value obtained by any model structure and NSE values of remaining model structures and divided by maximum NSE value and multiply by 100 for

every catchment. NSE values of model structures are obtained by moving means with 40 point - window size. Through visual inspection, 10% is selected as the most helpful threshold to show which model structure is performing differently in relation to a specific attribute. Therefore, bar plots of four model structures are created by taking 10% as the NSE difference. The range between two grey dashed vertical lines indicates the ranges where the smoothing is based on 20 left and right of the average calculated. Outside these ranges, points become increasingly biased by the points at the minimum and maximum signature values.

To ensure the robustness of our results to different performance metrics, we recreate Figure 3-3 and Figure 3-5 using KGE differences (Figure B7.1. and S7.2. in APPENDIX B). When we compare the results of NSE and KGE difference for the model structures (using the best NSE model), we find some differences between performance separations of both RRMT and FUSE model structures. For example, PEN+CRES seems to perform better for lower RR values when KGE is used, whereas PEN+2PAR seems to do worse in this region compared to using NSE. Moreover, while PEN/PDM+2PAR seem to outperform PEN/PDM+CRES in catchments with high slope of FDC values based on NSE values, this is not the case based on KGE values. When we look at FUSE model structures, we also observe some differences in their performance separations. For instance, ARNO/VIC does not outperform PRMS when KGE is used rather than NSE in high slope of FDC values (i.e. >4). Moreover, TOPMODEL and PRMS are within the 10% threshold in the range of 0<RR<0.2 and 0.4<RR<0.6, respectively, based on NSE, whereas this is not the case based for KGE values. These findings imply that using KGE instead of NSE makes some difference in the performance separation of model structures with respect to the signatures assessed. However, when checking the signature ranges that define specific catchment types (i.e. baseflow-dominated, leaky, wet), only PEN+2PAR (RRMT) and SACRAMENTO (FUSE) show different performance separations when using KGE instead of NSE in baseflow-dominated catchments (i.e. BFI >0.7). The other model structures from both frameworks show the same separation in all catchment types when using KGE or NSE. This result suggests that there is still some more to learn about the differences in assessing model performances between KGE and NSE, which is beyond this short technical note.

A final question is whether we can predict the hydrologic signatures used in this study (i.e. dRR, RR, BFI and slope of the FDC), so we could apply what we've learned to ungauged catchments. In the GB setting, BFI has been predicted from physical catchment properties in the BFI-HOST framework (Marsh and Hannaford, 2008). These BFI-HOST values indicate strong correlation with the BFI values that we use in our study as shown in Figure B8.1a while RR shows a strong dependence on AI as shown in Figure B8.1b (APPENDIX B). However, we could not identify a single physical attribute or a reasonable combination of attributes to predict dRR for ungauged catchments given that different physical properties and anthropogenic activities likely influence this deviation. Runoff of leaky (dRR< -0.2) and gaining (dRR > 0.2) catchments is affected by both geological differences and different water management practices such as abstractions, reservoirs, and effluent returns (see Figure B9.1.in APPENDIX B). However, the net effects of such practices across GB catchments have not been assessed so far.

3.4 Discussion

We compare two modular modelling frameworks to analyse the influence of priori model structure selection on performance separation in relation to catchment types across Great Britain. In a direct comparison of model performances, we find that the FUSE structures perform slightly better with respect to the NSE metric when this metric is larger than 0.5 for both frameworks (the result is the inverse for values below). It is generally not surprising that FUSE is slightly better given that its models have between 10-12 free parameters, while RRMT has between 3 and 7. Multiple studies found a link between model performance and the number of free calibration parameters (e.g. Perrin et al., 2001; Kollat et al., 2012; Höge et al., 2018). However, we also show that it is not just the number of parameters that matters for model performance, as for example found by Knoben et al. (2020), as models which include the leaky routing structure of RRMT work better than the FUSE structures in catchments with significant subsurface losses – even though they have

fewer parameters. Interestingly, the performance difference between the structures goes away when using the KGE metric. We have no straightforward explanation for this finding.

Figure 3-6 is a visual summary of what we find across the GB catchments studied here. There are 139 (14%), 62 (6%) and 391 (40%) catchments with BFI>0.7, dRR<-0.2 and RR>0.6, respectively. Slope of FDC did not provide additional information about separations between model structures because the flatter slopes are also the catchments that generally have higher BFI values. There is therefore a large mirroring of the BFI and FDC results which does not justify including both results Figure 3-6a shows that six model structures from RRMT are distinguished from each other across catchment types in line with our expectations regarding hydrological differences. In comparison, Figure 3-6b indicates that some of model structures from FUSE also outperform the others in some of the catchment types but it is challenging to explain why they differ (as was previously concluded by Lane et al., 2019). The reason is that there are no identifiable structural/behavioral differences which explain performance differences between these model structures. The six model structures chosen in the RRMT framework have evolved from experience in modelling diverse GB catchments (Moore, 2007; Lee et al., 2005; Wagener et al., 2004). Our results suggest that these model structures emerge as more suitable for specific catchment types, though we also find that they do not necessarily provide better performance than other model structures (except in the case of catchments with significant groundwater losses).

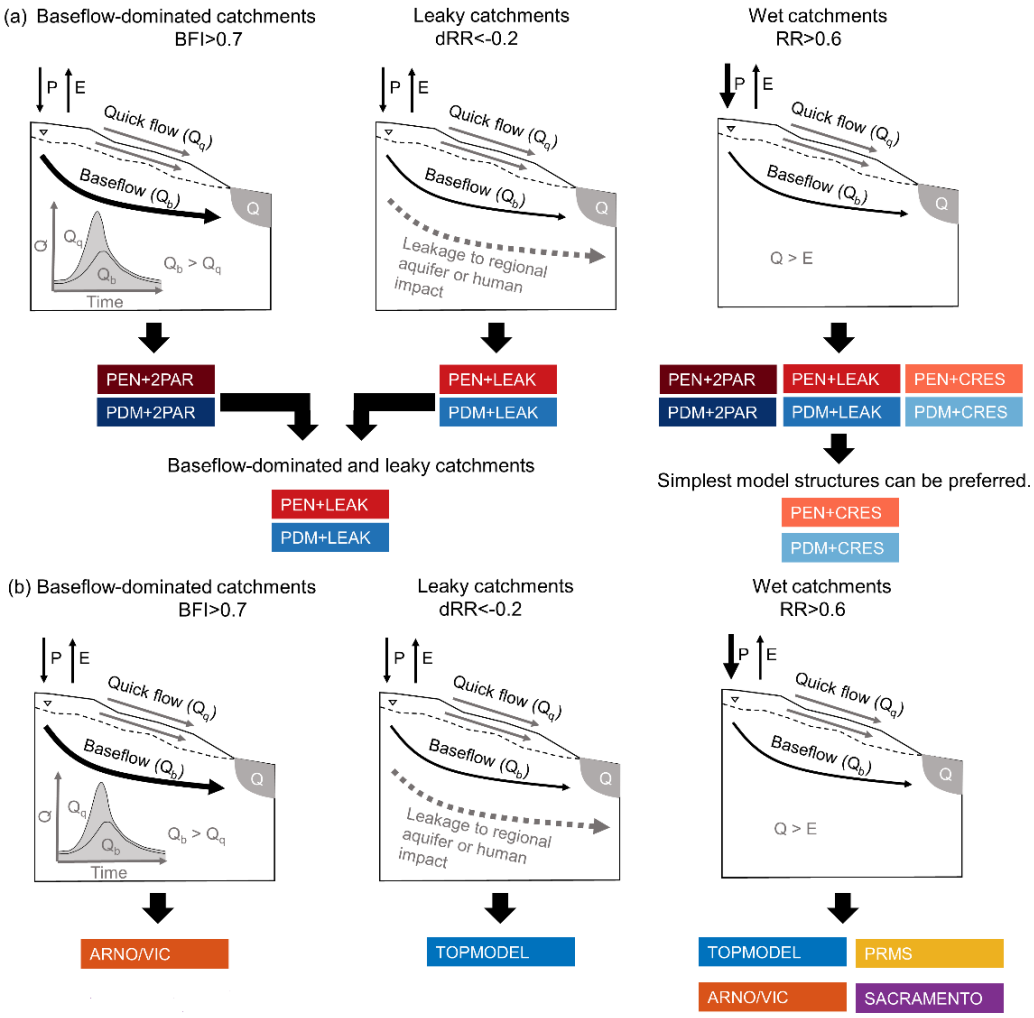


Figure 3-6. Illustration of (a) six model structures' separation (PEN+2PAR, PEN+LEAK, PEN+CRES, PDM+2PAR, PDM+LEAK, PDM+CRES) and (b) four model structures' separation (TOPMODEL, ARNO/VIC, PRMS, SACRAMENTO) for the catchments with different characteristics. Baseflow-dominated catchments are the ones containing a higher proportion of the river that derives from stored sources (i.e. having high BFI values). Leaky catchments are the ones most likely losing water (i.e. having very low negative dRR values). Wet catchments are the ones where the rainfall is most likely to become runoff (i.e. having high RR).

Some of these catchment types have also been found to produce distinguishable model performances elsewhere. Kavetski and Fenicia (2011) and David et al. (2022) also found baseflow dominated catchments to require routing structures with parallel reservoirs. Kavetski and Fenicia (2011) selected seven model structures from the SUPERFLEX framework and the fixed GR4H model and tested them on four catchments from New Zealand and Luxembourg. David et al. (2022) selected only four model structures also from SUPERFLEX and evaluated them across 508 Brazil catchments. Both studies selected model structures based on their prior knowledge and experience in their study domain.

Similarly, different studies found that wet catchments can be modelled well using a wide range of model structures (e.g. Atkinson et al., 2002; Kavetski and Fenicia, 2011; Coxon et al., 2014; Massmann, 2020; David et al., 2022). More specifically to GB, our findings are similar to Lee et al. (2005) who also found that a leaky routing component is needed in catchments with permeable aquifers across GB such as Chalk, Jurassic limestone, and Carboniferous/Devonian rock.

Nonetheless, some studies (e.g. Lee et al., 2005; Van Esse et al., 2013; Lane et al., 2019; Knoben et al., 2020) which are conducted in different countries (e.g. UK, France, US) and used model structures from different modular frameworks (e.g. RRMT, SUPERFLEX, FUSE, MaRRMoT) have not been able to identify clear model structure-catchment type relationships (beyond the aforementioned permeable catchments in the case of Lee et al.). Both Lee et al. (2005) who used 12 model structures from RRMT across 28 UK catchments, and Van Esse et al. (2013) who used 12 model structures from SUPERFLEX plus GR4H model across 237 French catchments, observed performances differences between the model structures that they used, but they could not establish a catchment type-model structure relationship. Both studies suggested that the catchment characteristics used were insufficient to reflect catchments' hydrological behaviors. Lee et al. (2005) stated some additional possible reasons for this such as the other choices made in their study (e.g. number of catchments, suitability criteria) and using observed rainfall-runoff data which is insufficient to represent the catchments. In addition, studies by Lane et al. (2019) who used 4 model structures from FUSE across 1013 GB and Knoben et al. (2020) who used 36 model structures from MARRMoT across 559 US catchments could not observe distinct separations between their model performances across catchment types due to selection of multiple model structures with similar process representations or complexities.

Our findings suggest that modular modelling frameworks might benefit from an adequate strategy for the inclusion of specific model structures, process modules or system components in their frameworks (tailored to a specific domain). It might be beneficial for

them to explicitly provide the conceptual differences and similarities between the process modules or components of model structures and to establish expectations regarding the type of catchments that they can potentially represent well or poorly. If these differences are unclear a priori, then it is unlikely that we can subsequently explain model performance differences. While some modular modelling frameworks such as SUPERFLEX (Fenicia et al., 2011) and MARRMoT (Knoben et al., 2019) provide detailed information about the differences/similarities between components/fluxes of model structures included and the hydrological processes that they can represent, this might not be enough. Knoben et al. (2020) investigated model suitability by pre-selecting 36 of 46 MARRMoT model structures for 559 US catchments. They ranked the model structures according to their performance in each catchment and then attempted to correlate these rankings with 52 catchment attributes (e.g. hydrologic, climatic and physical). However, the authors could not find clear relationships between model rankings and catchment attributes. The study stated that not using suitable hydrological signatures/catchment attributes to reflect distinct hydrologic behaviors across their study domains could possibly be a reason. Our results suggest that a stronger focus on pre-selecting model structures consisting of (as much as possible) distinct process-based components for the study domain might be a way forward to reduce this problem.

3.5 Summary

Modular modelling structures are widely popular although the best approach for selecting model structural components has remained unclear. Probably unsurprisingly, many studies have found it difficult to find meaningful separations between the model structures or structural components considered. Here we hypothesise that the long-term experience within a study domain (e.g. a region such as GB) can lead to the development of different model structures which provide a guide to a priori model inclusion. While rainfall-runoff models have often not explicitly evolved into modular frameworks, they nonetheless can contain at least some of the experiences made when trying to simulate diverse catchments

across a heterogeneous domain (e.g. Moore, 2007). We therefore use GB experience as a guide in our study.

Applying model structures selected in this manner, we find that these a priori chosen model structures more logically separate regarding their performance across catchments than those used in a previous multi-model study with non-UK focused model structures (Lane et al., 2019). The routing components of our framework separate based on the extent of baseflow contribution into single or parallel flow components, while a leaky component is required for catchments with significant subsurface losses. The two soil moisture accounting components do not separate as strongly, unless significant flexibility is required in which case the PDM structure is favoured (e.g. wetter catchments than expected based on climate alone).

Our results suggest that it might be helpful to first build perceptual models of the diverse catchments (or systems) encountered across a study domain such as Great Britain (e.g. Beven and Chappell, 2021; Wagener et al., 2021; McMillan et al., 2023). Here we conditioned our perceptions on previous experiences with different model structures applied across our study domain. Without consideration of different perceptual models which are reflected in the model structures included, the modular modelling exercise might reduce to a regression type analysis with limited knowledge gain.

Chapter 4 A Signature-based Hydrologic Efficiency Metric for Model Calibration and Evaluation in Gauged and Ungauged Catchments

This chapter has been submitted to Water Resources Research and has undergone slight modifications to align with the general layout of this thesis. The study was conceptualized by Melike Kiraz, Gemma Coxon, and Thorsten Wagener. Melike Kiraz conducted the data processing, model simulations and creation of figures under the guidance of Gemma Coxon and Thorsten Wagener. The manuscript was primarily written by Melike Kiraz, with input and comments from all co-authors.

Citation: Kiraz, M., Coxon, G. and Wagener, T. (2023). A signature-based hydrologic efficiency metric for model calibration and evaluation in gauged and ungauged catchments. Water Resources Research. (Under review).

4.1 Introduction

Statistical objective functions are widely used to quantify the difference between observed and simulated streamflow time series for rainfall-runoff model evaluation and calibration in situations where historical streamflow observations are available. Such objective functions integrate the differences between observed and simulated time series, i.e. the residuals. Many metrics are based on the mean squared error (MSE) which can be derived from basic statistical assumptions about the errors present (Gershenfeld, 1999). In hydrology, Nash and Sutcliffe (1970) suggested that this metric should be normalized to allow for a better comparison of model performances across catchments. Their unit-free objective function has become well known as the Nash Sutcliffe Efficiency (NSE).

Multiple authors subsequently pointed out that metrics based on MSE type assumptions can be broken up into several constituent components, i.e. bias, standard deviation and correlation (Murphy, 1988; Weglarczyk, 1998). However, these components are not equally weighted within the traditional NSE formulation. Gupta et al. (2009) therefore suggested to combine them using Euclidean distance, which weights them equally in their Kling Gupta Efficiency (KGE) (see also Kling et al., 2012). This KGE metric has been used widely since its introduction and some authors have suggested improvements. For example, Pool et al. (2018) proposed to make the constituent components non-parametric so that they are less dependent on underlying assumptions. They replaced Pearson's linear correlation with Spearman rank correlation, and they assessed discharge variability using a normalized flow duration curve (FDC) to remove volume information and retain information about distributions only.

These metrics are undoubtedly cornerstones of hydrologic modelling, but some underlying problems with their use have been the basis for an ongoing debate. First, it is difficult to interpret them and their constituent components hydrologically (Gupta et al., 2008). For example, what is hydrologically wrong with my model if the NSE value is only 0.5? This problem has led to the use of hydrologic signatures in model evaluation (e.g. Moges et al., 2022). Such signatures are indices of hydrologic function, such as the runoff ratio, which is an index that quantifies the fraction of precipitation that leaves the catchment as streamflow rather than evapotranspiration (McMillan, 2021). Second, the use of hard performance thresholds, though promoted by some (e.g. Moriasi et al., 2007; Rogelis et al., 2016; Towner et al., 2019), has been heavily criticized by others (e.g. Knoben et al., 2019; Clark et al., 2021). Flexible performance benchmarks have also been suggested to overcome this problem (e.g Seibert, 2001; Schaefli and Gupta, 2007; Seibert et al., 2018), while a more diagnostic evaluation of the underlying components has been proposed by others (Schwemmle et al., 2021).

Metrics like NSE and KGE are only applicable to gauged catchments because they require historical time series of observed streamflow to estimate residuals. However, previous studies have regionalized hydrologic signatures (e.g. Yadav et al., 2007; Hrachowitz et al., 2014; Pool and Seibert, 2021; Guo et al., 2021), and the statistical hydrology literature is rich with examples where streamflow statistics have been regionalized (e.g. Vogel et al., 1999). Therefore, at least some of the components that make up efficiency metrics, i.e., bias and variance, have already been estimated in ungauged basins. Indeed, there have been quite a few studies that have used (uncertain) regionalized hydrologic signatures as constraints for rainfall-runoff model ensembles (e.g. Zhang et al., 2008; Bulygina et al., 2009; Westerberg et al., 2011). However, there has been no attempt so far to build an efficiency metric for ungauged basins from these components.

In this chapter, we propose a signature-based hydrologic efficiency metric that builds upon the work that has been done previously with signatures in both gauged and ungauged catchments. Integration of hydrologic signatures in an evaluation metric will provide opportunity for hydrologic interpretation of model performance and being able to regionalize these signatures will provide hydrologic efficiency evaluation of models for ungauged catchments. We test our ideas across 633 catchments in Great Britain (GB) by using model simulations in a Monte Carlo framework for a 10-year time period.

4.2 Data

In this chapter, we analyse 633 catchments spread across Great Britain. Details about the general characteristics (e.g. climatic, topographic, geologic) of Great Britain are provided in Section 2.2. of Chapter 2. This study uses daily rainfall, streamflow, potential evapotranspiration time series for ten years (October 1, 1999 – September 30, 2009) and catchment attributes from the CAMELS-GB dataset to develop and demonstrate the new metric. More information regarding the CAMELS-GB dataset is available in Section 2.2. of Chapter 2. From the 671 CAMELS-GB catchments, we exclude 12 catchments from the analysis where (1) the runoff ratio or variance ratio value is higher than 1 – suggesting

significant and unexplained water balance issues, (2) there is no available BFI-HOST data or (3) there is insufficient streamflow data for the specified study years. In addition, we also exclude 26 catchments where water balance analysis (see Section C3 in APPENDIX C) shows that they are significantly losing water most likely through subsurface processes which is not captured by the hydrological model used in this study. Hence, 633 GB catchments are used in the subsequent analysis.

4.3 Methods

4.3.1 A Signature-based Hydrologic Efficiency (SHE) metric

We follow previous work discussed in the introduction section by adding a particular focus on signatures representing different hydrological dynamics as the individual components underlying hydrological efficiency metrics, as well as our ability to regionalize them (see Table 4-1).

4.3.1.1 Bias term: Runoff ratio

Runoff ratio (RR) is defined as the ratio of long-term average streamflow to long-term average precipitation. It is the long-term water balance separation between water being released from the catchment as streamflow and as evapotranspiration (Milly, 1994; Sankarasubramanian et al., 2001; Olden and Poff, 2003; Yadav, 2007). Higher runoff ratios identify catchments where a large amount of water leaves the catchment as streamflow with respect to precipitation and vice versa.

4.3.1.2 Variance (i.e. amplitude) term: Variance ratio

We define variance ratio as the ratio of standard deviation of streamflow to standard deviation of precipitation. The signature shows how variable (i.e. flashy) streamflow is with respect to precipitation drivers and is as such an indicator of the damping of precipitation variability through the catchment (a lower value indicating more damping).

4.3.1.3 Correlation term

Correlation is an aspect that is more difficult to capture in a signature. It could be represented as a function of the catchment response in relation to precipitation using the time of concentration of a catchment. However, estimates of time of concentration using the daily data we use in this study do not work very well for small and fast responding catchments in Great Britain (Giani et al., 2021). While exploration of this signature is beyond this technical note, we will return to the issue when we discussed ungauged basins. For now, we decided to use Spearman rank correlation between observed and simulated streamflow values as the correlation term of SHE like the non-parametric form of KGE developed by Pool et al. (2018). The components and formulation of SHE for gauged cases (i.e. SHE_g) are given in Table 4-1.

Table 4-1. Bias, variance and correlation components and formulations of evaluation metrics.

Objective function	Bias (β)	Variance (α)	Corre-lation (r)	Combination
NSE (Nash and Sutcliffe, 1970; Gupta et al., 2009)	$\frac{(\mu_S - \mu_O)}{\sigma_O}$	$\frac{(\sigma_S)}{(\sigma_O)}$	$r_{pearson}$	$2 * \alpha * r - \alpha^2 - \beta^2$
KGE (Gupta et al., 2009)	$\frac{(\mu_S)}{(\mu_O)}$			$1 - \sqrt{(\alpha - 1)^2 + (\beta - 1)^2 + (r - 1)^2}$
KGE* (modified version in Kling et al., 2012)		$\frac{[(\sigma_S)/(\mu_S)]}{[(\sigma_O)/(\mu_O)]}$		
NP (Pool et al., 2018)		$1 - ½ \sum_{i=1}^{n} \left\lvert \frac{x_{S,\,I(i)}}{n\mu_s} - \frac{x_{O,\,J(i)}}{n\mu_o} \right\rvert$	$r_{spearman}$	
SHE$_g$ (gauged situation)	$\frac{[(\mu_S)/(\mu_P)]}{[(\mu_O)/(\mu_P)]}$	$\frac{[(\sigma_S)/(\sigma_P)]}{[(\sigma_O)/(\sigma_P)]}$		
SHE$_u$ (ungauged situation with regionalized signatures)	$\frac{[(\mu_S)/(\mu_P)]}{[RR_{Pred}]}$	$\frac{[(\sigma_S)/(\sigma_P)]}{[VR_{Pred}]}$	$r^*_{spearman}$	

- S, O and P are simulated streamflow, and observed streamflow and precipitation, respectively.
- μ is the mean and σ is the standard deviation of streamflow.
- $x_{S,\,I(i)}$ is the simulated streamflow value where I(i) is the time step when the ith largest flow occurs within simulated time series and $x_{O,\,J(i)}$ is the observed streamflow value of target catchment where J(i) is the time step when the ith largest flow occurs within observed time series.
- VR_{Pred} and RR_{Pred} are regionalized variance ratio and runoff ratio for the target catchment derived using stepwise linear regression. Predictors of VR_{Pred} are aridity index, BFI-HOST and inland water percentage. Predictor of RR_{Pred} is only aridity index. Variance ratio is the ratio of standard deviation of streamflow to standard deviation of precipitation. Runoff ratio is the ratio of long-term mean of streamflow to long-term mean of precipitation.
- $r_{Pearson}$ = Pearson correlation between simulated and the observed streamflow in the target catchment
- $r_{Spearman}$ = Spearman rank correlation between simulated and the observed streamflow in the target catchment
- $r'_{Spearman}$= Spearman rank correlation between simulated streamflow of a catchment which is assumed to be ungauged and the streamflow values obtained by inverse distance weighting interpolation of this catchment's three closest catchments' observed streamflow.

4.3.2 Application of SHE metric in ungauged catchments

Applying the SHE metric in ungauged situations requires estimates all of three metric components for ungauged basins. We perform this regionalization step in two different ways. Bias and variance components, i.e. runoff ratio and variance ratio, or related signatures have been widely regionalized using different types of regressions (e.g. Yadav et al., 2007; for GB). We use the simplest and widely used strategy, stepwise linear regression, to establish the relationships between the catchment attributes and signatures (e.g. Almeida et al., 2016). In the MATLAB environment, we utilized the "stepwiselm" function to perform stepwise linear regression. The stepwise linear regression analysis function utilizes a combination of forward and backward stepwise regression techniques to derive the final model. Initially, it starts with a simple constant model, containing only the intercept term and no other variables. In each step of the process, the function examines whether to add or remove terms from the model based on a specified criterion (e.g., the p-value < 0.05 in this case). When a term is not yet included in the model, the null hypothesis assumes that the term's coefficient would be zero if added. The function tests this hypothesis and adds the term to the model if there is sufficient evidence to reject the null hypothesis, indicating its significance in improving the model fit. Conversely, if a term is already present in the model, the null hypothesis suggests that the term's coefficient is zero. The function then examines if there is enough evidence to reject this null hypothesis. If there is insufficient evidence, the term is removed from the model, as it does not significantly contribute to the model's performance. The p value (probability value) is a measure that helps determining the strength of evidence against a null hypothesis. It is calculated based on the observed data and the assumption that the null hypothesis is true. It represents the probability of obtaining the observed data if the null hypothesis is true. A low p-value suggest that the observed data is unlikely to have occurred under the assumption of the null hypothesis. A high p-value suggests that the observed data is

consistent with the null hypothesis. In this study, p-values are automatically produced when using “stepwiselm” function to perform stepwise linear regression.

We regionalize runoff ratio (RR) and variance ratio (VR) signatures for 633 GB catchments testing 64 catchment attributes from CAMELS-GB representing topography, climate, hydrology, land cover, soils, hydrogeology and human influences (see Table A1 in APPENDIX A). When the stepwise linear regression is performed using these catchment attributes, the regression equations producing the best estimations (i.e. having highest r2 values) for RR and VR consist of 10 and 19 predictors, respectively. To obtain simpler regression equations but still reasonable RR and VR estimations, we have tried other rounds of stepwise linear regression using multiple smaller groups of catchment attributes (see Table C1.1. in APPENDIX C). Based on stepwise linear regression analyses made using these groups of attributes, aridity index (AI) is selected as predictor of RR. AI, baseflow index (BFI-HOST) and inland water percentage (inwater_perc) are selected as predictors of VR. It is reasonable that RR and AI are highly correlated because both are influenced by similar climatic and hydrologic conditions related to the availability of water. Similarly, it is sensible that having BFI-HOST and inwater_perc in addition to aridity index as predictors of variance ratio because there are some damping effects due to subsurface storage and inland water storage. After selecting the predictors, 633 GB catchments are randomly divided into 5 groups. One group is left out each time and the remaining ones are used in the fitting of regression models for each signature (5-fold cross-validation). After obtaining regression models, the signature values are estimated for the catchments in omitted group each time. The regression equations, their r^2 and p values for each catchment group is listed for both RR and VR in Table 4-2 and Table 4-3, respectively.

Table 4-2. Regression equations, r^2 and p values for 5 catchment groups for runoff ratio (RR)

Catchment Group	Equation	r^2 value	p value
1	RR=0.98-0.79*AI	0.8	<< 0.05
2	RR =1-0.86* AI	0.79	<< 0.05
3	RR =0.99-0.79* AI	0.84	<< 0.05
4	RR y=0.97-0.79* AI	0.78	<< 0.05
5	RR =0.99-0.81* AI	0.78	<< 0.05

Table 4-3. Regression equations, r^2 and p values for 5 catchment groups for variance ratio (VR)

Catchment Group	Equation	r^2 value	p value
1	$VR=e^{0.21-0.07*inwater\_perc-2.4*AI*BFI-HOST}$	0.86	<< 0.05
2	$VR=e^{0.20-0.08*inwater\_perc-2.6*AI*BFI-HOST}$	0.87	<< 0.05
3	$VR=e^{0.21-0.08*AI*inwater\_perc-2.5*AI*BFI-HOST}$	0.87	<< 0.05
4	$VR=e^{0.03-0.09*AI*inwater\_perc-3*AI*BFI-HOST}$	0.88	<< 0.05
5	$VR=e^{0.34-0.08*inwater\_perc-2.1*AI*BFI-HOST}$	0.88	<< 0.05

The correlation term is more complicated; given that we have no simple approach to regionalize a single value as is the case with the other two signatures. However, Archfield and Vogel (2010) have demonstrated that it is feasible to estimate correlation for ungauged locations using a geostatistical strategy. They introduced their map correlation method which selects the strongest correlated gauge as the reference gauge for an ungauged catchment, given that the nearest gauge was not always the most correlated one in their study of US catchments. The approach by Archfield and Vogel (2010) follows the basic idea of directly transferring streamflow from gauged to ungauged locations (see wider review of such approaches by He et al., 2011). Drogue and Plasse (2014) tested four different distance-based regionalization methods including the strategy by Archfield and Vogel (2010) for European catchments. They found that using multiple reference catchments rather than one is preferrable for assessing daily streamflow hydrographs in a densely gauged study domain. The simplest strategy to directly transfer streamflow is likely the one by Patil and Stieglitz (2012), who used inverse distance weighted (IDW)

interpolation to transfer daily streamflow from multiple neighbouring gauged catchments to ungauged catchments in the US. Their approach is formulated as follows:

$$q(x) = \sum_{k=1}^{N} \frac{w_k(x)}{\sum_{k=1}^{N} w_k(x)} * q(x_k) \qquad (4\text{-}1)$$

$$and\ w_k(x) = \frac{1}{d(x,x_k)^p} \qquad (4\text{-}2)$$

where q(x) is daily streamflow (mm/day) at the ungauged catchment that is located at point x in the region, $q(x_k)$ is the daily streamflow of neighbouring reference catchment k located at point x_k in the region and N is the total number of neighbouring reference catchments for the interpolation. d is the distance between gauges of catchments and w is the interpolation weights of reference catchments. The exponent p is a positive real number, called a power parameter.

We adopt this approach for estimating streamflow to ungauged locations within our GB dataset because it works surprisingly well and because optimizing the regionalization performance is not our main concern. To identify a suitable number of reference catchments, we assume each catchment in turn to be ungauged, estimate the streamflow time series using IDW interpolation with different numbers of reference catchments (1-5 reference catchments), and calculate the Spearman Rank Correlation (SRC) between transferred and observed streamflow time series. We find that using three reference catchments provides optimum SRC estimate for the ungauged catchments in our sample (Figure C2.1 in APPENDIX C). We could actually use a similar streamflow transfer strategy to estimate the bias and variance terms but found this strategy to perform less well (see Figure C2.2 in APPENDIX C).

4.3.3 Rainfall-Runoff Model Implementation

We use a typical lumped parsimonious model structure widely used in Great Britain. The model structure, implemented in the Rainfall-Runoff Modelling Toolbox (RRMT; Wagener et al., 2001a) combines a probability-distributed soil moisture accounting component (i.e. PDM), which represents the variability in soil moisture storage across a

typical humid catchment using a distribution of storage depths (Moore, 2007), and a combination of two linear reservoirs in parallel for routing, one representing fast flow and the other representing slow flow (i.e. 2PAR), with a fixed split between them. Effective rainfall is produced as overflow from the PDM stores which are described as Pareto distribution based on two parameters, the maximum storage capacity, C_{max}, and parameter, b, describing the shape of the distribution. The effective rainfall (ER) is split with respect to parameter *a* describing the fraction of flow through the fast reservoir, while both reservoirs are defined by a single time constant (Wagener et al., 2001a). The reason of choosing PDM is that it represents a flexibility in soil moisture accounting through its distribution function to influence the runoff response and combining it with 2PAR flow routing module provides different flow pathways for catchments across GB with different levels of baseflow contribution.

To calibrate the model, 10,000 parameter sets are independently sampled using uniform random sampling. The first 5% of the ten-year study period is used as a warm-up period. The parameter set producing the best performance according to SHE metric is used to obtain simulated streamflow. These numbers have been widely used in previous studies.

4.4 Results

First, we compare the values estimated for our SHE metric in gauged situations with previous efficiency metric implementations, i.e. KGE (Kling et al., 2012), NSE (Gupta et al., 2009) and NP (Pool et al., 2018). Figure 4-1 shows scatter plots where SHE values are correlated with KGE, NSE and NP values with Pearson correlation (i.e. PC) and Spearman rank correlation (i.e. SRC) values to varying degrees. Correlations are highest for SHE-NP (above 0.8), then SHE-KGE (around 0.8) and then SHE-NSE (0.6 to 0.67). Our formulation is most closely related to that of Pool et al. (2018) and Gupta et al (2009) due to the equal weighting of the terms within the efficiency metric.

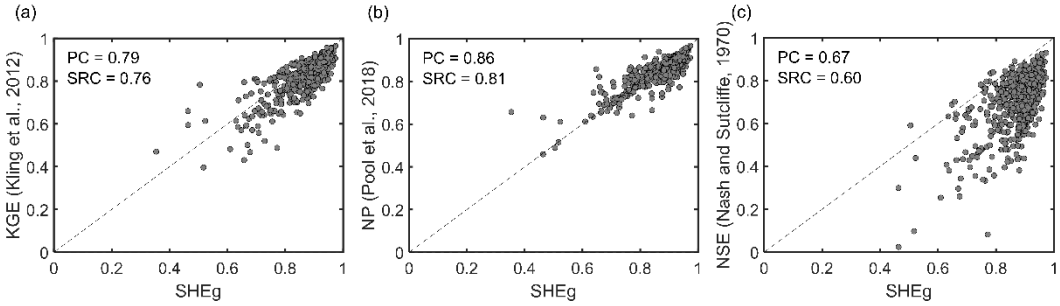


Figure 4-1. Scatter plots for (a) KGE vs. SHE, (b) NP vs. SHE and (c) NSE vs. SHE. x and y axes are limited to [0 1]. KGE, NP and NSE values are calculated using the best simulation values based on SHE metric values.

Second, we estimate the components of our metric for ungauged locations. The scatter plots in Figure 4-2a and Figure 4-2b show that the predicted RR and VR using stepwise linear regression correlate well with observed RR and VR values. We find PC and SRC correlation values above 0.9. The maps indicate that predicted RR and VR values have similar patterns with decreases from the north-west to south-east of GB. As shown in Figure 4-2c for an estimate of correlation for ungauged locations, SRC values between observed and transferred streamflow values are above 0.8 for 94% of all catchments (77% above 0.9), even when using the simple inverse distance method with the three closest catchments. All components of our SHE metric can therefore be estimated individually in ungauged catchments within our study domain.

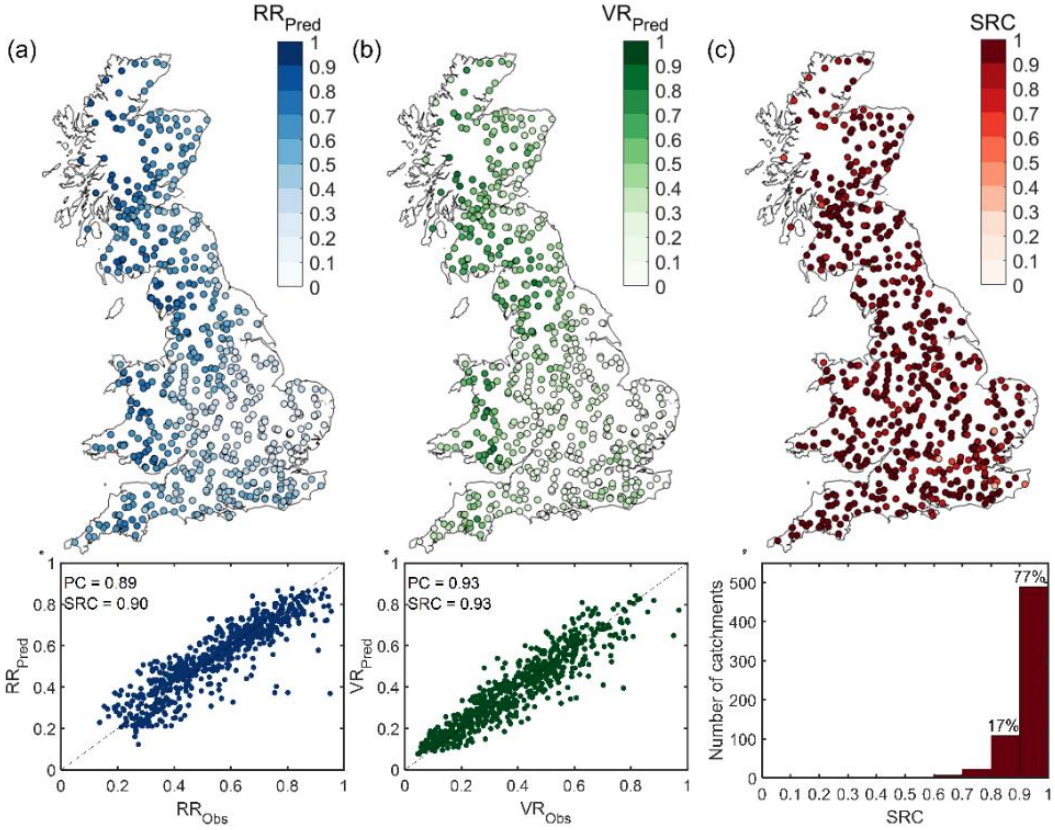


Figure 4-2. (a) Predicted RR map and scatter plot for predicted vs. observed RR, (b) predicted VR map and scatter plot for predicted vs. observed VR and (c) map illustrating SRC values between observed streamflow of catchments and the streamflow values calculated by taking inverse distance interpolation of their closest three catchments' observed streamflows and its histogram plot. Predictor of RR is aridity index and predictors of VR are aridity index, BFI-HOST and inland water percentage.

And third, we calculate the differences between SHE values for gauged and ungauged cases to evaluate how well we can estimate the performance of a model for ungauged catchments, in contrast to gauged catchments. Figure 4-3a, Figure 4-3b and Figure 4-3c shows histograms of the differences between SHE values for gauged and ungauged cases (i.e. $SHE_g - SHE_u$). Cumulative distribution functions (CDF) plots of the individual difference values are color-coded by (a) bias component difference (i.e. $\Delta\beta$), (b) variance component difference (i.e. $\Delta\alpha$) and (c) correlation component difference (i.e. Δr). The histograms (all three are identical) show that more than 50% of 633 catchments have difference values between -0.1 and 0.1, while 78% of them have difference values between -0.2 and 0.2. Low values of SHE difference are associated with small differences in the bias, variance, and correlation terms (see CDF plots in Figure 4-3). CDF plots also show that catchments with high positive differences (i.e. >0.3) have the highest positive and the lowest negative values

of the bias and variance component differences, respectively, suggesting the poor regionalization is a problem there. Figure 4-3c shows that correlation component differences are overall very small across catchments except for very few catchments with high positive differences. In summary, the results imply that when the regionalization of the bias and variance signatures works, we can obtain similar SHE values for both gauged and ungauged cases.

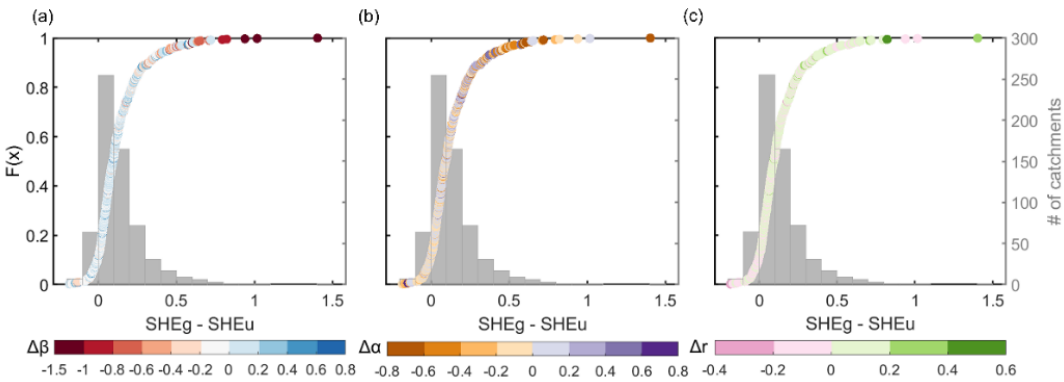


Figure 4-3. Cumulative distribution function (i.e. cdf) plot and histogram plot of difference between SHE for gauged and ungauged cases (i.e. $SHE_g - SHE_u$). Cdf plot is color-coded by (a) bias component difference (Δβ), (b) variance component difference (Δα) and (c) correlation component difference (Δr) between SHE formulations for gauged and ungauged cases summarized in Table 4-1.

4.5 Discussion and Summary

In summary, we introduced a new signature-based hydrologic efficiency (SHE) metric based on the idea that a model's fit to signatures will be easier to interpret hydrologically, and more importantly, that we can estimate it directly in ungauged basins. The SHE metric is correlated to different degree with existing metrics, and we show how its components, and hence the metric itself, can be estimated in ungauged catchments.

A flexible efficiency metric based on signatures provides significant opportunity for hydrologically relevant diagnostic model calibration and evaluation (Yadav et al., 2007; Yilmaz et al., 2008; Shafii and Tolson, 2015). Here, we simply replace the statistical components of the KGE (Gupta et al., 2009) with signatures suitable for our study domain, Great Britain. We chose to use runoff ratio and variance ratio as our signatures to represent

bias and variance aspects of the hydrograph. However, other signatures could and should be considered for different study domains. Hydrologists have investigated many signatures and found different ones to be useful to characterize major hydrologic functions or hydrograph aspects of catchments depending on the study domain (McMillan, 2020). Different aspects of the flow duration curve have for example been used to characterize the variability of flow through different signatures (e.g. Yilmaz et al., 2008; Sawicz et al., 2011; Westerberg et al., 2011; Pool et al., 2018; McMillan, 2021). It might be useful to use different signatures depending on whether study domains for example contain catchments with significant snow or those in arid domains.

We do not believe that SHE would be universally applicable in this form everywhere in the world. Actually, we believe that the different components should be replaced by appropriate signatures of a catchment's, water balance, its damping, and its translation of precipitation variability into streamflow variability and timing. Different signatures might be best suited to represent these components depending on whether the study domain is for example located in a temperate, dry or cold part of the world. Equally, existing regionalized streamflow indices correlated with these components might provide a baseline from which such a metric can be estimated in both gauged and ungauged catchments. An advantage of this opportunity and need for tailoring is that making these choices puts the discussion about suitable objective functions into the realm of hydrology, rather than just statistics.

The issue of signature choice is also linked to the ability for regionalising signatures or indices correlated with the components of the efficiency metric. Many regionalisation studies exist (e.g. He et al., 2011; Wagener and Montanari, 2011), though in how far these studies provide a regional basis to calculate efficiency metrics from in ungauged locations has so far been unexplored. One issue we did not tackle here in this context is that of uncertainty in these regionalisation estimates (e.g. Zhang et al., 2008; Kapangaziwiri et al., 2012; Westerberg et al., 2014). Uncertainties originate from the underlying measurements of physical catchment properties and of hydro-meteorological variables, from processing

of the original observations, and from choices made regarding space-time averaging etc. (McMillan et al., 2022a; Westerberg et al., 2016). There is opportunity for integrating uncertainty in a coherent statistical framework covering both gauged and ungauged situations, which should significantly increase the value of available regionalised information in the context of model calibration and evaluation.

Chapter 5 Conclusion and Outlook

5.1 Conclusions

In the Introduction Chapter, we highlight that some challenges still exist in rainfall-runoff modelling. These challenges briefly are: 1) Lack of using location specific information in the characterization of catchments with subsurface losses, 2) Lack of a coherent strategy for a priori selection of models to include in multi-model studies or modular modelling frameworks and 3) Lack of hydrologically diagnostic evaluation of models for both gauged and ungauged catchments. Our overarching objective is to address these challenges through a perceptual understanding of catchment functions in a nationally consistent framework. In conclusion, this thesis advances our knowledge of rainfall-runoff modelling in large-sample hydrology through three key contributions, tackling the three challenges we outlined:

- First, our study reveals that considering the catchment's relative location to the coast and its position within a wider river basin significantly contributes to explaining the water balance issues in highly permeable catchments affected by subsurface losses. We demonstrate the necessity of incorporating location indicators into large-sample hydrology datasets.
- Second, our research findings highlight the role of a priori model selection. By selecting model structures consistent with expected hydrologic variability, we demonstrate the possibility of observing meaningful performance differences between model structures in specific catchments.
- Finally, we have introduced a novel signature-based hydrologic efficiency (SHE) metric that provides a more hydrologically interpretable measure of how well a model fits certain functional aspects of a catchment's hydrology and can be estimated in ungauged catchments.

In this conclusions chapter, we summarize our analyses and findings for each challenge presented in each technical chapter (Chapter 2, 3, 4), give overarching remarks and outlook.

Challenge 1: Location, location, location – Considering relative catchment location to understand subsurface losses (Chapter 2)

The analysis of large samples of hydrologic catchments is regularly used to gain understanding of hydrologic variability and controlling processes. Several studies have pointed towards the problem that available catchment descriptors (such as mean topographic slope or average subsurface properties) are insufficient to capture hydrologically relevant properties. Here, we test the assumption that catchment location, i.e. the relative properties of catchments in relation to their surrounding neighbours, can provide additional information to reduce this problem. We test this idea in the context of Great Britain for a widely discussed problem, that of catchment water balance errors due to subsurface losses. We focus on three locational aspects (i.e. location to coast, location within a wider basin and location to a relevant neighbour), utilizing only basic and widely available geological and topographical information. We find that subsurface losses from catchments with a highly permeable geology connection to the coast are in order of 30% water balance error. We introduce a simple index to define location within a wider basin that is able to explain water balance issues of highly permeable headwater catchments. We attempt to quantify catchment location relevant to a neighbour but find that it does not increase water balance error predictability beyond using available catchment-scale geological information. The results imply that location, geology and topography combine to define the differences of water balances of Great Britain catchments compared to what we would expect from their climatic setting alone.

Challenge 2: A priori selection of hydrological model structures in modular modelling frameworks (Chapter 3)

Multi-model studies have become common in large sample hydrology. However, significant challenges remain in identifying connections between model structures and catchment characteristics, and thus in developing a coherent strategy for tailored multi-model ensembles. Here, we analyse and discuss the importance of selecting model structures that are consistent with the expected hydrologic variability across the modelling domain by comparing the results of two modular modelling frameworks across 998 Great Britain catchments. One framework is based on model structures which have historically evolved in the UK (RRMT), while the other is based on model structures originating from different parts of the world (FUSE). While both groups of model structures have members that achieve high performance, the historically evolved group members separate in their performances between catchments in a way that is more consistent with our expectation of hydrologic differences. We further find that four hydrological signatures organize these differences. Our results emphasize the importance of model structure selection based on explicit perceptual models and the need to go beyond statistical performance as sole criterion.

Challenge 3: A Signature-based Hydrologic Efficiency Metric for Model Calibration and Evaluation in Gauged and Ungauged Catchments (Chapter 4)

Rainfall-runoff models are commonly evaluated against statistical evaluation metrics. However, these metrics do not provide much insight into what is hydrologically wrong if a model fails to simulate observed streamflow well and they are also not applicable for ungauged catchments. Here, we propose a signature-based hydrologic efficiency (SHE) metric consisting of hydrologic signatures that can be regionalized for model evaluation in ungauged catchments. We test our new efficiency metric across 633 catchments from Great Britain. Strong correlations with Spearman rank and Pearson correlation values around 0.8 are found between our proposed metric and commonly used statistical evaluation metrics

(NSE, KGE, NP…) demonstrating that the proposed SHE metric is related to existing metrics as much as these metrics are related to each other. For ungauged catchments, we regionalise the three signatures included in SHE and find that 78% of catchments have an absolute difference of SHE values between gauged and ungauged cases of less than 0.2. This difference increases where the regionalized bias and variance signature values are different to the observed ones. It means that SHE metric is applicable for model evaluation in ungauged catchments if its signatures can be regionalized well.

5.2 Overarching remarks

The challenges addressed in this thesis are intrinsically inter-linked. Investigating water loss or gains in catchments due to regional subsurface connections by considering catchment location to its surrounding area is beneficial to better characterize catchments. This will advance our perception of catchment processes and ultimately contribute to the robust selection or development of suitable rainfall-runoff models (Beven, 2001). Moreover, model evaluation metrics and their intrinsic statistical components influence how hydrologically meaningful the performance of different models can be assessed and ultimately be linked to different catchment types. Hydrologically-based diagnostic model evaluation (i.e. using a signature-based evaluation metric) links the dominant hydrologic processes of catchments and certain process representations of model structures (Yilmaz et al., 2008). This will again help in selecting and developing suitable model structures for different catchment types.

Even though this thesis reduces some problems of rainfall-runoff modelling, there are aspects that continue to remain unknown. Certain model structures are able to be identified for certain catchment types (e.g. model structure with leaky routing module is able to simulate the water balance of leaky catchments). However, water balance errors are likely always combinations of multiple factors including errors in observations such as the precipitation (Montanari and Di Baldassarre, 2013). In addition, the net impact of water management activities, e.g. through abstractions or reservoirs, on the water balance of

catchments is often only poorly known. This challenge is amplified in large-sample studies where catchments are influenced by a wide variety of water management activities. Despite providing valuable opportunities for research, large-sample datasets currently lack sufficient characterization of these activities, which limits their use (Addor et al., 2020). Although recent efforts have been made to detect the effects of water management activities on flow regimes (e.g., Bloomfield et al., 2021; Van Loon et al., 2022; Salwey et al., 2023), there are still gaps that require further research to fully understand and address these impacts. Disentangling these different contributions to the overall water balance error will in many cases be very difficult given the current data available.

A priori selection of model structures has been analysed widely for gauged catchments due to a strong emphasis on the use of statistical performance metrics to distinguish model quality, but how to transfer such information to ungauged places is still challenging. The regionalization of hydrological models for streamflow prediction in ungauged catchments has been extensively explored, as for example highlighted in the comprehensive review by Guo et al. (2021). The authors' emphasized that the accuracy of hydrological simulations in ungauged catchments, utilizing parameter regionalization, heavily relies on the choice of model structures. While our study has provided some insights into the potential applicability of a signature-based evaluation metric in ungauged catchments, further research is needed to effectively transfer the a priori model selection from gauged to ungauged catchments. Additionally, addressing the uncertainties associated with signature regionalization remains an important task (Almeida et al., 2016; Westerberg et al., 2016). Lastly, working with a large number of catchments presents various challenges in hydrological research. These challenges include managing data quality and heterogeneity across catchments (Merheb et al., 2016; Hrachowitz et al., 2013), handling computational demands (Montanari et al., 2013), generalizing findings to other regions (Belvederesi et al., 2022), and addressing wider problems of uncertainties and variability (McMillan et al., 2012). Overcoming these challenges requires careful planning, robust methodologies, and

a comprehensive understanding of the limitations associated with large-sample hydrological studies.

5.3 Outlook

In this section, we briefly describe two promising areas of research for expanding and refining the work presented in this thesis. Rather than providing an extensive list of ideas, we have chosen to focus on these two ideas to provide a more detailed explanation and include some preliminary results for one of these ideas.

5.3.1 Quantification of hydrologic services using signatures

While addressing some challenges of rainfall-runoff modelling, we have shown the applicability of hydrologic signatures in multiple aspects (i.e. catchment characterization, model selection, hydrologically diagnostic model evaluation). The next step could be the quantification of hydrologic services using specifically defined signatures. Hydrologic services are defined as the benefits that human and nature can receive from the eco-hydrologic processes of a catchment (Brauman, 2007; Wagener et al., 2008). Water supply (extractive or in-stream), water damage mitigation, cultural and supporting services are different categories of hydrologic services (Brauman, 2007). Quantifying these services is important to understand their natural supply and the need for additional services through human activity if the natural supply is insufficient. However, such hydrologic services have mainly been assessed for small areas or individual catchments (e.g. Terrado et al., 2014; Carvalho-Santos et al., 2016; Casagrande et al., 2021), not for large samples of catchments in a comparative analysis. We believe that hydrologic signatures describing the functional behavior of catchments can be adjusted to quantify hydrologic services in large-sample hydrology.

As an example, we show some preliminary results regarding how to quantify hydropower generation in a large-sample of GB catchments (i.e. 671 CAMELS-GB catchments) using a 30-year climatic and hydrologic data (i.e. years between 1985 and 2015). UK has been

using water for energy generation since 1879 (International Hydropower Association, 2020). The installed capacity is mostly located in the wet and mountainous regions of Wales and Northwest Scotland. Hydropower generation corresponds to only 2 percent of total electricity generation in last 30 years (BEIS, 2021).

While quantifying hydrologic services, we consider their potential and consistency in a catchment. Potential (i.e. availability) refers to the capacity of a catchment to provide a hydrologic service. Consistency (i.e. sustainability) refers to the ability of a catchment to maintain a hydrologic service over time. If a catchment has high availability and consistency of a hydrologic service, it can provide high and sustainable supply of the service (Figure 5-1).

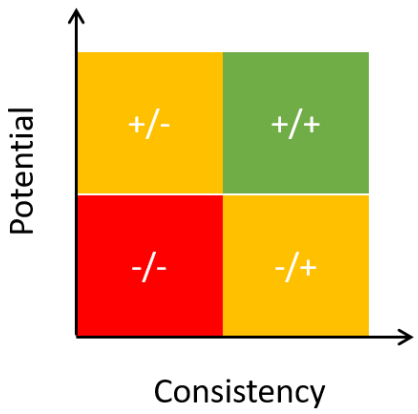


Figure 5-1. Illustration of how potential and consistency of a hydrologic service can affect the supply of a service. Green, yellow and red regions represent high, medium and low supply of a service.

Hydropower generation is one of the in-stream water supply services. In-stream water supply is a catchment's ability to provide water within the river system for different purposes (e.g. water recreation and transportation) or to be diverted temporarily for a certain use (e.g. hydropower generation) and returned back to the river system. Hydropower generation is a process of transforming the potential energy of water into kinetic energy and generating electricity by releasing water from high elevation to low elevation and passing it through a turbine (RenÖFÄLt and Nilsson, 2010; Guo and Peng, 2019). It is calculated as:

$$P = \eta * \rho * g * Q * H \qquad (5\text{-}1)$$

where P is the power generated (W), η is the efficiency (dimensionless), ρ is the density of water (kg/m^3), g is acceleration due to gravity (m/s^2), Q is the flow through turbine (m^3/s) and H is the head drop between the surface water level at the intake and the surface water level at the outfall (m) (Basso and Botter, 2012; Hatchard, 2021). The amount of water flowing through the river and the elevation that water drops on its way are the two main variables to quantify transformation of water's potential energy into kinetic energy, i.e. in order to quantify hydropower generation.

We create a formulation for a relative measure of which catchment is more or less likely to generate hydropower. In order to calculate hydropower generation potential, elevation difference between minimum and maximum elevation of a catchment is used instead of head by assuming that the beginning of the river is higher (i.e. higher head) in the catchments having higher elevation difference. Turbine efficiency coefficient is a constant of original hydropower generation formula in the literature but we do not consider it in our study because we will not be using any turbine to make a design calculation. Hence, the hydropower signature potential of a catchment that we formulate is;

$$\mathrm{HG}_{\mathrm{potential}} = \rho * \mathrm{g} * \mathrm{Q} * \mathrm{dE} \qquad (5\text{-}2)$$

where ρ is the density of water (i.e. 1000 kg/m^3), g is acceleration due to gravity (i.e. 9.81 m/s^2), Q is average daily streamflow of the catchment (in m^3/s) and dE is the difference between maximum and minimum elevation of the catchment (in m).

In addition to potential, we consider hydropower generation consistency. To ensure consistency of services across dry and wet years, it is important to consider the variability in yearly hydropower generation potential. This allows us to account for the fluctuations in hydropower generation and address the challenges associated with different hydrological conditions over time. That's why, and we formulate it as the inverse of coefficient of variation of yearly hydropower generation potential:

$$HG_{consistency} = standard\ deviation\ (HG_{yearly\ potential})/mean(HG_{yearly\ potential}) \quad (5\text{-}3)$$

where $$HG_{yearly\ potential} = \rho * g * Q_{yearly\ average} * dE \quad (5\text{-}4)$$

and $Q_{yearly\ average}$ is yearly average streamflow (in m^3/s).

We create interpolated maps from 671 CAMELS-GB catchments of hydropower generation potential and consistency maps of Great Britain (Figure 5-2). Both maps (a) and (b) in Figure 5-2 show similar patterns with higher values in the North-West decreasing to South-East. The highest hydropower generation potential values are mostly observed in Scotland, in some regions of North England, Midlands and Wales. Similarly, the highest hydropower generation consistency values are mostly observed in Scotland, in some regions of Northern west of England and Wales. Moreover, the scatter plot (Figure 5-2c) indicates that most of the catchments which currently have hydropower generation are within the 90th percentile of both hydropower generation potential and consistency. This implies that the defined signature is able to quantify hydropower generation service in the catchments because we assume that the catchments currently having hydropower generation are chosen as suitable because of their potential and consistency for hydropower generation.

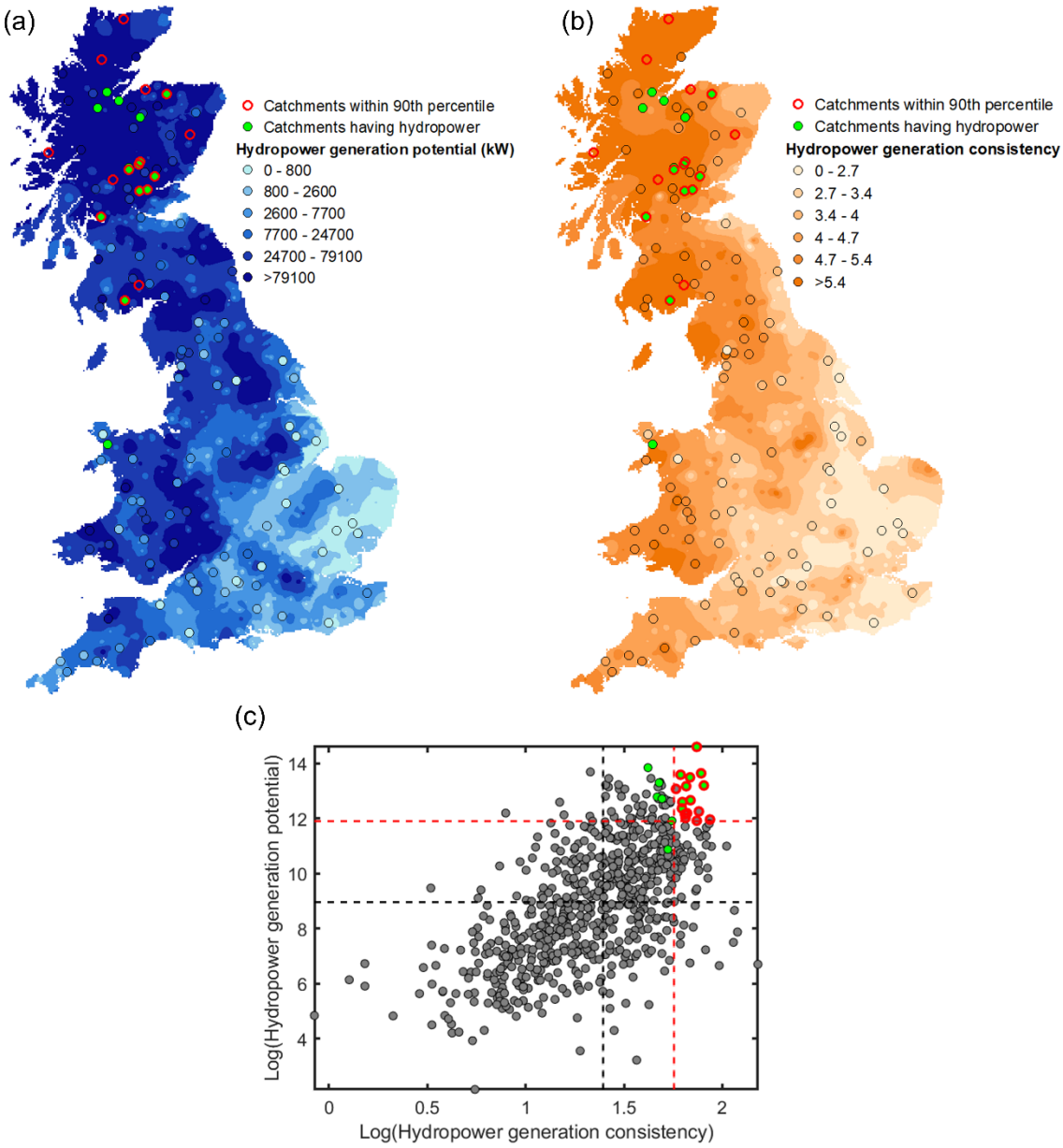


Figure 5-2. Hydropower generation potential (a) and consistency (b) maps of Great Britain maps and scatter plot (c) of hydropower generation potential vs. consistency. 105 benchmark catchments are shown as circles on the maps. Catchments that already have a hydropower reservoir in them are shown on maps as green circles. The catchments that have hydropower generation potential and consistency values higher than the 90th percent of all catchments are also shown as red circles. The black and red dashed lines on the scatter plot indicate 50th and 90th percentile values of potential and consistency, respectively. Logarithmic values are used while creating the scatter plots of hydropower generation potential vs. consistency to make the outliers less extreme so we can see all catchments better while preserving their order.

To summarise, hydrologic services can be quantified across large samples of catchments in a comparative approach using hydrologic service signatures. Both potential and consistency of services should be considered while quantifying hydrologic services. We

can quantify other hydrologic services such as water supply and flood mitigation services across GB. In future studies, it might also be interesting to investigate if regionalization of hydrologic signatures enables us to robustly quantify hydrologic services in ungauged catchments.

5.3.2 A national-scale groundwater modelling across GB

In Chapter 2, we could not solve the problem of how the mechanisms of groundwater transfers between neighbouring catchments work by investigating the catchments themselves (i.e. only using their locational information and available physical descriptors). Surface-subsurface interactions and hydrogeological connectivity between neighbouring catchments and the net effects of water management activities can be complex and not easily captured in simple descriptors. Therefore, we believe that a national-scale groundwater modelling could be a way forward so that we can estimate inter-catchment groundwater flow between neighbouring catchments. To date, most groundwater modelling efforts have taken place at regional scales in GB (e.g. Shepley et al., 2012; Jackson et al., 2016; Collins et al., 2020). There are several national-scale hydrological modelling studies (e.g. Bell et al., 2018; Coxon et al., 2019), but groundwater flow is not explicitly represented in them. Recently, Rahman et al. (2023) conducted a study to develop an explicit groundwater flow model for England and Wales using observed groundwater head data and local hydrogeological information. As their proposed conceptual (or perceptual) model for groundwater flow indicates (Figure 5-3), they used two-dimensional representation of hydrogeological properties using depth-average transmissivity. Their model could be a basis for coupling with hydrologic models that would add surface and unsaturated zone processes to gain a more complete picture of the water cycle. Integrated modelling of hydrologic fluxes both within and across catchments might provide a better way to assess subsurface losses and gains than currently possible.

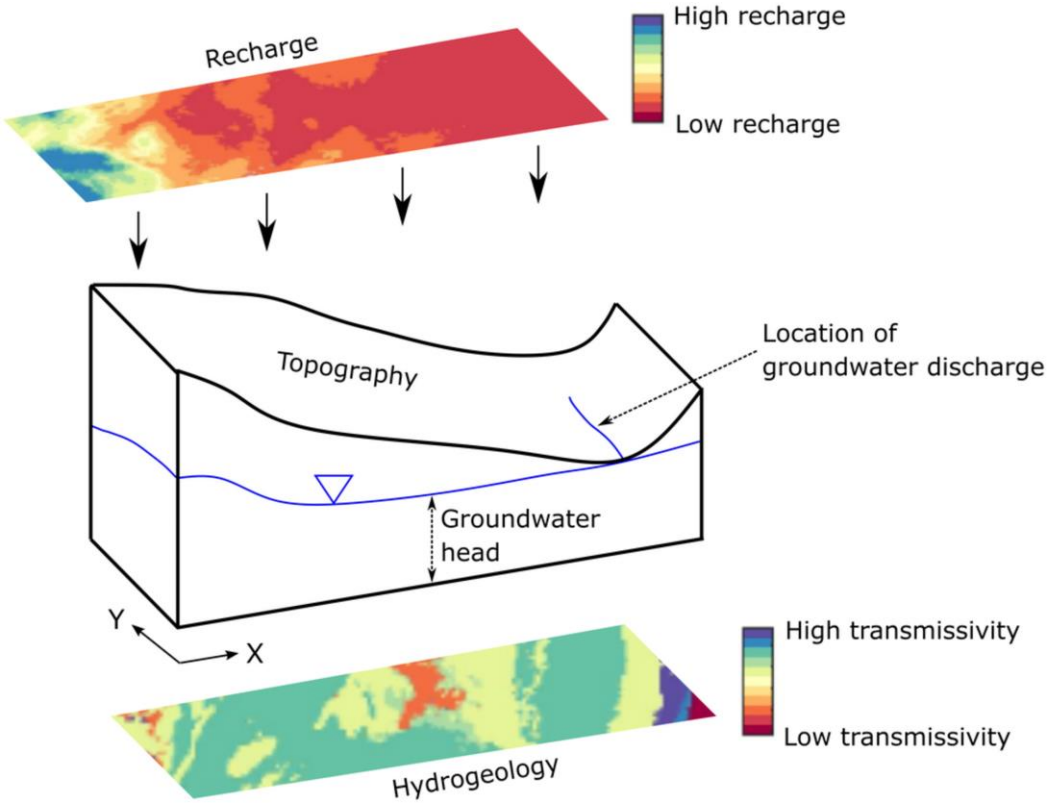


Figure 5-3. Conceptualization of the groundwater flow model proposed by Rahman et al. (2023). The blue line in the subsurface represents the location of the groundwater table. Groundwater discharge occurs when groundwater table intersects the surface. The two-dimensional representation of hydrogeology in this figure considers depth-averaged transmissivity in the model. Taken from Rahman et al. (2023).

APPENDIX A Supplemental Material Chapter 2

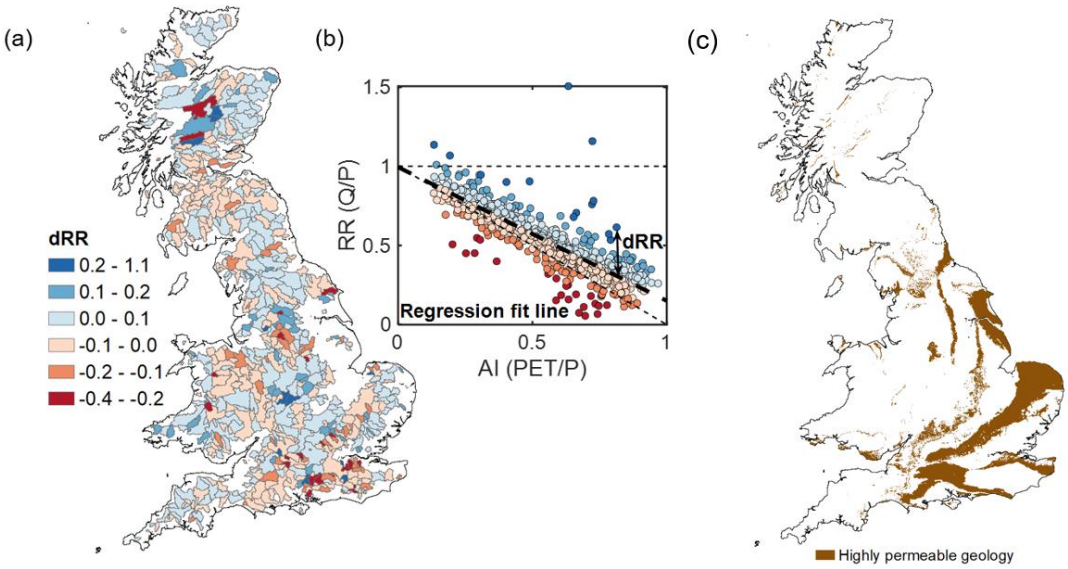


Figure A1. (a) Map of dRR values and (b) Scatter plot of RR vs. AI for 660 CAMELS-GB catchments. The bold black dashed line in Figure (b) is the regression fit line based on linear regression using AI as the predictor.

*Table A1. List of catchment attributes from CAMELS-GB dataset. These attributes are used both in Chapter 2 (i.e. Figure 2-4 and Figure 2-5) and Chapter 4. "**baseflow_index# (BFI-HOST)**" is only used in Chapter 4.*

Attribute Class	Attribute Name	Description	Unit
Topography	area	catchment area	km^2
	dpsbar	catchment mean drainage path slope	m km^{-1}
	elev_mean	catchment mean elevation	m.a.s.l
	elev_min	catchment minimum elevation	m.a.s.l
	elev_10	catchment 10th percentile elevation	m.a.s.l
	elev_50	catchment median elevation	m.a.s.l
	elev_90	catchment 90th percentile elevation	m.a.s.l
	elev_max	catchment maximum elevation	m.a.s.l
Climatic Indices	p_mean	mean daily precipitation	mm day^{-1}
	pet_mean	mean daily PET (Penman-Monteith equation without interception correction)	mm day^{-1}
	Aridity_index	aridity, calculated as the ratio of mean daily potential evapotranspiration to mean daily precipitation	-
	p_seasonality	seasonality and timing of precipitation (estimated using sine curves to represent the annual temperature and precipitation cycles; positive (negative) values indicate that precipitation peaks in summer (winter) and values close to zero indicate uniform precipitation throughout the year)	-
	frac_snow	fraction of precipitation falling as snow (for days colder than 0°C)	-
	high_prec_freq	frequency of high precipitation days (≥ 5 times mean daily precipitation)	days yr^{-1}
	high_prec_dur	average duration of high precipitation events (number of consecutive days ≥ 5 times mean daily precipitation)	days
	low_prec_freq	frequency of dry days (< 1mm day-1)	days yr^{-1}
	low_prec_dur	average duration of dry periods (number of consecutive days < 1mm day-1)	days
#	**baseflow_index# (BFI-HOST)**	**A base-flow index derived from the 29-class Hydrology Of Soil Types (HOST) classification**	-
Land Cover Attributes	dwood_perc	percentage cover of deciduous woodland	%
	ewood_perc	percentage cover of evergreen woodland	%
	grass_perc	percentage cover of grass and pasture	%
	shrub_perc	percentage cover of medium scale vegetation (shrubs)	%
	crop_perc	percentage cover of crops	%
	urban_perc	percentage cover of suburban and urban	%
	inwater_perc	percentage cover of inland water	%
	bares_perc	percentage cover of bare soil and rocks	%
Soil Attributes	sand_perc	percentage sand	%
	silt_perc	percentage silt	%
	clay_perc	percentage clay	%
	organic_perc	percentage organic content	%
	bulkdens	bulk density	g cm^{-3}
	tawc	total available water content	mm
	porosity_cosby	volumetric porosity (saturated water content estimated using a pedotransfer function based on sand and clay fractions)	-
	porosity_hypres	volumetric porosity (saturated water content estimated using a pedotransfer function based on silt, clay and organic fractions, bulk density and topsoil)	-
	conductivity_cosby	saturated hydraulic conductivity (estimated using a pedotransfer function based on sand and clay fractions)	cm h^{-1}
	conductivity_hypres	saturated hydraulic conductivity (estimated using a pedotransfer function based on silt, clay and organic fractions, bulk density and topsoil)	cm h^{-1}
	root_depth	depth available for roots	M
	soil_depth_pelletier	depth to bedrock (maximum 50m)	M
Hydrogeology Attributes	inter_high_perc	significant intergranular flow – high productivity	%
	inter_mod_perc	significant intergranular flow – moderate productivity	%
	inter_low_perc	significant intergranular flow – low productivity	%
	frac_high_perc	flow through fractures – high productivity	%
	frac_mod_perc	flow through fractures – moderate productivity	%
	frac_low_perc	flow through fractures – low productivity	%
	no_gw_perc	rocks with essentially no groundwater	%
	low_nsig_perc	generally low productivity (intergranular flow) but some not significant aquifer	%
	nsig_low_perc	generally not significant aquifer but some low productivity (intergranular flow)	%
Human Influences	surfacewater_abs	mean surface water abstraction	mm day^{-1}

	groundwater_abs	mean groundwater abstraction	mm day^{-1}
	discharges	mean discharges (daily discharges into water courses from water companies and other discharge permit holders reported to the Environment Agency)	mm day^{-1}
	abs_agriculture_perc	percentage of total (groundwater and surface water) abstractions in catchment for agriculture	%
	abs_amenities_perc	percentage of total (groundwater and surface water) abstractions in catchment for amenities	%
	abs_energy_perc	percentage of total (groundwater and surface water) abstractions in catchment for energy production	%
	abs_environmental_perc	percentage of total (groundwater and surface water) abstractions in catchment for environmental purposes	%
	abs_industry_perc	percentage of total (groundwater and surface water) abstractions in catchment for industrial, commercial and public services	%
	abs_watersupply_perc	percentage of total (groundwater and surface water) abstractions in catchment for water supply	%
	num_reservoir	number of reservoirs in the catchment	-
	reservoir_cap	total storage capacity of reservoirs in the catchment in megalitres	ML
	reservoir_he	percentage of total reservoir storage in catchment used for hydroelectricty	%
	reservoir_nav	percentage of total reservoir storage in catchment used for navigation	%
	reservoir_drain	percentage of total reservoir storage in catchment used for drainage	%
	reservoir_wr	percentage of total reservoir storage in catchment used for water resources	%
	reservoir_fs	percentage of total reservoir storage in catchment used for flood storage	%
	reservoir_env	percentage of total reservoir storage in catchment used for environmental	%

#: BFI-HOST of each catchment is obtained from NRFA website (https://nrfa.ceh.ac.uk/data/search) where detailed information of each stream gauges is given. In the Hydrology Of Soil Types (HOST) dataset, the percentage of each HOST soil class present for each grid square is estimated for each 1 km square (Boorman et al., 1995). Firstly, BFI (i.e. a measure of catchment responsiveness) values was calculated using daily mean flow data with a baseflow separation method developed in the Low Flow Studies (Institute of Hydrology, 1980). After that, BFI-HOST values were derived by regionalizing these BFI values with multiple linear regressions by using the fractions of HOST soil classes within the topographic boundaries of catchments (Boorman et al., 1995).

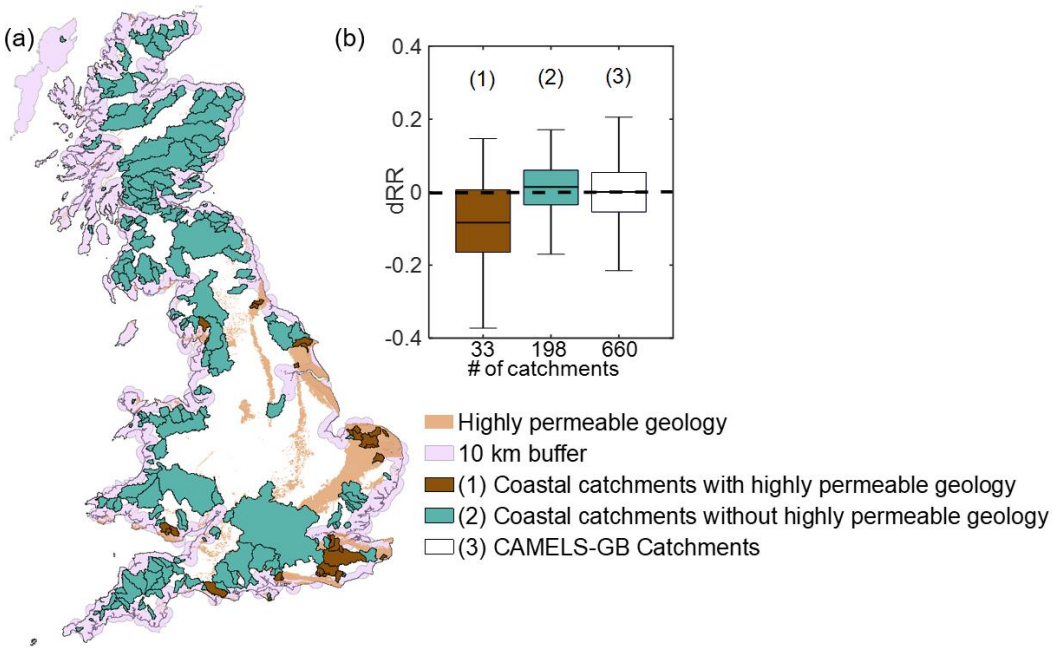


Figure A2. (a) Map of and (b)Whisker-Box plot of coastal catchment groups. Map indicates the catchment groups analyzed in the hypothesis. The central horizontal line of the whisker plot represents the median dRR value of each group. The bottom and top edges of the whisker box indicate 25th and 75th percentiles, respectively. The bottom and top edges of the vertical line demonstrate the lowest and highest data point in the dataset excluding any outliers, respectively. dRR values are calculated according to Turc-Mezentsev Curve.

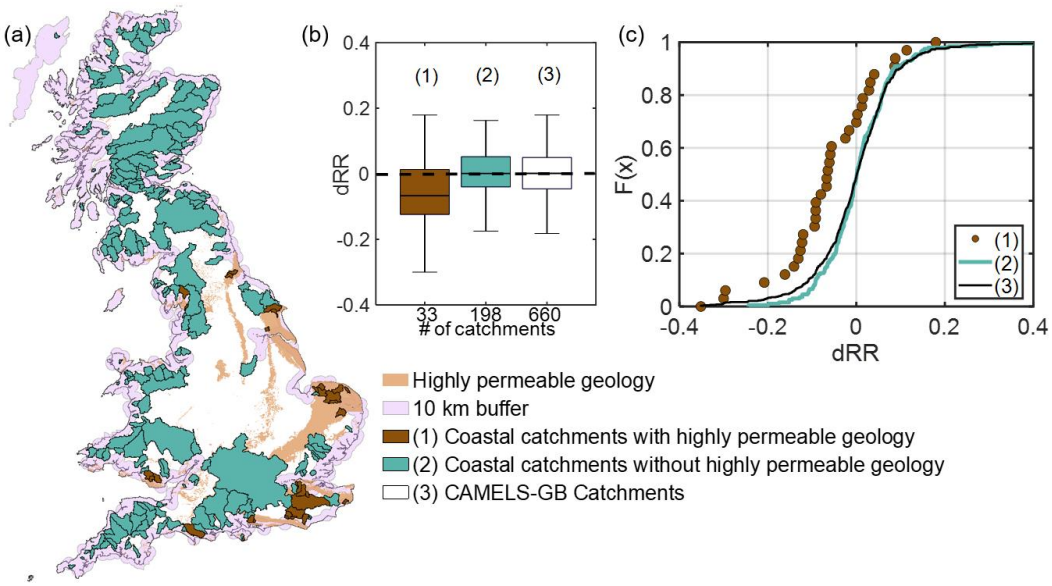


Figure A3. (a) Map of, (b)Whisker-Box plot of and (c) cdf plot of coastal catchment groups. Map indicates the catchment groups analyzed in the hypothesis. The central horizontal line of the whisker plot represents the median dRR value of each group. The bottom and top edges of the whisker box indicate 25th and 75th percentiles, respectively. The bottom and top edges of the vertical line demonstrate the lowest and highest data point in the dataset excluding any outliers, respectively. dRR values are calculated based on linear regression using AI as the predictor.

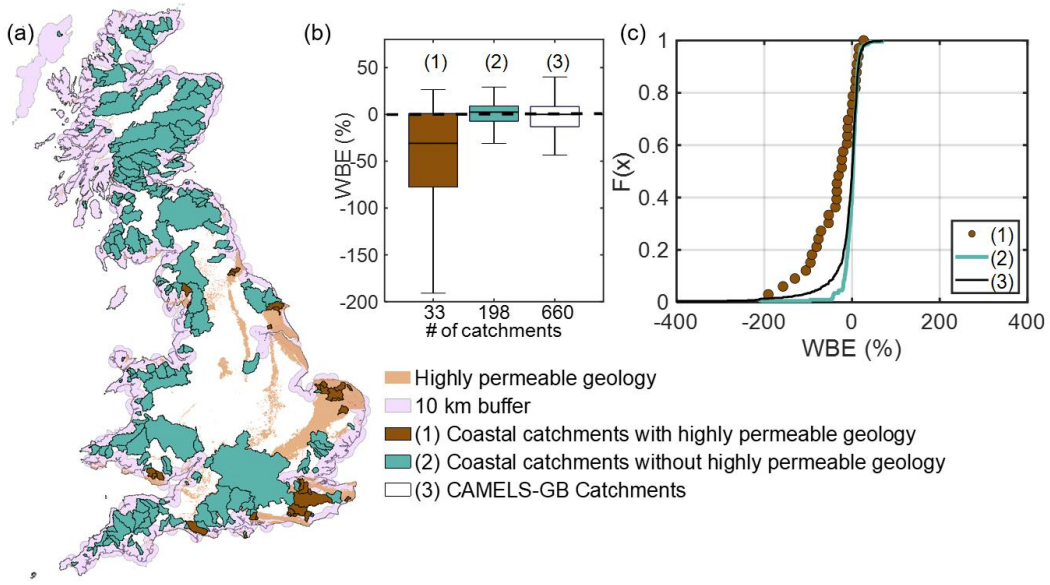


Figure A4. a) Map of, (b)Whisker-Box plot of and (c) cdf plot of coastal catchment groups with water balance error values in %. Water balance error (WBE) is calculated as dRR values divided by observed RR values and multiplied by 100. Map indicates the catchment groups analyzed in the hypothesis. The central horizontal line of the whisker plot represents the median WBE value of each group. The bottom and top edges of the whisker box indicate 25th and 75th percentiles, respectively. The bottom and top edges of the vertical line demonstrate the lowest and highest data point in the dataset excluding any outliers, respectively. dRR values are calculated according to Turc-Mezentsev Curve.

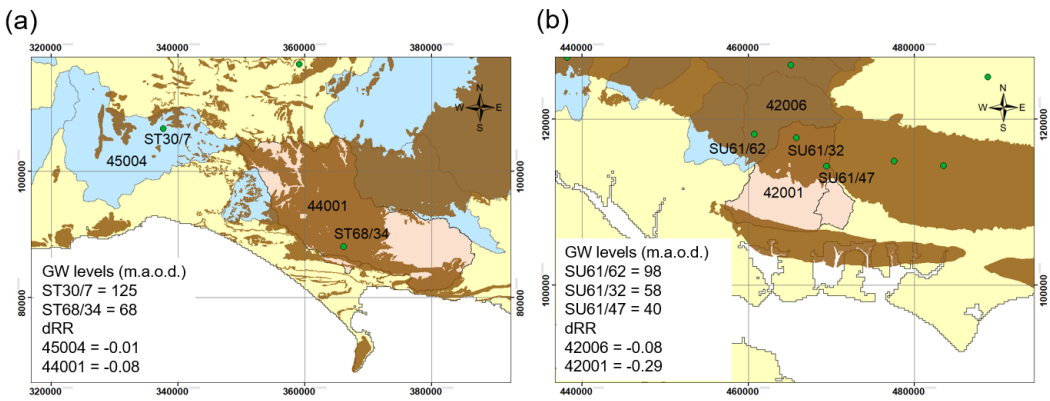


Figure A5. Maps of catchments 45004 and 44001 (a) and catchments 42001 and 42006 (b). Light pink and blue catchments are highly permeable and not highly permeable, respectively. Yellow color represents the land of Great Britain. Highly permeable geology is also shown as brown regions on the map. The wells located in these catchments are shown as green-filled circles. Groundwater (GW) levels of wells and dRR values of catchments are given on the bottom left corner in each map.

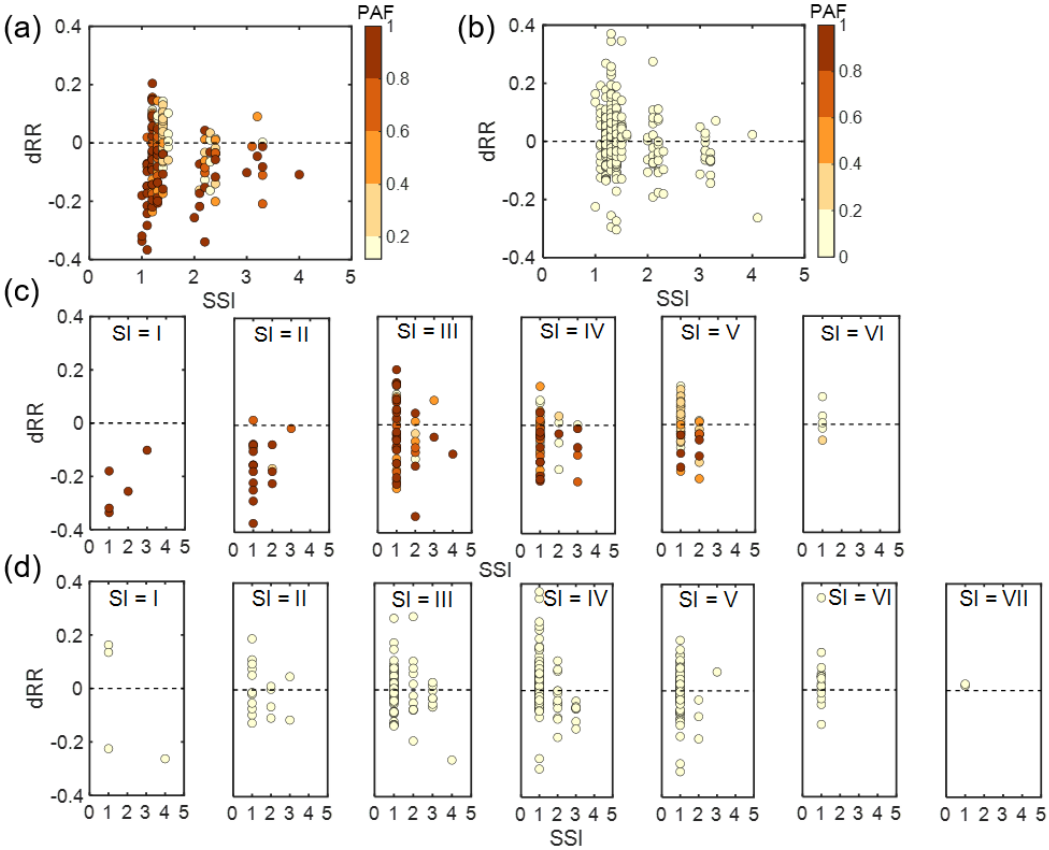


Figure A6. (a) Scatter plot of dRR vs. SSI (Strahler Sequence Index) of 174 CAMELS-GB catchments with highly permeable geology (PAF (permeable area fraction)>0.1), (b) Scatter plot of dRR vs. SSI of 320 CAMELS-GB catchments without highly permeable geology (PAF<0.1). The subgrouped version of Figure (a) based on catchments' SI (Strahler Index) values and the subgrouped version of Figure (a) based on catchments' SI values are also given as (c) and (d), respectively. dRR values are calculated according to Turc-Mezentsev Curve.

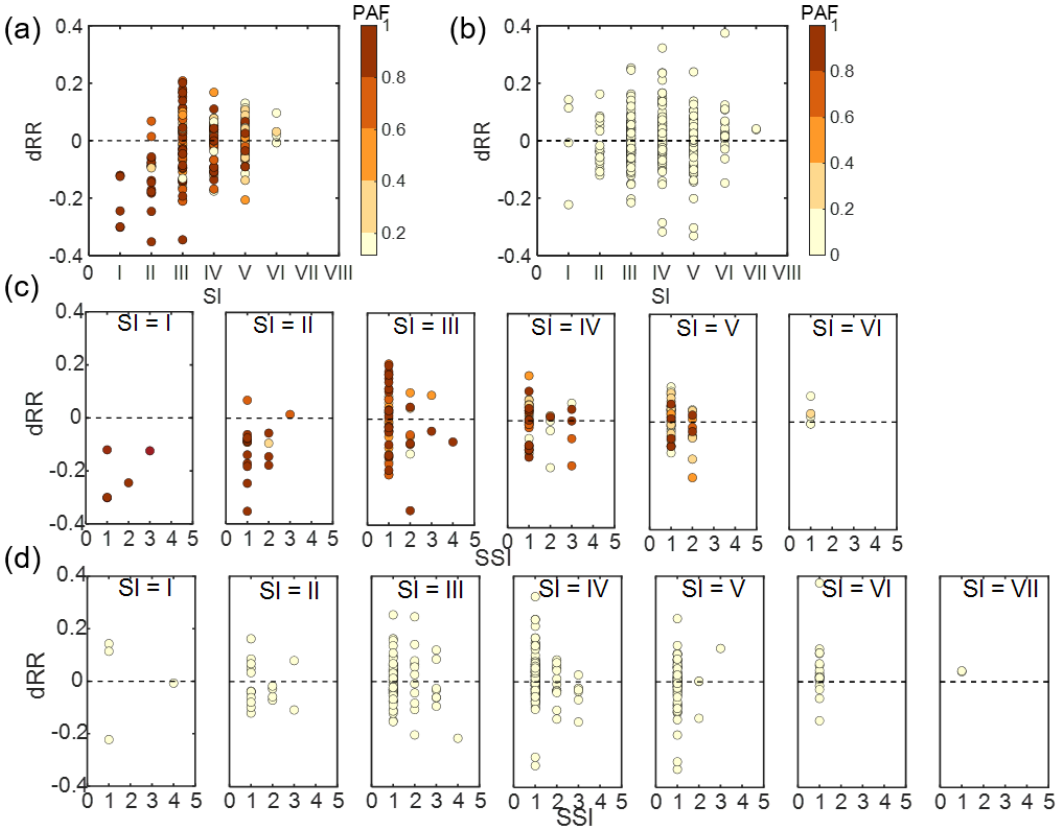


Figure A7. (a) Scatter plot of dRR vs. SI (Strahler Index) of 174 CAMELS-GB catchments with highly permeable geology (PAF (permeable area fraction)>0.1), (b) Scatter plot of dRR vs. SI of 320 CAMELS-GB catchments without highly permeable geology (PAF<0.1), (c) Scatter plots of dRR vs. SSI (Strahler Sequence Index) of catchment in (a) based on their SI values and (d) Scatter plots of dRR vs. SSI of catchment in (b) based on their SI values. dRR values are calculated based on linear regression using AI as the predictor.

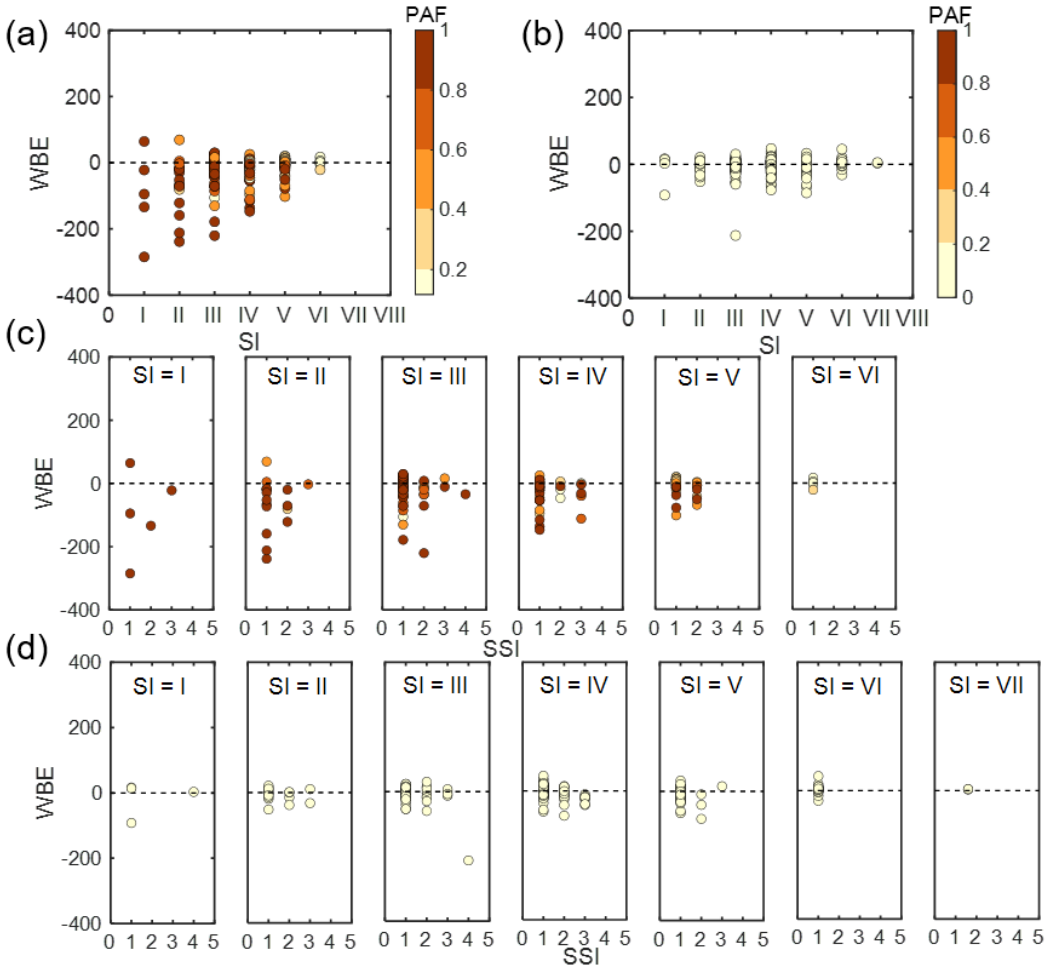


Figure A8. (a) Scatter plot of WBE vs. SI (Strahler Index) of 174 CAMELS-GB catchments with highly permeable geology (PAF (permeable area fraction)>0.1), (b) Scatter plot of dRR vs. SI of 320 CAMELS-GB catchments without highly permeable geology (PAF<0.1), (c) Scatter plots of dRR vs. SSI (Strahler Sequence Index) of catchment in (a) based on their SI values and (d) Scatter plots of dRR vs. SSI of catchment in (b) based on their SI values. dRR values are calculated according to Turc-Mezentsev Curve.

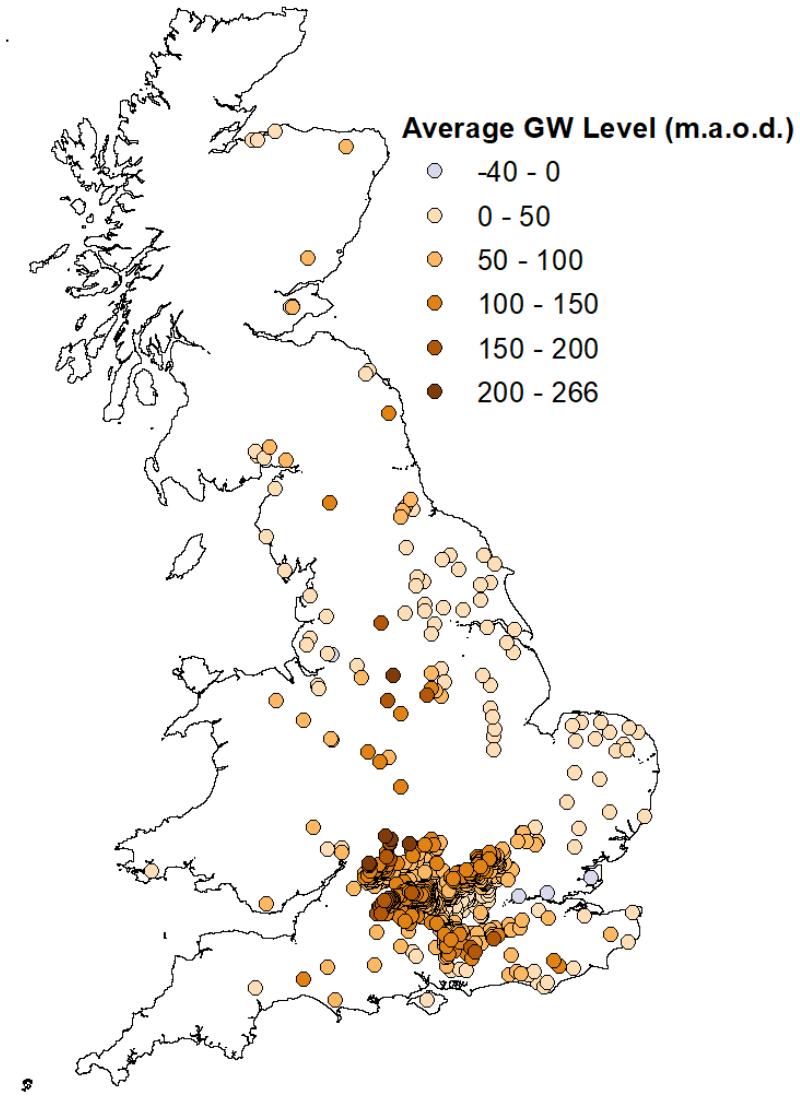


Figure A9. Map of average groundwater levels (m.a.o.d.) for 878 GB wells.

APPENDIX B Supplemental Material Chapter 3

B1. Different window sizes for smoothing process

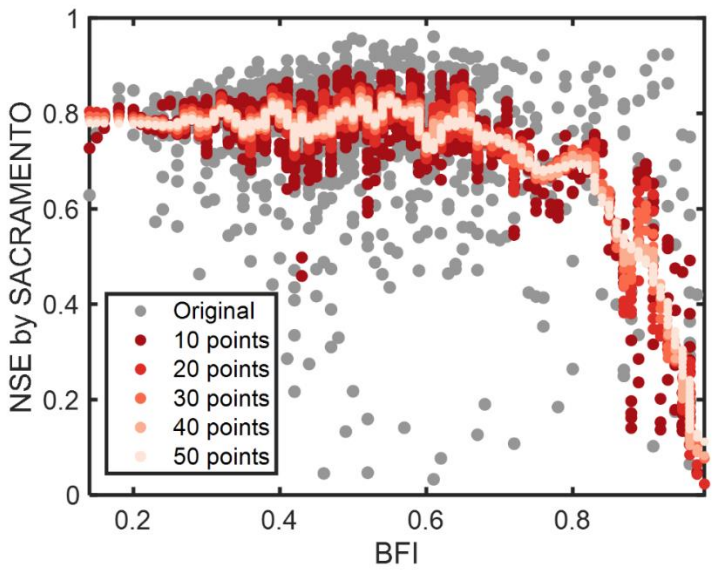


Figure B1.1. NSE values of SACRAMENTO model structure as original values and smoothed values (moving from left to right) by LOWESS approach with different window sizes including between 10 to 50 catchments (points). These different window sizes are checked for every model structure based on different catchment attributes. The most appropriate window size is selected as 40 points.

B2. Tables of parameters for RRMT and FUSE model structures

Table B2.1. Parameters of soil moisture accounting modules (i.e. PEN and PDM) and flow routing modules (i.e. 2PAR, LEAK and CRES) and their value ranges

Module	Parameter	Unit	Value range
PEN	Root constant	mm	10 - 200
	Bypass (fraction)	-	0 - 0.25
PDM	Maximum storage capacity	mm	0 - 500
	Degree of variability	-	0 - 2.5
2PAR	Residence time of first reservoir	T	1 - 15
	Residence time of second reservoir	T	15 - 100
	Fraction of effective rainfall going through first reservoir	-	0 - 1
LEAK	Residence time of upper part	T	1 - 200
	Residence time of middle part	T	1 - 100
	Residence time of lower part	T	1 - 25
	Lower threshold	mm	0 -175
	Upper threshold	mm	0 - 275
CRES	Residence time	T	1 - 15

Table B2.2. Parameters of FUSE model structures, units, value ranges and model(s) using parameters (1=TOPMODEL, 2=ARNO, 3=PRMS, 4=SACRAMENTO)

Parameter	Description	Unit	Value range	Model(s)
MAXWATER 1	Depth of upper soil layer	mm	25 - 500	1,2,3,4
MAXWATER 2	Depth of lower soil layer	mm	50 - 5000	1,2,3,4
FRACTEN	Fraction total storage in tension storage	-	0.05 - 0.95	1,2,3,4
FRCHZNE	Fraction tension storage in recharge zone	-	0.05 - 0.95	3
FPRIMQB	Fraction storage in first baseflow reservoir	-	0.05 - 0.95	4
RTFRAC1	Fraction of roots in the upper layer	-	0.05 - 0.95	2
PERCRTE	Percolation rate	$mm\ d^{-1}$	0.01 - 1000	1,2,3
PERCEXP	Percolation exponent	-	1 - 20	1,2,3
SACPMLT	SACRAMENTO model percolation multiplier for dry soil layer	-	1 - 250	4
SACPEXP	SACRAMENTO model percolation exponent for dry soil layer	-	1 - 5	4

PERCFRAC	Fraction of percolation to tension storage	-	0.5 – 0.95	4
FRACLOWZ	Fraction of soil excess to lower zone	-	0.5 – 0.95	3
IFLWRTE	Interflow rate	mm d^{-1}	0.1 - 1000	3,4
BASERTE	Baseflow rate	mm d^{-1}	0.001 - 1000	1,2
QB_POWR	Baseflow exponent	-	1 – 10	1,2
QB_PRMS	Baseflow depletion rate	d^{-1}	0.001 – 0.25	3
QBRATE_2A	Baseflow depletion rate first reservoir	d^{-1}	0.001 – 0.25	4
QBRATE_2B	Baseflow depletion rate second reservoir	d^{-1}	0.001 – 0.25	4
SAREAMAX	Maximum saturated area	-	0.05 - 0.95	3,4
AXV_BEXP	ARNO/VIC b exponent	-	0.001 - 3	2
LOGLAMB	Mean value of the topographic index	m	5 - 10	1
TISHAPE	Shape parameter for the topographic index gamma distribution	-	2 - 5	1
TIMEDELAY	Time delay in runoff	d	0.01 - 7	1,2,3,4

B3. Process of LOWESS smoothing approach

In LOWESS (locally weighted scatterplot smoothing) approach, the best NSE or KGE value of each catchment is taken in turn as the central point x in a set of 2k+1 data points after sorting the catchments by selected catchment attributes. 2k of 2k+1 includes the k number of former points and k number of after points for each of central NSE or KGE point. The smoothed (i.e. estimated) NSE or KGE value for the central point and its variance is produced by fitting a weighted linear regression to the selected set of 2k+1 data points. After repeating this process for all data points, a single LOWESS fit is obtained. Weights are calculated using the following function:

$$w_i = (1 - \left|\frac{(x - x_i)}{\max(x - x_i)}\right|^3)^3 \qquad \textit{(B3.1)}$$

Where x is central NSE or KGE point, x_i are the other NSE or KGE points within the selected 2k+1 data points defined by span (i.e. half of a window size). While the most weight is given to the data points nearest to the point of estimation and the least weight to the data points that are furthest away. To account for outliers, a first LOWESS fit is produced using the original data, followed by a calculation of the residuals r from this initial fit. Then, NSE or KGE point is weighted according to its distance from the fitted line using following bi-square weight function:

$$w_i = (1 - (r_i - 6\mathrm{MAD})^2)^2 \qquad \textit{(B3.2)}$$

$$\mathrm{MAD} = \mathrm{median}(|r|) \qquad \textit{(B3.1)}$$

where r_i is the residual of the ith residual data point and MAD is the median absolute deviation of the residuals. Data points with large residual values are down-weighted by this function (adapted from Coxon et al., 2015).

B4. Maps for differences between highest KGE and NSE values of RRMT and FUSE frameworks and maps for the highest KGE values of RRMT and FUSE frameworks

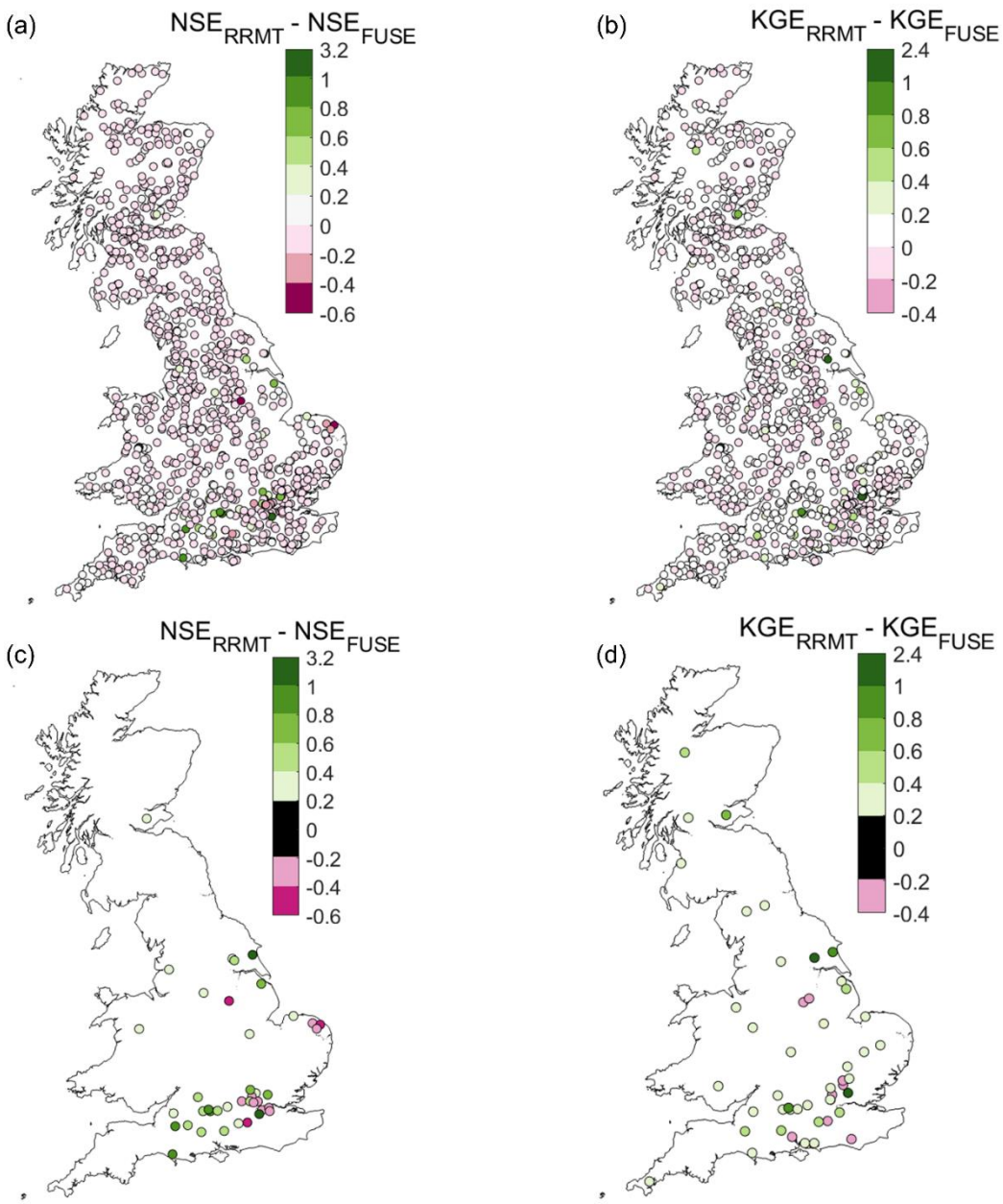


Figure B4.1: Map of highest (a) NSE and (b) KGE differences between RRMT and FUSE frameworks. For visual clarity, maps are recreated as (c) and (d) by eliminating catchments with NSE and KGE differences values ranging between -0.2 and 0.2 (i.e. indicated as black range in the colorbar).

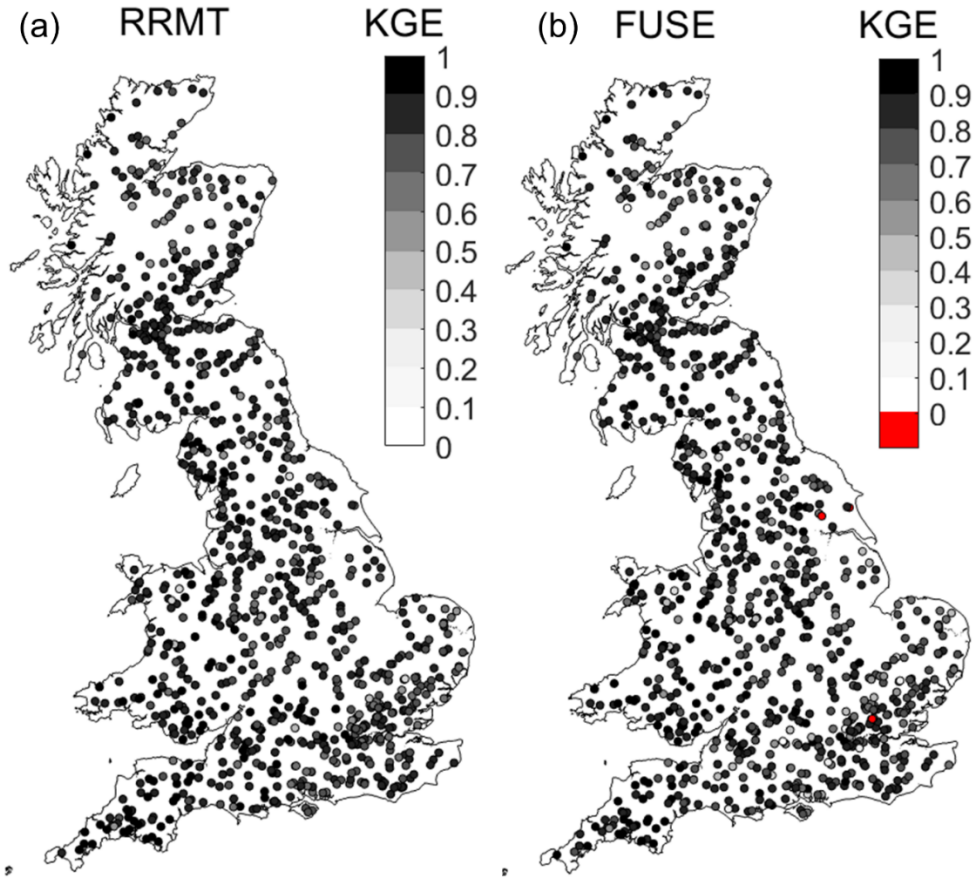


Figure B4.2. KGE values of best simulations performed by any of model structures which are selected from a) RRMT and b) FUSE frameworks. Best simulations are obtained based on NSE.

B5. Information regarding the reservoirs in our study catchments

According to the UK reservoir inventory (Durant and Counsell 2018) and reservoir data for Scottish catchments from the Scottish Environment Protection Agency (SEPA), 252 of the 989 catchments studied have reservoirs located in the catchment. Contributing area upstream of the reservoir (%) and normalized upstream capacity are two catchment descriptors suggested by Salwey et al. (2023) to quantify the influence of reservoirs on streamflow characteristics. Contributing area is defined as the percentage of catchment area which drains into the reservoir. A contributing area close to 100% means that the location of the reservoir is close to the catchment outlet. Normalized upstream capacity is the ratio of total reservoir capacity to average annual precipitation volume received by a catchment. It is an indicator of relative reservoir storage size. We investigate the relationship between each of these two reservoir-related descriptors and highest NSE scores obtained by RRMT and FUSE model structures for 252 catchments (Figure S5.1). Spearman rank correlations between highest NSE values obtained by the RRMT (FUSE) model structures and contributing area is -0.32 (-0.45), it is only -0.18 (-0.34) between NSE and normalized upstream capacity values. So, there is a small decline in model performance the closer a reservoir exists to the catchment outlet, and to a lesser degree the larger it is. However, the variability in performance change is very large and it would take consideration of additional aspects such as reservoir management to add reservoirs to the models used here (e.g. Payan et al. 2008). This effort is beyond our study.

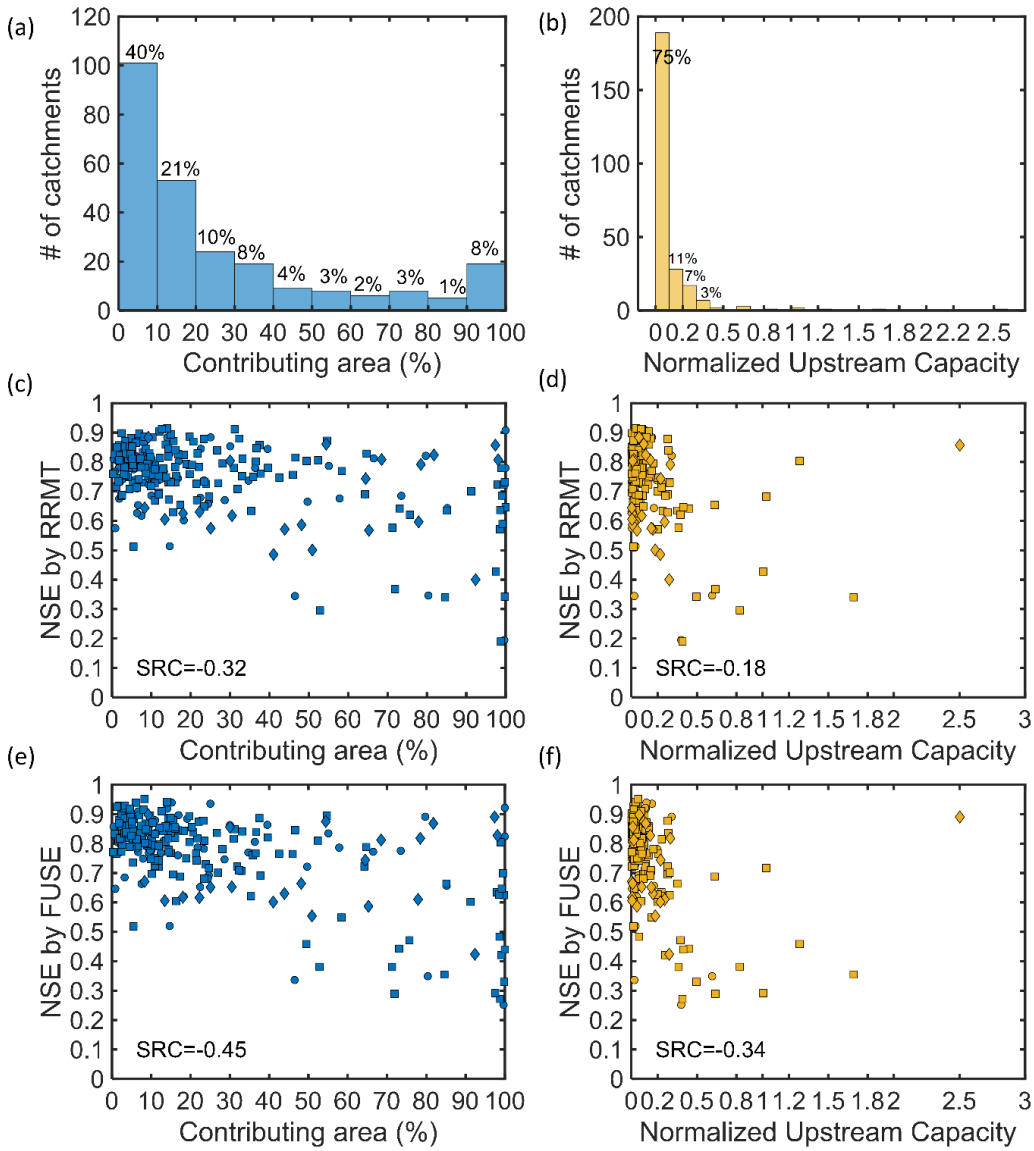


Figure B5.1. Histogram plots of (a) contributing area (%) and (b) normalized upstream capacity of 252 catchments. Scatter plot of (c) and (d) indicates highest NSE values obtained by RRMT model structures vs. contributing area and vs. normalized upstream, respectively. Same scatter plots with highest NSE values obtained by FUSE model structures are given in (e) and (f).

B6. Different thresholds for separation between RRMT model structures

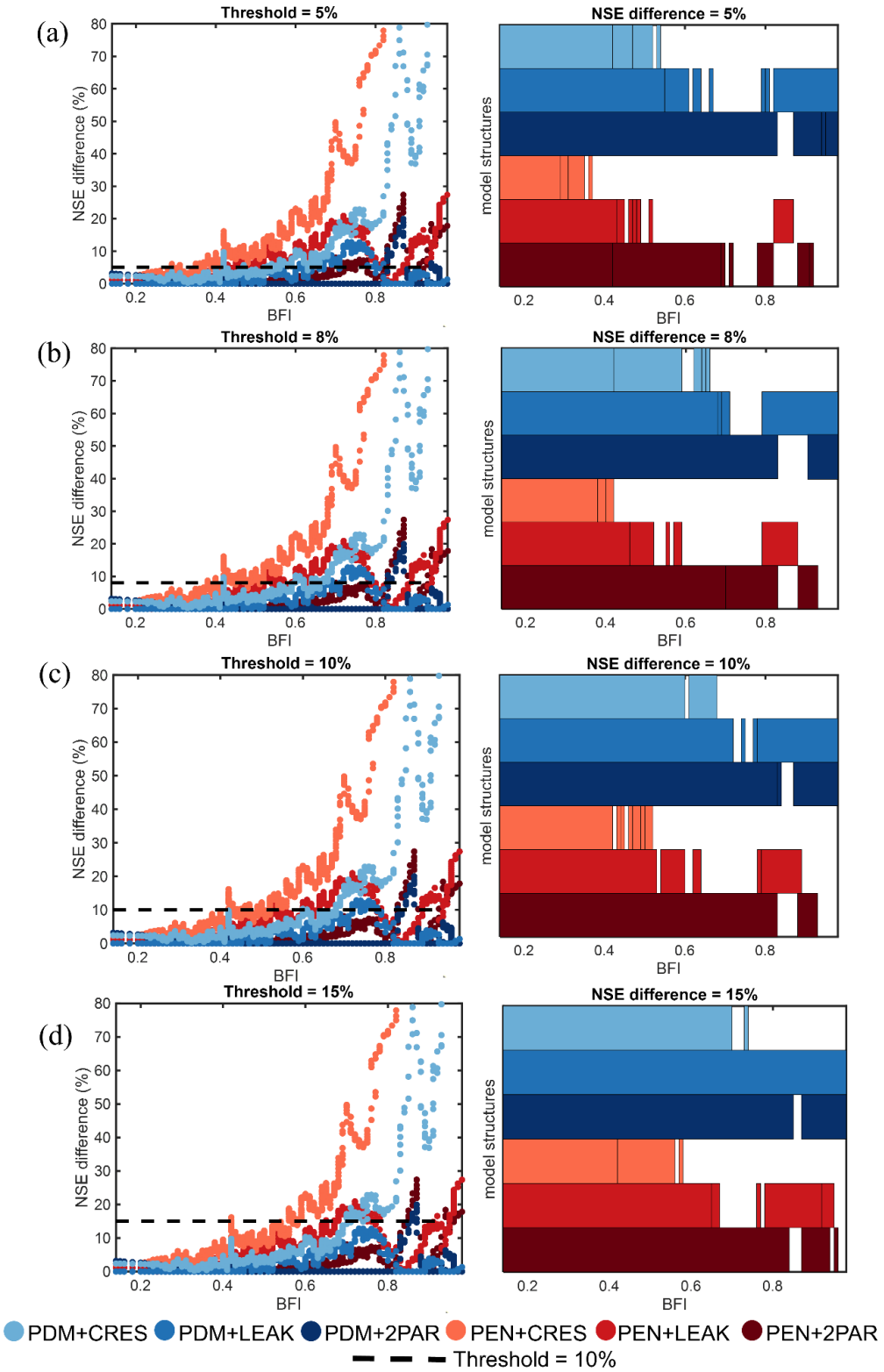


Figure B6.1. NSE difference (%) values and bar plots of six model structures (PEN+2PAR, PEN+LEAK, PEN+CRES, PDM+2PAR, PDM+LEAK, PDM+CRES) plotted against their BFI. Different thresholds (i.e. 5%, 8%, 10%, 15%) are tried and 10% is selected as the most reasonable threshold to decide which model structures performing enough in specific ranges of attributes by checking for every model structure.

B7. Separations between model structures of RRMT and FUSE frameworks based on KGE

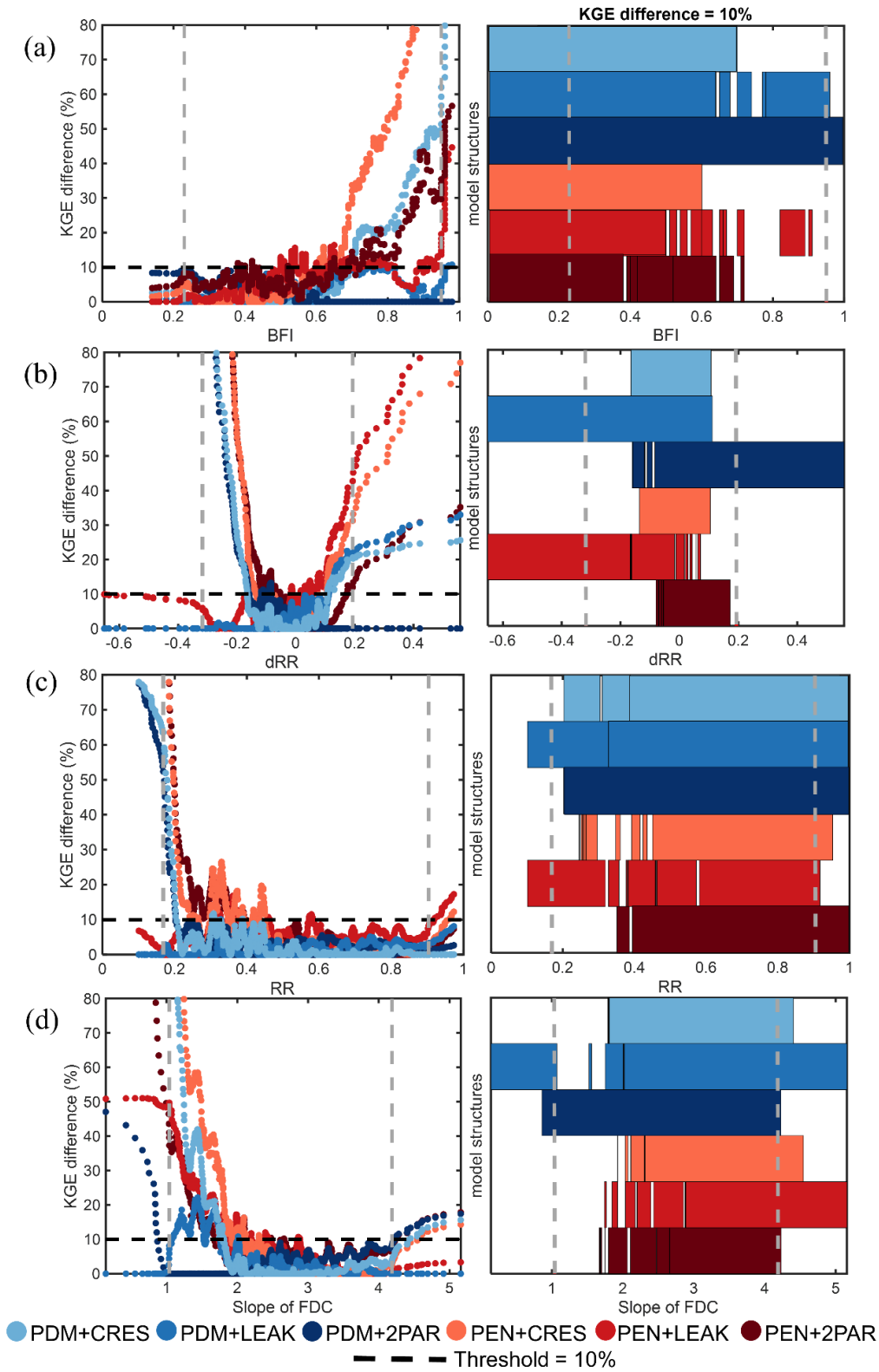


Figure B7.1. KGE difference (%) values and bar plots of six model structures (PEN+2PAR, PEN+LEAK, PEN+CRES, PDM+2PAR, PDM+LEAK, PDM+CRES) plotted against their

BFI (a), dRR (b), RR (c) and slope of FDC (d) attributes. KGE difference values are calculated by taking the difference between maximum KGE value obtained by any model structure and KGE values of remaining model structures and divided by maximum KGE value and multiply by 100 for every catchment. KGE values of model structures are obtained by moving means with 40 point - window size. Through visual inspection, 10% is selected as the most helpful threshold to show which model structure is performing differently in relation to a specific attribute. The range between two grey dashed vertical lines indicates the ranges where the smoothing is based on 20 left and right of the average calculated. Outside these ranges, points become increasingly biased by the points at the minimum and maximum signature values.

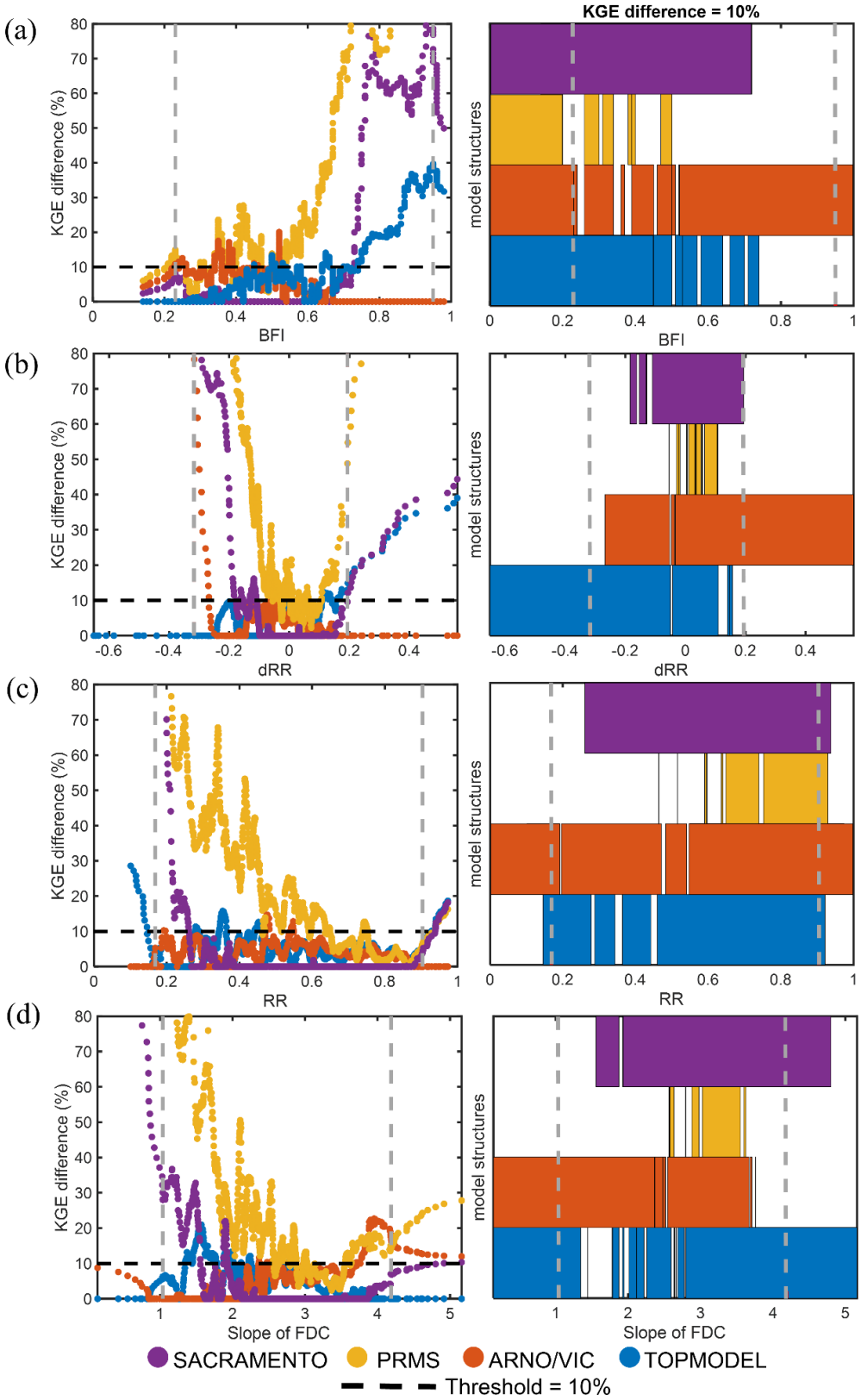


Figure B7.2. KGE difference (%) values and bar plots of four model structures (TOPMODEL, ARNO/VIC, PRMS, SACRAMENTO) plotted against their BFI (a), dRR (b), RR (c) and slope of FDC (d) attributes. KGE difference values are calculated by taking the difference between maximum KGE value obtained by any model structure and KGE values

of remaining model structures and divided by maximum KGE value and multiply by 100 for every catchment. KGE values of model structures are obtained by moving means with 40 point - window size. Through visual inspection, 10% is selected as the most helpful threshold to show which model structure is performing differently in relation to a specific attribute. Therefore, bar plots of four model structures are created by taking 10% as the KGE difference. The range between two grey dashed vertical lines indicates the ranges where the smoothing is based on 20 left and right of the average calculated. Outside these ranges, points become increasingly biased by the points at the minimum and maximum signature values.

B8. Relationship of BFI and RR with related catchment attributes (i.e. baseflow index values based on HOST classification (BFI-HOST) and Aridity Index (AI))

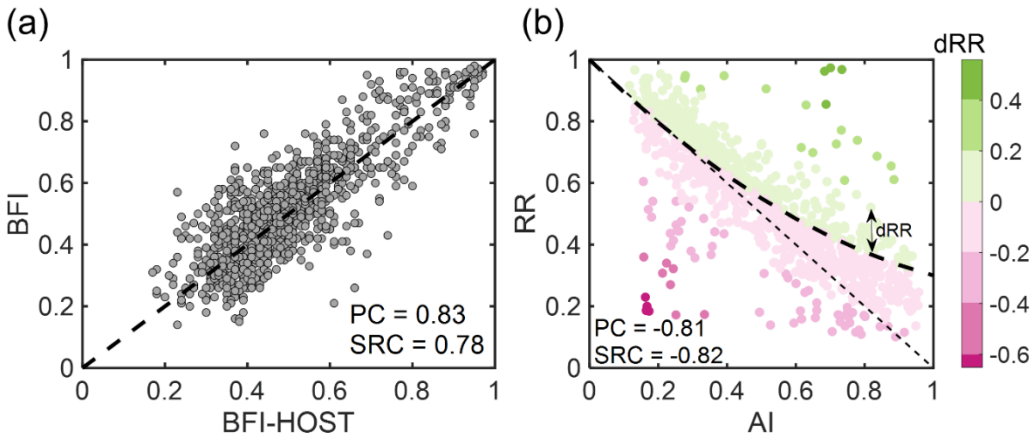


Figure B8.1. Scatter plots of (a) BFI vs BFI-HOST values and (b) RR vs AI values for 998 catchments. Dashed line in (a) is y=x. In (b), the thick black dashed curve is the Turc-Mezentsev Curve and dRR values for each catchment are calculated as the vertical difference between the observed RR and their corresponding points on the Turc-Mezentsev Curve. The thin dashed lines reflect energy and water limits While Pearson (PC) and Spearman rank correlation (SRC) values between BFI and BFI-HOST are 0.83 and 0.78, respectively, these are -0.81 and -0.82 between RR and AI.

B9. Water management activities for catchments with water balance issues

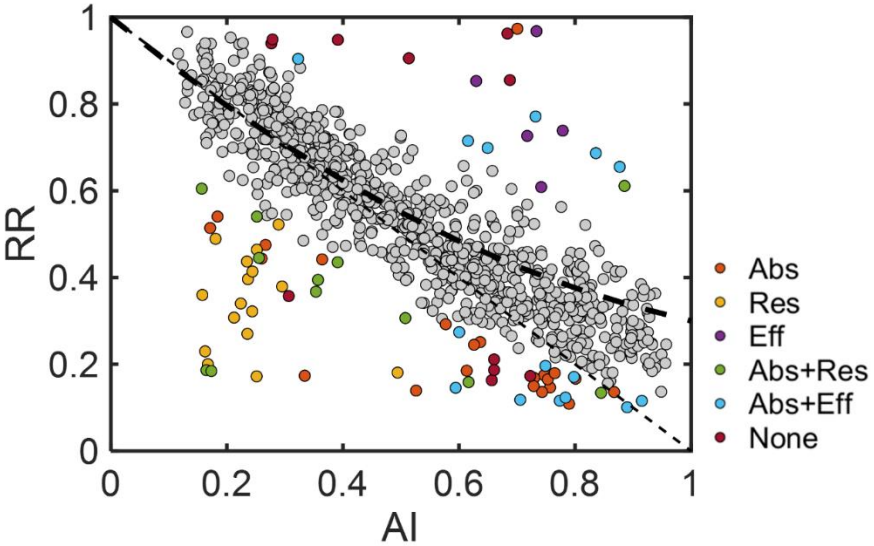
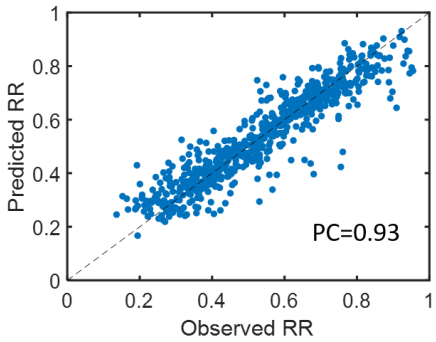
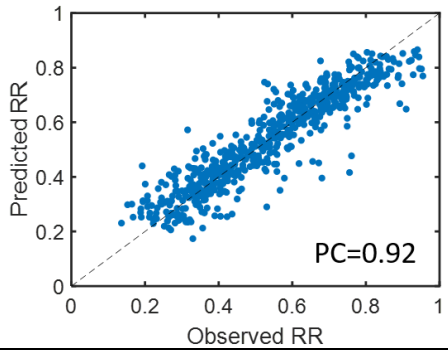
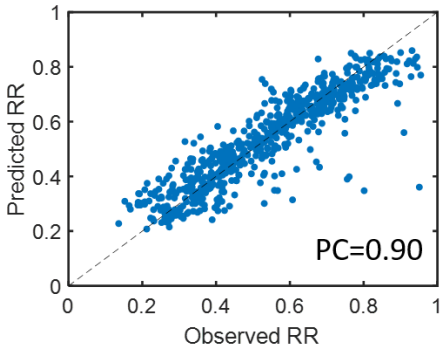
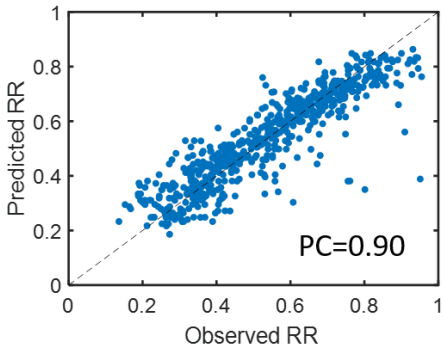


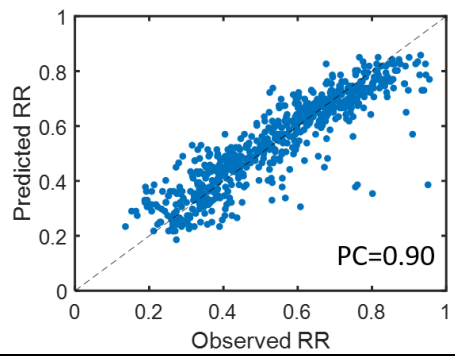
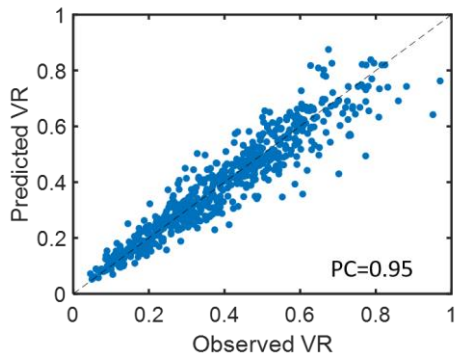
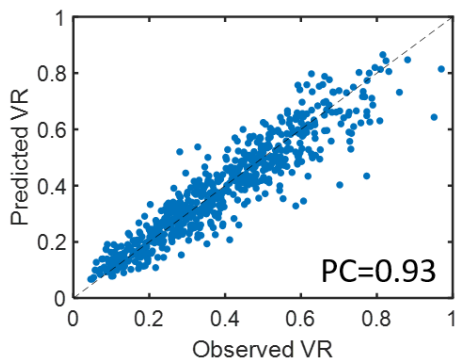
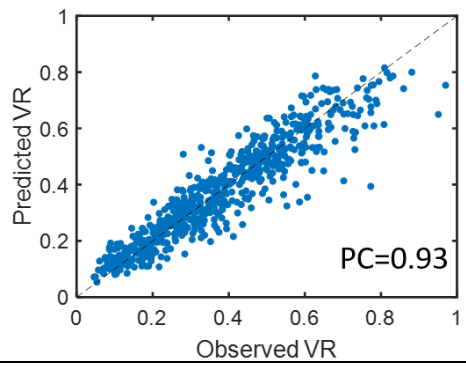
Figure B9.1. Water management practices effecting runoff of (a) 62 leaky catchments (i.e. dRR< -0.2) and (b) 19 gaining catchments (i.e. dRR >0.2). Red edged circles indicate Chalky catchments. Abs, res and eff represent abstraction (i.e. taking water out of surface water or groundwater for water supply or industrial, agricultural purposes), reservoir (i.e. the effect on river flow due to water storage or release in or above gauged catchment) and effluent returns (i.e. outflow from sewage treatment works augmenting river flow if effluent originates from outside of the catchment), respectively. None represents negligible artificial influence. The thick black dashed curve is the Turc-Mezentsev Curve. The thin dashed lines reflect energy and water limits.

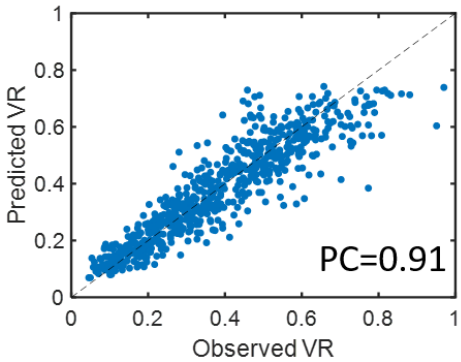
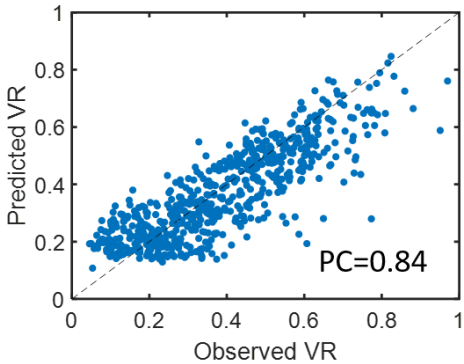
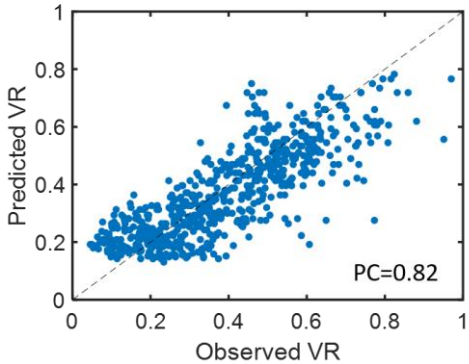
APPENDIX C Supplemental Material Chapter 4

C1. Stepwise Linear Regression Analysis

Table C1.1. *Different groups of catchment attributes used as predictors to predict RR and VR and the correlations between the predicted and observed values. Selected groups of predictors for RR and VR are shown as bold.*

Signature as response	Predictors	Correlation of predicted vs. observed signature values
RR	10 predictors (elev_min, aridity_index, frac_snow, dwood_perc, silt_perc, inter_mod_perc, low_nsig_perc, reservoir_he)	
	6 predictors (aridity_index, frac_snow, dwood_perc, BFI-HOST, urban_perc, frac_high_perc)	
	3 predictors (aridity_index, frac_snow and BFI_HOST)	
	2 predictors (aridity_index and frac_snow)	

	1 predictor (aridity_index)	
VR	19 predictors (elev_min, elev_50, pet_mean, aridity_index, frac_snow, high_prec_frequency, low_prec_dur, dwood_perc, inwater_perc, inter_high_perc, BFI-HOST etc.)	
	7 predictors (pet_mean, aridity_index, frac_snow, low_prec_dur, inwater_perc, inter_high_perc, BFI-HOST)	
	3 predictors (aridity_index, inwater_perc and BFI-HOST)	

	2 predictors (aridity_index and BFI-HOST)	
	2 predictors (aridity_index and inwater_perc)	
	1 predictor (aridity_index)	

C2. Some decisions made for quantification of correlation term for ungauged catchments

We checked how many nearest gauges should be taken as reference to estimate streamflow time series for ungauged catchments to calculate correlation term (see Section 3.2.). Since not much changed is observed in the percentage of catchments having Spearman Rank Correlation (SRC)>0.8 after three catchments, taking three nearest gauges is chosen as optimum option.

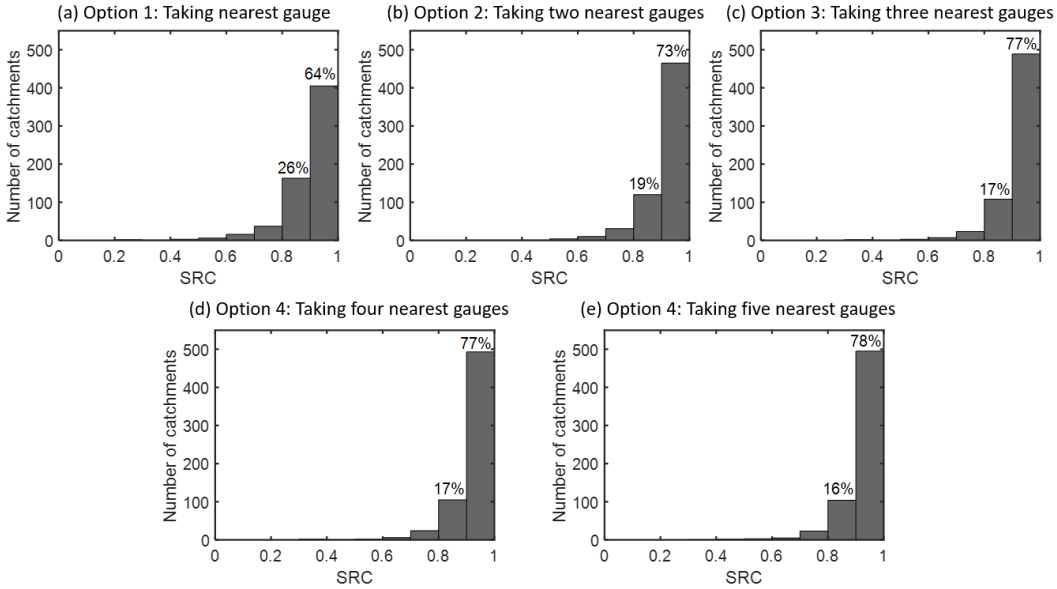


Figure C2.1. Five histogram graphs of Spearman Rank Correlation (SRC) between estimated and observed streamflow time series of 633 catchments. Estimated streamflow time series are calculated by taking inverse distance weighted interpolation of taking (a) nearest gauge, (b) nearest two gauges, (c) nearest three gauges, (d) nearest four gauges and (e) nearest five gauges as reference for ungauged cases.

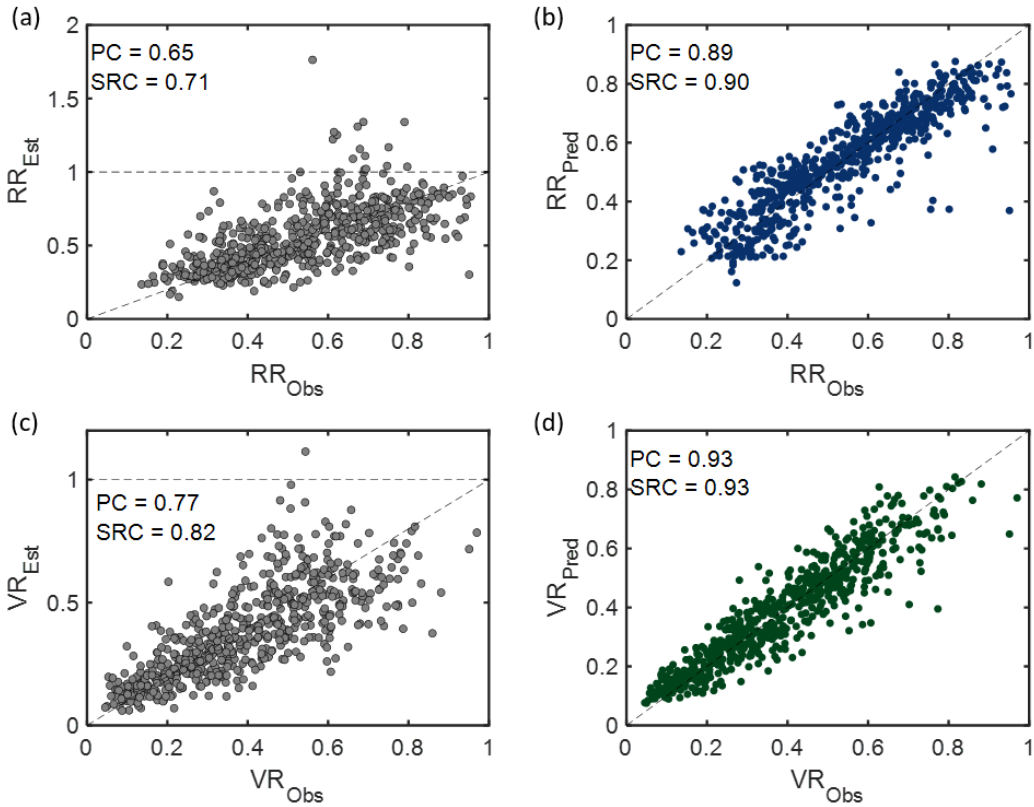


Figure C2.2. Scatter plots of observed RR (i.e. RR_{Obs}) vs. (a) estimated RR (i.e. RR_{Est}) using estimated streamflow time series by inverse distance weighted interpolation of nearest three catchments' streamflow, (b) predicted RR values (RR_{Pred}) by stepwise linear regression analysis using aridity index as predictor and scatter plot of observed VR (i.e. VR_{Obs}) vs. (a) estimated VR (i.e. VR_{Est}) using estimated streamflow time series by inverse distance weighted interpolation of nearest three catchments' streamflow, (b) predicted VR values (VR_{Pred}) by stepwise linear regression analysis using aridity index, baseflow index (BFI-HOST) and inland water percentage as predictors.

C3. Inclusion of leaky catchments

There are 26 leaky catchments (i.e. ones having high water balance errors) that are eliminated from study. In order to specify these catchments, we calculate their expected water balance (i.e. expected runoff ratio (RR)) based on only climate using the Turc-Mezentsev curve. Turc-Mezentsev curve provides the relationship between long term the long-term average evaporation to long-term average precipitation (Turc, 1955; Mezentsev, 1955). Since measurements of actual evapotranspiration are not available at the catchment scale, we adjust the formula using 1- (Q/P) as a response term instead of AET/P as used in the original formulation. The formula that we use to create Turc-Mezentsev Curve is therefore;

$$1 - \frac{Q}{P} = \frac{1}{\left[1+\left(\frac{P}{PET}\right)^2\right]^{\frac{1}{2}}} \qquad (C3.1)$$

The water balance errors, i.e. delta runoff ratio (dRR) values, of catchments are calculated as the difference between their observed and expected RR (i.e. Q/P) derived from the Turc-Mezentsev Curve formula. If catchments have negative dRR values, it means that their observed RR is less than expected RR and they are likely losing water.

In addition to PDM+2PAR, we also conducted model simulation using PDM+LEAK model structure for all 659 catchments. PDM+LEAK consists of same soil moisture accounting component with a leaky aquifer routing component, which allows the model to consider the situation when the water balance of a catchment is not closed. The flow from the bottom outlet represents leakage from the catchment, while the middle and upper outlets contribute to routing the effective rainfall. Model structures are visualized in Figure C3.5. SHE values obtained by PDM+2PAR are compared to ones obtained by PDM+LEAK based on dRR values as shown in Figure C3.3. It demonstrates that PDM+LEAK model structures works better than PDM+2PAR mainly in catchments where dRR values less than –0.2. Therefore, 26 catchments having dRR<-0.2 (Figure C3.4) are assumed to be leaky catchments and eliminated from the results in the main chapter to simply the study by using only one model

structure. Even though 26 leaky catchments are eliminated, the results including them (i.e. 659 GB catchments) are given in Figure C3.6, C3.7 and C3.8.

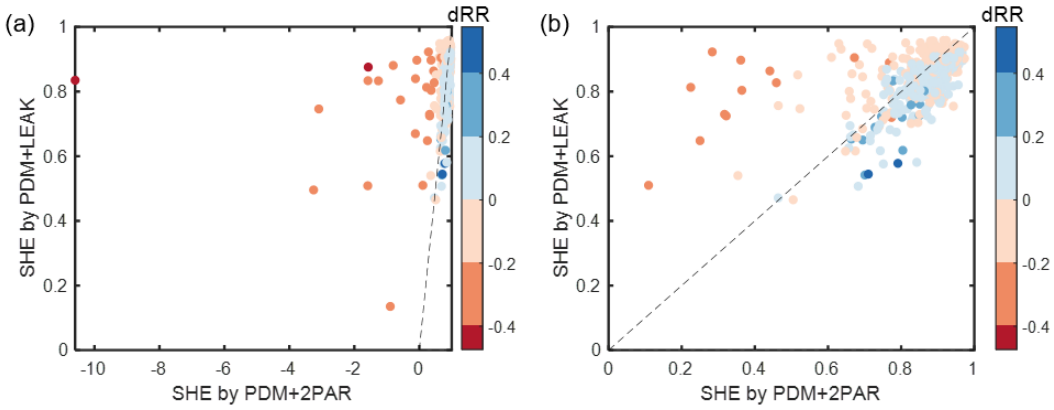


Figure C3.3. Scatter plots of SHE values obtained by PDM+2PAR and PDM+LEAK and color-coded by dRR values. The original plot and its version of being x-axis limited to [0 1] are given in (a) and (b), respectively.

Figure C3.4. 659 CAMEL-GB catchments shown in GB map. While 26 leaky catchments are shown in dark blue filled circles, 633 remaining catchments indicated as grey circles. While 633 GB catchments are used to simply the study by using only one model structure

to produce results, results with 659 catchments are also presented in supplemental material by using another model structure for 26 leaky catchments.

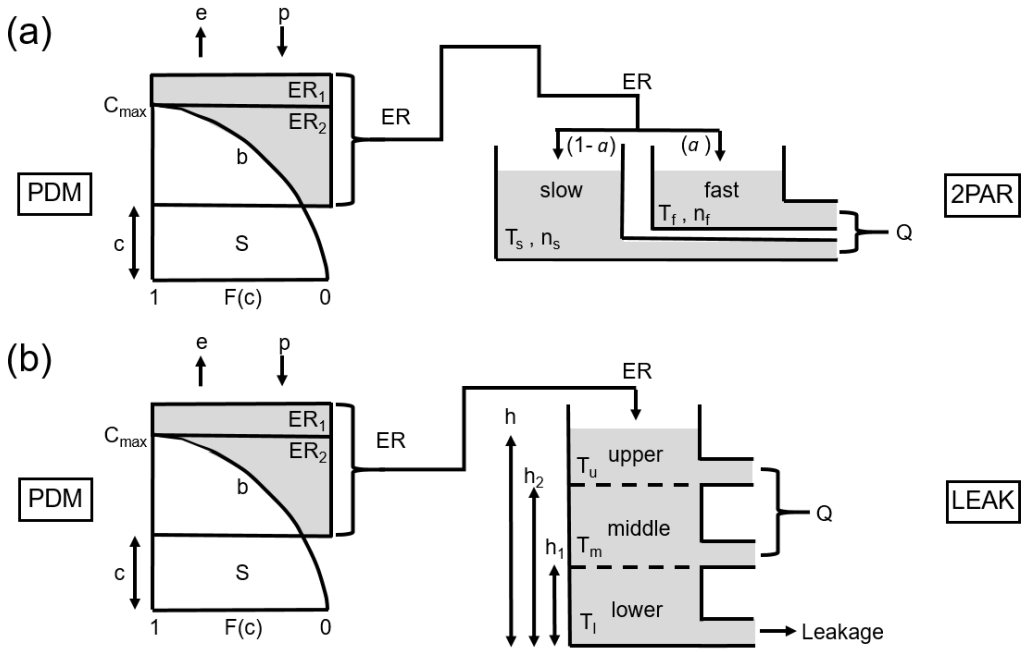


Figure C3.5. Visualization of model structures used: a) PDM+2PAR and b) PDM+LEAK

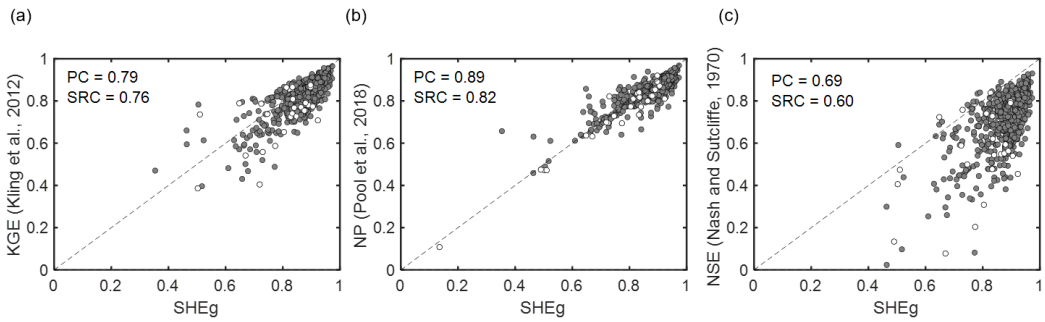


Figure C3.6. Scatter plots for (a) KGE vs. SHE, (b) NP vs. SHE and (c) NSE vs. SHE. x and y axes are limited to [0 1]. 659 GB catchments are used in these results.

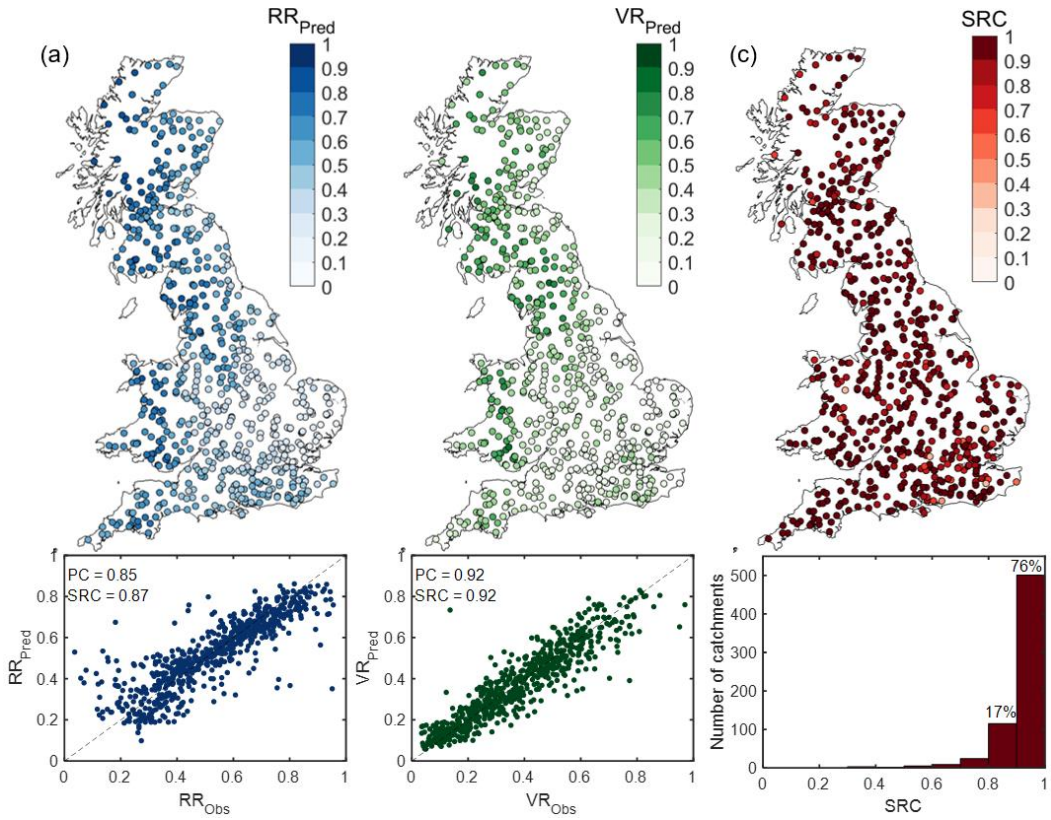


Figure C3.7. (a) Predicted RR map and scatter plot for predicted vs. observed RR, (b) predicted VR map and scatter plot for predicted vs. observed VR and (c) map illustrating SRC values between observed streamflow of catchments and the streamflow values calculated by taking inverse distance interpolation of their closest three catchments' observed streamflows and its histogram plot. Predictor of RR is aridity index and predictors of VR are aridity index, BFI-HOST and inland water percentage. 659 GB catchments are used in these results.

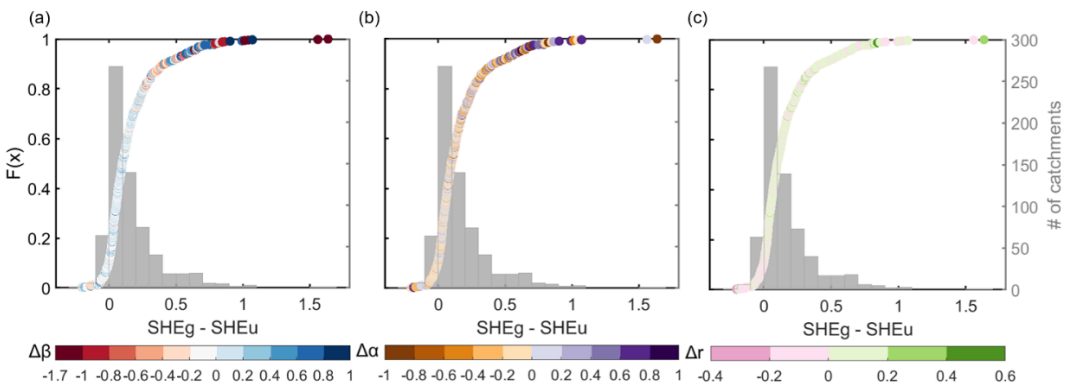


Figure C3.8. Cumulative distribution function (i.e. cdf) plot and histogram plot of difference between SHE(gauged) and SHE(ungauged) values (i.e. SHE(gau) – SHE(ung)). Cdf plot is color-coded by (a) bias component difference (B), (b) variance component difference (V) and (c) correlation component difference (C) between SHE(gauged) and SHE(ungauged) formulations summarized in Table 1. Histogram of SHE(gau)-SHE(ung) is also shown on the figure. 659 GB catchments are used in these results.

APPENDIX D Curriculum Vitae

PERSONAL INFORMATION

Melike Kiraz

(+90) 530 545 06 99

melike.kiraz@bristol.ac.uk

OUTLOOK

My primary research focus is centered around addressing the challenges related to planning and managing the sustainable provision of water for societies. I aim to achieve this by employing modelling techniques and gaining a deep understanding of the processes involved in large-sample hydrology. Through my work, I aspire to support decision-makers in developing comprehensive and coherent strategies for sustainable water management that cater to the needs of both society and the environment. By combining my expertise in hydrological modelling and a holistic approach to water resource management, I strive to contribute to the development of effective and sustainable solutions in this field.

EDUCATION

09/2018 – Present

Doctor of Philosophy – Hydrology

University of Bristol, Bristol (United Kingdom)

- Thesis Title: Addressing the Challenges of Catchment Characterisation, Model Selection and Evaluation in Large-Sample Hydrology: Application to Great Britain

09/2016 – 08/2018

Master Degree - Graduate School of Applied Sciences - Department of Environmental Engineering

Middle East Technical University, Ankara (Turkey)

- Thesis Title: Sustainable Water and Stormwater Management for METU Campus

09/2011 – 06/2016

Bachelor Degree - Environmental Engineering

Middle East Technical University, Ankara (Turkey)

- Key Modules: Environmental Engineering Chemistry Lab, Thermodynamics, Environmental Microbiology, Water Supply and Urban Drainage, Water Quality Management, Treatment Disposal of Water Wastewater Sludge

WORK EXPERIENCE

09/2016 – 03/2018	Project Assistant Middle East Technical University, Ankara (Turkey) ▪ Department of Environmental Engineering
08/2015 – 09/2015	Intern Engineer Turkish Scientific and Technological Research Council (TUBITAK) (Turkey)
06/2014 – 07/2014	Intern Engineer ICDAS Iron - Steel Industry Inc. (Turkey)

SKILLS

Software	Microsoft, ArcMap, WEAP, SWAT, SWMM
Programming	MATLAB (proficient)
Machine Learning	Regression, Rainfall-runoff modelling, Monte-Carlo sampling
Communication	Academic writing, figures, presentations, posters

PUBLICATIONS

Under review	Kiraz, M., Coxon, G. and Wagener, T. (2023). A priori selection of hydrological model structures in modular modelling frameworks: Application to Great Britain. Hydrological Sciences Journal.
	Kiraz, M., Coxon, G. and Wagener, T. (2023). A signature-based hydrologic efficiency metric for model calibration and evaluation in gauged and ungauged catchments. Water Resources Research.
In preparation	Kiraz, M., Coxon, G., Rahman, M. and Wagener, T. (2023). Location, location, location – Considering relative catchment location to understand subsurface losses.
Published	Alp, E., Erdoğan, E. K. and Kiraz, M. (2023). The NEXUS graduate. In Zamel, D., Constantianos, V., Gawlik, B., Laspidou, C., Abadi, A., Easton, P., Berman, R., Kazezyilmaz-Alhan, C., Głowacka, N., Elelman, R., (Eds.), The gateway to the future of the Mediterranean – Water, energy, food and the environment. (pp. 138-139). Publications Office of the European Union. https://data.europa.eu/doi/10.2760/38718

ADDITIONAL INFORMATION

Honors and awards

- Honor Student as ranked 2nd amongst 71 students, METU Environmental Engineering Department,06/2016

- Best 2nd Design of METU Environmental Engineering Department Dissertation- Bolu Gerede OIZ Treatment Plant Design, 06/ 2016

- 1st Prize of German Water Partnership Award Turkey 2016 - Sustainable Water Management and the Water-Energy Nexus in Middle East Technical University, 10/ 2016

Projects

- Preparation of Sludge Management Plan and Action Plan in Turkey, Ministry of Environment and Urbanization, 2016-2019, Principal Investigator: Prof. Dr. Dilek Sanin

- A Green Campus Application: Sustainable Stormwater Management in METU Campus, METU Research Fund (Project No: BAP - 08 -1 - KB2014K120600-2), 2014-2016, 47,000 TL, PrincipalInvestigator: Assoc. Prof. Dr. Emre Alp

- Evaluation of Agricultural Diffuse Pollution and its Control Alternatives with SWAT model in Lake Mogan Watershed, The Scientific and Technological Research Council of Turkey (Project No: 1 1Y284), 2012-2014, $78,000, Principal Investigator: Assoc. Prof. Dr. Emre Alp

- Water Quality Evaluation and Determination of Pollution Load Discharging into Lake Mogan, METU Research Fund (Project No: BAP-08-1 -2013-008), 2013-2014, 20,000 TL, Principal Investigator: Assoc. Prof. Dr. Emre Alp

Conferences

–IWA Balkan Young Water Professionals, 05/2015, Thessaloniki, Greece

- The 6th Turkish-German Water Partnership-Day, 10/ 2016, Mersin,Turkey
- EGU General Assembly, 04/2019, Vienna, Austria
- Vienna Catchment Science Symposium, 04/2019, Vienna, Austria
- EGU General Assembly, 05/2021, Online

- Nordic Hydrologic Conference, 08/2022, Tallinn, Estonia

Teachings	– Assisted as Teaching Assistant in module Scheme Design 3, 10/2019 - Assisted with marking MATLAB in module Numerical Methods, 10/ 2019 - Assisted as Teaching Assistant in module Water Resources Project 3, 10/2020
Voluntary Activities	- Being a voluntary member of Legambiente, international voluntary workcamp in Pian Di Spagna Reserve, 08/2015, Sorico, Italy - Being a member of organization team of "Quality through equality – tackling gender issues in hydrology" workshop, 02/2019 - Being a guest author of the blog titled "Quality through equality – tackling gender issues in hydrology" in EGU HS Blog, 07/2019 - Participation in organization team of Equality in Engineering Workshop, 11/2020

REFERENCES

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293-5313.

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. *Water Resources Research*, *54*(11), 8792-8812.

Addor, N., & Melsen, L. A. (2019). Legacy, rather than adequacy, drives the selection of hydrological models. *Water resources research*, *55*(1), 378-390.

Addor, N., 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.

Addor, N., 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.

Almeida, S., Le Vine, N., McIntyre, N., Wagener, T., & Buytaert, W. (2016). Accounting for dependencies in regionalized signatures for predictions in ungauged catchments. *Hydrology and Earth System Sciences*, *20*(2), 887-901.

Allen, D. J., Brewerton, L. J., Coleby, L. M., Gibbs, B. R., Lewis, M. A., MacDonald, A. M., Wagstaff, S. J., & Williams, A. T. (1997). The physical properties of major aquifers in England and Wales. *British Geological Survey Technical Report WD/97/34*, pp. 312, Environment Agency R&D Publication 8.

Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. *Fao, Rome*, *300*(9), D05109.

Allen, D. J., & Crane, E. J. (2019). The chalk aquifer of the Wessex Basin. British Geological Survey Research Report No. RR/11/02. 118pp.

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

Archfield, S. A., & Vogel, R. M. (2010). Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments. *Water resources research*, *46*(10).

Arnold, J. G., Srinivasan, R., Muttiah, R. S. & Williams, J. R. (1998). Large Area Hydrologic Modeling and Assessment Part 1: Model Development. *Journal of American Water Resources Association*, 34 (1), pp.73-89

Atkinson, S. E., Woods, R. A., & Sivapalan, M. (2002). Climate and landscape controls on water balance model complexity over changing timescales. *Water Resources Research*, *38*(12), 50-1.

Bai, Y., Wagener, T., & Reed, P. (2009). A top-down framework for watershed model evaluation and selection under uncertainty. *Environmental Modelling & Software*, *24*(8), 901-916.

Basso, S., & Botter, G. (2012). Streamflow variability and optimal capacity of run-of-river hydropower plants. *Water Resources Research*, *48*(10).

Bathurst, J. C., Ewen, J., Parkin, G., O'Connell, P. E., & Cooper, J. D. (2004). Validation of catchment models for predicting land-use and cli-mate change impacts. 3. Blind validation for internal and outlet responses. *Journal of Hydrology*,287(1–4), 74–94.

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I., & Wood, E. F. (2020). Global fully distributed parameter regionalization based on observed streamflow from 4,229

headwater catchments. *Journal of Geophysical Research: Atmospheres*, 125(17), e2019JD031485

BEIS. (2021). UK Energy in Brief 2021. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1032260/UK_Energy_in_Brief_2021.pdf

Bell, V. A., Kay, A. L., Rudd, A. C., & Davies, H. N. (2018). The MaRIUS-G2G datasets: Grid-to-Grid model estimates of flow and soil moisture for Great Britain using observed and climate model driving data. *Geoscience Data Journal*, *5*(2), 63-72.

Belvederesi, C., Zaghloul, M. S., Achari, G., Gupta, A., & Hassan, Q. K. (2022). Modelling river flow in cold and ungauged regions: A review of the purposes, methods, and challenges. *Environmental Reviews*, *30*(1), 159-173.

Berghuijs, W. R., Aalbers, E. E., Larsen, J. R., Trancoso, R., & Woods, R. A. (2017). Recent changes in extreme floods across multiple continents. *Environmental Research Letters*, *12*(11), 114035.

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological sciences journal*, *24*(1), 43-69.

Beven, K. J. (2000). Uniqueness of place and process representations in hydrological modelling. *Hydrology and earth system sciences*, *4*(2), 203-213.

Beven, K. J. (2001). *Rainfall-runoff modelling: the primer*. John Wiley & Sons.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of hydrology*, *320*(1-2), 18-36.

Beven, K. (2019). How to make advances in hydrological modelling. *Hydrology Research*, *50*(6), 1481-1494.

Beven, K. J., & Chappell, N. A. (2021). Perceptual perplexity and parameter parsimony. *Wiley Interdisciplinary Reviews: Water*, *8*(4), e1530.

BGS. (2022). National Groundwater Level Archive. Retrieved from https://www2.bgs.ac.uk/groundwater/datainfo/levels/ngla.html.

Birkel, C., & Barahona, A. C. (2019). Rainfall-runoff modeling: a brief overview. *Reference Module in Earth Systems and Environmental Sciences*.

Blair, G. S., Beven, K., Lamb, R., Bassett, R., Cauwenberghs, K., Hankin, B., ... & Towe, R. (2019). Models of everywhere revisited: A technological perspective. *Environmental Modelling & Software*, *122*, 104521.

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., & Savenije, H. H. G. (2013). Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales, Cambridge University Press, Cambridge.

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., ... & Renner, M. (2019). Twenty-three unsolved problems in hydrology (UPH)–a community perspective. *Hydrological sciences journal*, *64*(10), 1141-1158.

Bloomfield, J. P., Gong, M., Marchant, B. P., Coxon, G., & Addor, N. (2021). How is Baseflow Index (BFI) impacted by water resource management practices?. *Hydrology and Earth System Sciences*, *25*(10), 5355-5379.

Bouaziz, L., Weerts, A., Schellekens, J., Sprokkereef, E., Stam, J., Savenije, H., & Hrachowitz, M. (2018). Redressing the balance: quantifying net intercatchment groundwater flows. *Hydrology and Earth System Sciences*, *22*(12), 6415-6434.

Bradford, R. B. (2002). Controls on the discharge of Chalk streams of the Berkshire Downs, UK. *Science of the Total Environment*, *282*, 65-80.

Brauman, K. A., Daily, G. C., Duarte, T. K. E., & Mooney, H. A. (2007). The nature and value of ecosystem services: an overview highlighting hydrologic services. *Annu. Rev. Environ. Resour.*, *32*, 67-98.

Brunner, M. I., Melsen, L. A., Newman, A. J., Wood, A. W., & Clark, M. P. (2020). Future streamflow regime changes in the United States: assessment using functional classification. *Hydrology and Earth System Sciences*, *24*(8), 3951-3966.

Brunner, M. I., Slater, L., Tallaksen, L. M., & Clark, M. (2021). Challenges in modeling and predicting floods and droughts: A review. *Wiley Interdisciplinary Reviews: Water*, *8*(3), e1520.

Budyko, M. I. (1961). The heat balance of the earth's surface. *Soviet Geography*, 2(4), pp. 3-13.

Bulygina, N., McIntyre, N., & Wheater, H. (2009). Conditioning rainfall-runoff model parameters for ungauged catchments and land management impacts analysis. *Hydrology and Earth System Sciences*, *13*(6), 893-904.

Burn, D. H., & Elnur, M. A. H. (2002). Detection of hydrologic trends and variability. *Journal of hydrology*, *255*(1-4), 107-122.

Burnash, R. J., & Ferral, R. L. (1973). *A generalized streamflow simulation system: Conceptual modeling for digital computers*. US Department of Commerce, National Weather Service, and State of California, Department of Water Resources.

Carvalho-Santos, C., Nunes, J. P., Monteiro, A. T., Hein, L., & Honrado, J. P. (2016). Assessing the effects of land cover and future climate conditions on the provision of hydrological services in a medium-sized watershed of Portugal. *Hydrological processes*, *30*(5), 720-738.

Casagrande, E., Recanati, F., Rulli, M. C., Bevacqua, D., & Melia, P. (2021). Water balance partitioning for ecosystem service assessment. A case study in the Amazon. *Ecological Indicators*, *121*, 107155

Centre for Ecology and Hydrology. (1999). Flood Estimation Handbook, Volume 5, Institute of Hydrology.

Centre for Ecology and Hydrology. (2020). *National River Flo*w Archive. Available from: http://nrfa.ceh.ac.uk/ [Accessed 16 August 2021].

Chagas, V. B., Chaffe, P. L., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C., & 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.

Cheng, L., Yaeger, M., Viglione, A., Coopersmith, E., Ye, S., & Sivapalan, M. (2012). Exploring the physical controls of regional patterns of flow duration curves–Part 1: Insights from statistical analyses. *Hydrology and Earth System Sciences*, *16*(11), 4435-4446.

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

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, 47(9), 1–16.

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

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., ... & Marks, D. G. (2015b). A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies. *Water resources research*, *51*(4), 2515-2542.

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, e2020WR029001. https:// doi.org/10.1029/2020WR029001

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, *74*(368), 829-836.

Collins, S. L., Christelis, V., Jackson, C. R., Mansour, M. M., Macdonald, D. M., & Barkwith, A. K. (2020). Towards integrated flood inundation modelling in groundwater-dominated catchments. *Journal of Hydrology*, *591*, 125755.

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), pp. 6602-6621.

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., & Clark, M. (2014). Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments. *Hydrological Processes*, *28*(25), 6135-6150.

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., & Smith, P. J. (2015). A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water resources research*, *51*(7), 5531-5546.

Coxon, G.; Freer, J.; Lane, R.; Dunne, T.; Knoben, W.J.M.; Howden, N.J.K.; Quinn, N.; Wagener, T.; Woods, R. (2019). DECIPHeR model estimates of daily flow for 1366 gauged catchments in Great Britain (1962-2015) using observed driving data. NERC Environmental Information Data Centre. (Dataset). https://doi.org/10.5285/d770b12a-3824-4e40-8da1-930cf9470858

Coxon, G.; Addor, N., Bloomfield, J.P., Freer, J., Fry, M., Hannaford, J., Howden, N.J.K., Lane, R., Lewis, M., Robinson, E.L., Wagener, T., & Woods, R. (2020a). Catchment attributes and hydro-meteorological timeseries for 671 catchments across Great Britain

(CAMELS-GB). *NERC Environmental Information Data Centre*. https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9.

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., ... & Woods, R. (2020b). 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., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., ... & Tolson, B. A. (2020). Flexible watershed simulation with the Raven hydrological modelling framework. *Environmental Modelling & Software*, *129*, 104728.

Crawford, N.H. and R.K. Linsley, 1966. Digital Simulation in Hydrology. In: Contemporary Hydrology, Wilby, R. (Ed.). John Wiley and Sons, England, pp: 157-158.

Dal Molin, M. D., Kavetski, D., Albert, C., & Fenicia, F. (2023). Exploring Signature-Based Model Calibration for Streamflow Prediction in Ungauged Basins. *Water Resources Research*, e2022WR031929.

David, P. C., Chaffe, P. L., Chagas, V. B., Dal Molin, M., Oliveira, D. Y., Klein, A. H., & Fenicia, F. (2022). Correspondence Between Model Structures and Hydrological Signatures: A Large-Sample Case Study Using 508 Brazilian Catchments. *Water Resources Research*, *58*(3), e2021WR030619.

Dawdy, D. R., & O'Donnell, T. (1965). Mathematical models of catchment behavior. *Journal of the Hydraulics Division*, *91*(4), 123-137.

Dawson, C. W., Abrahart, R. J., Shamseldin, A. Y., & Wilby, R. L. (2006). Flood estimation at ungauged sites using artificial neural networks. *Journal of Hydrology*, 319(1-4), pp. 391-409.

de la Vega-Leinert, A. C., & Nicholls, R. J. (2008). Potential implications of sea-level rise for Great Britain. *Journal of Coastal Research*, *24*(2), 342-357.

Dey, P., & Mishra, A. (2017). Separating the impacts of climate change and human activities on streamflow: A review of methodologies and critical assumptions. *Journal of Hydrology*, *548*, 278-290.

Drogue, G. P., & Plasse, J. (2014). How can a few streamflow measurements help to predict daily hydrographs at almost ungauged sites? *Hydrological Sciences Journal*, *59*(12), 2126-2142.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Wood, E. F. (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.

Durant, M. J., & Counsell, C. J. (2018). Inventory of reservoirs amounting to 90% of total UK storage. *NERC Environmental Information Data Centre, Wallingford*. (Dataset). https://doi.org/10.5285/f5a7d56c-cea0-4f00-b159-c3788a3b2b38

Dutta, R., & Maity, R. (2021). Time-varying network-based approach for capturing hydrological extremes under climate change with application on drought. *Journal of Hydrology*, *603*, 126958.

Ebel, B. A., & Loague, K. (2006). Physics-based hydrologic-response simulation: Seeing through the fog of equifinality. *Hydrological Processes: An International Journal*, *20*(13), 2887-2900.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S. & Savenije, H. H. G. (2013). A framework to assess the realism of model structures using hydrological signatures. Hydrology and Earth System Sciences, 17(5), pp.1893-1912.

Falkenmark, M., & Chapman, T. (1989). *Comparative hydrology: An ecological approach to land and water resources*. Unesco.

Fan, Y. (2019). Are catchments leaky? *Wiley Interdisciplinary Reviews: Water*, *6*(6), e1386.

Fang K., Sivakumar B., & Woldemeskel F. M. (2017). Complex networks, community structure, and catchment classification in a large-scale river basin. *Journal of Hydrology*, pp. 545:478–493, doi: https://doi.org/10.1016/j.jhydrol.2016.11.056

Fenicia, F., Kavetski, D., & Savenije, H. H. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research*, *47*(11).

Fenicia, F., & McDonnell, J. J. (2022). Modeling streamflow variability at the regional scale:(1) perceptual model development through signature analysis. *Journal of Hydrology*, *605*, 127287.

Frisbee, M. D., Tysor, E. H., Stewart-Maddox, N. S., Tsinnajinnie, L. M., Wilson, J. L., Granger, D. E., & Newman, B. D. (2016). Is there a geomorphic expression of interbasin groundwater flow in watersheds? Interactions between interbasin groundwater flow, springs, streams, and geomorphology. *Geophysical Research Letters*, *43*(3), pp. 1158-1165.

Fowler, K. J., 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.

Gale, I., & Rutter, H. (2006). *The chalk aquifer of Yorkshire*. British Geological Survey Research Report, RR/06/04. 68pp.

Gebrechorkos, S. H., Taye, M. T., Birhanu, B., Solomon, D., & Demissie, T. (2023). Future changes in climate and hydroclimate extremes in East Africa. *Earth's Future*, *11*(2), e2022EF003011.

Genereux, D. P., Jordan, M. T., & Carbonell, D. (2005). A paired-watershed budget study to quantify interbasin groundwater flow in a lowland rain forest, Costa Rica. *Water Resources Research*, *41*(4).

Genereux, D. P., & Jordan, M. (2006). Interbasin groundwater flow and groundwater interaction with surface water in a lowland rainforest, Costa Rica: a review. *Journal of Hydrology*, *320*(3-4), 385-399.

Gershenfeld, N., (1999). The nature of mathematical modelling. Cambridge University Press.

Giani, G., Rico-Ramirez, M. A., & Woods, R. A. (2021). A practical, objective, and robust technique to directly estimate catchment response time. *Water Resources Research*, *57*(2), e2020WR028201.

Gnann, S. J., Woods, R. A., & Howden, N. J. (2019). Is there a baseflow Budyko curve?. *Water Resources Research*, *55*(4), 2838-2855.

Grayson, R. B., Moore, I. D., & McMahon, T. A. (1992). Physically based hydrologic modeling: 2. Is the concept realistic?. *Water resources research*, *28*(10), 2659-2666.

Griffiths, J., Binley, A., Crook, N., Nutter, J., Young, A., & Fletcher, S. (2006). Streamflow generation in the Pang and Lambourn catchments, Berkshire, UK. *Journal of Hydrology*, *330*(1-2), 71-83.

Gupta, H. V., Wagener, T., and Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. Hydrological Processes, 22(18), 3802– 3813.

Gupta, H.V. et al., (2009). Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. Journal of Hydrology, 377 (1–2), 80–91. doi: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.

Guo, W., & Peng, Z. (2019). Hydropower system operation stability considering the coupling effect of water potential energy in surge tank and power grid. *Renewable energy*, *134*, 846-861.

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), e2019WR026752.

Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. *Wiley Interdisciplinary Reviews: Water*, *8*(1), e1487.

Höge, M., Wöhling, T., & Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. *Water Resources Research*, *54*(3), 1688-1715.

Hatchard, S. (2021). Hydropower and its Environmental Impacts: Quantifying Trade-offs in Data Scarce Regions. University of Bristol, Bristol, UK.

He, Y., Bárdossy, A., & Zehe, E. (2011). A review of regionalisation for continuous streamflow simulation. *Hydrology and Earth System Sciences*, *15*(11), 3539-3553.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., ... & Gascuel-Odoux, C. (2014). Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. Water resources research, 50(9), 7445-7469.

Hogue, T., Wagener, T., Shaake, J., Duan, Q., Hall, A., Gupta, H., ... & Andreassian, V. (2004). Model parameter experiment begins new phase.

Horton, R. E. (1945), "Erosional development of streams and their drainage basins: hydro-physical approach to quantitative morphology", *Geological Society of America Bulletin*, 56 (3): 275–370.

Horton, P., Schaefli, B., & Kauzlaric, M. (2022). Why do we have so many different hydrological models? A review based on the case of Switzerland. *Wiley Interdisciplinary Reviews: Water*, *9*(1), e1574.

Höge, M., Scheidegger, A., Baity-Jesi, M., Albert, C., & Fenicia, F. (2022). Improving hydrologic models for predictions and process understanding using Neural ODEs. *Hydrology and Earth System Sciences Discussions*, 1-29.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., ... & Cudennec, C. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. Hydrological sciences journal, 58(6), 1198-1255.

International Hydropower Association. (2020). United Kingdom. Retrieved from https://www.hydropower.org/country-profiles/united-kingdom.

IPCC. (2001). Climate Change 2001: Impacts, adaptation and vulnerability – Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK

Jackson, C. R., Wang, L., Pachocka, M., Mackay, J. D., & Bloomfield, J. P. (2016). Reconstruction of multi-decadal groundwater level time-series using a lumped conceptual model. *Hydrological Processes*, *30*(18), 3107-3125.

Jehanzaib, M., Shah, S. A., Yoo, J., & Kim, T. W. (2020). Investigating the impacts of climate change and human activities on hydrological drought using non-stationary approaches. *Journal of Hydrology*, *588*, 125052.

Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., & Houska, T. (2020). Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrology and Earth System Sciences*, *24*(3), 1081-1100.

Johnston, L. H., Dunnington, D. W., Greenwood, M. C., Kurylyk, B. L., & Jamieson, R. C. (2022). Identifying Hydrologic Regimes and Drivers in Nova Scotia, Canada: Catchment Classification Efforts for a Data-Limited Region. *Journal of Hydrologic Engineering*, *27*(11), 05022017.

Kapangaziwiri, E., Hughes, D. A., & Wagener, T. (2012). Incorporating uncertainty in hydrological predictions for gauged and ungauged basins in southern Africa. *Hydrological Sciences Journal*, *57*(5), 1000-1019.

Kavetski, D., & Fenicia, F. (2011). Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. *Water Resources Research*, *47*(11).

Kay, A. L., Davies, H. N., Bell, V. A., & Jones, R. G. (2009). Comparison of uncertainty sources for climate change impacts: flood frequency in England. *Climatic change*, 92(1), pp. 41-63.

Kelleher, C., McGlynn, B., & Wagener, T. (2017). Characterizing and reducing equifinality by constraining a distributed catchment model with regional signatures, local observations, and process understanding. *Hydrology and Earth System Sciences*, *21*(7), 3325-3352.

Khatami, S., Peel, M. C., Peterson, T. J., & Western, A. W. (2019). Equifinality and flux mapping: A new approach to model evaluation and process representation under uncertainty. *Water Resources Research*, 55, 8922–8941

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3).

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of hydrology*, *424*, 264-277.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., ... & Nearing, G. (2022). Uncertainty estimation with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, *26*(6), 1673-1693.

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.

Kneis, D. (2015). A lightweight framework for rapid development of object-based hydrological model engines. *Environmental Modelling & Software*, *68*, 110-121.

Knoben, W. J., Freer, J. E., Fowler, K. J., 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.

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). 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, e2019WR025975. https://doi.org/10.1029/2019WR025975

Kollat, J. B., Reed, P. M., & Wagener, T. (2012). When are multiobjective calibration trade-offs in hydrologic models meaningful?. *Water Resources Research*, *48*(3).

Kraft, P., Vaché, K. B., Frede, H. G., & Breuer, L. (2011). CMF: a hydrological programming language extension for integrated catchment models. *Environmental Modelling & Software*, *26*(6), 828-830.

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.

Lamontagne, J. R., Barber, C. A., & Vogel, R. M. (2020). Improved estimators of model performance efficiency for skewed hydrologic data. *Water Resources Research*, *56*(9), e2020WR027101.

Lane, R. A., Coxon, G., Freer, J. E., Wagener, T., Johnes, P. J., Bloomfield, J. P., ... & Reaney, S. M. (2019). Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain. *Hydrology and Earth System Sciences*, 23(10), 4011-4032.

Lane, R. A., & Kay, A. L. (2021). Climate change impact on the magnitude and timing of hydrological extremes across Great Britain. *Frontiers in Water*, *3*, 684982.

Leavesley, G. H. (1984). *Precipitation-runoff modeling system: User's manual* (Vol. 83, No. 4238). US Department of the Interior.

Leavesley, G. H., Restrepo, P. J., Markstrom, S. L., Dixon, M., & Stannard, L. G. (1996). The modular modeling system (MMS): User's manual. *US Geological Survey Open-File Report*, *96*(151,142).

Lebecherel, L., Andréassian, V., & Perrin, C. (2013). On regionalizing the Turc-Mezentsev water balance formula. *Water Resources Research*, *49*(11), 7508-7517.

Lee, H., McIntyre, N., Wheater, H., & Young, A. (2005). Selection of conceptual models for regionalisation of the rainfall-runoff relationship. *Journal of Hydrology*, 312(1-4), 125-147.

Le Moine, N., Andréassian, V., Perrin, C., & Michel, C. (2007). How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments. *Water resources research*, *43*(6).

Lee, G., Tachikawa, Y., Sayama, T., & Takara, K. (2012). Catchment responses to plausible parameters and input data under equifinality in distributed rainfall-runoff modeling. *Hydrological Processes*, *26*(6), 893-906.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., & Dadson, S. J. (2021). Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*, *25*(10), 5517-5534.

Lehner, B., Döll, P., Alcamo, J., Henrichs, T., & Kaspar, F. (2006). Estimating the impact of global change on flood and drought risks in Europe: a continental, integrated analysis. *Climatic Change*, *75*, 273-299.

Ley, R., Hellebrand, H., Casper, M. C., & Fenicia, F. (2016). Is catchment classification possible by means of multiple model structures? A case study based on 99 catchments in Germany. *Hydrology*, *3*(2), 22.

Li, L. J., Zhang, L., Wang, H., Wang, J., Yang, J. W., Jiang, D. J., ... & Qin, D. Y. (2007). Assessing the impact of climate variability and human activities on streamflow from the Wuding River basin in China. *Hydrological Processes: An International Journal*, *21*(25), 3485-3491.

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research: Atmospheres*, *99*(D7), 14415-14428.

Liu, Y., Wagener, T., Beck, H. E., & Hartmann, A. (2020). What is the hydrologically effective area of a catchment?. *Environmental Research Letters*, *15*(10), 104024.

Liu, Y., Wagener, T., & Hartmann, A. (2021). Assessing streamflow sensitivity to precipitation variability in karst influenced catchments with unclosed water balances. *Water Resources Research*, 57, e2020WR028598. https://doi.org/10.1029/2020WR028598

Luijendijk, E., Gleeson, T., & Moosdorf, N. (2020). Fresh groundwater discharge insignificant for the world's oceans but important for coastal ecosystems. Nature communications, 11(1), pp. 1-12.

Malede, D. A., Alamirew, T., & Andualem, T. G. (2023). Integrated and Individual Impacts of Land Use Land Cover and Climate Changes on Hydrological Flows over Birr River Watershed, Abbay Basin, Ethiopia. *Water*, *15*(1), 166.

Marsh, T.J. and Hannaford, J., eds., (2008). UK Hydrometric Register. Hydrological data UK series. *Centre for Ecology & Hydrology*, pp. 210.

Massmann, C. (2020). Identification of factors influencing hydrologic model performance using a top-down approach in a large number of US catchments. *Hydrological Processes*, *34*(1), 4-20.

McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrological Processes*, *26*(26), 4078-4111.

McMillan, H., Westerberg, I., & Branger, F. (2017). Five guidelines for selecting hydrological signatures. *Hydrological Processes*, *31*(26), 4757-4761.

McMillan, H. (2020). Linking hydrologic signatures to hydrologic processes: A review. *Hydrological Processes*, *34*(6), 1393-1409.

McMillan, H. K. (2021). A review of hydrologic signatures and their applications. *Wiley Interdisciplinary Reviews: Water*, *8*(1), e1499.

McMillan, H. K., Coxon, G., Sikorska-Senoner, A. E., & Westerberg, I. K. (2022a). Impacts of observational uncertainty on analysis and modelling of hydrological processes: Preface. Hydrological Processes, 36 (2),[e14481]. https://doi. org/10.1002/hyp. 14481.

McMillan, H. K., Gnann, S. J., & Araki, R. (2022b). Large scale evaluation of relationships between hydrologic signatures and processes. *Water Resources Research*, *58*(6), e2021WR031751.

McMillan, H., Araki, R., Gnann, S., Woods, R., & Wagener, T. (2023). How do hydrologists perceive watersheds? A survey and analysis of perceptual model figures for experimental watersheds. *Hydrological Processes*, *37*(3), e14845.

Merheb, M., Moussa, R., Abdallah, C., Colin, F., Perrin, C., & Baghdadi, N. (2016). Hydrological response characteristics of Mediterranean catchments at different time scales: a meta-analysis. *Hydrological Sciences Journal*, *61*(14), 2520-2539.

Merz, R. and Blöschl., G. 2004. Regionalisation of catchment model parameters. Journal of Hydrology, 287(1-4), 95-123.

Mezentsev, V. (1955). Back to the computation of total evaporation. *Meteorologia i Gidrologia*, 5, pp. 24–26.

Milly, P. C. D. (1994). Climate, soil water storage, and the average annual water balance. *Water Resources Research*, *30*(7), 2143-2156.

Modarres, R., Sarhadi, A., & Burn, D. H. (2016). Changes of extreme drought and flood events in Iran. *Global and Planetary Change*, *144*, 67-81.

Moges, E., Ruddell, B. L., Zhang, L., Driscoll, J. M., Norton, P., Perez, F., & Larsen, L. G. (2022). HydroBench: Jupyter supported reproducible hydrological model benchmarking and diagnostic tool. *Frontiers in Earth Science*, 1469.

Moliere, D. R., Lowry, J. B. C., & Humphrey, C. L. (2009) Classifying the flow regime of data-limited streams in the wet-dry tropical region of Australia. *Journal of Hydrology*, 367(1–2), pp. 1–13

Montanari, A., & Di Baldassarre, G. (2013). Data errors and hydrological modelling: The role of model structure to propagate observation uncertainty. *Advances in Water Resources*, *51*, 498-504.

Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L., ... & Belyaev, V. (2013). "Panta Rhei—everything flows": change in hydrology and society—the IAHS scientific decade 2013–2022. *Hydrological Sciences Journal*, *58*(6), 1256-1275.

Monteith, J. L. (1965). Evaporation and environment. In *Symposia of the society for experimental biology* (Vol. 19, pp. 205-234). Cambridge University Press (CUP) Cambridge.

Moore, R. J. (2007). The PDM rainfall-runoff model. *Hydrology and Earth System Sciences*, *11*(1), 483-499.

Moradkhani, H., & Sorooshian, S. (2009). General review of rainfall-runoff modeling: model calibration, data assimilation, and uncertainty analysis. *Hydrological modelling and the water cycle*, 1-24.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885-900.

Muñoz, E., Arumí, J. L., Wagener, T., Oyarzún, R. & Parra, V. (2016). Unraveling complex hydrogeological processes in Andean basins in south-central Chile: An integrated assessment to understand hydrological dissimilarity, *Hydrological Processes*, 30, pp. 4934–4943

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly weather review*, *116*(12), 2417-2424.

Nash, J., Sutcliffe, J.V. (1970). River flow forecasting through conceptual models part I: discussion of principles. J. Hydrology 10 (3), 282e290.

Neri, M., Coulibaly, P., & Toth, E. (2022). Similarity of catchment dynamics based on the interaction between streamflow and forcing time series: Use of a transfer entropy signature. *Journal of Hydrology*, *614*, 128555.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., ... & Duan, Q. (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.

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., ... & Morice, E. (2014). Benchmarking hydrological models for low-flow simulation and forecasting on French catchments. *Hydrology and Earth System Sciences*, *18*(8), 2829-2857.

Olden, J. D., & Poff, N. L. (2003). Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. *River research and applications*, *19*(2), 101-121.

Oldfield, F., & Dearing, J. A. (2003). The role of human activities in past environmental change. In *Paleoclimate, global change and the future* (pp. 143-162). Springer, Berlin, Heidelberg.

Oldham, L. D., Freer, J., Coxon, G., Howden, N., Bloomfield, J. P., & Jackson, C. (2023). Evidence-based requirements for perceptualising intercatchment groundwater flow in hydrological models. *Hydrology and Earth System Sciences*, *27*(3), 761-781.

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J., & Zappa, M. (2015). Does model performance improve with complexity? A case study with three hydrological models. *Journal of Hydrology*, *523*, 147-159.

Pan, Z., Liu, P., Gao, S., Xia, J., Chen, J., & Cheng, L. (2019). Improving hydrological projection performance under contrasting climatic conditions using spatial coherence through a hierarchical Bayesian regression framework. *Hydrology and Earth System Sciences*, *23*(8), 3405-3421.

Patil, S., & Stieglitz, M. (2012). Controls on hydrologic similarity: role of nearby gauged catchments for prediction at an ungauged catchment. *Hydrology and Earth System Sciences*, *16*(2), 551-562.

Paul, P. K., Zhang, Y., Ma, N., Mishra, A., Panigrahy, N., & Singh, R. (2021). Selecting hydrological models for developing countries: Perspective of global, continental, and country scale models over catchment scale models. *Journal of Hydrology*, *600*, 126561.

Payan, J. L., Perrin, C., Andréassian, V., & Michel, C. (2008). How can man-made water reservoirs be accounted for in a lumped rainfall-runoff model?. *Water Resources Research*, *44*(3).

Peel, M. C., & McMahon, T. A. (2020). Historical development of rainfall-runoff modeling. *Wiley Interdisciplinary Reviews: Water*, *7*(5), e1471.

Penman, H. L. (1950). The dependence of transpiration on weather and soil conditions. *Journal of Soil Science*, *1*(1), 74-89.

Perrin, C., Michel, C., & Andréassian, V. (2001). Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of hydrology*, *242*(3-4), 275-301.

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of hydrology*, *279*(1-4), 275-289.

Prieto, C., Le Vine, N., Kavetski, D., García, E., & Medina, R. (2019). Flow prediction in ungauged catchments using probabilistic random forests regionalization and new statistical adequacy tests. *Water Resources Research*, 55(5), pp. 4364-4392.

Prudhomme, C., Jakob, D., & Svensson, C. (2003). Uncertainty and climate change impact on the flood regime of small UK catchments. *Journal of Hydrology*, 277(1-2), pp. 1-23.

Pool, S., Vis, M. and Seibert, J. (2018). Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. Hydrological Science Journal, 63, pp.13-14.

Pool, S., Vis, M., & Seibert, J. (2021). Regionalization for ungauged catchments—lessons learned from a comparative large-sample study. *Water Resources Research*, *57*(10), e2021WR030437.

Rahman, M., Pianosi, F., & Woods, R. (2023). Simulating spatial variability of groundwater table in England and Wales. *Hydrological Processes*, *37*(3), e14849.

Rakovec, O., Mizukami, N., Kumar, R., Newman, A. J., Thober, S., Wood, A. W., ... & Samaniego, L. (2019). Diagnostic evaluation of large-domain hydrologic models calibrated across the contiguous United States. *Journal of Geophysical Research: Atmospheres*, *124*(24), 13991-14007.

RenÖFÄLt, B. M., Jansson, R., & Nilsson, C. (2010). Effects of hydropower generation and opportunities for environmental flow management in Swedish riverine ecosystems. *Freshwater Biology*, *55*(1), 49-67

Robinson, E.L., Blyth, E., Clark, D.B., Finch, J. & Rudd, A.C. (2015a). Climate hydrology and ecology research support system potential evapotranspiration dataset for Great Britain (1961-2012) [CHESS-PE]. *NERC Environmental Information Data Centre*. (Dataset). https://doi.org/10.5285/d329f4d6-95ba-4134-b77a-a377e0755653.

Robinson, E.L., Blyth, E., Clark, D.B., Finch, J. & Rudd, A.C. (2015b). Climate hydrology and ecology research support system meteorological dataset (1961-2012) [CHESS-met]. *NERC-Environmental Information Data Centre* doi:10.5285/80887755-1426-4dab-a4a6-250919d5020c.

Rogelis, M. C., Werner, M., Obregón, N., & Wright, N. (2016). Hydrological model assessment for flood early warning in a tropical high mountain basin. *Hydrology and Earth System Sciences Discussions*, 1-36.

Saavedra, D., Mendoza, P. A., Addor, N., Llauca, H., & Vargas, X. (2022). A multi-objective approach to select hydrological models and constrain structural uncertainties for climate impact assessments. *Hydrological Processes*, *36*(1), e14446.

Sadegh, M., AghaKouchak, A., Flores, A., Mallakpour, I., & Nikoo, M. R. (2019). A multi-model nonstationary rainfall-runoff modeling framework: analysis and toolbox. *Water Resources Management*, *33*, 3011-3024.

Saft, M., Peel, M. C., Western, A. W., Perraud, J.-M., & Zhang, L. (2016). Bias in streamflow projections due to climate-induced shifts in catchment response. *Geophysical Research Letters*, 43, 1574–1581

Safeeq, M., Bart, R. R., Pelak, N. F., Singh, C. K., Dralle, D. N., Hartsough, P., & Wagenbrenner, J. W. (2021). How realistic are water-balance closure assumptions? A demonstration from the southern sierra critical zone observatory and kings river experimental watersheds. *Hydrological Processes*, *35*(5), e14199.

Sahraei, S., Asadzadeh, M., & Unduche, F. (2020). Signature-based multi-modelling and multi-objective calibration of hydrologic models: Application in flood forecasting for Canadian Prairies. *Journal of Hydrology*, *588*, 125095.

Salwey, S., Coxon, G., Pianosi, F., Singer, M. B., & Hutton, C. (2023). National-Scale Detection of Reservoir Impacts Through Hydrological Signatures. *Water Resources Research*, e2022WR033893.

Sankarasubramanian, A., Vogel, R. M., & Limbrunner, J. F. (2001). Climate elasticity of streamflow in the United States. *Water Resources Research*, *37*(6), 1771-1781.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A. and Carrillo, G. (2011). Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA. Hydrology and Earth System Sciences, 15(9), 2895–2911. https://doi.org/10.5194/hess-15-2895-2011

Schaefli, B. and Gupta, H. V. (2007). Do Nash value have value? *Hydrological processes*, 21(15), 2075–2080.

Schaller, M. F., & Fan, Y. (2009). River basins as groundwater exporters and importers: Implications for water cycle and climate modeling. *Journal of Geophysical Research: Atmospheres*, *114*(D4).

Schwamback, D., Gesualdo, G. C., Sone, J. S., Kobayashi, A. N. A., Bertotto, L. E., Garcia, M. V. S., ... & Oliveira, P. T. S. (2022). Are Brazilian catchments gaining or losing water? The effective area of tropical catchments. *Hydrological Processes*, *36*(3), e14535.

Schwemmle, R., Demand, D., & Weiler, M. (2021). Diagnostic efficiency–specific evaluation of model performance. *Hydrology and Earth System Sciences*, *25*(4), 2187-2198.

Seibert, J. (2001). On the need for benchmarks in hydrological modelling. *Hydrological Processes*, *15*(6), 1063-1064.

Seibert, J., & McDonnell, J. J. (2002). On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research*, *38*(11), 23-1.

Seibert, J., Vis, M. J., Lewis, E., & Meerveld, H. V. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological processes*, 32(8), 1120-1125.

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

Shafii, M., Basu, N., Craig, J. R., Schiff, S. L. & Van Cappellen, P. (2017). A diagnostic approach to constraining flow partitioning in hydrologic models using a multiobjective optimization framework. Water Resources Research, 53(4), pp.3279-3301.

Shepley, M. G., Whiteman, M. I., Hulme, P. J., & Grout, M. W. (2012). Introduction: groundwater resources modelling: a case study from the UK. *Geological Society, London, Special Publications*, *364*(1), 1-6.

Singh, N. K., & Basu, N. B. (2022). The human factor in seasonal streamflows across natural and managed watersheds of North America. *Nature Sustainability*, *5*(5), 397-405.

Sivapalan M. (2005) Pattern, processes and function: elements of a unified theory of hydrology at the catchment scale. In Encyclopedia of Hydrological Sciences, Anderson M. (Ed.), John Wiley & Sons Ltd: London, pp. 193–219.

Sivapalan, M. (2006). Pattern, process and function: elements of a unified theory of hydrology at the catchment scale. Encyclopedia of hydrological sciences.

Stein, L., Clark, M. P., Knoben, W. J., Pianosi, F., & Woods, R. A. (2021). How do climate and catchment attributes influence flood generating processes? A large-sample study for 671 catchments across the contiguous USA. *Water Resources Research*, *57*(4), e2020WR028300.

Strahler, A. N. (1952), "Hypsometric (area-altitude) analysis of erosional topology", *Geological Society of America Bulletin*, 63 (11): 1117–1142.

Strahler, A. N. (1957), "Quantitative analysis of watershed geomorphology", *Transactions of the American Geophysical Union*, 38 (6): 913–920.

Tanguy, M., Dixon, H., Prosdocimi, I., Morris, D.G. & Keller, V.D.J. (2021). *Data from: Gridded estimates of daily and monthly areal rainfall for the United Kingdom (1890-2019)* [CEH-GEAR dataset]. NERC EDS Environmental Information Data Centre. Available

from: https://doi.org/10.5285/dbf13dd5-90cd-457a-a986-f2f9dd97e93c [Accessed 18 August 2021].

Terrado, M., Acuña, V., Ennaanay, D., Tallis, H., & Sabater, S. (2014). Impact of climate extremes on hydrological ecosystem services in a heavily humanized Mediterranean basin. *Ecological indicators*, *37*, 199-209.

Thirel, G., Andréassian, V., & Perrin, C. (2015). On the need to test hydrological models under changing conditions. *Hydrological Sciences Journal*, *60*(7-8), 1165-1173.

Thompson, S. A. (2017). *Hydrology for water management*. CRC Press.

Triana, J. S. A., Chu, M. L., Guzman, J. A., Moriasi, D. N., & Steiner, J. L. (2019). Beyond model metrics: The perils of calibrating hydrologic models. *Journal of Hydrology*, *578*, 124032.

Troch, P. A., Martinez, G. F., Pauwels, V. R. N., Durcik, M., Sivapalan, M., Harman, C., et al. (2009), Climate and vegetation water use efficiency at catchment scales. Hydrological Processes, 23(16), 2409–2414. https://doi.org/10.1002/hyp.7358

Todini, E. (1996). The ARNO rainfall—runoff model. *Journal of hydrology*, *175*(1-4), 339-382.

Todorović, A., Grabs, T., & Teutschbein, C. (2022). Advancing traditional strategies for testing hydrological model fitness in a changing climate. *Hydrological Sciences Journal*, 67:12, 1790-1811, DOI: 10.1080/02626667.2022.2104646

Toth, J. (1963). A theoretical analysis of groundwater flow in small drainage basins. *Journal of geophysical research*, *68*(16), 4795-4812.

Toth, E. (2013). Catchment classification based on characterisation of streamflow and precipitation time series. Hydrology and Earth System Sciences. 17, pp.1149-1159.

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., Coughlan de Perez, E., & Stephens, E. M. (2019). Assessing the performance of global hydrological models for capturing peak river flows in the Amazon basin. *Hydrology and Earth System Sciences*, *23*(7), 3057-3080.

Tumiran, S. A., & Sivakumar, B. (2021). Catchment classification using community structure concept: application to two large regions. *Stochastic Environmental Research and Risk Assessment*, 35(3), pp. 561-578.

Turc, L. (1955). Le bilan d'eau des sols: relations entre les précipitations, l'évaporation et l'écoulement. *Journées de l'hydraulique*, 3(1), pp. 36-44.

Uhlenbrook, S., Seibert, J. A. N., Leibundgut, C., & Rodhe, A. (1999). Prediction uncertainty of conceptual rainfall-runoff models caused by problems in identifying model parameters and structure. Hydrological Sciences Journal, 44(5), 779-797.

Van Esse, W. R., Perrin, C., Booij, M. J., Augustijn, D. C., Fenicia, F., Kavetski, D., & Lobligeois, F. (2013). The influence of conceptual model structure on model performance: a comparative study for 237 French catchments. *Hydrology and Earth System Sciences*, *17*(10), 4227-4239.

Van Loon, A. F., Rangecroft, S., Coxon, G., Werner, M., Wanders, N., Di Baldassarre, G., ... & Van Lanen, H. A. (2022). Streamflow droughts aggravated by human activities despite management. *Environmental Research Letters*, *17*(4), 044059.

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, 447–457.

Vaze, J., Jordan, P., Beecham, R., Frost, A., Summerell, G. (2012). Guidelines for Rainfall Runoff Modelling: Towards best practice model application, pp. 47.

Vega-Briones, J., de Jong, S., Galleguillos, M., & Wanders, N. (2023). Identifying driving processes of drought recovery in the southern Andes natural catchments. *Journal of Hydrology: Regional Studies*, *47*, 101369.

Vogel, R. M., Wilson, I., & Daly, C. (1999). Regional regression models of annual streamflow for the United States. *Journal of Irrigation and Drainage Engineering*, *125*(3), 148-157.

Vogel, R. M. (2011). Hydromorphology. *Journal of Water Resources Planning and Management*, *137*(2), 147-149.

Wagener, T., Lees, M. J., & Wheater, H. S. (2001a). A toolkit for the development and application of parsimonious hydrological models. *Mathematical models of large watershed hydrology*, *1*, 87-136.

Wagener, T., Lees, M. J., & Wheater, H. S. (2001b). Rainfall-runoff modelling toolbox user manual. *Civil and Environmental Engineering Department, Imperial College London, London*.

Wagener, T., Wheater, H., & Gupta, H. V. (2004). *Rainfall-runoff modelling in gauged and ungauged catchments*. World Scientific.

Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment classification and hydrologic similarity. *Geography compass*, *1*(4), 901-931.

Wagener, T., Sivapalan, M., & McGLYNN, B. R. I. A. N. (2008). Catchment classification and services. Toward a new paradigm for catchment hydrology driven by societal needs. *Encyclopedia of Hydrological Sciences, John Wiley, Chichester, UK*.

Wagener, T., Sivapalan, P. A. Troch, B. L. McGlynn, C. J. Harman, H. V. Gupta, P. Kumar, P. S. C. Rao, N. B. Basu, & J. S. Wilson. (2010). The future of hydrology: An evolving science for a changing world, *Water Resources Research*, 46,W05301, doi:10.1029/2009WR008906

Wagener, T., & Montanari, A. (2011). Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. *Water Resources Research*, 47(6).

Wagener, T., Dadson, S. J., Hannah, D. M., Coxon, G., Beven, K., Bloomfield, J. P., ... & Old, G. (2021). Knowledge gaps in our perceptual model of Great Britain's hydrology. *Hydrological Processes*, *35*(7), e14288.

Wang, S., Ancell, B., Yang, Z. L., Duan, Q., & Anagnostou, E. N. (2022). Hydroclimatic extremes and impacts in a changing environment: Observations, mechanisms, and projections. *Journal of Hydrology*, *608*, 127615.

Watson, A., Miller, J., Fink, M., Kralisch, S., Fleischer, M., & De Clercq, W. (2019). Distributive rainfall–runoff modelling to understand runoff-to-baseflow proportioning and its impact on the determination of reserve requirements of the Verlorenvlei estuarine lake, west coast, South Africa. *Hydrology and Earth System Sciences*, *23*(6), 2679-2697.

Weiler, M., and K. Beven (2015), Do we need a Community Hydrological Model?, *Water resources research*,51, 7777–7784, doi:10.1002/2014WR016731.

Węglarczyk, S. (1998). The interdependence and applicability of some statistical quality measures for hydrological models. *Journal of Hydrology*, *206*(1-2), 98-103.

Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., ... & Xu, C. Y. (2011). Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, 15(7), pp. 2205-2227.

Westerberg, I. K., Gong, L., Beven, K. J., Seibert, J., Semedo, A., Xu, C. Y., & Halldin, S. (2014). Regional water balance modelling using flow-duration curves with observational uncertainties. *Hydrology and Earth System Sciences*, *18*(8), 2993-3013.

Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A., & Freer, J. (2016). Uncertainty in hydrological signatures for gauged and ungauged catchments. *Water Resources Research*, *52*(3), 1847-1865.

Wu, L., Liu, X., Chen, J., Yu, Y., & Ma, X. (2022). Overcoming equifinality: Time-varying analysis of sensitivity and identifiability of SWAT runoff and sediment parameters in an arid and semiarid watershed. *Environmental Science and Pollution Research*, *29*(21), 31631-31645.

Yadav, M., Wagener, T., and Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, *Advances in Water Resources*, 30, 1756G1774

Yan, H., Sun, N., Eldardiry, H., Thurber, T. B., Reed, P. M., Malek, K., ... & Rice, J. S. (2023). Characterizing uncertainty in Community Land Model version 5 hydrological applications in the United States. *Scientific Data*, *10*(1), 187.

Yang, D., Yang, Y., & Xia, J. (2021). Hydrological cycle and water resources in a changing world: A review. *Geography and Sustainability*, *2*(2), 115-122.

Yang, W., Xia, R., Chen, H., Wang, M., & Xu, C. Y. (2022). The impact of calibration conditions on the transferability of conceptual hydrological models under stationary and nonstationary climatic conditions. *Journal of Hydrology*, *613*, 128310.

Yilmaz, K. K., Gupta, H. V. & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrological model. Water Resources Research, 44, 9.

Young, A. R. (2006). Stream flow simulation within UK ungauged catchments using a daily rainfall-runoff model. *Journal of Hydrology*, 320(1-2), pp. 155-172.

Zhang, Z., Wagener, T., Reed, P., & Bhushan, R. (2008). Reducing uncertainty in predictions in ungauged basins by combining hydrologic indices regionalization and multiobjective optimization. Water Resources Research, 44(12).