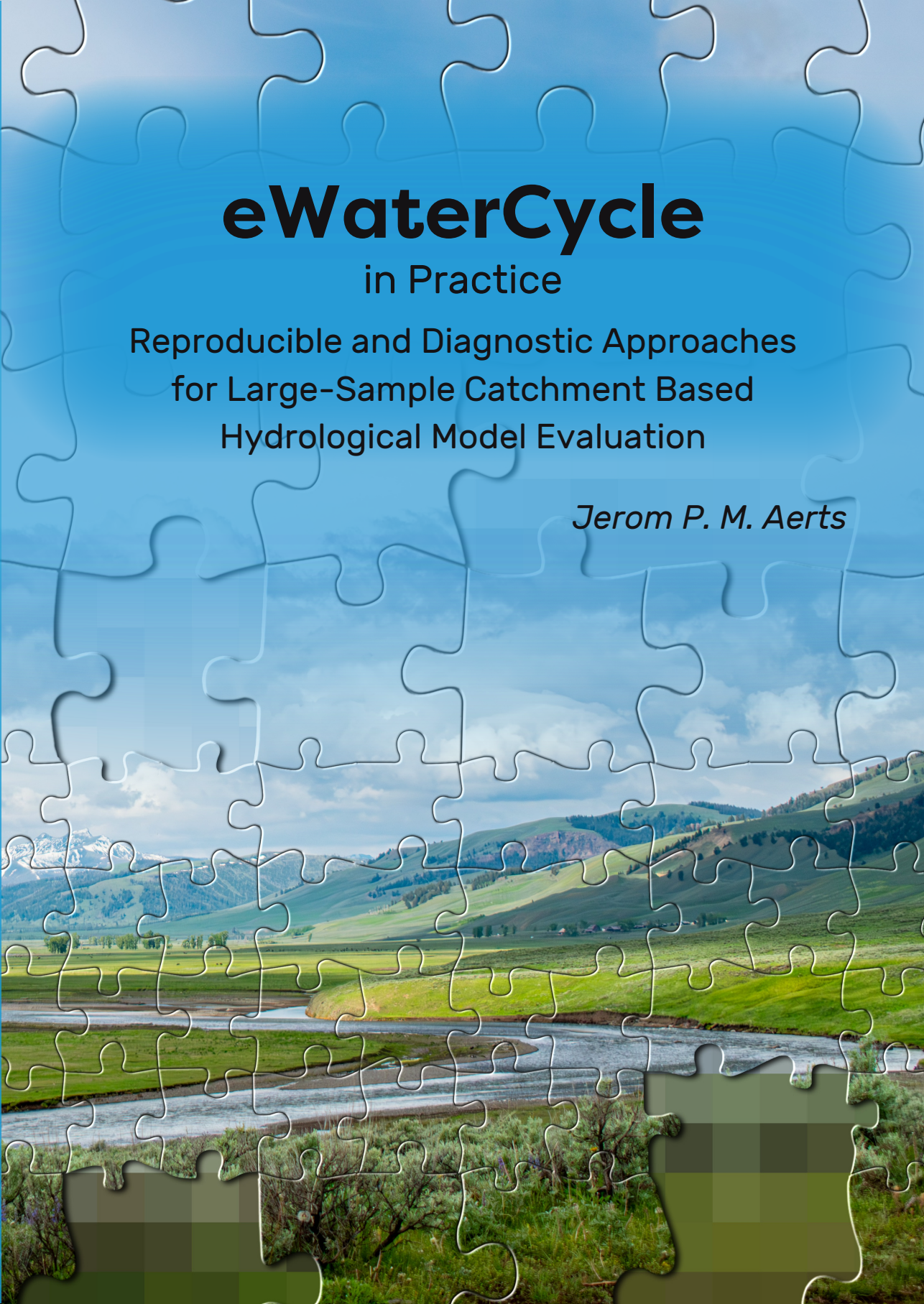
eWaterCycle

in Practice

Reproducible and Diagnostic Approaches for Large-Sample Catchment Based Hydrological Model Evaluation

Jerom P. M. Aerts



Propositions

accompanying the dissertation

eWaterCycle in Practice

Reproducible and Diagnostic Approaches for Large-Sample Catchment Based Hydrological Model Evaluation

by

Jerom Pierre Marie AERTS

1. Higher spatial resolution in hydrological models guarantees higher computational costs, not higher accuracy. *This thesis, Chapter 3.*
2. Machine learning cannot replace process understanding, but it can expose where it is absent. *This thesis, Chapter 4.*
3. Claimed improvements in hydrological model performance and reproducibility vanish when confronted with observational uncertainty. *This thesis, Chapter 5.*
4. The fact that we still call it "validation" rather than "falsification" shows how poorly Popper's lessons have been absorbed.
5. Standardizing model interfaces advances hydrological science more than improving any single model.
6. The biggest obstacle to reproducible hydrological science is not technical complexity but incentive structures.
7. Peer review would improve greatly if reviewers had to reproduce the results before recommending either acceptance or rejection.
8. The most valuable skills acquired during a PhD are not found in the thesis; they are found in the committees, conferences, and corridors.
9. The duration of a PhD is inversely proportional to the clarity of the research question at its start.

These propositions are regarded as opposable and defendable, and have been approved as such by the promotor Prof.dr.ir. N.C. van de Giesen and the copromotor Dr.ir. R.W. Hut.

Stellingen

behorende bij het proefschrift

eWaterCycle in Practice

Reproducible and Diagnostic Approaches for Large-Sample Catchment Based Hydrological Model Evaluation

door

Jerom Pierre Marie AERTS

1. Hogere ruimtelijke resolutie in hydrologische modellen garandeert meer rekenkosten, niet meer nauwkeurigheid. *Dit proefschrift, Hoofdstuk 3.*
2. Machine learning kan procesbegrip niet vervangen, maar kan wel blootleggen waar het ontbreekt. *Dit proefschrift, Hoofdstuk 4.*
3. Beweerde verbeteringen in de prestaties en reproduceerbaarheid van hydrologische modellen verdwijnen wanneer ze worden geconfronteerd met observatieonzekerheid. *Dit proefschrift, Hoofdstuk 5.*
4. De waterbalans forceren om te sluiten vermindert de onzekerheid niet; het verbergt deze.
5. Het feit dat we het nog steeds “validatie” noemen in plaats van “falsificatie” laat zien hoe slecht de lessen van Popper zijn doorgedrongen.
6. Het standaardiseren van modelinterfaces brengt de hydrologische wetenschap meer vooruit dan het verbeteren van welk individueel model dan ook.
7. Het grootste obstakel voor reproduceerbare hydrologische wetenschap is niet technische complexiteit maar het beloningssysteem.
8. Peer review zou sterk verbeteren als reviewers de resultaten moesten reproduceren alvorens acceptatie of afwijzing aan te bevelen.
9. De meest waardevolle vaardigheden die tijdens een promotietraject worden opgedaan, zijn niet te vinden in het proefschrift; ze zijn te vinden in de commissies, conferenties en wandelgangen.
10. De duur van een promotietraject is omgekeerd evenredig met de helderheid van de onderzoeksvraag bij aanvang.

Deze stellingen worden opponeerbaar en verdedigbaar geacht en zijn als zodanig goedgekeurd door de promotor Prof.dr.ir. N.C. van de Giesen en de copromotor Dr.ir. R.W. Hut.

eWaterCycle in Practice

Reproducible and Diagnostic Approaches for Large-Sample Catchment Based Hydrological Model Evaluation

eWaterCycle in Practice

Reproducible and Diagnostic Approaches for Large-Sample Catchment Based Hydrological Model Evaluation

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus,
Prof.dr.ir. H. Bijl,
chair of the Board for Doctorates
to be defended publicly on
Thursday 1 October 2026 at 12:30 hours

by

Jerom Pierre Marie AERTS

This dissertation has been approved by the (co)promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof.dr.ir. N.C. van de Giesen,	Delft University of Technology, promotor
Dr.ir. R.W. Hut,	Delft University of Technology, copromotor

Independent members:

Prof.dr. R. Maxwell,	Princeton University, USA
Dr. L.A. Melsen,	Wageningen University & Research
Dr. W. Berghuijs,	Vrije Universiteit Amsterdam
Prof.dr.ir. F.C. Vossepoel,	Delft University of Technology
Prof.dr.ir. R. Uijlenhoet,	Delft University of Technology







Cover design: Judith Jongen-Boekee

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

CONTENTS

Summary 1

Samenvatting 3

1 Introduction 5
- 1.1 From Plains to Proofs . . . 6
- 1.2 Hydrological Models . . . 7
- 1.3 Competing Modeling Philosophies . . . 9
- 1.4 Scaling Problems in Hydrological Modeling . . . 10
- 1.5 Large-Sample Datasets and Multifaceted Evaluation . . . 12
- 1.6 Observational Uncertainty and Statistical Significance . . . 13
- 1.7 eWaterCycle: Challenges and Solutions for Model Hypothesis Testing . . . 13
- 1.8 Scope of the Thesis . . . 15

2 The eWaterCycle platform for Open and FAIR hydrological collaboration 17
- 2.1 Introduction . . . 18
  - 2.1.1 Glossary . . . 21
- 2.2 The eWaterCycle Platform . . . 22
  - 2.2.1 Explorer . . . 23
  - 2.2.2 Experiment . . . 23
  - 2.2.3 Analyze . . . 23
  - 2.2.4 Share . . . 24
  - 2.2.5 Pre-processing . . . 24
  - 2.2.6 Available Models . . . 24
- 2.3 Technical Design of the eWaterCycle Platform . . . 25
  - 2.3.1 The Explorer . . . 26
  - 2.3.2 The Notebook Generator . . . 26
  - 2.3.3 Downloading ERA5 with era5cli . . . 28
  - 2.3.4 ESMValTool-based Model Input Pre-processor . . . 28
  - 2.3.5 Interfacing Models through grpc4bmi . . . 29
- 2.4 Use-Cases (Case Studies, Application Examples) . . . 30
  - 2.4.1 Hello World: One Model, One Catchment, One Forcing . . . 30
  - 2.4.2 Model Comparison: Two Models, One Catchment, One Forcing . . . 32
  - 2.4.3 Model Coupling: Add Output of One Model as Input for Downstream Model . . . 34
  - 2.4.4 Model Interaction: Change State of Model during Runtime . . . 35
  - 2.4.5 Calibrate Model Using an Ensemble of Model Runs . . . 37
- 2.5 Conclusions . . . 38

3 Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model **43**
3.1 Introduction . . . 44
3.2 Methodology . . . 46
3.2.1 Input data . . . 46
3.2.2 Model Experiment Setup . . . 48
3.2.3 Benchmark Selection . . . 50
3.2.4 Analysis of results . . . 51
3.2.5 eWaterCycle platform . . . 52
3.3 Results . . . 52
3.3.1 Calibration period results . . . 52
3.3.2 Evaluation period results . . . 55
3.3.3 The effect of spatial scale on terrain characteristics . . . 61
3.4 Discussion . . . 63
3.4.1 Benchmarks . . . 63
3.4.2 Streamflow estimates and uncertainty . . . 65
3.4.3 Relative model instance differences . . . 66
3.4.4 Outlook . . . 68
3.4.5 Conclusions . . . 68
Supplementary Analyses . . . 70

4 Advancing Hydrological Model Diagnostics: An Exploratory Approach Using Random Forest Models and a Large-Sample Catchment Data Set **71**
4.1 Introduction . . . 72
4.2 Methods and data . . . 74
4.2.1 Hydrological models . . . 74
4.2.2 Hydrological model runs . . . 75
4.2.3 Hydrological model performance . . . 76
4.2.4 Random forest models . . . 77
4.3 Results . . . 81
4.3.1 Streamflow . . . 81
4.3.2 Actual evaporation . . . 84
4.3.3 Soil moisture . . . 89
4.4 Discussion . . . 93
4.4.1 The effectiveness of random forest models . . . 93
4.4.2 The added value of exploratory diagnostics . . . 93
4.5 Conclusion . . . 95
Supplementary Analyses . . . 97

5 On the importance of discharge observation uncertainty when interpreting hydrological model performance **103**
5.1 Introduction . . . 104
5.2 Methodology . . . 106
5.2.1 Use cases . . . 106
5.2.2 Data . . . 108
5.2.3 Hydrological models . . . 109
5.2.4 Model runs . . . 110
5.2.5 Analyses . . . 111

5.3 Results . . . 114
5.3.1 Discharge-based model performance . . . 114
5.3.2 Discharge observation uncertainty estimates . . . 115
5.3.3 Use Cases . . . 115
5.3.4 Temporal sampling uncertainty . . . 120
5.4 Discussion . . . 121
5.4.1 Performance interpretation under discharge observation uncertainty . . . 121
5.4.2 Temporal robustness of model performance . . . 122
5.4.3 Practical implications for model users . . . 123
5.4.4 Limitations & Outlook . . . 123
5.5 Conclusions . . . 124
Supplementary Analyses . . . 125

6 Synthesis **127**
6.1 Introduction . . . 127
6.2 Summary of Main Findings . . . 128
6.2.1 Cross-Cutting Themes . . . 131
6.3 Main Conclusions . . . 135

7 Discussion **137**
7.1 Types of Model Users and Model Selection . . . 137
7.2 Model Purpose: Research Tool versus Practical Tool . . . 138
7.3 Reproducibility in Hydrological Sciences . . . 139
7.4 Testing Model Hypotheses with Field Measurements . . . 140
7.5 Can Modeling Improve Hydrological Understanding? . . . 141

8 Outlook **143**

Bibliography **147**

Acknowledgements **175**

Curriculum Vitæ **179**

List of Publications **181**

SUMMARY

This dissertation addresses multiple challenges in hydrological modeling, focusing on reproducibility, large-sample catchment diagnostics, and the role of observational uncertainty in interpreting model performance. Through a series of applied studies, it proposes practical tools and methods to improve the transparency, comparability, and interpretability of hydrological models, without overstating what current data or models can reliably deliver. Throughout, the thesis argues that apparent limitations in model evaluation, including the persistence of equifinality, can be reduced by expanding the observational basis against which models are tested, rather than accepting them as inherent properties of the system.

Central to this work are the development and application of the eWaterCycle platform: a modular, community-driven system that promotes reproducible modeling through standardized interfaces and containerized environments. By building on existing models through the Basic Model Interface (BMI) and deploying experiments at scale, eWaterCycle enables model intercomparison and refinement with fewer technical barriers. Although it does not eliminate all integration complexity, the platform provides a practical foundation for reproducible, FAIR-compliant workflows and has already enabled new forms of collaborative experimentation.

Using this platform, the thesis evaluates the effects of spatial resolution on hydrological model performance. Results from 454 CAMELS-US catchments show that a finer spatial resolution (down to 200 m) does not universally improve streamflow predictions at catchment outlets. Although resolution matters in some contexts, particularly where topography or land cover is complex, the overall gains are limited and often fall within uncertainty bounds. These findings suggest that investing in higher spatial resolution without a corresponding investment in observational density is a questionable allocation of research effort. "Hyper-resolution" modeling, though technically feasible, must be approached with clear use cases and a realistic assessment of whether the available data can actually support it.

Expanding the scope of model diagnostics, the research applies machine learning (specifically, a random forest) to link catchment attributes to model performance across 617 CAMELS-GB catchments. The analysis shows that model suitability often varies with catchment scale and physical characteristics, with different models (wflow_sbm and PCR-GLOBWB) excelling under different conditions. This underscores the "reality of locality" in hydrology: every catchment presents a unique combination of dominant processes, making the pursuit of a universal model impractical. Although the results expose structural model limitations, they also demonstrate the potential of machine learning as a diagnostic tool, an application that deserves greater emphasis alongside its more common use in prediction.

A key contribution of the dissertation is the integration of uncertainty in discharge observations into model evaluation. Using quantified uncertainty estimates from 299

CAMELS-GB catchments, the analysis shows that a significant share of perceived performance improvements, whether achieved through parameter tuning or arising from structural differences, may fall within the uncertainty bounds of the observations. Notably, no clear relationship was found between model complexity and model performance, challenging the assumption that more complex models are inherently more capable. Together, these findings call into question how model results are interpreted and underline the need to report and account for uncertainty more systematically.

The concluding chapters synthesize these results, arguing for a more critical yet constructive modeling practice: one that embraces reproducibility and large-sample evaluation while recognizing the limits of model resolution, structural complexity, and observational quality. They make the case that advancing hydrology depends as much on investing in new and better observations—from rehabilitating existing gauging networks to deploying novel sensing technologies and satellite products—as it does on improving models. By embedding uncertainty diagnostics into the evaluation process and facilitating open, repeatable workflows, this work contributes to a more robust and realistic foundation for model improvement and application.

Looking ahead, the thesis outlines clear directions for future research: expanding eWaterCycle to include more models and domains; exploiting underutilized satellite missions and opportunistic data sources; and bridging the cultural divide between modelers and field scientists so that data collection and model experimentation become two sides of the same scientific process. Although challenges remain, particularly in the community-wide adoption of reproducible practices, the tools and insights developed here offer a tangible step forward.

In summary, this dissertation contributes to hydrological science by combining reproducible modeling tools with diagnostic depth. It promotes a more transparent and critical modeling practice, one that acknowledges uncertainty, challenges complacency about data scarcity, and encourages the responsible use of models. Rather than seeking breakthroughs, the work focuses on enabling more reproducible, interpretable, and broadly applicable hydrological modeling, thereby helping build a stronger foundation for both scientific progress and practical decision-making. The path forward is not to accept the limitations we have inherited, but to systematically reduce them.

SAMENVATTING

Dit proefschrift behandelt meerdere uitdagingen in hydrologische modellering, met een focus op reproduceerbaarheid, diagnostiek en de rol van observatieonzekerheid bij het interpreteren van modelprestaties. Aan de hand van een reeks toegepaste studies worden praktische instrumenten en methoden voorgesteld om de transparantie, vergelijkbaarheid en interpreteerbaarheid van hydrologische modellen te verbeteren, zonder te overschatten wat met de huidige data en modellen betrouwbaar kan worden bereikt. Het proefschrift betoogt consequent dat ogenschijnlijke beperkingen in modelevaluatie, waaronder het blijven bestaan van equifinaliteit, kunnen worden verkleind door de observationele basis waarop modellen worden getoetst te verbreden, in plaats van deze beperkingen als inherente eigenschappen van het systeem te beschouwen.

De ontwikkeling en toepassing van het eWaterCycle-platform staan centraal in dit werk. Het is een modulair, gemeenschapsgedreven systeem dat reproduceerbare modellering stimuleert via gestandaardiseerde interfaces en containergebaseerde omgevingen. Door via de Basic Model Interface (BMI) voort te bouwen op bestaande modellen en experimenten op schaal uit te voeren, maakt eWaterCycle modelvergelijking en -verbetering mogelijk met minder technische drempels. Hoewel het platform niet alle integratiecomplexiteit wegneemt, biedt het een praktische basis voor reproduceerbare, FAIR-conforme workflows.

Met behulp van dit platform onderzoekt het proefschrift de effecten van ruimtelijke resolutie op de prestaties van hydrologische modellen. Resultaten van 454 CAMELS-US-stroomgebieden laten zien dat een fijnere ruimtelijke resolutie (tot 200 m) niet altijd leidt tot betere afvoerschattingen bij de uitstroompunten van stroomgebieden. Hoewel resolutie in bepaalde contexten belangrijk is, met name bij complexe topografie of landbedekking, blijven de algehele prestatiewinsten beperkt en vallen ze vaak binnen de onzekerheidsmarges. Deze bevindingen suggereren dat investeren in een hogere ruimtelijke resolutie zonder een evenredige investering in observatiedichtheid een twijfelachtige besteding van onderzoeksmiddelen is. Hyperresolutiemodellering is weliswaar technisch haalbaar, maar moet worden benaderd vanuit duidelijke toepassingsdoelen en een realistische beoordeling van de mate waarin de beschikbare data deze kunnen ondersteunen.

Het onderzoek verbreedt de reikwijdte van modeldiagnostiek door machinelearning (specifiek een random forest) toe te passen om stroomgebiedseigenschappen te koppelen aan modelprestaties voor 617 CAMELS-GB-stroomgebieden. De analyse laat zien dat de geschiktheid van modellen vaak varieert met de schaal en fysieke kenmerken van een stroomgebied, waarbij verschillende modellen (wflow_sbm en PCR-GLOBWB) onder verschillende omstandigheden beter presteren. Dit benadrukt de “realiteit van lokaliteit” in de hydrologie: elk stroomgebied kent een unieke combinatie van dominante processen, waardoor het streven naar een universeel model onpraktisch is. Hoewel de resultaten structurele modelbeperkingen blootleggen, tonen ze ook het potentieel van

machinelearning als diagnostisch instrument, een toepassing die meer aandacht verdient naast het gangbaardere gebruik ervan voor voorspellingen.

Een belangrijke bijdrage van dit proefschrift is de integratie van onzekerheid in afvoerwaarnemingen in de modelevaluatie. Met behulp van gekwantificeerde onzekerheidsschattingen voor 299 CAMELS-GB-stroomgebieden wordt aangetoond dat een aanzienlijk deel van de ogenschijnlijke prestatieverbeteringen, hetzij door parameteroptimalisatie, hetzij door structurele modelverschillen, binnen de onzekerheidsmarges van de waarnemingen valt. Opvallend is dat er geen duidelijk verband werd gevonden tussen modelcomplexiteit en modelprestaties, wat de aanname dat complexere modellen inherent beter presteren ter discussie stelt. Deze bevindingen stellen ook de interpretatie van modelresultaten ter discussie en onderstrepen de noodzaak om onzekerheid systematischer te rapporteren en mee te nemen.

De afsluitende hoofdstukken vatten deze resultaten samen en pleiten voor een kritischere, maar ook constructievere modelleerpraktijk: een praktijk die reproduceerbaarheid en grootschalige evaluatie omarmt en tegelijkertijd de grenzen van modelresolutie, structurele complexiteit en waarnemingskwaliteit erkent. Deze hoofdstukken stellen dat vooruitgang in de hydrologie evenzeer afhangt van investeringen in nieuwe en betere waarnemingen als van verbeteringen aan modellen. Die waarnemingen lopen uiteen van het herstellen van bestaande meetnetwerken tot het inzetten van nieuwe meettechnieken en satellietproducten. Door onzekerheidsdiagnostiek te integreren in het evaluatieproces en open, herhaalbare workflows te faciliteren, levert dit werk een robuustere en realistischere basis voor modelverbetering en toepassing.

Met het oog op de toekomst schetst het proefschrift duidelijke richtingen voor vervolgonderzoek: het uitbreiden van eWaterCycle met meer modellen en toepassingsdomeinen; het benutten van onderbenutte satellietmissies en opportunistische databronnen; en het overbruggen van de culturele kloof tussen modelleurs en veldwetenschappers, zodat dataverzameling en modelexperimentatie twee zijden van hetzelfde wetenschappelijke proces worden. Hoewel er uitdagingen blijven bestaan, met name bij de brede adoptie van reproduceerbare werkwijzen, vormen de hier ontwikkelde instrumenten en inzichten een concrete stap vooruit.

Samenvattend levert dit proefschrift een bijdrage aan de hydrologische wetenschap door reproduceerbare modelleringsinstrumenten te combineren met diepgaande diagnostiek. Het bevordert een transparantere en kritischere modelleerpraktijk die onzekerheid erkent, berusting in dataschaarste ter discussie stelt en verantwoord modelgebruik stimuleert. In plaats van doorbraken na te streven, richt dit werk zich op het mogelijk maken van reproduceerbare, interpreteerbare en breed toepasbare hydrologische modellering, waarmee een steviger fundament wordt gelegd voor zowel wetenschappelijke vooruitgang als praktische besluitvorming. De weg vooruit bestaat niet uit het accepteren van bestaande beperkingen, maar uit het systematisch verkleinen ervan.

1

INTRODUCTION

All models are wrong, but some are useful.

George E. P. Box

This thesis aims to improve the reproducibility and rigor of hydrological modeling by systematically comparing a range of hydrological model structures, from lumped conceptual to hyper-resolution physically based approaches, across large-sample catchment datasets. It builds on and contributes to the eWaterCycle framework, a community-driven platform that facilitates open, transparent, and FAIR-compliant model experimentation by standardizing interactions across different hydrological models. By integrating multiple data sources, including hydrometeorological measurements and derived products, this research evaluates how models of varying complexity and resolution perform under diverse hydroclimatic conditions. The findings demonstrate how scale, heterogeneity, and data limitations influence streamflow and other hydrological outputs, offering guidance for more robust and transferable hydrological modeling practices.

1.1. FROM PLAINS TO PROOFS

In the 1930s, unsustainable agricultural practices and severe drought ravaged the Great Plains of the United States, creating the infamous "Dust Bowl." Severe dust storms disrupted livelihoods and agricultural productivity, and 1934 proved especially devastating, underscoring the need for sound agricultural and water management practices. In that same year, British-Austrian philosopher Karl Popper formulated the modern scientific method by introducing his theory of falsification in his seminal work *Logik der Forschung* (later published in English as *The Logic of Scientific Discovery*). Popper argued that scientific theories can never be definitively proven, but they can be corroborated through attempts at falsification. The method of falsification became central to the scientific approach: a hypothesis should yield predictions that can be tested and potentially falsified.

Within the hydrological sciences, models serve as key tools for predicting and evaluating how catchments function. A catchment is the topographic area within which any raindrop ultimately flows into the same stream or river[1]. Understanding how a catchment responds to precipitation, temperature, (reference) evapotranspiration, and other inputs is central to water resource management. The process of building hydrological models typically begins with perceptual models—mental or schematic representations grounded in field observations and expert knowledge. These perceptual models capture the primary pathways of water movement, from rainfall infiltration to runoff, as shaped by each hydrologist's unique training and experience (Beven, 2001a; Beven and Chappell, 2021; Gupta et al., 2008, 2012; Seibert and McDonnell, 2002; Tetzlaff et al., 2008).

As understanding of the system grows, perceptual models evolve into conceptual models, which employ mathematical constructs to represent interconnected reservoirs (e.g., unsaturated and saturated soil zones) and fluxes (the water exchanges between them) (Beven, 2001b; Wagener, 2004). In keeping with Popper's principle of falsification, each hydrological model can be treated as a hypothesis regarding catchment behavior and can be refined or refuted by comparing its predictions with observed data (Clark, Kavetski, and Fenicia, 2011).

In practice, however, applying falsification to hydrological models is far from straightforward. The inverse problem in hydrology is statistically ill-posed (Vrugt et al., 2002): the available observations are insufficient to uniquely determine model parameters, so multiple parameter sets or even structurally different models can reproduce the same observed behavior. This situation has been formalized as the concept of equifinality (Beven and Freer, 2001), which, while highlighting uncertainty, risks undermining the very principle of falsification that should drive model improvement. If many models are deemed equally acceptable, meaningful hypothesis rejection becomes difficult, impeding scientific progress. This is fundamentally at odds with the Popperian approach to science outlined above. This thesis argues that this impasse can be broken not by accepting equifinality as an inherent property of the system, but by expanding the observational basis against which models are tested, thereby reducing the number of acceptable model configurations, ideally to one.

[1] With the exception of groundwater seeping out of the catchment (Bouaziz, Weerts, et al., 2018).

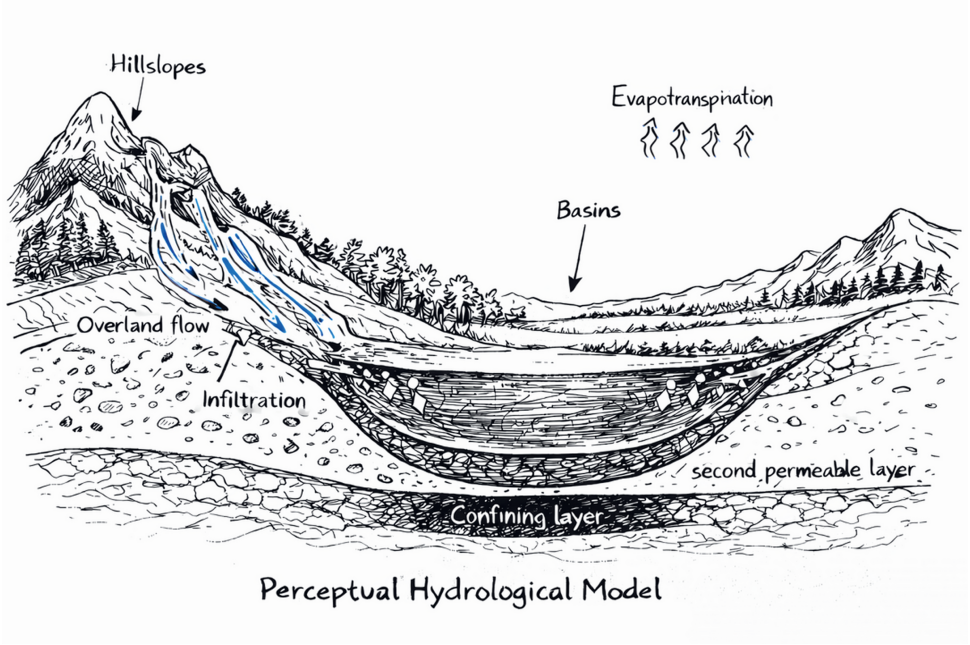


Figure 1.1: Illustration of a perceptual model.

1.2. HYDROLOGICAL MODELS

Early hydrological research laid the groundwork for our present understanding of catchment processes, shaping today's diverse range of hydrological models. During the 17th and 18th centuries, pioneers such as Pierre Perrault, Edme Mariotte, and Bernard Palissy systematically collected data and used analytical methods to advance scientific knowledge of how water flows through landscapes (i.e., the hydrological cycle). In the 19th century, the seminal work of Darcy, Buckingham, Navier, and Stokes established physical principles that would become central to hydrology. By the 1930s, pivotal breakthroughs concerning infiltration and overland flow (Horton, 1932) and the movement of water in unsaturated soils (Richards, 1931) further refined the field's theoretical foundation. These insights allowed hydrologists to move from initial *perceptual models* (i.e., mental sketches of catchment behavior) to more formal *conceptual models* of how water moves from precipitation to runoff.

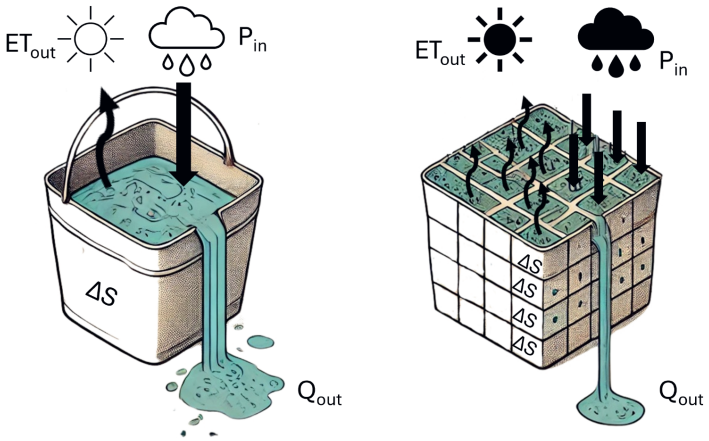


Figure 1.2: Two "leaky bucket" models: a lumped conceptual representation on the left and a physically based distributed representation on the right.

Conceptual hydrological models consist of relationships that describe, through mathematical equations, the connections between a hydrological system's reservoirs (states) and flows (fluxes). These relationships adhere to the principles of mass balance, conservation of mass, and the energy balance (e.g., Clark, Kavetski, and Fenicia, 2008; Savenije, 2009). Together, they form the water balance, which in its simplest form[2] can be written as:

$$P - Q - ET = \Delta S,$$

where P is precipitation entering the system over time (input), Q is streamflow leaving the system over time (output), ET is evapotranspiration, including both evaporation from the surface and transpiration from plants over time (output), and ΔS is the change in the system's storage over time. The storage state, S, is often represented in hydrological models as a "leaky bucket," with the input and output fluxes P, Q, and ET filling and emptying the "bucket" over time.

The apparent simplicity of the water balance equation conceals a complexity that hydrologists have struggled with for decades (Grayson et al., 1997; Kirchner, 2006; Or, 2019; Peter-Lidard et al., 2017). The processes underlying the movement of water vary significantly across landscapes, climate zones, and timescales and are also affected by human influences (Beven et al., 1988; de Marsily et al., 2005; Gao et al., 2018; McDonnell et al., 2007; Troch et al., 2009). This system heterogeneity is difficult to resolve mathematically and underscores the reality that no single model structure can adequately represent all hydroclimatic conditions. This challenge is part of the broader "curse of locality" (Bierkens et al., 2015), which highlights the difficulty of generalizing hydrological models across regions.

[2] Inter-catchment groundwater flow is not included in this formulation.

Over time, a wide variety of hydrological models has emerged, each reflecting different assumptions about how best to represent catchment-scale complexity (Abbott et al., 1986; Bergström and Forsman, 1973; Burnash, 1973; Kollet and Maxwell, 2008; Perrin et al., 2003; Refsgaard and Storm, 1995; van Beek and Bierkens, 2008).

Initially, models were often *lumped conceptual*, treating the catchment as a single mathematical unit with aggregated parameters (Beven and Chappell, 2021). In hydrological modeling, a parameter is a measurable or adjustable value that defines specific characteristics or behaviors of the hydrological system and represents complex processes that are too intricate to model explicitly. Later, increasing computational power led to the development of spatially explicit, *distributed* (or "physically based") models such as the Stanford Hydrological Model (Crawford and Linsley, 1966), which can simulate both surface and subsurface processes at finer resolutions. Freeze and Harlan (1969) expanded on these ideas by outlining a systematic, physically based framework for simulating hydrological processes across a catchment.

For those unfamiliar with hydrological modeling, the diversity of approaches can be perplexing given the maturity of the science. Much of this diversity stems from differing views on the appropriate balance between empirical observation and physical representation, the treatment of spatial variability, and the desired complexity of model structures.

1.3. COMPETING MODELING PHILOSOPHIES

Building on the historical progression from lumped conceptual to distributed models, hydrological modeling approaches can be broadly categorized into three main groups: **empirical**, **top-down**, and **bottom-up** (Hrachowitz and Clark, 2017). Although each approach shares the fundamental goal of describing how water moves through a catchment, they differ in the degree to which they rely on physical theory versus observed data and in their treatment of spatial variability.

Empirical models emphasize data-driven relationships between inputs (e.g., precipitation) and outputs (e.g., streamflow) without necessarily representing the underlying physical processes. Classic examples include the unit hydrograph (Sherman, 1932) and the Budyko framework (e.g., Bouaziz, Duan, et al., 2018). The unit hydrograph represents the runoff time series resulting from one unit of rainfall excess occurring uniformly over a catchment for a specific duration. The Budyko framework describes how precipitation is partitioned between evapotranspiration and runoff within a catchment. Modern applications of empirical models include machine-learning-based descriptions of the rainfall–runoff relationship (e.g., Kratzert et al., 2018). These methods have yielded promising results for streamflow prediction. Nevertheless, they are often criticized for being merely descriptive and for offering limited insight into the underlying hydrological processes. Moreover, their skill does not always extend to conditions outside the range of the training data, underscoring the potential risk of misapplication in changing or data-scarce environments.

Top-down models offer a more process-based perspective and are specifically developed at the catchment or subcatchment scale. They typically model the domain as a single unit (lumped models) that can be further discretized into subcatchments (semi-distributed models) (e.g., Arnold et al., 1998; Fenicia et al., 2011; Gharari et al., 2014; Lindström et al., 2010; Wagener et al., 2001).

By "zooming out," this approach integrates small-scale heterogeneity at a larger scale through flux parameterizations at the modeling scale (Figure 1.2). This results in a parsimonious representation of catchment heterogeneity, requiring fewer observations while retaining key dynamics. Top-down models are refined by deductively testing alternative model structures as competing hypotheses (Clark, Kavetski, and Fenicia, 2011; Popper, 1959). Top-down models are often referred to as "conceptual" models, although this term is misleading because all models are conceptual to some degree, having progressed from initial perceptual models.

Bottom-up models use the opposite approach by "zooming in" to capture small-scale heterogeneity. These types of process-based models are developed at or near the scale of observations and typically discretize the modeling domain into grid cells (e.g., Kollet and Maxwell, 2008; Refsgaard and Storm, 1996; Sutanudjaja et al., 2018; van Verseveld et al., 2022; Figure 1.2). By "zooming in," this approach uses a "physically based" or "continuum" description of small-scale surface and subsurface heterogeneity to represent the relationships between states and fluxes within the catchment (flux parameterization; Figure 1.2). This approach estimates larger-scale fluxes by aggregating the individual small-scale process descriptions (Beven, 2006a). This is an inductive approach to model refinement because it assumes that the individual parts provide insight into the system as a whole (Clark et al., 2015). Despite the promise of high fidelity, bottom-up models tend to be computationally demanding and highly sensitive to data quality. In regions lacking detailed observations, it can be difficult to constrain the many parameters that govern local-scale processes.

In practice, each of these philosophies has strengths and weaknesses, prompting ongoing debate over their relative merits (Clark et al., 2017). Empirical models excel at rapid prediction when robust data exist; top-down models balance simplicity with physical realism; and bottom-up models offer fine-grained insights but may impose high demands on data and computing power. Consequently, model selection often depends on practical considerations such as data availability and the intended application, in addition to philosophical preferences about the representation of hydrological systems. Therefore, model users need to test competing hypotheses to select the model structure that best aligns with their research question, data constraints, and intended use.

1.4. SCALING PROBLEMS IN HYDROLOGICAL MODELING

A key challenge in hydrology is determining how well the "average" parameterization of small-scale processes represents large-scale catchment behavior. This mismatch, referred to as the "catchment-scale closure relationship" (Beven, 2006a), remains a major theoretical hurdle. It reflects the inherent difficulty of upscaling fine-scale dynamics into reliable large-scale predictions. This challenge is rooted in what Bierkens et al. (2015)

termed the "curse of locality." In this thesis, however, we refer to it as the "reality of locality" because the local character of hydrological processes is not merely an inconvenience to be overcome but a fundamental and defining characteristic of the discipline.

This locality poses such a challenge for modeling because hydrological processes are nonlinear. For example, when two nonlinear reservoirs—such as two grid cells with threshold-driven runoff generation—are aggregated into a single unit, the effective parameter of the combined system is not simply a weighted average of the original parameter values. Because the underlying processes are nonlinear, the relative contribution of each cell to the aggregated response depends on the antecedent conditions, such as how close each cell is to its saturation threshold. As a result, the effective parameter varies with the hydrological state, making it impossible to derive it solely from the original parameters. Higher-resolution models have the advantage of operating at scales closer to those at which field observations are typically made, where parameters can, in principle, be directly measured or constrained. However, when processes are aggregated to coarser scales, this connection to direct observation is lost. This has a practical consequence: new observations are needed at the scale of the aggregated unit, and, ideally, these observations should constrain a unique parameter set rather than admit multiple solutions.

This challenge is amplified by the so-called "curse of locality," whereby localized processes do not scale linearly, resulting in inaccuracies when they are aggregated to the catchment scale or vice versa (Bierkens, 2015a). By comparing multiple models, researchers can explore whether large-scale patterns emerge naturally from fine-scale discretization or whether upscaling introduces biases (Clark et al., 2017).

Advances in computational power have driven the development of "hyper-resolution" hydrological models, capable of grid resolutions finer than 1 km (Beven and Cloke, 2012; Bierkens et al., 2015; Wood, Roundy, Troy, van Beek, Bierkens, Blyth, de Roo, et al., 2011). Although these models promise to capture spatial heterogeneity in land cover, topography, and hydrological processes in greater detail, they can also amplify uncertainties. These uncertainties stem from data quality, parameterization, and scaling assumptions (Blöschl and Sivapalan, 1995; Gupta et al., 1986). Higher-resolution models demand detailed inputs and extensive computational resources, making their application impractical in many cases. For instance, a model grid resolution of 200 m requires much greater computational capacity than a 3 km grid, which is 15 times coarser in each spatial dimension, raising questions about the cost-effectiveness of such fine resolutions for broader hydrological assessments.

Furthermore, as alluded to above, a critical issue arises from the mismatch between the scales at which observations are made and simulations are conducted. This discrepancy can lead to significant errors in model calibration and validation because models may not accurately capture processes observed at different scales.

Scaling challenges are particularly evident in the representation of upstream–downstream interactions and temporal dynamics within catchments. Misrepresentations of localized processes, such as infiltration rates in headwaters, can propagate errors downstream, resulting in inaccurate flood forecasts or water resource estimates (Clark et al., 2015). Moreover, discrepancies between observed and modeled behaviors often arise from inconsistent or sparse datasets, which exacerbate overparameterization issues in high-resolution models (Beven, 2019). These challenges emphasize the need for a careful balance between model complexity, data availability, and computational feasibility.

The contrast between empirical and physically based modeling frameworks adds further complexity to scaling. Empirical models, while computationally efficient and data-driven, often struggle to generalize beyond the conditions for which they were calibrated. In contrast, physically based models aim to represent fundamental hydrological processes but require extensive datasets and careful calibration to ensure accuracy across spatial scales (e.g., Wagener, 2004).

In Chapter 3, this thesis conducts a systematic evaluation of scaling issues by investigating the performance of the "wflow_sbm" model across varying spatial resolutions using the CAMELS-US dataset. This analysis explores whether hyper-resolution models deliver meaningful improvements in streamflow prediction and identifies potential biases introduced during the upscaling process.

1.5. LARGE-SAMPLE DATASETS AND MULTIFACETED EVALUATION

To better evaluate hydrological models, the hydrological community has increasingly turned to large-sample catchment datasets, such as MOPEX (Schaake et al., 1996), CAMELS-US (Addor et al., 2017), CAMELS-GB (Coxon et al., 2020b), and CARAVAN (Kratzert et al., 2022). These datasets compile standardized hydrometeorological time series, catchment attributes (e.g., land cover, topography, and soils), and metadata for hundreds of catchments within a consistent framework. By providing both forcing and evaluation data across diverse hydroclimatic regions, large-sample datasets allow for the systematic testing of hydrological models under a wide range of environmental conditions. This has enabled comparative studies that assess model robustness, explore the transferability of model parameters, and benchmark different model structures or calibration strategies (Andréassian et al., 2006; Gupta et al., 2014). Their availability also facilitates open, reproducible hydrological research.

Recent efforts in large-sample hydrology have expanded the scope of model evaluation by integrating new data sources and diagnostic approaches to better understand model behavior. Studies have used satellite-derived soil moisture to constrain model simulations (Albergel et al., 2018; van Dijk et al., 2011), incorporated evapotranspiration observations to assess model partitioning of water fluxes (Pan et al., 2020; Zhang et al., 2019), and applied multi-objective calibration to balance trade-offs across different hydrological variables (Vrugt and ter Braak, 2013; Vrugt et al., 2003). Others have investigated the propagation of input data uncertainties into model outputs (Bárdossy et

al., 2022; McMillan, Westerberg, and Trosset, 2010) or explored how discharge observation uncertainty affects model calibration and evaluation (Coxon et al., 2015; McMillan, Westerberg, and Krueger, 2018; Westerberg et al., 2020). Together, these studies demonstrate the potential of large-sample datasets not only for benchmarking but also for advancing diagnostic hydrology and uncertainty-aware model evaluation.

Despite their utility, large-sample datasets present unique challenges. Diagnosing poor model performance in specific catchments often requires additional data sources or novel diagnostic tools. This thesis adopts a multifaceted evaluation strategy that extends beyond traditional streamflow metrics. By incorporating satellite-based soil moisture, evapotranspiration estimates, and event-based diagnostics, this approach provides a comprehensive assessment of model reliability (Beven, 2023; Clark et al., 2005; Vrugt and Sadegh, 2013). Moreover, machine learning techniques are employed to link catchment attributes with model performance, revealing nonlinear interactions that inform future model refinements.

Chapter 4 details this exploratory approach, showing how multifaceted evaluation can generate testable hypotheses and improve understanding of model and catchment behavior.

1.6. OBSERVATIONAL UNCERTAINTY AND STATISTICAL SIGNIFICANCE

Hydrological model evaluation is constrained by observational uncertainties arising from measurement errors, interpolation, and data processing (McMillan, Westerberg, and Rahman, 2018). These uncertainties propagate through performance metrics, influencing model comparisons and the perceived adequacy of different approaches. Accounting for these uncertainties is critical for ensuring robust conclusions about model performance.

This thesis introduces a novel methodology for quantifying and integrating observational uncertainties into performance evaluations. By explicitly addressing how these uncertainties affect statistical significance, the proposed method ensures that model comparisons are not only rigorous but also reproducible. Chapter 5 elaborates on this approach, demonstrating its application in testing competing hydrological hypotheses.

1.7. EWATERCYCLE: CHALLENGES AND SOLUTIONS FOR MODEL HYPOTHESIS TESTING

Despite the diversity of hydrological models described above, many are difficult to compare directly because they have been developed in different programming languages over long periods (Hut et al., 2017; Hutton, Piper, and Tucker, 2016). The resulting disparity poses challenges for users comparing different models and competing hypotheses (e.g., Addor et al., 2018; Cunderlik and Burn, 2013). The consequences are limited knowledge exchange and suboptimal collaboration on model refinement and improvement. This has often led to pragmatic model selection driven by accessibility and familiarity

rather than suitability for the research question at hand (Addor and Melsen, 2019). Consequently, this practice has considerably impeded progress in hydrological science (Hut, Drost, Alidoost, et al., 2021).

Even more critically, practical challenges have contributed to a lack of reproducibility in modeling studies, undermining scientific confidence and transparency. Reproducibility depends on the availability of data, source code, and documentation. All three are dynamic and therefore require careful documentation and archiving at the time of publication. For example, catchment heterogeneity often leads researchers to calibrate (tune) models using local data for a specific region. If the calibration procedure is poorly described, the software or documentation is incomplete, or the underlying data are not fully available or archived, it becomes impossible to replicate the original study.

In recent years, **Open Science** and the **FAIR** (Findable, Accessible, Interoperable, Reusable) principles have been embraced as guiding frameworks to address these challenges (Wilkinson et al., 2016). The Open Science initiative stipulates that researchers should share the data and software associated with their publications. The FAIR principles state that data and software must be Findable, Accessible, Interoperable, and Reusable.

One approach involves creating integrated modeling environments, such as wflow (Schellekens et al., 2020a), SUMMA (Clark et al., 2015), MARRMoT (Knoben, Freer, Fowler, et al., 2019), and Raven (Craig and Horritt, 2020), that consolidate diverse conceptual or physically based models within a single software suite. While these frameworks streamline the process of switching between model structures, they may require extensive modifications to align distinct models under a single "umbrella."

An alternative approach, central to this thesis, is the eWaterCycle framework. The contributions of this thesis include software development and the design of model experiments (Hut et al., 2022; Wiersma et al., 2022; see also Chapter 2). The eWaterCycle platform is community-driven and designed according to the principles of Open Science and FAIR. Rather than merging different models into a monolithic codebase, eWaterCycle uses a Basic Model Interface (BMI) (Hutton, Piper, and Tucker, 2016; Hutton et al., 2020) to standardize interactions with each model. This interface preserves the models' original code and structure, allowing researchers to manage version control and data dependencies for each experiment while maintaining full reproducibility. In addition, eWaterCycle's flexible architecture permits researchers to explore novel model couplings, such as integrating a hydrological model with a glacier model (Wiersma et al., 2022) or linking it to agricultural systems (Chevuru et al., 2024), thereby broadening the platform's scope and versatility.

A key advantage of eWaterCycle is its capacity to standardize model intercomparison experiments (e.g., Kauffeldt et al., 2016; Schellekens et al., 2017). Researchers can rely on a common set of input–output protocols, calibration methods, and evaluation metrics, thereby reducing inconsistencies that arise when tailoring each model configuration separately. By simplifying these steps, eWaterCycle lowers barriers to comprehensive hypothesis testing and model refinement. New users can experiment with multiple models side by side, gaining insight into which model best suits their research objectives, all within a common open-source environment that supports collaborative science.

1.8. SCOPE OF THE THESIS

This thesis seeks to improve methodologies for hydrological model comparison and refinement, with a strong emphasis on reproducibility through the eWaterCycle framework. Using large-sample catchment datasets, the thesis addresses several key research topics:

- **Chapter 2** introduces the core principles of eWaterCycle and demonstrates how these principles foster robust, transparent model experimentation. It provides multiple application examples that highlight the platform's adaptability and showcase the research opportunities it enables.
- **Chapter 3** investigates the impact of increasing spatial resolution in physically based distributed models within the "hyper-resolution" domain. The objective is to determine whether higher-resolution simulations result in meaningful improvements in streamflow prediction accuracy.
- **Chapter 4** demonstrates an exploratory approach to model comparison, incorporating diverse observational products and multiple competing models. Applying machine learning uncovers nonlinear relationships between catchment attributes and model performance, offering new insights into model and catchment behavior.
- **Chapter 5** introduces and tests a novel, straightforward method for estimating the statistical significance of model performance differences in the context of observational uncertainty. This approach provides a more robust basis for evaluating competing hypotheses and interpreting model results.
- **Chapter 6** synthesizes the main findings, discussing how these advances contribute to a better understanding of the interplay between small-scale processes and large-scale patterns. Finally, the chapter outlines future research directions and potential next steps in hydrological modeling.

2

THE EWATERCYCLE PLATFORM FOR OPEN AND FAIR HYDROLOGICAL COLLABORATION

Big wheel keep on turnin'.

Creedence Clearwater Revival, *Proud Mary*

eWaterCycle is a platform that separates the experiment done on the model from the model code. In eWaterCycle, hydrological models are accessed through a common interface (BMI) in Python and run inside of software containers. In this way, all models are accessed in a similar manner facilitating easy switching between models, model comparison and model coupling. Currently the following models and model suites are available through eWaterCycle: PCR-GLOBWB 2.0, wflow, HYPE, LISFLOOD, TopoFlex HBV, MARRMoT and WALRUS. While these models are written in different programming languages, they can all be run and interacted with from the Jupyter notebook environment within eWaterCycle. Furthermore, the pre-processing of input data for these models has been streamlined by making use of ESMValTool. Forcing for the models available in eWaterCycle from well-known datasets such as ERA5 can be generated with a single line of code. To illustrate the type of research that eWaterCycle facilitates, this manuscript includes five case studies: from a simple 'Hello World' where only a hydrograph is generated to a complex coupling of models in different languages.

In this chapter, we describe the design choices made in building eWaterCycle and provide all the technical details to understand and work with the platform. With the eWaterCycle platform, we are providing the hydrological community with a platform on which to conduct research that is fully compatible with the principles of Open Science and FAIR science.

2.1. INTRODUCTION

In hydrology, scientists try to better quantify the movement of water into, out of, and through the land surface and rivers in order to better predict droughts, floods, navigation hazards, and reservoir operations (Wood, Roundy, Troy, van Beek, Bierkens, Blyth, de Roo, et al., 2011). In addition, a better understanding of hydrological processes will allow determination of anthropogenic and climate change impacts on the hydrological cycle (McMillan et al., 2020).

From a hydrological point of view, every field, every street, and every part of the world is different. We may have adequate descriptions of how water moves through plants and soils at small scales, but the medium is never the same from one spot to the next. This is the curse of locality (Bierkens, 2015b). Nonlinear processes need to be integrated over time and space in the presence of tremendous natural and manmade heterogeneity.

Traditionally, the hydrological community has dealt with this curse by building custom models for small natural watersheds (Beven, 2001a). Hydrologists often work with *effective parameters* (Bárdossy and Singh, 2008; Kirchner, 2006), parameters that cannot be measured directly but describe aggregated movement of water through the environment. Typical models are partially based on first-principles physics (mass, momentum, and energy balances), and partially on (statistical or heuristic) assumptions that simplify the complex reality.

Local process knowledge is used to conceptualize many local models by hydrologists from all over the world. This leads to a plethora of local models exhibiting great diversity in the exact methodologies applied, competing hypotheses of hydrologic behavior, technology stacks, and programming languages used in these models (Hut et al., 2017; Hutton, Wagener, et al., 2016). Although a wealth of knowledge is encoded in these models, this knowledge is hardly shared due to the technical difficulties of working with models created by other researchers. As a result, models are often chosen based on availability and familiarity, rather than suitability for the research performed (Addor and Melsen, 2019), severely hindering scientific progress in the hydrological community.

Fortunately, the problem of limited sharing of scientific results has recently been picked up as an important topic by the wider academic community. Open Science, where data, publications, and software are shared publicly, is deemed more and more important, both by researchers and funders (Hall et al., 2021). In addition, the FAIR principles (Wilkinson et al., 2016) describe that all data resulting from research must be made Findable, Accessible, Interoperable, and Reusable (FAIR). Recently, FAIR has also been extended to software[1] and other academic output.

To enable significant progress in the sharing of hydrological knowledge, we introduce the concept of FAIR Hydrological Models. FAIR Hydrological Models make it possible for other researchers to use a model to generate novel scientific results without needing extensive support from the original authors. Any preexisting hydrological model can

Parts of this chapter have been published verbatim in Hut, R., Drost, N., van de Giesen, N., van Werkhoven, B., Abdollahi, B., **Aerts, J.**, Albers, T., Alidoost, F., Andela, B., Camphuijsen, J., Dzigan, Y., van Haren, R., Hutton, E., Kalverla, P., van Meersbergen, M., van den Oord, G., Pelupessy, I., Smeets, S., Verhoeven, S., de Vos, M., & Weel, B. (2022). The eWaterCycle platform for open and FAIR hydrological collaboration. Geoscientific Model Development, 15(12), 5371–5390. https://doi.org/10.5194/gmd-15-5371-2022. Personal contributions have been made to the conceptualization, software code, use cases, and text of this publication.

[1] https://fair-software.eu/

be made FAIR by adding open interfaces and documentation. With a model that has been made FAIR, it is not only possible to re-create the experiments done by the authors of that model, but also to perform novel research by applying the model to different locations, settings, or input data. For a model to be FAIR, not only must the software and required data be available, but the model must also be properly documented and have well-defined interfaces. FAIR Hydrological models have a lower barrier to entry and create scientific results that are as open and FAIR as possible, thereby truly enabling researchers to build on each other's results.

In this paper, we introduce the eWaterCycle Platform, a platform where hydrologists can work with each other or add their own FAIR models. To the best of our knowledge eWaterCycle is the first platform for hydrological modelling that focuses on providing access to pre-existing models and data sources in a way in which the platform handles the computer science and data science aspects to allow the hydrologists to focus on the hydrology. The goal of the eWaterCycle Platform is to be a trailblazer for making hydrological modeling Open and FAIR. The eWaterCycle platform is designed to support researchers including graduate students (MSc and PhD levels) to run hydrological experiments with ease, focusing on the science rather than the technology. The platform is designed to enable users to run a simple experiment, such as generating a hydrograph for a certain catchment with a certain model, within minutes of getting started, while also giving users the freedom to perform very advanced experiments, such as multi-model coupling and intervening in model states during run-time. With this, we aim to reduce the cycle time in going from idea to experiment, from months to days, and to support fully reproducible experiments. We embrace what has been done already. Rather than rewriting all models from scratch, we share, re-use, couple, and build on existing models. The methods and technology developed within the eWaterCycle platform are re-usable, both inside and outside of Hydrology.

To illustrate how to use the eWaterCycle platform as a hydrologist and demonstrate the types of experiments one can conduct on the platform (coupling, calibrating, comparing scenarios, etc.), a series of Jupyter notebooks that showcase the platform is provided with this paper. For scientists who want to work with the platform as users, a separate video in which a hands-on demonstration of the models is given is available on YouTube at https://youtu.be/eE75dtIJ1lk and, for archiving purposes, also on Zenodo. The remainder of this paper covers the rationale and the technology behind the platform.

In the eWaterCycle project, we strongly believe that the best approach for generating impact is to build on existing efforts as much as possible. We make use of container technology (specifically Docker[2] and Singularity[3]) that allows capture and preservation of software environments. Pre-processing of forcing data is done using the ESMValTool (Righi et al., 2020), originally developed in the climate sciences. We use Jupyter[4] as the main user interface to our system. The Basic Model Interface (BMI, Hutton et al., 2020)

[2] https://www.docker.com
[3] https://sylabs.io/singularity/
[4] https://jupyter.org

provides a stable, easy-to-implement interface to an existing model. For sharing software, data and results in a FAIR manner we rely on GitHub[5], HydroShare (Horsburgh et al., 2015) and Zenodo[6]. Any software contributions that we create in the eWaterCycle project are purposely small and independent, so there is a high chance our components are, in turn, reusable by other projects.

The eWaterCycle platform can incorporate with ease any existing model, be it conceptual, semi-distributed or distributed, in any commonly used programming language. An alternative and often used approach to making hydrological models FAIR is to create a single model framework that incorporates as many model concepts as possible. This usually requires significant modifications to a model code, and sometimes even a complete rewrite, to fit within the model framework. The result is a coherent set of models, that can be interchanged relatively easily. Examples of such model frameworks include wflow (Schellekens et al., 2020b), SUMMA (Clark et al., 2015), MARRMoT (Knoben, Freer, Fowler, et al., 2019), and Raven (Craig et al., 2020). In a way, this approach can be seen as orthogonal to the eWaterCycle platform, as both approaches can be combined. In the eWaterCycle platform, we have incorporated the wflow and MARRMoT frameworks, making it possible to use any of the models within these frameworks within eWaterCycle. Incorporating the SUMMA framework is on the long-term list of objectives for the eWaterCycle platform as well, but has not been finished at the time of writing.

There are several other platforms that support Open and FAIR hydrology that eWaterCycle connects to. HydroShare (Horsburgh et al., 2015) focuses on making hydrological data FAIR. It offers a service to publish datasets, and access datasets, among others in a Jupyter notebook environment. Unlike eWaterCycle, HydroShare offers no support in using datasets for an experiment, and simply provides data as is. HydroShare can be used to store data resulting from eWaterCycle experiments in a FAIR manner. A more structured way of acquiring data for use in a Hydrological model is used in HydroDS (Gichamo et al., 2020). HydroDS provides a web service where users can call upon the service to download data in a format suitable for their model. HydroShare and HydroDS can also be combined to generate and store data needed to run models (Gan et al., 2020).

The Community Surface Dynamics Modeling System (CSDMS) (Tucker et al., 2021) community gathers a large number of hydrological models in a model repository. This repository contains metadata on models, as well as the source code. In addition, model developers are encouraged to add a BMI interface to a model to facilitate cooperation on models. BMI simplifies the use of codes but often the installation and compilation of scientific code are non-trivial and prove a practical bottleneck. The eWaterCycle platform builds on the BMI interface using containerized models offering an easily reproducible model software environment. This includes the support for generating forcing and other needed input for each model, allowing scientists to build on all data and models eWaterCycle provides access to.

The example Jupyter notebooks provided with this paper demonstrate how hydrologists can, for example, couple two models written in different programming languages, calibrate models, or run "what if?" scenarios, using existing models from research groups

[5] https://github.com
[6] https://zenodo.org

all over the world, forced with datasets from different data providers, all without having to install a single package on their own laptops. To make full use of the knowledge created in hydrology, we do need to be able to stand on each other's shoulders, which eWaterCycle facilitates.

The rest of this paper is organized as follows. Section 2.2 introduces the eWaterCycle platform and describes the different functionalities it provides. It explains how each part of the platform contributes to Open and FAIR hydrological science. Section 2.3 presents the technical design and implementation of the platform. Section 2.4 presents a number of case studies that demonstrate the capabilities of the eWaterCycle platform and highlights the diverse set of supported use-cases. Section 2.5 concludes and discusses future work.

2.1.1. GLOSSARY

In this paper, we use the following terminology. We acknowledge that different fields of science and even different scientists within fields may use different definitions of these terms (Venhuizen et al., 2019) and provide these definitions solely to clarify how we use those terms in this paper and within the eWaterCycle platform.

- Hydrological model. A piece of software code that calculates stores and fluxes of water in, through, and on the surface of the Earth. Most hydrological models need forcing (input) such as precipitation and produce outputs such as river discharge. Many hydrological models require parameters.
- Parameter. A constant input (in time, not necessarily in space) that a model needs to calculate the outputs. Parameters can either be calibrated based on (often historical) forcing and output data, or parameters can be derived from third-party sources, including, but not limited to, digital elevation maps, soil maps, etc. A parameter does not change over the runtime of the experiment.
- Forcing. A time-varying input that a model needs to calculate the outputs. In hydrological models, the most common forcing is precipitation data.
- Input(s). Forcings and parameters.
- State. All variables calculated by the model that are needed to calculate the next time step.
- Output(s). Any variable calculated by the model that is stored or shared with the experimenter and can be used for analysis. Outputs can be either state variables or variables derived from state variables (and inputs). For example, in the PCR-GLOBWB model 'channel storage' is a state variable that is updated every timestep; 'river discharge' is a value calculated using 'channel storage'. Both 'river discharge' and 'channel storage' are outputs of the PCR-GLOBWB model.
- Observations. Any data derived from observations, direct or indirect, of the Earth system. Observations can be used as forcing, such as precipitation, as parameters, for example soil maps, or as validation for a model output, such as the often used river discharge.

- FAIR hydrological model. A hydrological model that is Findable, Accessible, Interoperable, and Reusable. See Wilkinson et al. (2016). Note that 'Accessible' in this context means that it must be clear to anyone how the model can be accessed and not necessarily that it is openly available to anyone.

- Open Science. The principle of openly sharing all aspects of the scientific endeavour within ethical and legal limits. See Hall et al. (2021).

- Experiment. A set of hydrological model runs that use inputs and generate and analyse outputs. An experiment can include actively intervening in the state of the model during runtime. Within eWaterCycle, an experiment is described in a Jupyter notebook. Experiments can be as simple as generating a single hydrograph for a single catchment, or as complicated as coupling multiple global models and forcing them with different datasets. See Section 2.4 for example experiments in the eWaterCycle platform.

- eWaterCycle platform. The combination of the core eWaterCycle software, all models contributed to eWaterCycle, and all available input datasets. At the time of writing, the platform is hosted on demonstration infrastructure provided by SURF at lab.ewatercycle.org. The platform is designed to be deployed by system administrators on any sufficiently high-performance infrastructure. Depending on future funding streams, eWaterCycle will be made available on publicly accessible infrastructure for the entire hydrological community.

2.2. THE EWATERCYCLE PLATFORM

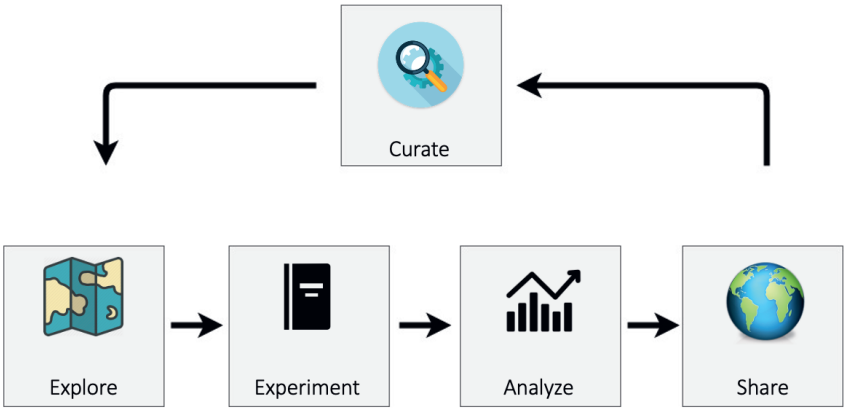


Figure 2.1: An overview of the components of the eWaterCycle platform. The design of the eWaterCycle platform closely follows the typical workflow of running hydrological experiments.

Figure 2.1 presents an overview of the eWaterCycle platform. The design of the eWaterCycle platform closely follows the typical workflow for running hydrological experiments. A hydrological researcher (henceforth called the user) accesses the eWaterCycle platform using only a web browser.

2.2.1. EXPLORER

The user starts at the data and model explorer, which shows a geographic map with data sets and associated models that can be instantiated. Once a data set, model, parameter set, and forcing have been selected, the notebook generator generates a Jupyter Notebook containing a basic hydrological experiment for the chosen selection. Jupyter notebooks combine cells where Python code can be run and cells where text can be added (in Markdown). This combination makes notebooks ideally suited to execute and communicate experiments. While the experiment code is written in Python, model code in eWaterCycle can be written in most commonly used languages; see Section 2.3.5 below for how this is achieved.

2.2.2. EXPERIMENT

Notebooks have the advantage of being easy to access and present the user with an interface that is flexible enough to be adapted by the user to their specific experiment. The user accesses the notebooks through a browser. The notebooks are hosted (executed) on a designated server. The model can, but need not, be executed on the same system. Since models are accessed from the notebook through remote procedure calls (see Section 2.3.5), models can be launched remotely on, for example, a dedicated HPC system. At the time of writing, the eWaterCycle platform is hosted on infrastructure from SURF, the infrastructure provider of the Dutch Academic Community (surf.nl). Anyone collaborating with a Dutch partner can access this infrastructure. This is demonstration infrastructure intended to show the capabilities of the platform. Anyone with a budget on the SURF Research Cloud can start up an instance of eWaterCycle there. Those without access to this resource can install the software of the eWaterCycle platform on their own infrastructure; see Verhoeven et al. (2021) for details. See section 2.5 for future plans regarding making the platform more broadly accessible to the hydrological community.

2.2.3. ANALYZE

The code in the generated notebook, when run, results in a hydrograph for the chosen combination of model, area and forcing dataset and includes discharge observation data obtained through the Global Runoff Data Centre (GRDC)[7]. This notebook calculating a hydrograph is an excellent starting point for hydrologists to conduct their research without having to set up the model, preprocessing, observation, and evaluation pipeline all from scratch. In addition to the generated notebooks, a list of tutorial notebooks is also available for different typical use-cases, including common calibration methods. The Jupyter notebook environment provides an excellent platform for data analysis, and there are many different libraries and tools available, including, for example, HydroStats (Roberts et al., 2018).

[7] The Global Runoff Data Centre, 56068 Koblenz, Germany, https://www.bafg.de/GRDC

2.2.4. SHARE

Once a user is ready to share the results created within the platform, to be completely FAIR and Open, the experiment, data, and results should be published in a data repository such as Zenodo, HydroShare, Figshare and ESGF. Currently, this is done by manually uploading the notebook and outputs to those services. In a future version of eWaterCycle, this process will be automated. Finally, to allow even greater re-use of models and datasets within the eWaterCycle platform, after curation, models and data sets can be added to the set of available items in the platform and the associated software package by users making pull requests on the GitHub repository of the eWaterCycle platform.

2.2.5. PRE-PROCESSING

For pre-processing of forcing data, ESMValTool (Righi et al., 2020), a community diagnostic and performance metrics tool for evaluation of Earth system models, has been adapted (see Section 2.3 below). Any dataset can be made available to the hydrological community by making the dataset compatible with the CMOR format, a *de facto* standard from the climate science community natively supported by ESMValTool (see paragraph 4.2 of Righi et al. (2020)). Any dataset already made ready for ESMValTool ('CMORized') can be added to eWaterCycle by having the eWaterCycle installation point to the location of those data sources upon installation of the platform, as explained in Verhoeven et al. (2021). Currently, both ERA-Interim (Dee et al., 2011) and ERA5 (Hersbach et al., 2020) are available as forcing datasets within eWaterCycle.

2.2.6. AVAILABLE MODELS

Any pre-existing hydrological model can be made more Interoperable (part of FAIR, Wilkinson et al., 2016) by adding the BMI interface as explained in Section 2.3.5. The following hydrological models or model suites are currently integrated into, or being integrated into, the online eWaterCycle platform for use by all hydrologists:

- PCR-GLOBWB 2.0 (Sutanudjaja et al., 2017)
- wflow (Schellekens et al., 2020b)
- HYPE (Lindström et al., 2010)
- LISFLOOD (Van Der Knijff et al., 2010)
- TopoFlex HBV (Gao et al., 2014)
- MARRMoT (Knoben, Freer, Fowler, et al., 2019)
- WALRUS (Brauer et al., 2014).

Future models and frameworks that will be added by the eWaterCycle project team include GlobWat (Hoogeveen et al., 2015), SUMMA (Clark et al., 2015) and mHM (Samaniego et al., 2021). The models already in eWaterCycle and those still to be added differ greatly in their underlying hypotheses of hydrological processes and methodologies applied, and they are implemented by different communities using different programming languages. The eWaterCycle platform removes technological barriers for these communities to work together more easily.

2.3. TECHNICAL DESIGN OF THE EWATERCYCLE PLATFORM

The eWaterCycle platform uses the following design philosophy for building its software stack. Our first choice is to build upon existing software. When we cannot re-use existing software directly for our needs we try to contribute new features to existing projects, and only as a last resort we develop new reusable standalone tools.

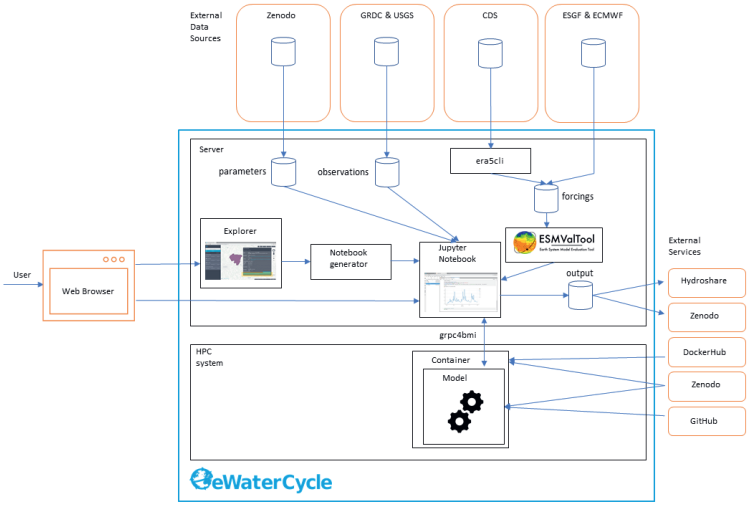


Figure 2.2: The technological design of the eWaterCycle platform, including external services. This overview shows the different components that are, on purpose, built as separate entities to facilitate re-use in other projects. The different components of this technical overview are discussed in detail in the main text of this publication.

Figure 2.2 shows the technical design of the eWaterCycle platform. The interface to the system is through a web interface. As the user only needs a web browser to access the eWaterCycle platform, there is no need to install any other software locally, which ensures a low barrier to entry to the system.

The core of the eWaterCycle platform is a collection of services. These are generally deployed on a dedicated server with ample storage for datasets required to run hydrological models and output. A guide for system administrators on how to install the platform is available in the documentation (Verhoeven et al., 2021). To ensure all data are FAIR, all data in the eWaterCycle platform are downloaded from stable repositories such as Zenodo (general datasets and parameters), the GRDC (runoff observations), ECMWF (ERA5 and ERA-Interim forcing), and ESGF (Climate Model Data). In addition, required (model) software is downloaded from external sources such as Docker Hub, GitHub and Zenodo.

In turn, the results of experiments done on the eWaterCycle platform can be exported to FAIR data repositories such as HydroShare and Zenodo. This, together with the FAIR input data, ensures that results generated using the eWaterCycle platform do not depend on the specific server to be sustained indefinitely.

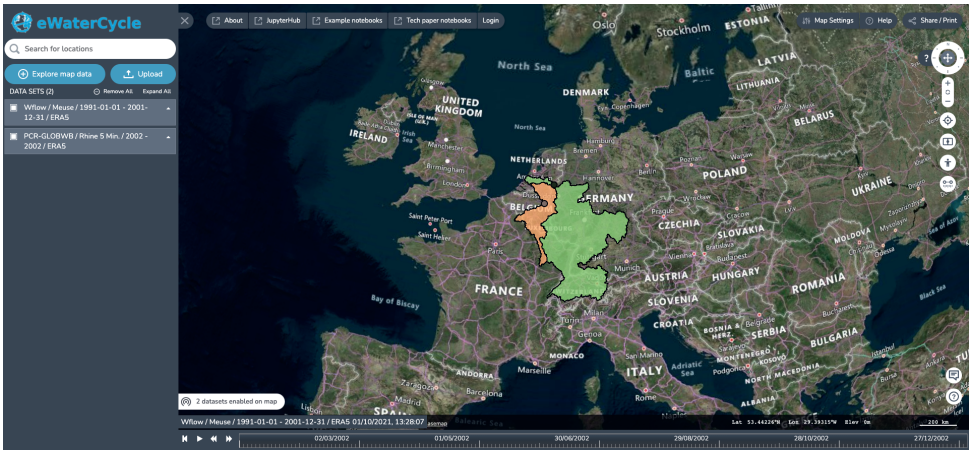


Figure 2.3: Screenshots of the Explorer, which allows users to explore and select models available based on regions for which they are available (i.e., suitable forcing and parameter sets are available). Background map screenshots ©Microsoft Bing Maps 2018, reprinted with permission from Microsoft Corporation.

The rest of this section discusses the individual components of the eWaterCycle platform that required significant development effort from the eWaterCycle team. The core of the software stack that runs the eWaterCycle platform has been released as a Python package. Information on how to install this package and all dependencies is provided in the documentation.

2.3.1. THE EXPLORER

The Explorer is a web-based geospatial data explorer for hydrological models and data sets. Using this graphical user interface, users get an overview of the hydrological models available for various regions or catchments and the available data sets at various resolutions. The user can select a combination of models and datasets and configure the experiment they want to set up through this interface.

2.3.2. THE NOTEBOOK GENERATOR

Notebook environments are increasingly popular. They are used to conduct research, teach the next generation of hydrologists, analyze data, and provide advanced access to large data sets and HPC resources. The notebook generator creates a notebook specifically for the settings chosen by the user, as shown in Figure 2.4. The notebook can subsequently be run by the user to perform the requested computation or analysis. In the default use-case, the notebook contains the code for running a hydrological model and creating a first hydrograph. Because it is a notebook and not a ready-made user interface with limited options, the user is free to modify the notebook. This method thus allows novice users to get going quickly, while allowing advanced users all the freedom they require. The notebook also forms the perfect basis for a user to start tinkering with the experiment.

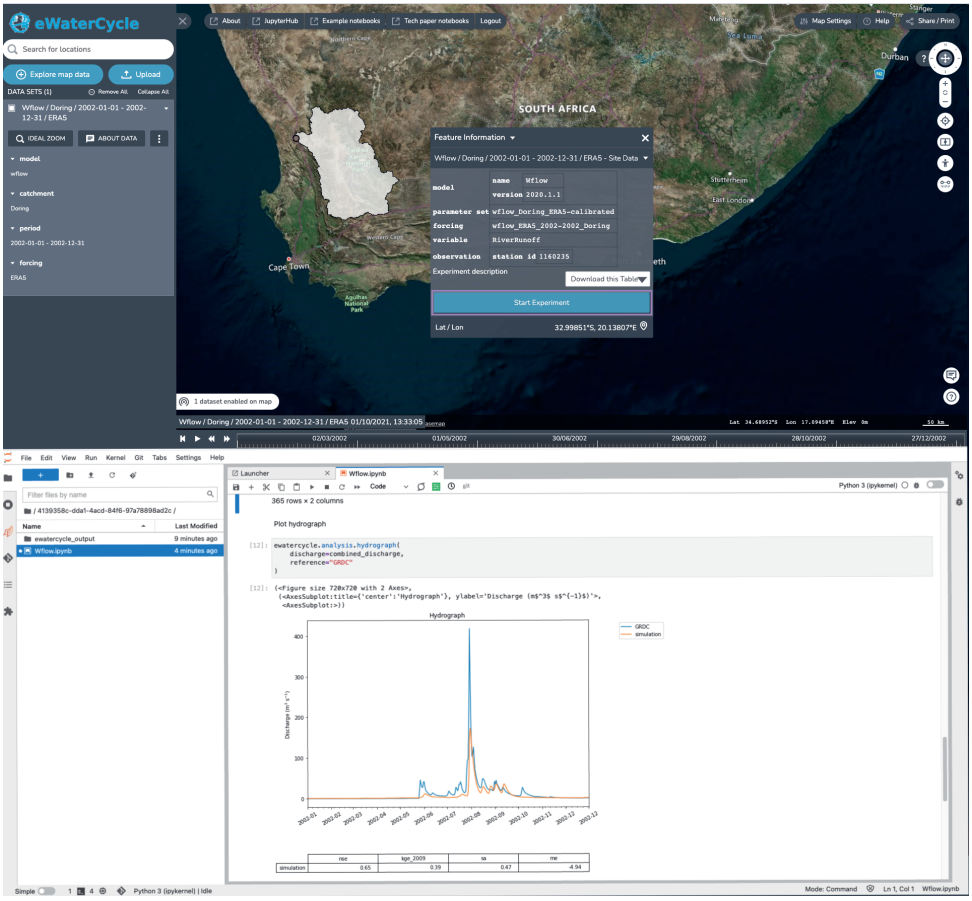


Figure 2.4: The Notebook Generator, which generates a notebook based on the experimental setup configured in the Explorer. After selecting a model and clicking 'start experiment', a Jupyter notebook is generated and launched with all the code needed to generate a hydrograph for the selected model in the selected region. By providing this working notebook, researchers have a good starting point from which to change the notebook to answer their own research questions. Background map screenshots ©Microsoft Bing Maps 2018, reprinted with permission from Microsoft Corporation.

2.3.3. DOWNLOADING ERA5 WITH ERA5CLI

With the release of the ERA5 dataset (Hersbach et al., 2020), worldwide high-resolution reanalysis data became available with open access for public use. The Copernicus CDS (Climate Data Store) offers two options for accessing the data: a web interface and a Python API. Consequently, automated downloading of the data requires advanced knowledge of Python. Following our design philosophy of building reusable standalone tools, we have created era5cli to simplify the process of downloading ERA5 data (Haren et al., 2023). The command-line interface tool era5cli enables automated downloading of ERA5 using a single command. All variables and options available in the CDS web form are now available for download in an efficient way. Both monthly and hourly ERA5 datasets are supported. Besides automation, era5cli adds several useful functionalities to the download pipeline, such as spreading a single download over multiple CDS requests, and saving files in either GRIB or netCDF. Within the eWaterCycle platform era5cli is used by administrators to download relevant selections from ERA5. Users working in the notebook environment need not work with this command-line tool to be able to use eWaterCycle for their research. Source code is available through Haren et al., 2023.

2.3.4. ESMVALTOOL-BASED MODEL INPUT PRE-PROCESSOR

A large barrier to using a new model or a new dataset for any hydrologist is preparing model input data. In general, this is different for every model as, in general, data requirements and data preparations differ for most models. The preparation steps are often performed by various sets of scripts that may or may not be included with the model code, which hamper reproducible science. However, there generally is a lot of overlap between the data preparation steps for different models, and as such it would be a valuable asset to the hydrological community if the pre-processing of the input data is done in an open and FAIR manner.

To that end, we decided to extend ESMValTool (Righi et al., 2020), a community diagnostic and performance metrics tool for evaluation of Earth system models in CMIP, instead of writing model-specific preprocessing scripts. The ESMValTool pre-processing functions cover a broad range of operations on data before diagnostics or metrics are applied; for example, vertical interpolation, land-sea masking, re-gridding, multi-model statistics, temporal and spatial manipulations, variable derivation, and unit conversion. The pre-processor performs these operations in a centralized, documented and efficient way. The current pre-processing pipeline of the eWaterCycle using ESMValTool consists of hydrological model-specific recipes and supports ERA5 and ERA-Interim data provided by the ECMWF (European Centre for Medium-Range Weather Forecasts) through the Climate Data Store (CDS). The pipeline starts with the downloading and CMORization (Climate Model Output Rewriter) of input data. CMORization standardizes the data to make sure that the data is CF-compliant data and follows the CMOR tables. See the ESMValTool documentation for more information on CMORization (https://docs.esmvaltool.org/en/latest/develop/dataset.html). Following CMORization, a recipe is prepared to find the data and run the preprocessors. An ESMValTool recipe contains model-specific code to derive forcing variables required by the model and stores provenance information to ensure transparency and reproducibility. CMORization is dataset-specific and recipes are model-specific. This means that after CMORization a dataset is available

for all models and a model-specific recipe does not have to be adjusted for a different forcing dataset. Most recipes take a shapefile as input, so, once created, the recipe can derive forcing data for any region where data is available in the dataset.

ESMValTool can also be used to pre-process datasets other than ERA5 and ERA-Interim. Examples include climate model results created as part of the Coupled Model Intercomparison Project (CMIP; Eyring et al., 2020), distributed through the ESGF platform. Our recipes and other additions to ESMValTool have been merged into the ESMValTool main version. This approach ensures that data preprocessing routines are openly available, documented, and re-used.

If hydrologists want to work with their own data sources as forcing for their models within the eWaterCycle platform, they can follow the steps in the ESMValTool documentation (https://docs.esmvaltool.org/en/latest/input.html) to make their data available. Making their data available for one model, given the way ESMValTool is set up, makes it immediately available for other models as well.

2.3.5. INTERFACING MODELS THROUGH GRPC4BMI

Hydrological models are written in many different programming languages and often require specific versions of supporting software packages to function as intended by the original model developer. To give hydrologists access to these models without having to learn different programming languages, a common interface to the models is needed. We use the Basic Model Interface (BMI; Hutton et al., 2020) as our main API for models. BMI exposes functions for controlling the model time stepping and retrieving and manipulating the model's state at any given moment. BMI is designed specifically to make it easy to implement the interface in any given model. BMI is very forgiving of model structure and is defined for many languages; see https://csdms.colorado.edu/wiki/BMI for more details. By using grpc4bmi, we can call any model that has implemented BMI from the eWaterCycle platform. We are currently supporting C/C++, Fortran, Octave (an open-source version of MATLAB) and R but can add support for any language that has a gRPC library.

To make sure that models are always run using the correct additional libraries and other dependencies, in eWaterCycle, models are run inside software containers. A software container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably on any compute environment. Where model code might break down if a dependency it relies on is no longer supported, packaging a model with its dependencies guarantees that the model can be run on any infrastructure that supports running of containers, thus prolonging the lifetime of the model's codebase. Communication with a container, for example to instruct a model to run for one time step, happens through well-defined channels that can pass procedure calls. The containers are openly available on Docker Hub or Zenodo, to promote re-use by others. In eWaterCycle, models written in different programming languages can be added to the platform inside a container, while the Jupyter notebook environment that runs the experiments runs Python. There is, therefore, a need for a tool to translate BMI calls from Python to other programming languages.

To 'translate' between different BMI versions we have used Google's protocol buffer framework (gRPC) and developed grpc4bmi as a standalone software package that allows us to interface with code written in any programming language from our Jupyter Notebook environment. grpc4bmi wraps a BMI-enabled model into a server process, possibly executed within a container and/or on a remote system, and transfers client-side BMI calls to the running model instance. Using gRPC, grpc4bmi establishes the communication between the Jupyter notebook and the hydrological model.

grpc4bmi allows the user to address a model via a standard Python BMI, irrespective of the model's programming language and installation requirements, and allows the coupling of models and the running of multiple instances of the same model. grpc4bmi serves as a key component of the eWaterCycle platform and is a valuable tool for reproducible analysis and online coupling of BMI-enabled models.

A complete overview of the interface built in eWaterCycle can be found in the documentation by Verhoeven et al. (2021).

2.4. USE-CASES (CASE STUDIES, APPLICATION EXAMPLES)

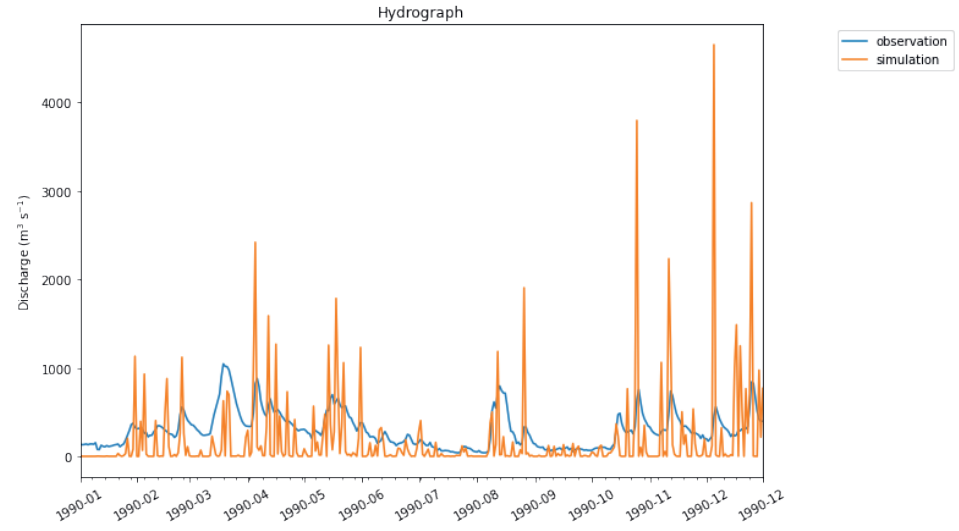
In this section, we present a number of case studies that illustrate the capabilities of the eWaterCycle platform and the use-cases it supports. These illustrate the application of the Explorer and Notebook Generator (Section 2.4.1), how grpc4bmi allows users to switch between different models (Section 2.4.2), coupling models in different programming languages (Section 2.4.3), performing experiments on a model's internal state while the model is running (Section 2.4.4), and calibrating a model using a remotely executed ensemble (Section 2.4.5).

The case studies presented here are chosen to demonstrate the features of the platform in a clear way. They all feature only a single catchment. Current hydrological modeling studies often include many catchments in their analysis. The platform fully supports those more complex studies, as shown by Aerts et al. (2021), who studied the impact of different spatial resolutions of the same model for 454 catchments from the CAMELS dataset.

All figures presented with these use-cases are generated with the eWaterCycle platform and have not been optimized for printing, to show what the output of experiments done with eWaterCycle looks like. The Jupyter notebooks for these use-cases are provided in a separate GitHub repository, to facilitate adding additional notebooks at later stages. For the exact notebooks used in this study, a release with associated DOI has been made (Hut, Drost, van de Giesen, van Werkhoven, et al., 2021).

2.4.1. HELLO WORLD: ONE MODEL, ONE CATCHMENT, ONE FORCING

Our first application is the most basic notebook that is produced by the Notebook generator after the experiment has been configured in the Explorer. This notebook includes all the code and configuration needed to download and pre-process model input data, and forcing data from ERA5. After pre-processing is finished, the code to set up, initialize, and run the model while capturing discharge estimates is given. And finally, there is code to download observation data from GRDC for the corresponding GRDC station,



	nse	kge_2009	sa	me
simulation	-4.33	-0.56	1.09	-117.53

Figure 2.5: A first hydrograph generated with eWaterCycle. This shows the output of the MARRMoT M01 model (a single bucket representing the entire Merrimack basin) when ERA5 is used as forcing data. While the over-simplification of representing the entire basin with a single bucket is clear, the timing of high-flow periods is still rather well represented. This figure is generated using the hydrograph() function, a standard method within the eWaterCycle package. In addition to plotting a hydrograph of the provided time series, it calculates often used metrics like the Nash–Sutcliffe efficiency and the Kling–Gupta efficiency. This example is used as a first 'hello world' use case illustrating with a simple model how the eWaterCycle platform supports standard workflows often used in computational hydrological research.

and plot a hydrograph of both the modeled and observed discharge over the simulated period, as shown in Figure 2.5.

For this specific example, we have chosen to use the MARRMoT (Knoben, Freer, Fowler, et al., 2019) model suite. MARRMoT is a suite of conceptual catchment models written in the MATLAB programming language in which the user can specify the model structure through setting files. In eWaterCycle, these setting files are generated when the model instance is created. Specific settings, such as the maximum soil moisture storage in this case, can be set by the user in the notebook. Here we have chosen the most basic model available: a single bucket. Figure 2.6 shows the central part of the notebook where the model is run. The Merrimack basin is chosen as the catchment and ERA5 is used as forcing dataset. While the Merrimack is too large and complex a basin to be represented by such a simple model, this 'Hello World' example shows that in eWaterCycle any model, lumped or distributed, can be run using any available forcing dataset for any

```
In [12]: simulated_discharge = []
         timestamps = []

         while (model.time < model.end_time):
             model.update()
             timestamps.append(model.time_as_datetime.date())
             # Marrmot M01 is a lumped model, so only single value is returned
             value_in_mmday = model.get_value(discharge_variable)[0]
             # Convert from mm/day to m3/s
             value = value_in_mmday * merrimack_area * conversion_mmday2m3s
             simulated_discharge.append(value)
```

Figure 2.6: Snippet from the code that generated the hydrograph in Figure 2.5. Each time model.update() is called, the eWaterCycle platform instructs the hydrological model running in a container to run for one timestep. After each update the user can interact with the state of the model, in this simple example to extract the calculated discharge at every timestep.

region. To run this model for any other region, the only change a user needs to make is to provide another shapefile and select another GRDC observation station.

This first use-case shows that a model written in a completely different programming language can be run in eWaterCycle without having to understand, install, or even be aware of the model code. In this example, the parameters for the single-bucket model in MARRMoT were chosen by the user. Normally, conceptual models, such as MARRMoT, need to be calibrated; see Section 2.4.5 below for a use-case demonstrating how to calibrate a hydrological model within eWaterCycle.

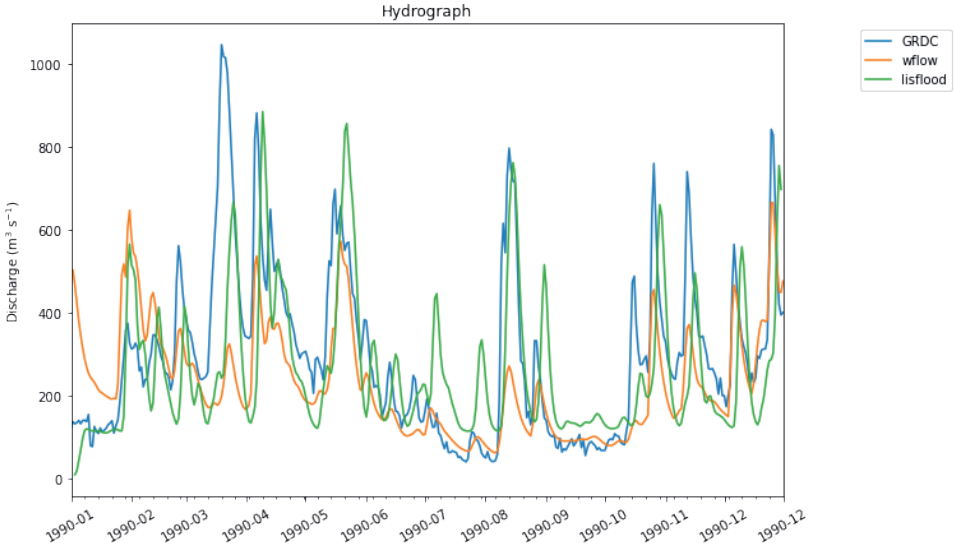
2.4.2. MODEL COMPARISON: TWO MODELS, ONE CATCHMENT, ONE FORCING

In hydrological modelling, an important decision is which model to use for answering a specific research question[8]. To this end, hydrologists want to compare two or more models, with identical forcing, to evaluate the differences in model behavior and performance.

Figure 2.7 shows the result of such an experiment, where simulated discharge estimates produced by wflow and LISFLOOD are evaluated against GRDC observations at the Merrimack basin outlet. As Figure 2.8 shows, the code used in the notebook to run the two models is nearly identical to that of the first use case from Section 2.4.1 thanks to the uniform way in which models are interfaced within the eWaterCycle platform. This significantly reduces the effort involved in comparing different models, and stimulates objective model selection based on adequacy for answering a particular research question.

The code in this notebook can be adapted to run a selection of available models by changing which container is started, on any region supported by those models, using any input forcing dataset available on the platform. Changing a forcing dataset is done by passing another forcing object to the model upon creation of the model instance. In

[8] An analysis of all the considerations that go into choosing the best available hydrological model for a given experiment is beyond the scope of this paper. We recommend Addor and Melsen (2019) and references therein as a starting point.



	nse	kge_2009	sa	me
wflow	0.25	0.42	0.49	-66.68
lisflood	0.19	0.50	0.54	-36.90

Figure 2.7: Two hydrographs generated by the wflow and the LISFLOOD models compared to GRDC observations for the Merrimack basin. While the code used to run these two models, given in Figure 2.8, is similar to that in the first use case, the models are far more complex and generate better predictions of streamflow.

In [11]:
```
discharge = []
time_range = []

while model.time < model.end_time:
    model.update()
    # Get Discharge at one location
    discharge_at_pixel = model.get_value_at_coords('RiverRunoff', lon=station_lon, lat=station_lat)[0]
    discharge.append(discharge_at_pixel)
    time_range.append(model.time_as_datetime.date())
    print(model.time_as_isostr, end="\r")
```

Figure 2.8: Snippet from the code that generated the hydrographs in Figure 2.7 showing how much this looks like the code of Figure 2.6 while the models involved are far more complex. This shows that the user of the eWaterCycle platform can work with different, complex models from the same environment.

practice this means as little as changing a string in the Jupyter notebook describing the experiment from for example 'ERA-Interim' to 'ERA5'. Hydrologists use this to determine which model to use to answer their research question. Consultants and policy makers can use this to, for example, determine which model to use for future projections of the impact of climate change on local hydrology. Operational water managers can use this to determine which model to use in their operational forecasting systems.

2.4.3. MODEL COUPLING: ADD OUTPUT OF ONE MODEL AS INPUT FOR DOWNSTREAM MODEL

As a consequence of the curse of locality, mentioned in the introduction, different regions often have a different model that behaves best for a particular research goal. However, technological differences between the implementations of these models currently prevent researchers from coupling different models for different regions into a larger patchwork of multi-models. The eWaterCycle platform removes such technological barriers and enables users to couple models written in entirely different programming languages.

As an example, we demonstrate how discharge calculated by the simplest one-bucket model from the MARRMoT suite of models for the Moselle River (a tributary of the Rhine) is inserted into the PCR-GLOBWB 2.0 model, which is simulating the rest of the Rhine basin. Figure 2.9 shows a map of the calculated discharge in the Rhine basin with the Moselle subcatchment on top of it. We have cut out the Moselle subcatchment from the PCR-GLOBWB 2.0 model and, at every time step, the discharge calculated by MARRMoT is added to the channel storage variable where the Moselle enters the Rhine. As a reference, we also run the PCR-GLOBWB 2.0 model for the whole Rhine basin without coupling to MARRMoT.

Figure 2.10 shows the central loop of this coupling experiment. Within the loop, three models are run: MARRMoT for the Moselle sub-basin and two instances of PCR-GLOBWB 2.0. The first, called 'reference', runs the entire Rhine basin without alteration. The second, called 'experiment', runs the Rhine basin with the Moselle subcatchment cut out (by altering the landmask file). At every time step, the discharge calculated by MARRMoT for the Moselle is added to the 'channel storage' variable of PCR-GLOBWB 2.0 at the mouth where the Moselle enters the Rhine (at Koblenz). This code snippet shows how the eWaterCycle platform supports model coupling without having to modify the code of the models being coupled. While this example shows the coupling of two hydrological models, the eWaterCycle platform facilitates the coupling of any models that implement a Basic Model Interface (BMI), including those not describing hydrology. Work where hydrological models are connected to, for example, models describing human behavior (Elshafei et al., 2015) or geomorphological processes (Hancock and Willgoose, 2001) can be done using the eWaterCycle platform.

Figure 2.11 compares the coupled PCR-GLOBWB 2.0–MARRMoT experiment with the uncoupled PCR-GLOBWB 2.0 reference run and GRDC observations. The effect of using MARRMoT for the Moselle subcatchment is visible in the differing simulated discharge. Note that, in this example, the two models are written in different programming

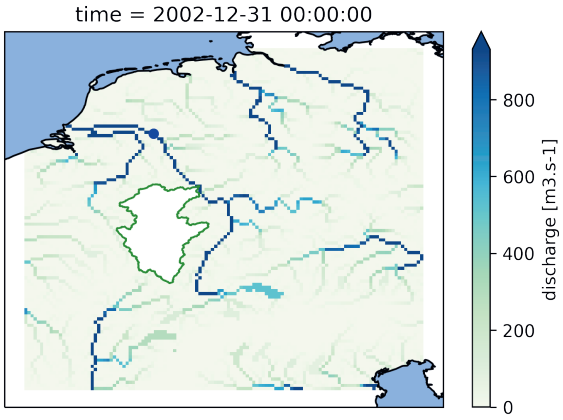


Figure 2.9: Calculated discharge in the Rhine basin at the end of the model run using the PCR-GLOBWB 2.0 model for the Rhine coupled with the simplest one-bucket model from the MARRMoT suite of models for the Moselle subcatchment. The Moselle subcatchment has been cut out of the PCR-GLOBWB 2.0 model (by altering its 'landmask' setting). MARRMoT calculates the discharge for the Moselle and, at each time step, this is added to the 'channel storage' variable of PCR-GLOBWB 2.0 at the location where the Moselle flows into the Rhine (at Koblenz). The blue dot on the map indicates the location of the observation station at Lobith, which is used for the hydrographs shown in Figure 2.11.

languages (MATLAB for MARRMoT and Python for PCR-GLOBWB 2.0). While the experiment formulated in the Jupyter notebook is also written in Python, it does not interact directly with the code of the models; all interaction happens through grpc4bmi.

2.4.4. MODEL INTERACTION: CHANGE STATE OF MODEL DURING RUNTIME

Hydrological models are great tools for 'what if' type research questions such as "What is the impact on river discharge downstream if the land use upstream is changed?". These types of research questions often require direct intervention in the state of the model[9]. Previously, this meant having to adapt the model code to reflect the required research question, leading to a new version of the model. Using the eWaterCycle platform, model variables are exposed through BMI and can be queried or set without changing the model source code itself. This separates the 'experiment' from the hydrological model used to conduct the experiment with.

In this example use-case, we answer the research question: 'What would happen to the predicted discharge if we used observations of evaporation instead of calculating evaporation?'. At every time step, the soil storage variables in the PCR-GLOBWB 2.0 model are changed based on FLUXNET (Pastorello et al., 2020) observations. Hydrologically relevant choices for this experiment, such as 'what to do if the observations indicate that more water has evaporated than is available in the soil storage?' are now implemented as part of the experiment, and not as changes to the model source code, as shown in Figure 2.12. This increased transparency and separation of experiment and

[9] with 'state' as defined in the glossary

In [25]:
```
print(f"Running experiment", flush=True)

# the experiment model is used as central 'time keeper'
while pcrglob_exp_model.time < pcrglob_exp_model.end_time:
    print(f"Current time: {pcrglob_exp_model.time_as_isostr}", end="\r")
    time_range.append(pcrglob_exp_model.time_as_datetime.date())

    # run the reference model for one timestep, store the output at grdc station
    pcrglob_ref_model.update()
    pcrglob_ref_discharge = pcrglob_ref_model.get_value_at_coords(
        "discharge", lat=[gridded_grdc_lat], lon=[gridded_grdc_lon]
    )[0]
    pcrglob_ref_output.append(pcrglob_ref_discharge)

    # run MARRMoT and store the output, note that MARRMoT output is in mm!
    marrmot_model.update()
    marrmot_output.append(marrmot_model.get_value(marrmot_output_variable)[0])

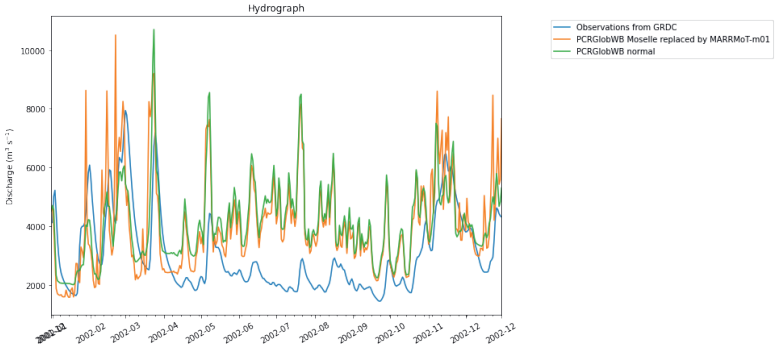
    # add the output of MARRMoT to PCRGlob Experiment model. Note that channel storage
    # is in m3, while MARRMoT output is in mm, so we need to convert
    water_to_add_to_pcrglob = marrmot_output[-1] * moselle_area / 1000
    current_value_in_pcrglob = pcrglob_exp_model.get_value_at_coords(
        "channel_storage", lat=[lat_moselle_mouth], lon=[lon_moselle_mouth]
    )
    value_to_set_in_pcrglob = water_to_add_to_pcrglob + current_value_in_pcrglob
    pcrglob_exp_model.set_value_at_coords(
        "channel_storage",
        lat=[lat_moselle_mouth],
        lon=[lon_moselle_mouth],
        values=value_to_set_in_pcrglob,
    )

    # Run the experiment PCRGlobWB model one timestep, store the output at grdc station
    pcrglob_exp_model.update()
    pcrglob_exp_discharge = pcrglob_exp_model.get_value_at_coords(
        "discharge", lat=[gridded_grdc_lat], lon=[gridded_grdc_lon]
    )[0]
    pcrglob_exp_output.append(pcrglob_exp_discharge)

print("")
```

```
Running experiment
Current time: 2002-12-30T00:00:00Z
```

Figure 2.10: The central loop of the experiment presented in Section 2.4.3. Within the loop, three models are run: MARRMoT for the Moselle sub-basin and two instances of PCR-GLOBWB 2.0. The first, called 'reference', runs the entire Rhine basin without alteration. The second, called 'experiment', runs the Rhine basin with the Moselle subcatchment cut out (by altering the landmask file). At every time step, the discharge calculated by MARRMoT for the Moselle is added to the 'channel storage' variable of PCR-GLOBWB 2.0 at the mouth where the Moselle enters the Rhine (at Koblenz). This code snippet shows how the eWaterCycle platform supports model coupling without having to modify the code of the models being coupled.



	nse	kge_2009	sa	me
PCRGlobWB Moselle replaced by MARRMoT-m01	-0.86	0.32	0.42	993.64
PCRGlobWB normal	-0.74	0.30	0.41	1057.08

Figure 2.11: The calculated discharge for the Rhine basin from PCR-GLOBWB 2.0 when run normally (i.e., the reference) and when coupled with MARRMoT for the Moselle subcatchment. The calculated discharge is compared to GRDC observations of discharge.

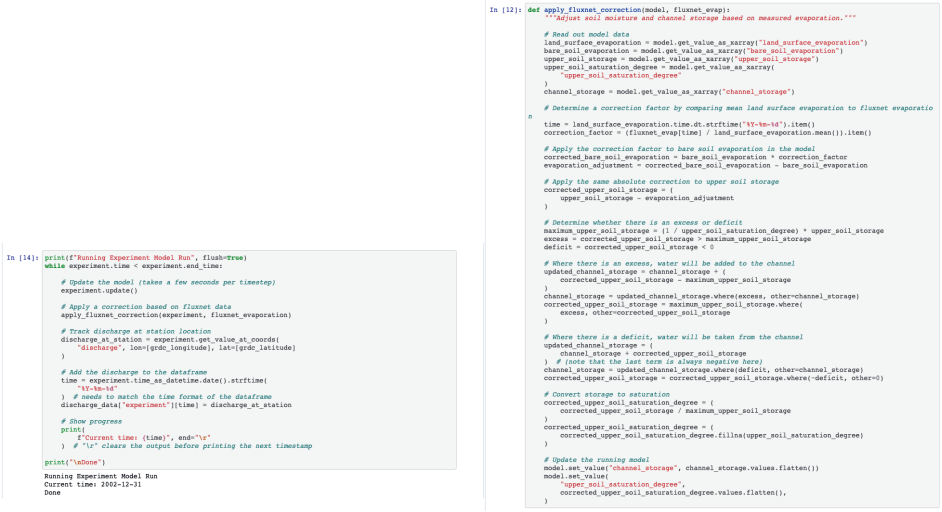
model make it easier to repeat the experiment with different models and to understand and build upon each other's work.

Figure 2.13 shows the calculated river discharge at the basin outlet for both the reference run, with no intervention in the model, as well as for the experiment run. The impact is clearly visible; the discharge estimates of PCR-GLOBWB 2.0 become worse when the model is constrained with observations of evaporation. This leads to a host of follow-up questions: Is the evaporation measurement used representative of the (entire) catchment? Does the calculated evaporation compensate for missing fluxes in other parts of the model? Etcetera. Finally, this experiment is part of the thesis work of BSc student Thomas Albers (Albers, 2020), illustrating that, by using the eWaterCycle platform, students can focus on hydrologically interesting research questions without having to invest a large amount of time to learn and modify a particular model's source code.

2.4.5. CALIBRATE MODEL USING AN ENSEMBLE OF MODEL RUNS

Many hydrological models have parameters that need to be calibrated before a model can be used. A whole subfield of hydrology has formed around the research question of what the best, most efficient, or most robust way is to calibrate a model (Bárdossy, 2007; Bárdossy and Singh, 2008). These calibration methods typically require an ensemble of model runs as part of the optimization process. The eWaterCycle platform supports transparent and efficient calibration of hydrological models through the separation of model and experiment, as well as the ability to run models in containers on remote machines.

In this example, we calibrate the MARRMoT m14 model (an implementation of TOPMODEL), which has seven free parameters, for the Savannah basin. We used the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) calibration scheme, which is provided by the CMA-ES/pycma software package (Hansen et al., 2021). The model runs



```
In [14]: print(f"Running Experiment Model Run", flush=True)
         while experiment.time < experiment.end_time:

             # Update the model (takes a few seconds per timestep)
             experiment.update()

             # Apply a correction based on fluxnet data
             apply_fluxnet_correction(experiment, fluxnet_evaporation)

             # Track discharge at station location
             discharge_at_station = experiment.get_value_at_coords(
                 "discharge", lon=[grdc_longitude], lat=[grdc_latitude]
             )

             # Add the discharge to the dataframe
             time = experiment.time_as_datetime.date().strftime(
                 "%Y-%m-%d"
             )  # needs to match the time format of the dataframe
             discharge_data["experiment"][time] = discharge_at_station

             # Show progress
             print(
                 f"Current time: {time}", end="\r"
             )  # "\r" clears the output before printing the next timestamp

         print("\nDone")

         Running Experiment Model Run
         Current time: 2002-12-31
         Done
```

```
In [12]: def apply_fluxnet_correction(model, fluxnet_evap):
             """Adjust soil moisture and channel storage based on measured evaporation."""

             # Read out model data
             land_surface_evaporation = model.get_value_as_xarray("land_surface_evaporation")
             bare_soil_evaporation = model.get_value_as_xarray("bare_soil_evaporation")
             upper_soil_storage = model.get_value_as_xarray("upper_soil_storage")
             upper_soil_saturation_degree = model.get_value_as_xarray(
                 "upper_soil_saturation_degree"
             )
             channel_storage = model.get_value_as_xarray("channel_storage")

             # Determine a correction factor by comparing mean land surface evaporation to fluxnet evaporatio
         n
             time = land_surface_evaporation.time.dt.strftime("%Y-%m-%d").item()
             correction_factor = (fluxnet_evap[time] / land_surface_evaporation.mean()).item()

             # Apply the correction factor to bare soil evaporation in the model
             corrected_bare_soil_evaporation = bare_soil_evaporation * correction_factor
             evaporation_adjustment = corrected_bare_soil_evaporation - bare_soil_evaporation

             # Apply the same absolute correction to upper soil storage
             corrected_upper_soil_storage = (
                 upper_soil_storage - evaporation_adjustment
             )

             # Determine whether there is an excess or deficit
             maximum_upper_soil_storage = (1 / upper_soil_saturation_degree) * upper_soil_storage
             excess = corrected_upper_soil_storage > maximum_upper_soil_storage
             deficit = corrected_upper_soil_storage < 0

             # Where there is an excess, water will be added to the channel
             updated_channel_storage = channel_storage + (
                 corrected_upper_soil_storage - maximum_upper_soil_storage
             )
             channel_storage = updated_channel_storage.where(excess, other=channel_storage)
             corrected_upper_soil_storage = maximum_upper_soil_storage.where(
                 excess, other=corrected_upper_soil_storage
             )

             # Where there is a deficit, water will be taken from the channel
             updated_channel_storage = (
                 channel_storage + corrected_upper_soil_storage
             )  # (note that the last term is always negative here)
             channel_storage = updated_channel_storage.where(deficit, other=channel_storage)
             corrected_upper_soil_storage = corrected_upper_soil_storage.where(~deficit, other=0)

             # Convert storage to saturation
             corrected_upper_soil_saturation_degree = (
                 corrected_upper_soil_storage / maximum_upper_soil_storage
             )
             corrected_upper_soil_saturation_degree = (
                 corrected_upper_soil_saturation_degree.fillna(upper_soil_saturation_degree)
             )

             # Update the running model
             model.set_value("channel_storage", channel_storage.values.flatten())
             model.set_value(
                 "upper_soil_saturation_degree",
                 corrected_upper_soil_saturation_degree.values.flatten(),
             )
```

Figure 2.12: The code used to intervene in the state of PCR-GLOBWB 2.0 during runtime. At each time step, the function apply_fluxnet_correction() is called. The right-hand side of this figure shows the code of that function. The model.get_value_as_xarray() calls at the start of the function extract information on the state from the model. The model.set_value() calls at the end of the function update the state with the newly calculated information. This example shows that numerical hydrologists can interact and experiment with the state of a hydrological model without having to interact with the code of the model. In this way, a clear separation between model and experiment is achieved.

were done on the Cartesius supercomputer. Figure 2.14 shows a snippet of the code that shows the clear separation between model and calibration routine.

The results for this calibration experiment are shown in Figure 2.15. The increase in the objective function shows the convergence of the algorithm. This use case shows that separation of model and experiment makes calibration of hydrological models easier to set up with eWaterCycle.

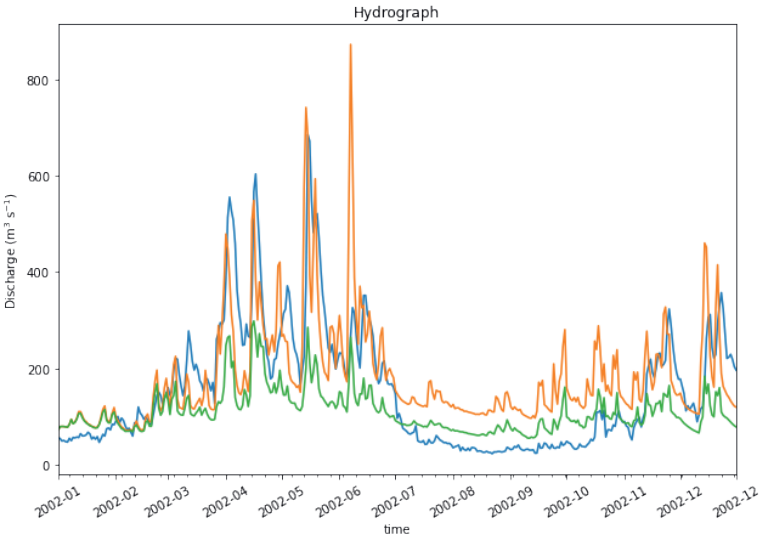
2.5. CONCLUSIONS

We have introduced the eWaterCycle platform for FAIR and Open hydrological modeling. Using eWaterCycle, hydrologists can easily build on each other's work by using each other's models, data, and experiments in a manner that is 'FAIR by design'. Using the online eWaterCycle platform, hydrologists do not have to install or download software or pre-process data before they can start computational hydrological experiments.

The eWaterCycle platform implements the 'FAIR' principles as follows:

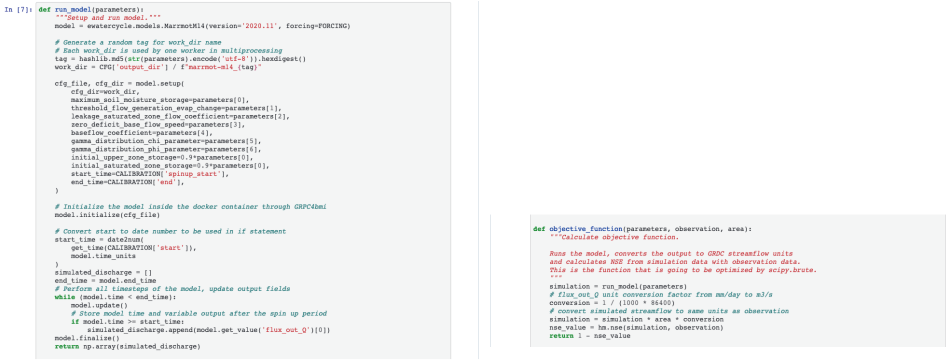
- **F**indable: Through the Explorer all available models and datasets are exposed in an easily findable manner. The documentation (Verhoeven et al., 2021) further specifies all available models, datasets and their properties.

- **A**ccessible: By making the entire software-stack of the eWaterCycle platform Open Source and by including only models and datasets that are also openly available,



	nse	kge_2009	sa	me
reference	0.40	0.62	0.48	27.70
experiment	0.34	0.27	0.42	-43.53

Figure 2.13: This graph shows the effect of intervening in the state of the PCR-GLOBWB 2.0 model. The calculated discharge for the Merrimack basin from two model runs is shown: one reference run in which the model is run without intervention and one experiment in which the model state is changed at every time step to incorporate evaporation observation data from FLUXNET (Pastorello et al., 2020). The calculated discharge for both model runs is compared to GRDC observations of discharge.



```
In [7]: def run_model(parameters):
            """Setup and run model."""
            model = ewatercycle.models.MarrmotM14(version='2020.11', forcing=FORCING)

            # Generate a random tag for work_dir name
            # Each work_dir is used by one worker in multiprocessing
            tag = hashlib.md5(str(parameters).encode('utf-8')).hexdigest()
            work_dir = CFG['output_dir'] / f'marrmot-m14_{tag}'

            cfg_file, cfg_dir = model.setup(
                cfg_dir=work_dir,
                maximum_soil_moisture_storage=parameters[0],
                threshold_flow_generation_evap_change=parameters[1],
                leakage_saturated_zone_flow_coefficient=parameters[2],
                zero_deficit_base_flow_speed=parameters[3],
                baseflow_coefficient=parameters[4],
                gamma_distribution_chi_parameter=parameters[5],
                gamma_distribution_phi_parameter=parameters[6],
                initial_upper_zone_storage=0.9*parameters[0],
                initial_saturated_zone_storage=0.9*parameters[0],
                start_time=CALIBRATION['spinup_start'],
                end_time=CALIBRATION['end'],
            )

            # Initialize the model inside the docker container through GRPC4bmi
            model.initialize(cfg_file)

            # Convert start to date number to be used in if statement
            start_time = date2num(
                get_time(CALIBRATION['start']),
                model.time_units
            )
            simulated_discharge = []
            end_time = model.end_time
            # Perform all timesteps of the model, update output fields
            while (model.time < end_time):
                model.update()
                # Store model time and variable output after the spin up period
                if model.time >= start_time:
                    simulated_discharge.append(model.get_value('flux_out_Q')[0])
            model.finalize()
            return np.array(simulated_discharge)
```

```
def objective_function(parameters, observation, area):
    """Calculate objective function.

    Runs the model, converts the output to GRDC streamflow units
    and calculates NSE from simulation data with observation data.
    This is the function that is going to be optimized by scipy.brute.
    """
    simulation = run_model(parameters)
    # flux_out_Q unit conversion factor from mm/day to m3/s
    conversion = 1 / (1000 * 86400)
    # convert simulated streamflow to same units as observation
    simulation = simulation * area * conversion
    nse_value = hm.nse(simulation, observation)
    return 1 - nse_value
```

Figure 2.14: A snippet of the code used to run the calibration experiment described in Section 2.4.5. The entire workflow of the first use case presented in Section 2.4.1 is wrapped in the function run_model() (left part of the figure). This function runs the model for a given set of parameters. This function is called in objective_function(), which is passed as an argument to the calibration scheme of the CMA-ES/pycma package. This package, based on settings provided, runs multiple instances of the objective_function() (and thus of the hydrological model) in parallel to find the optimum set of parameters that optimizes the objective function.

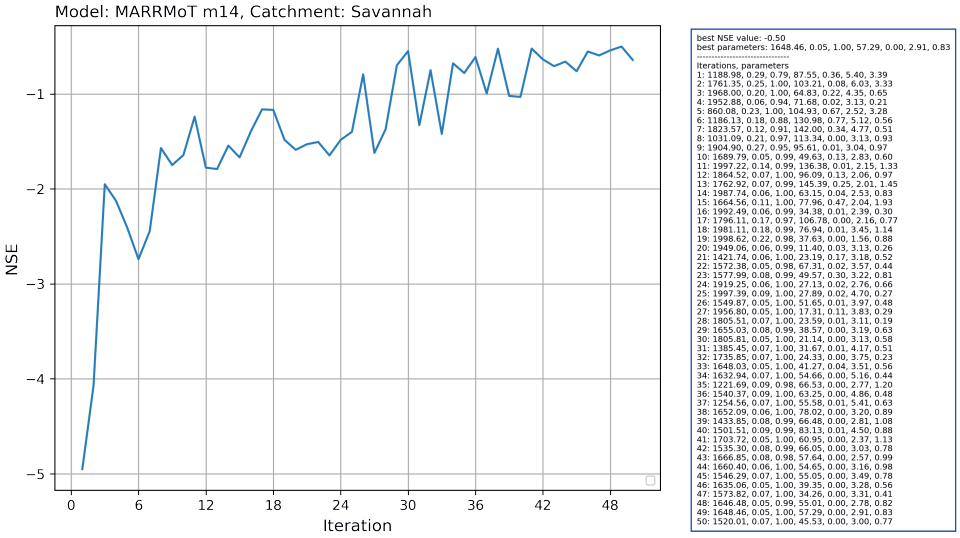


Figure 2.15: Output of the running the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) calibration scheme from the CMA-ES/pycma software package on the MARRMoT m14 model (an implementation of TOPMODEL) for the Savannah basin.

the platform makes it possible for anyone with sufficient computational infrastructure to install and run an instance of the eWaterCycle platform. For those without these resources, the eWaterCycle team (i.e., the authors of this paper) currently hosts an instance of the eWaterCycle platform on demonstration hardware provided by SURF. Access to this version can be obtained through contact with the authors. In the near future, the eWaterCycle team is hoping to acquire funds to more sustainably offer an instance of eWaterCycle to the entire hydrological community.

- **I**nteroperable: The eWaterCycle platform uses open interfaces between the different parts of the platform; for example, grpc4bmi is used to communicate between experiments in Jupyter notebooks and the hydrological models in containers. The use of open interfaces and containers makes it easy to connect to a dataset or a model and use it for a different study.

- **R**eusable: The experiments in eWaterCycle, as contained in Jupyter notebooks, are separate entities from the models and datasets. Because of this separation and the open interfaces between components, re-use of data, models, or parts of experiments is facilitated.

As laid out in the introduction, currently an ecosystem of services is emerging that makes it easier for hydrologists to do computational research, with each service focusing on different parts of the hydrological research cycle (Tarboton et al., 2014; Tucker et al., 2021). In this ecosystem, eWaterCycle is developed as a platform on which hydrologists can execute their computational hydrological experiments. In this paper, we have

presented the core components of the eWaterCycle platform, the Explorer, the notebook environment, and the underlying technology to deal with models and datasets in a FAIR manner. The hydrological community can install the openly and freely available eWaterCycle platform on their own infrastructure. The eWaterCycle team (i.e., the authors of this paper) is seeking sustainable funding to provide an online place where more researchers will be able to execute their computational hydrological research and education.

For future development, integration with other platforms that facilitate hydrological research is foreseen, most notably coupling to models from CSDMS, sharing and retrieving data from HydroShare, integrating higher-level interfaces to models, such as PyMT, and giving access to libraries that facilitate additional types of research such as the data assimilation software OpenDA.

The use-cases presented in this paper give an overview of the type of research that the eWaterCycle platform can facilitate, from model selection to coupling, and calibration. eWaterCycle is set up as a modular collection of services that together form a complete platform. By making sure that the individual modules contain as few assumptions about hydrology as possible (those are represented in the models and experiments) we are facilitating that the technologies developed for the eWaterCycle platform are portable to other domains of (geo)science where researchers work with each other's models and datasets.

Hutton, Wagener, et al. (2016) argued that computational hydrology can only be a proper science if the hydrological community makes sure that hydrological model studies are executed and presented in a reproducible manner. We replied that, to improve current practices, hydrologists using hydrological models in their work should not 're-invent the water wheel' but rather use existing technology from other fields, such as containers and ESMValTool, and open interfaces, such as BMI, to do their computational science (Hut et al., 2017). With this paper and the release of the eWaterCycle platform, we are putting our money where our mouth is and providing the hydrological community with a 'FAIR by design' platform to do science.

3

LARGE-SAMPLE ASSESSMENT OF VARYING SPATIAL RESOLUTION ON THE STREAMFLOW ESTIMATES OF THE WFLOW_SBM HYDROLOGICAL MODEL

These go to eleven.

Nigel Tufnel, *This Is Spinal Tap*

Distributed hydrological modelling is moving into the realm of hyper-resolution modelling. This results in a plethora of scaling-related challenges that remain unsolved. In light of model result interpretation, finer-resolution output might imply to the user an increased understanding of the complex interplay of heterogeneity within the hydrological system. Here, we investigate spatial scaling in the form of varying spatial resolution by evaluating the streamflow estimates of the distributed wflow_sbm hydrological model based on 454 basins from the large-sample CAMELS data set. Model instances are derived at 3 spatial resolutions, namely 3 km, 1 km, and 200 m. The results show that a finer spatial resolution does not necessarily lead to better streamflow estimates at the basin outlet. Statistical testing of the objective function distributions (KGE scores) of the 3 model instances identified a statistically significant difference only between the 3 km and 200 m streamflow estimates. However, an assessment of sampling uncertainty shows high uncertainty surrounding the KGE scores throughout the domain. This makes the conclusion based on the statistical testing inconclusive. The results do indicate strong local variation between model instances, with differences in KGE scores averaging 0.22 and values greater than 0.5. The results of this study open up research paths for investigating the changes in flux and state partitioning due to spatial scaling. This will help further our understanding of the challenges that need to be resolved for hyper-resolution hydrological modelling.

Parts of this chapter have been published verbatim in Aerts, J. P. M., Hut, R. W., van de Giesen, N. C., Drost, N., van Verseveld, W. J., Weerts, A. H., and Hazenberg, P.: Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model, Hydrol. Earth Syst. Sci., 26, 4407–4430, https://doi.org/10.5194/hess-26-4407-2022, 2022.

3.1. INTRODUCTION

Hydrological model development follows competing model philosophies (Hrachowitz and Clark, 2017). These range from high-resolution, small-scale, process-resolving distributed models to spatially lumped conceptual models. All model structures developed within the competing philosophies have their own limitations and advantages. All models face overlapping challenges, including parameter estimation and the representation of spatial heterogeneity (Clark et al., 2017).

The parameter identifiability problem stems from the inability to obtain unique and realistic parameters at the modelling scale due to structural model deficiencies and applied calibration techniques (Sorooshian and Gupta, 1983). Multiple studies have extensively researched calibration techniques to overcome the parameter identifiability problem (e.g. Guse et al., 2020; Sorooshian and Gupta, 1983; Vrugt et al., 2002). The identifiability problem is emphasized in distributed modelling, the focus of this study, by the incompatibility of parameter measurements with the modelling scale (Grayson et al., 1992). This results in the need for transferring parameters in space and time. Multiple studies have looked into parameter transferability (e.g., Finnerty et al., 1997; Haddeland et al., 2006; Wagener and Wheater, 2006), and Melsen et al. (2016) discussed how the inadequacy of transferring parameters in space and time may indicate a lack of representation of spatial heterogeneity and temporal variability in the models. Methods such as the multi-scale parameter regionalization technique (MPR; Samaniego et al., 2010) emerged to increase the transferability of parameters from the data resolution to the hydrological model resolution. Imhoff et al. (2020) used a different method that applied pedotransfer functions to derive parameters at the finest available data resolution and upscale them to the various spatial model resolutions of the wflow_sbm hydrological model.

The effects of spatial heterogeneity have been studied at a catchment scale using the representative elementary watershed (REW) theory developed by Reggiani and Schellekens (2003), Reggiani et al. (1998, 1999), and Wood et al. (1988). In Wood et al. (1988), the basin was divided into sub-basins and then aggregated to basin level to study scaling behavior. This type of research is still very relevant as in recent years, the hydrologic modelling community has surpassed the REW scale threshold of 1 km^2 with the move towards so-called hyper-resolution modelling. The discussion following this move revealed that in addition to the many benefits, e.g. applicability for stakeholders, there are multiple challenges to address (Beven and Cloke, 2012; Bierkens et al., 2015; Wood, Roundy, Troy, van Beek, Bierkens, Blyth, de Roo, et al., 2011). These challenges include scaling issues (Blöschl and Sivapalan, 1995; Gupta et al., 1986), such as (1) the need to explicitly model processes that are parameterized at coarser resolutions; (2) lateral connections between compartments of the hydrological system that are averaged out or ignored at coarser resolutions; and (3) an increase in uncertainty due to insufficient process and parameter knowledge resulting from insufficient data quality at finer resolutions (Bierkens et al., 2015).

The scaling issues arise when the (often unconscious) assumption is made that a hydrological model used at various spatial and temporal resolutions should estimate similar states and fluxes independently of scale. A utopian model has scale-invariant model parameterization and hydrological process descriptions. The development of

scale-invariant hydrological models is, however, very challenging as most hydrological processes do not scale in a linear manner (e.g. Bras, 2015; Rouholahnejad Freund et al., 2020). Instead, processes at one length scale influence those at other scales (Horritt and Bates, 2001).

Because of the complex nature of scaling issues and a shift in distributed modelling towards hyper-resolution modelling, it is important to continuously assess the effects of scaling. Without investigating what this move entails the hydrological modelling community risks communication problems with the users of model results. In the case of spatial model resolution, the increase in level of detail in model output might imply to the user an increased understanding of the complex interplay of heterogeneity within the hydrological system. We can only determine whether this implication is warranted by continuously assessing how models behave under various spatial (and temporal) resolutions.

Multiple studies have looked into spatial scaling effects by varying spatial model resolution. Booij (2005) found that increasing the spatial resolution of a semi-lumped HBV hydrological model only marginally increased model performance based on streamflow estimates. The coupled ParFlow-CLM model was evaluated with various grid cell sizes by Shrestha et al. (2015) and they found, among other effects, that soil moisture estimates were grid cell size dependent. Sutanudjaja et al. (2018) introduced the transition from 30-arcminute to 5-arcminute grid-cell simulations of the distributed PCR-GLOBWB model. Results showed a general increase in model performance compared to streamflow observations. However, regional scaling issues were present. In some of the basins model performance was lower at a finer spatial model resolution. This study made it apparent that a large sample of hydrologically diverse basins should be considered when investigating the effects of varying spatial resolution. To our knowledge, no studies have investigated varying spatial resolution within the hyper-resolution realm using a large sample of basins. This will be the focus of this research.

The distributed conceptual wflow simple bucket model (wflow_sbm; Schellekens et al., 2020b) utilizes high resolution datasets to derive model instances globally at varying spatial resolutions. Parameter estimates are based on the work of Imhoff et al. (2020) to ensure consistency across scale. Remotely sensed soil and land cover data sets are sources for estimating parameters through pedotransfer functions (PTFs) (e.g. Brakensiek et al., 1984; Cosby et al., 1984). PTFs are a collection of predictive functions, so-called super-parameters (Tonkin and Doherty, 2005), derived at the point scale that estimate soil parameters where underlying data are scarce. For most wflow_sbm model parameters, a priori estimates are available. This is not (yet) the case for the horizontal conductivity fraction (KsatHorFrac) parameter, making it a logical parameter for calibration as it is also one of the more sensitive parameters in the model (Imhoff et al., 2020). The flexible setup of wflow_sbm can be used to assess spatial scaling issues due to quasi-scale-invariant parameters whilst maintaining similar hydrological process descriptions across scales. This setup includes the recent improvements by Eilander et al. (2021), who developed a scale-invariant method for upscaling river networks (one of the suggested causes of the inconsistent streamflow estimates across scales as shown in Imhoff et al. (2020)).

In this study, we quantify the effects of varying spatial resolution on the wflow_sbm streamflow estimates for a large sample of hydrologically diverse basins in the CAMELS dataset (Addor et al., 2017; Newman et al., 2015). By conducting this research on a large sample of basins, we can assess the results on consistency and locality. The assessment is conducted by creating 3 model instances at varying spatial resolutions for each basin: 3 km, 1 km, and 200 m spatial grid resolutions. These instances cover a broad range of large- and small-scale dynamics. For example, snow accumulation at the mountain range (> 1 km) and the mountain ridge scale (< 1 km) (e.g. Houze Jr., 2012; Mott et al., 2018; Vionnet et al., 2021), or closing in on the hillslope scale (< 100 m) (e.g. Fan et al., 2019; Tromp-van Meerveld and McDonnell, 2006). The parameters for the wflow_sbm model instances are estimated at the highest available data resolution and aggregated to the modelling grid using the upscale rules as defined in Imhoff et al., 2020.

Our hypothesis is that the differences in streamflow estimates at various spatial resolutions will be small due to the parameters being quasi-scale-invariant and hydrological process descriptions in the model remaining the same across spatial scales. We will reject this hypothesis when the results show significantly different streamflow estimates across the studied resolutions. In addition, with this study, we showcase how the eWaterCycle platform (Hut, Drost, van de Giesen, van Werkhoven, et al., 2021) can be utilized for computationally intensive large-sample modelling studies.

3.2. METHODOLOGY

3.2.1. INPUT DATA

THE CAMELS DATA SET

The CAMELS data set is a collection of hydrologically relevant data on 671 basins located in the Contiguous United States (CONUS) (Newman et al., 2015). The basins were selected based on a minimum amount of human influence on the hydrological system, e.g. the absence of large reservoirs. The data set includes 20 years of continuous streamflow records from 1990 to 2009 from the United States Geological Survey (USGS). The CAMELS data set covers a hydrologically and climatologically diverse selection of basins. The sample size, hydrological diversity of the basins, and common use of this data set in other hydrological modelling studies (e.g. Gauch et al., 2021; Knoben et al., 2020) are the reasons for selecting this case study area.

Of the 671 basins, we successfully ran models for 567 basins at each of the 3 resolutions (i.e., 3 km, 1 km, and 200 m). Reasons for excluding basins from our analyses were missing streamflow observations (7 basins) or errors during parameter estimation (97 basins). Parameter estimation errors occurred mainly during drainage network delineation, either when the basin outlet consisted of a single grid cell, resulting in a model coding error or when inconsistencies occurred in the local drainage direction layer. When one of the 3 model instances failed, the basin was excluded from further analyses. Figure 3.1 shows the locations of the included and excluded basins as well as the reason for exclusion.

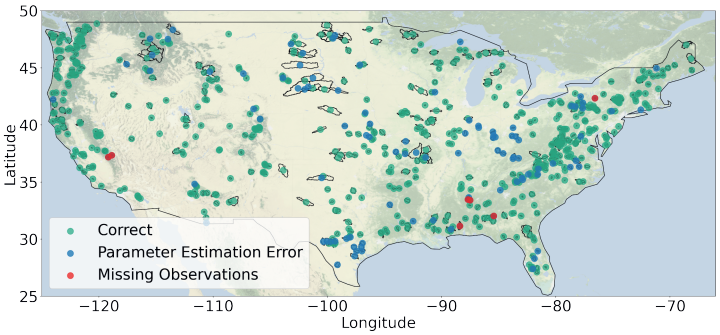


Figure 3.1: Basin locations of the CAMELS data set with the included basins shown in green, basins excluded because of parameter-estimation errors in blue, and basins excluded because of missing streamflow observations in red. The basemap was made with Natural Earth.

STREAMFLOW OBSERVATIONS

The United States Geological Survey (USGS) streamflow observation records were downloaded to match our model simulation period from 1996 to 2016. The data were resampled to daily values and the units were converted to $m^3\,s^{-1}$. We ensured consistency in time zones between the observations and the model simulations by converting the USGS streamflow timestamps to UTC. The tools used for downloading, resampling, unit conversion, and shifting of time zones might be of interest to others in the hydrological community and are available in the GitHub repository (https://github.com/jeromaerts/eWaterCycle_example_notebooks).

METEOROLOGICAL INPUT AND PRE-PROCESSING

The meteorological input requirements of the wflow_sbm model are precipitation, temperature, and potential evapotranspiration. Precipitation data were obtained from the Multi-Source Weighted-Ensemble Precipitation Version 2.1 (MSWEP; Beck et al., 2019). The data set was constructed using bias-corrected gauge, satellite, and reanalysis data. The data are available at 0.1 degrees spatial (11 km) and 3-hourly temporal resolution for the period from 1979 to 2017. The temperature variable was obtained from the ERA5 reanalysis data set (Hersbach et al., 2020). The data are available at 0.25 degrees (31 km) spatial and a 1-hourly temporal resolution. In addition to temperature, we used ERA5 variables to calculate potential evapotranspiration using the De Bruin method (Bruin et al., 2016).

We conducted a preliminary analysis for 6 basins in which we compared model simulations based on streamflow estimates that use ERA5 precipitation to those that use

MSWEP precipitation. Results indicated that use of ERA5 precipitation did not produce desirable streamflow estimates compared to MSWEP precipitation. Switching to the MSWEP precipitation product improved streamflow estimates throughout the case study area. Figure 3.12 in the supplementary analyses at the end of this chapter compares the ERA5 and MSWEP forcing for these basins; the corresponding objective-function results are available in the data repository. Notably, some of the basins are very sensitive to changes due to different forcing data sets as shown by the streamflow-based objective functions.

The meteorological inputs are pre-processed within the eWaterCycle platform using the Earth System Model Evaluation Tool (ESMValTool) V2.0 (Righi et al., 2020; Weigel et al., 2021). Before further processing, the data are aggregated to daily values. The precipitation variable is disaggregated to the modelling grid using the second-order conservative method to ensure consistency of the total volume of precipitation across spatial scales. The temperature variable is disaggregated using the environmental lapse rate and the Digital Elevation Model (DEM) used by the hydrological model. The variables required by the De Bruin method (Bruin et al., 2016) are disaggregated using the bilinear method and subsequently used to calculate potential evapotranspiration. The code used for pre-processing is included in the Jupyter notebooks made available with this manuscript (DOI:10.5281/zenodo.5724512).

PARAMETER ESTIMATION FROM EXTERNAL DATA SOURCES

The parameter sets used in this study were derived using the HydroMT software package (Eilander and Boisgontier, 2021). The data sources for deriving parameter sets are open-source global data sets. These include topography, surface water, land cover & land use, soil, meteorological data, and river gauge data. The pedotransfer function to estimate soil properties is based on Brakensiek et al., 1984. In 25 of the 567 basins, lakes and/or reservoirs were included in the model parameters given a threshold area of 1 km^2 and 10 km^2, respectively.

An overview of the data and references is provided in Table 3.1.

3.2.2. MODEL EXPERIMENT SETUP

THE WFLOW_SBM MODEL (V.2020.1.2)

The wflow_sbm model is available as part of the wflow open-source modeling framework (Schellekens et al., 2020b), which is based on PCRaster (Karssenberg et al., 2010) and Python. Figure 3.2 shows the different processes and fluxes that are part of the wflow_sbm hydrological concept. The soil part of the wflow_sbm model is largely based on the Topog_SBM model (Vertessy and Elsenbeer, 1999), which considers the soil as a "bucket" with saturated and unsaturated stores. This model was developed for simulating small-scale hydrology. For channel, overland, and lateral subsurface flow a kinematic wave approach is used, similar to TOPKAPI (Benning, 1995; Ciarapica and Todini, 2002), G2G (Bell et al., 2007), 1K-DHM (Tanaka and Tachikawa, 2015), and Topog_SBM (Vertessy and Elsenbeer, 1999). Wflow_sbm has a simplified physical basis with parameters that represent physical characteristics, leading to (theoretically) an easy linkage of the parameters to actual physical properties. Topog_SBM is mainly used to simulate fast

Table 3.1: Overview of data sources for parameter estimation with categories, references, and version.

Data set	Category	Reference	Version
MERIT Hydro	Topography	Yamazaki et al., 2019	1.0
GRAND (hydro_reservoirs)	Surface water	Lehner et al., 2011	1.0
HydroLAKES (hydro_lakes)	Surface water	Messager et al., 2016	1.0
Randolph Glacier Inventory	Surface water	Pfeffer et al., 2014	6.0
CHELSA	Meteorology	Karger et al., 2017	1.2
Köppen-Geiger	Meteorology	Kottek et al., 2006	2017
VITO	Land use & land cover	Buchhorn et al., 2020	v2.0.2
MODIS LAI	Land use & land cover	Myneni et al., 2015	MCD15A3H V006
SoilGrids	Soil	Hengl et al., 2017	2017

runoff processes during discrete storm events in small basins (< 10 km^2), while evapotranspiration losses are ignored. Since evapotranspiration losses and capillary rise were added to wflow_sbm, the derived wflow_sbm approach can be applied to a wider variety of basins. The main differences between wflow_sbm and Topog_SBM are:

- The addition of evapotranspiration and interception losses using the Gash model (Gash, 1979) on daily time steps or a modified Rutter model on subdaily time steps (Rutter et al., 1971, 1975).
- The addition of a root water uptake reduction function (Feddes et al., 1978).
- The addition of capillary rise.
- The addition of glacier and snow build-up and melting processes.
- Wflow_sbm routes water over an eight-direction (D8) network, instead of the element network based on contour lines and trajectories used by Topog_SBM.
- The option to divide the soil column into any number of different layers.
- Vertical transfer of water is controlled by the saturated hydraulic conductivity at the water table or bottom of a layer, the relative saturation of the layer, and a power coefficient depending on the soil texture (Brooks and Corey, 1964).

MODEL RUNS & CALIBRATION

We derived 3 model instances at varying spatial resolutions covering 3 km, 1 km, and 200 m grids. While for most parameters of the wflow_sbm model a priori estimates can

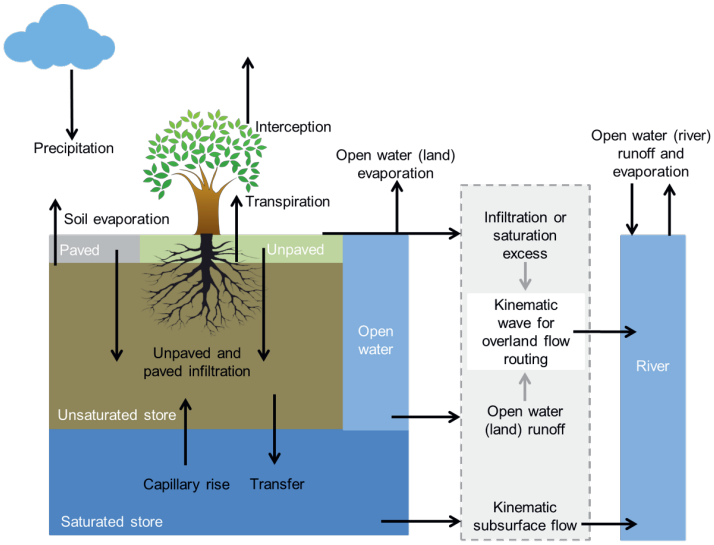


Figure 3.2: Overview of the different processes and fluxes in the wflow_sbm model (Schellekens et al., 2020b).

be derived from external sources, a single nondistributed parameter needs to be calibrated for each basin: the saturated horizontal conductivity often expressed as a fraction (KsatHorFrac) of the vertical conductivity. This parameter cannot be derived from external data sources because it compensates for anisotropy, unrepresentative point measurements of the saturated vertical conductivity, and model resolution (Schellekens et al., 2020b). A sensitivity analysis conducted by the model developers concluded that the KsatHorFrac is the most effective parameter when it comes to calibration based on streamflow estimates, as also briefly discussed by Imhoff et al. (2020). Increasing the value of this parameter leads to an increased base flow component and reduced peak flow and flashiness.

We calibrated the models to match typical model setups used by users of the hydrological model. The model instances are calibrated using the modified Kling-Gupta efficiency score (KGE) (Gupta et al., 2009; Kling et al., 2012) by comparing the simulated streamflow estimates with the streamflow observations at the basin outlet. 21 runs are evaluated based on an interval of KsatHorFrac values ranging between 1 and 1000. The best-performing model run and its corresponding KsatHorFrac parameter are selected for further analysis during the evaluation period. The model calibration is conducted from 1997 to 2005 and the model evaluation from 2007 to 2016. The years 1996 and 2006 are regarded as spin-up years and are not included in the analyses.

3.2.3. BENCHMARK SELECTION

To select basins with model performance considered at least reasonable, we applied a statistical benchmark to beat. The use of a benchmark allows for better interpretation of objective-function-based results (Garrick et al., 1978; Knoben et al., 2020; Pappenberger et al., 2015; Schaefli and Gupta, 2007; Seibert, 2001; Seibert et al., 2018). We adopt, in

part, the same methodology for statistical benchmark creation as Knoben et al. (2020). The benchmark is created by calculating the mean and median of streamflow observations per calendar day of the 10-year evaluation period and the results are compared to streamflow predictions from the model. KGEs are calculated for the observed streamflow versus these mean and median values. The mean is expected to better represent larger basins with more stable flow regimes whilst the median better covers the flashiness of smaller headwater basins. This benchmark serves as a lower boundary for the model predictions: if the model at any of the 3 resolutions has a lower KGE than either the median or mean flow benchmark, it is considered unsuitable for this study and removed from further analyses.

3.2.4. ANALYSIS OF RESULTS

OBJECTIVE FUNCTION AND SAMPLING UNCERTAINTY

We use the modified Kling-Gupta efficiency metric for the analyses of results. Ideal model performance has a KGE score of 1 and a KGE score of -0.41 is equal to taking the mean flow as a benchmark (Knoben, Freer, and Woods, 2019). The KGE score is defined by its 3 components: the Pearson correlation, the mean bias, and the variability bias. The 3 components collectively determine the KGE score. The KGE score and its components are defined as follows:

$$KGE' = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \tag{3.1}$$

$$\beta = \left(\frac{\mu_s}{\mu_o}\right) \tag{3.2}$$

$$\gamma = \left(\frac{\sigma_s/\mu_s}{\sigma_o/\mu_o}\right) \tag{3.3}$$

where r is the Pearson correlation, β is the mean bias, and γ is the variability bias. μ_s and σ_s are the mean and standard deviation of the simulations. μ_o and σ_o are the mean and standard deviation of the observations.

We quantify the sampling uncertainty of the KGE score for the selected basins based on the statistical benchmark, following the methodology of Clark et al. (2021). This applies the non-overlapping bootstrap method (Efron and Tibshirani, 1986) to calculate tolerance intervals and the jackknife-after-bootstrap method (Efron, 1992) for the standard error calculation of those tolerance intervals. This method applies bootstrap and Jackknife methods to estimate the standard errors and tolerance interval of KGE uncertainty. The tolerance interval is defined as the difference between the 5th and 95th percentiles. We ran the gumboot analyses package (Clark et al., 2021) with a sample size of 500 during the evaluation period to calculate the results.

COMPARISON OF STREAMFLOW ESTIMATES

To provide more context for the results in terms of general model performance, we compared the streamflow estimates from wflow_sbm to those of the study by Knoben et al. (2020). The study by Knoben et al. (2020) ran 36 conceptual models using the Modular Assessment of Rainfall-Runoff Models Toolbox V1.0 (MARRMoT) (Knoben, Freer, Fowler,

et al., 2019) on the CAMELS data set. First, we calculated the mean of the 36 models for each basin. Next, we matched the basins investigated in both studies. Due to differences in time periods, forcing, and numerical solvers, the results cannot be directly compared to those of this study. They do, however, provide context for the results.

The inter-model (instance) comparison of the streamflow estimates in this study is assessed using cumulative distribution functions (CDFs). We applied the Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1933) to test whether the differences between the KGE score distributions of the model instances are statistically significant. This allows the acceptance or rejection of the hypothesis stating that the differences in streamflow estimates at various spatial resolutions will be small.

3.2.5. EWATERCYCLE PLATFORM

This research was conducted within the eWaterCycle platform (Hut, Drost, van de Giesen, van Werkhoven, et al., 2021). eWaterCycle follows by design the FAIR principles of data science (Wilkinson et al., 2018) and allows high-level communication with models regardless of programming language through the Basic Model Interface (Hutton et al., 2020). This study showcases how eWaterCycle handles the setup of extensive modelling studies. A Jupyter notebook with the model experiments of this study is provided in the GitHub repository. As notebooks are not ideal for long-running experiments on high-performance computing (HPC) machines, we exported the notebooks to regular Python code which we ran directly on an HPC system. The calibration and evaluation procedures totalled 41,025 model runs on the Dutch national supercomputer Cartesius, hosted by SURF.

3.3. RESULTS

The results in this section are based on the modified KGE 2012 objective function applied to the streamflow estimates at the basin outlet. The Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970) results are available in the repository (DOI:10.5281/zenodo.5724576).

3.3.1. CALIBRATION PERIOD RESULTS

THE EFFECT OF CALIBRATION ON STREAMFLOW ESTIMATES

To illustrate how model calibration affects the streamflow estimates of each model instance, we first show the calibration curves of a single basin (ID:14301000). To avoid presentation bias we selected a basin with moderate performance and only show the last year of calibration. Figures 3.3a–c show the calibration curves (yellow to red) for each of the instances that were generated by tuning the horizontal conductivity (i.e., KsatHorFrac) parameter. The parameter ranges from 1 to 1000 and has a single value for each basin. Not all calibration interval values are shown in the figure for visualization purposes. The results depict the effect that increasing the KsatHorFrac value has on the hydrograph: base flow increases whilst peak flow reduces. In addition, large KsatHorFrac values reduce the flashiness of the streamflow estimates as is visible in Figures 3.3b and 3.3c during the second week of November 2005. Of note is that the selection of best calibration parameter values is strongly dependent on the chosen objective function as for example the NSE score would be more favourable towards flashiness and less towards

base flow. As shown in Figure 3.3d the streamflow estimates of the model instance are similar (KGE scores of 0.58–0.66) while the parameter values deviate (KsatHorFrac values of 125–1000). No strong apparent trends for KsatHorFrac values in relation to model resolution or geographic location were found after calibration.

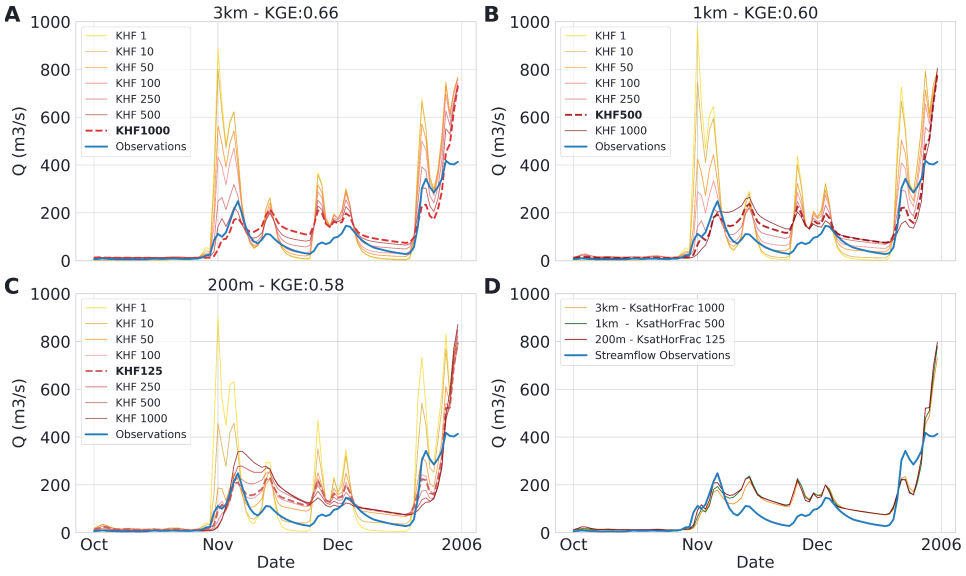


Figure 3.3: Calibration-period results. The calibration interval of the KsatHorFrac parameter (KHF) for the 3 model instances at the basin outlet (ID:14301000) of the final calibration period year. (a) 3 km model instance. (b) 1 km model instance. (c) 200 m model instance. Values range from 1 to 1000 (yellow to red); only a sub-selection is shown. Best-performing calibration values are indicated by a dotted red line and streamflow observations in blue. (d) Best streamflow estimates of the 3 km (orange), 1 km (green), and 200 m (red) model instances.

Figure 3.4 shows the CDFs of the KGE-score distributions based on the best-performing calibration model run of the 3 model instances. Starting with the modified KGE score distributions in Figure 3.4a, we find that the model instance distributions are very similar for the whole domain, especially the 1 km and 200 m instances. The 3 km instance has lower scores for approximately 60 % of the distribution between 0.2 and 0.8 CDF. Approximately 18 % of each distribution is below a KGE score of -0.41 which corresponds to taking the mean flow. All three model instances show a similar Pearson r correlation CDF (Figure 3.4b) and gamma variability ratio (Figure 3.4c). The largest differences are visible in the beta bias component (Figure 3.4d), which shows similar biases for the 1 km and 200 m instances, with approximately 60 % of each distribution below 1 and 40 % above that value. Agreement is highest in the lower and upper 5 % of the distributions. The 3 km instance agrees with the 1 km and 200 m instances only in the upper 10 % of the distribution, with 70 % of the 3 km distribution below 1 and 30 % above that value. The bias term of the KGE score has the largest weight in determining the shape of the KGE score CDF as is shown by the 20 % of each distribution with bias values larger than 2.0.

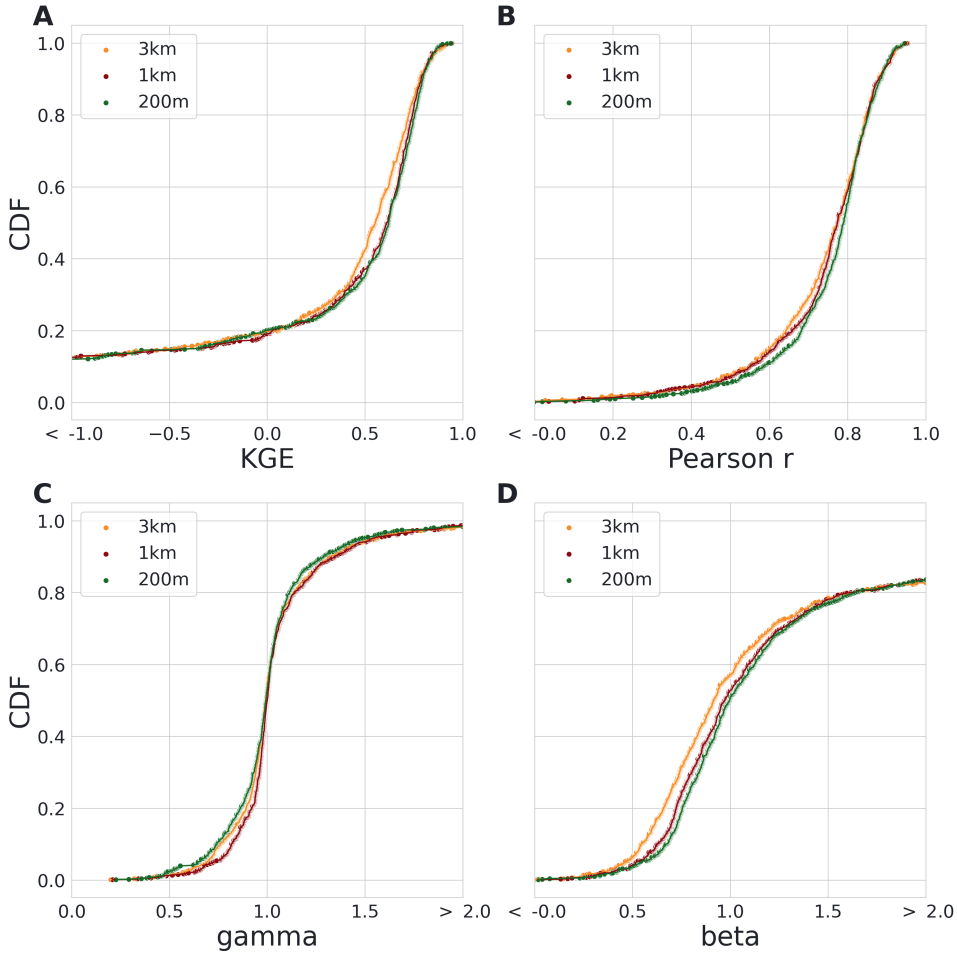


Figure 3.4: Calibration-period results. CDFs of the modified KGE scores and their individual components for the best-performing model runs of the 3 model instances during the calibration period. The 3 km model instance is shown in orange, the 1 km instance in red, and the 200 m instance in green. (a) The CDF of the modified KGE score. (b) The CDF of the Pearson r correlation component. (c) The CDF of the gamma (variability ratio) component. (d) The CDF of the beta (bias ratio) component.

3.3.2. EVALUATION PERIOD RESULTS

BENCHMARK SELECTION

The statistical benchmark is applied to determine which basins contain the streamflow estimates of the model instances that are deemed adequate for further analyses. The statistical benchmark is based on the best-performing type of calendar-day climatology, either mean or median, during the model evaluation period. Figure 3.5a shows spatially which type of calendar day benchmark is best performing per basin. Of the 567 simulated basins, the results for 454 basins exceed the benchmark for each model instance. This is the case for all basins in the Midwest of the United States. Poor performance in comparison to the benchmark is mainly present in the Southwest. Based on the KGE scores, 83 % of the benchmarks favour the climatological mean and 27 % favour the median. An overview is provided in Figure 3.5c, and the distribution of the benchmark KGE scores is shown in Figure 3.5d. The distribution ranges from -0.62 to 0.71 KGE score and is skewed towards values lower than 0.0 with a mean of 0.02 and a median of -0.02.

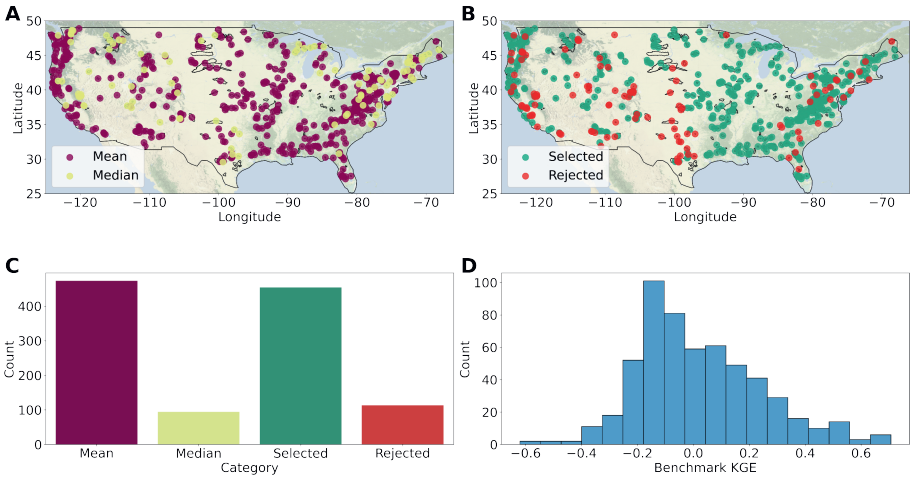


Figure 3.5: Evaluation-period results. (a) The best-performing type of 10-year calendar-day climatology, either mean (purple) or median (yellow). (b) The spatial distribution of accepted (green) and rejected (red) basins based on the benchmark. (c) Overview of the number of basins that are accepted or rejected and the best-performing benchmark type. (d) The KGE 2012 score distribution of the best-performing benchmark type. Basemaps are made with Natural Earth.

The statistical benchmarks (mean and median during the evaluation period) are plotted as CDFs based on the KGE score and its 3 individual components in Figure 3.6. For most of the KGE distributions in Figure 3.6a, the mean benchmark outperforms the median with the upper 10 % being the exception. The Pearson correlation coefficient (Figure 3.6b) is slightly higher for the median at 70 % of the distribution. In the case of the gamma variability ratio (Figure 3.6c) the bottom 38 % is lower for the median than the mean and the upper 55 % slightly higher for the median than the mean. The determining

component of the KGE score is the beta component shown in Figure 3.6d. As expected, the mean benchmark is for the most part close to a value of 1, meaning that it is close to the mean of the observations. Of interest are the points of the median that are not close to 1, the lower 10 % and the upper 1 % of the distribution. These basins have flow regimes that greatly differ per year compared to the climatology. Considering basins for which the mean statistical benchmark has a beta bias score below 0.75 or above 1.75, we find that the model instances outperform the statistical benchmark in 35 out of the 44 basins.

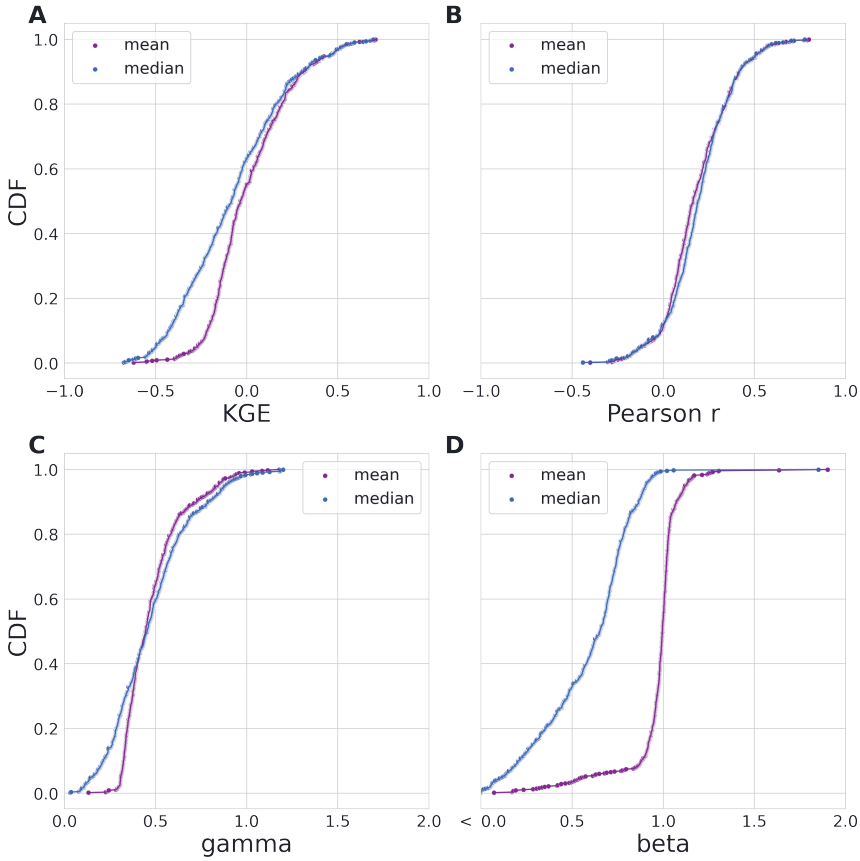


Figure 3.6: Evaluation-period results. The CDFs of the modified KGE score and its 3 individual components for the statistical benchmarks during the evaluation period. The mean statistical benchmark is shown in purple and the median benchmark in blue. (a) The CDF of the modified KGE score. (b) The CDF of the Pearson r correlation KGE component. (c) The CDF of the Gamma (variability ratio) KGE component. (d) The CDF of the Beta (bias ratio) KGE component.

THE EFFECT OF SPATIAL RESOLUTION ON STREAMFLOW ESTIMATES

The same 3 example basins used for the calibration period are used to illustrate the differences in streamflow estimates between model instances for the evaluation period. Only the last year of the evaluation period is shown in Figure 3.7.

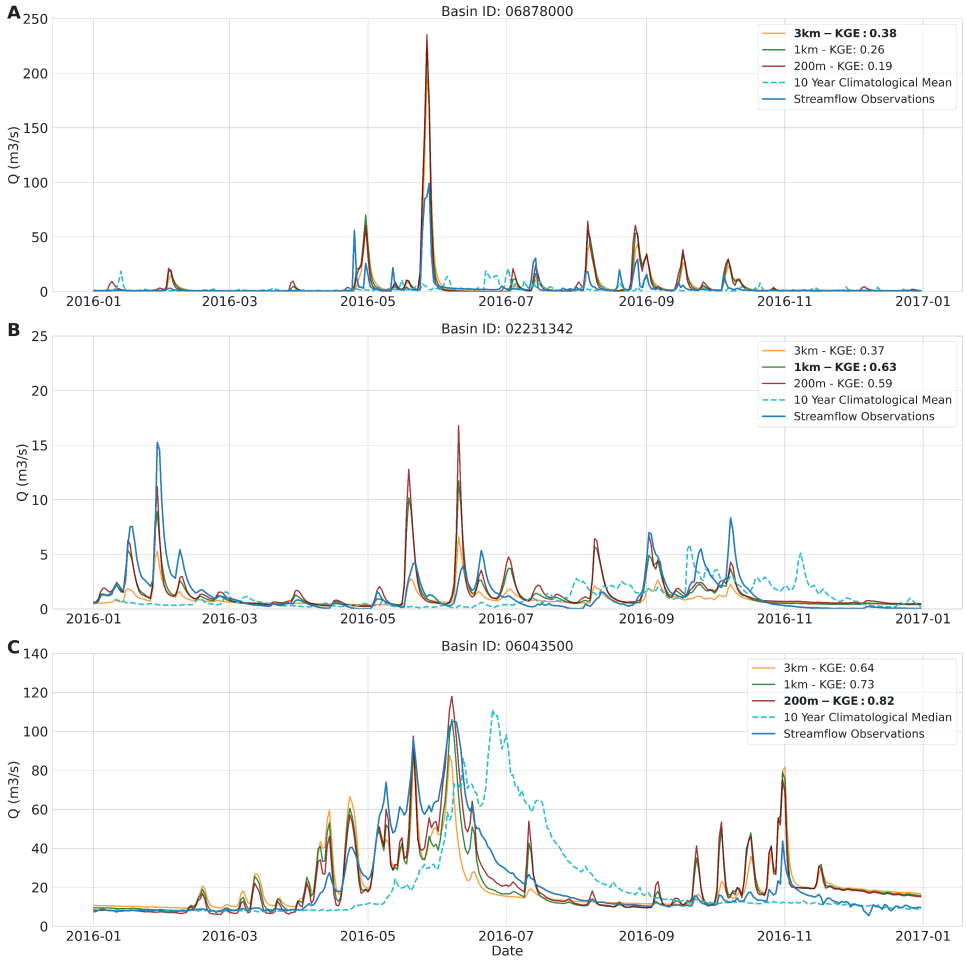


Figure 3.7: Evaluation-period results. 3 example hydrographs showing the last year of the evaluation period. The 3 km (orange), 1 km (red), and 200 m (green) model-instance streamflow estimates at the basin outlet are shown. The streamflow observations are shown in blue and the 10-year calendar-day climatology as a dotted cyan line of the statistical benchmark. (a) Basin ID: 06878000. (b) Basin ID: 02231342. (c) Basin ID: 06043500.

In cases of poor performance, such as that shown in Figure 3.7a, the model instances are overestimating streamflow during peak flow. The best-performing model instance has the smallest peak-flow estimates, which in many cases is the coarsest spatial resolution instance (3 km). Of the 454 basins, the 3 km instance has the lowest peak-flow estimates in 279 cases and is the best-performing instance in 148 of those cases. The

example in Figure 3.7b shows that when the 1 km model instance performs best, the 200 m model instance often performs similarly. In 78.7 % of the cases in which the 1 km instance performs best, its KGE score differs from that of the 200 m instance by less than 0.1. The final example in Figure 3.7c illustrates a case in which the finest spatial model resolution (200 m) performs best and the coarsest (3 km) performs worst. The 200 m model instance best captures peak flow and the receding limbs of the hydrograph.

STREAMFLOW ESTIMATES OF MODEL INSTANCES

The KGE score results for the evaluation period are shown in Figure 3.8a. The KGE score distribution of the mean of 36 hydrological models from Knoben et al. (2020) is included and is referred to as "MARRMoT mean results.". Of note is that comparison between studies is not one-on-one due to differences in model run periods, forcing, and numerical solvers. We can, however, obtain information about general model performance across the two studies.

The mean KGE score distribution of the MARRMoT models (Figure 3.8, blue) of Knoben et al. (2020) is close to the mean of the distributions of the 3 model instances. Differences between study results are mainly present in the tails of the distributions. Below a CDF value of 0.17 (worst 17 % of the results) the MARRMoT mean results KGE score distribution is higher than the 1 km model instance. The MARRMoT mean results for the lower 5 % of the CDF perform better than the 3 model-instance distributions. Here, the range of KGE scores is smaller for the MARRMoT mean (-1.55 to 0.09) than for the 3 model instances (-13.56 to 0.00). Above a CDF value of 0.17 (83 % of the results) the distributions of the 3 model instances are higher in KGE scores than those of the MARRMoT mean.

When we consider only the wflow_sbm instances, approximately 64 % of the model-instance results have KGE scores above 0.50, and 18 % have scores above 0.75. The distributions cross at multiple points, for example at the bottom 10 % of the distribution the 3 km instance has the highest and the 1 km instance the lowest KGE score. At 40 % of the distribution and lower the 200 m instance has the highest KGE scores, followed by the 1 km and 3 km instances. The Pearson r component of the KGE score CDF in Figure 3.8b and the gamma variability ratio component in Figure 3.8c show small differences between model instances. The beta bias component in Figure 3.8d shows the largest differences between model instances, especially between the 3 km instance and the 1 km and 200 m model instances. The bias component is the main factor for the differences in the overall KGE score.

Next, we apply the Kolmogorov-Smirnov (KS) statistic to test whether the CDFs of the model instances differ significantly from each other at a p-value threshold of 0.05. The KS test results in Table 3.2 show that the difference between the 3 km and 200 m model instances is statistically significant at a p-value of 0.02. Across the large sample, this means that increasing the spatial model resolution from 3 km to 1 km or from 1 km to 200 m does not lead to significant differences in streamflow performance. When changing resolution from 3 km to 200 m, the distribution of KGE scores is significantly different ($p < 0.05$) according to the KS test.

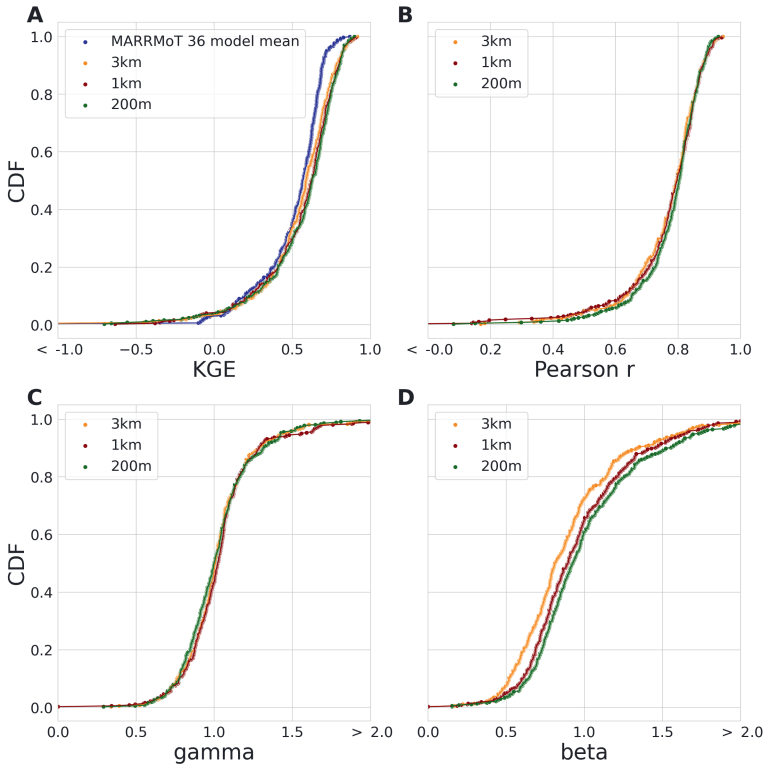


Figure 3.8: Evaluation-period results. The CDFs are based on the modified KGE scores and their 3 individual components for the 454 selected basins. (a) The modified KGE score with the MARRMoT 36-model mean results shown in blue, the 3 km instance in yellow, the 1 km instance in red, and the 200 m instance in green. (b) The Pearson r correlation component. (c) The gamma variability component. (d) The beta bias component.

Table 3.2: Kolmogorov-Smirnov statistics and the corresponding p-values. The results are based on differences between the KGE distributions of the 3 model instances: 3 km, 1 km, and 200 m.

CDFs	**Kolmogorov-Smirnov Statistic**	**p-value**
3 km - 1 km	0.08	0.14
1 km - 200 m	0.05	0.82
3 km - 200 m	0.11	0.02

OBJECTIVE FUNCTION UNCERTAINTY

In addition to the streamflow evaluation, we conducted a sampling uncertainty assessment of the KGE objective function using bootstrap and jackknife-after-bootstrap methods. The results of this assessment for each of the model instances are shown in Figure 3.9. The results for the 3 model instances are very similar. The tolerance interval results denoted with the black lines show that approximately 100 basins have a KGE sampling uncertainty of 0.1 or lower and approximately half of the basins have an uncertainty of 0.2 or lower. Half of the basins show KGE uncertainties greater than 0.2 with approximately 80 basins exceeding 0.5 for all model instances.

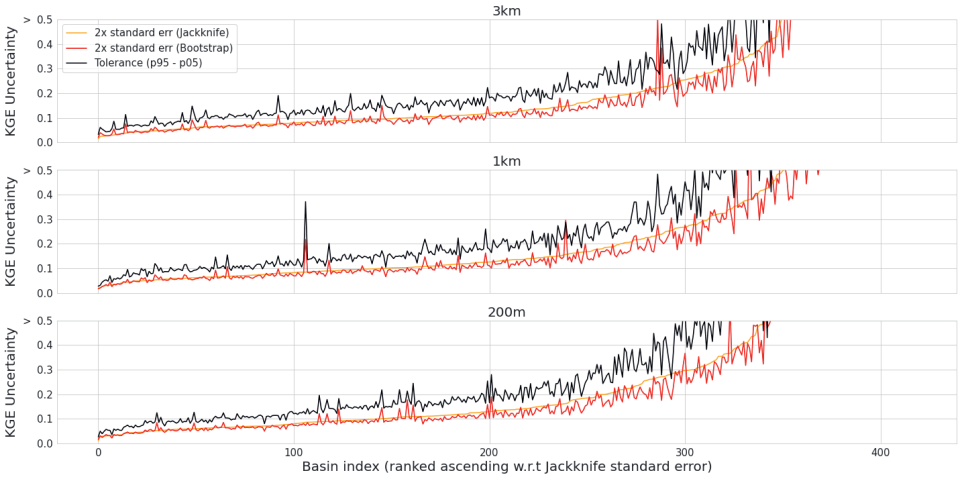


Figure 3.9: Evaluation-period results. The bootstrap and jackknife-after-bootstrap results of the sampling uncertainty of the KGE score. The 2× standard error of the jackknife method is shown in orange, the 2× standard error of the bootstrap method in red, and the tolerance interval obtained by subtracting the 5th percentile from the 95th percentile is shown in black. The horizontal axis shows the basin index ranked by the jackknife standard error and the vertical axis shows the modified KGE score sampling uncertainty. (a) 3 km model instance. (b) 1 km model instance. (c) 200 m model instance.

We project the sampling-uncertainty results onto the evaluation-period CDF in Figure 3.8 by calculating the mean of the tolerance interval, jackknife standard error, and bootstrap standard error for each of the 3 model instances per quarter of the total CDF. The results in Table 3.3 show slightly higher results for quarters of the CDF for the tolerance interval relative to the jackknife and bootstrap standard errors. The lower tail of the CDF contains the highest average values for the three sampling-uncertainty statistics and the upper tail the lowest average values, indicating that sampling uncertainty is high at low KGE values and vice versa.

SPATIAL DISTRIBUTION OF EVALUATION PERIOD RESULTS

The CDF does not provide information at the basin level. To gain insight into the spatial distribution of the KGE scores of the model instances, Figure 3.10 shows the KGEs of the streamflow estimates plotted on a map of the CONUS domain. The minimum KGE scores of 0.50 to 0.89, shown in Figure 3.10a, are found in the Pacific Northwest,

Table 3.3: Sample uncertainty analysis results per quarter of the KGE cumulative distribution function (CDF) during the evaluation period. The mean across the three model instances is shown for each method.

CDF Range	Mean Tolerance (p95 – p05)	Mean 2× Std. Err. (Jackknife)	Mean 2× Std. Err. (Bootstrap)
0.00 – 0.25	6.35	5.11	4.00
0.26 – 0.50	0.29	0.18	0.21
0.51 – 0.75	0.19	0.12	0.13
0.76 – 1.00	0.14	0.09	0.10

Atlantic South, Appalachia, and Northeast of the CONUS. KGE scores lower than -0.41 are found throughout the CONUS. The highest KGE scores in Figure 3.10b are located in the Northwest, Rocky Mountains, and Appalachia. These regions are characterised as steep sloping headwater basins. Figure 3.10c shows that there are large local streamflow discrepancies of more than 1.00 KGE score. These are mainly found in the Pacific Southwest, the South, and the Midwest. These regions span a wide range of hydroclimatically diverse basins. The average KGE score difference is 0.22. Figure 3.10d shows the best-performing model instance for each of the 454 selected basins. Although regions show clusters of best-performing model instances, there is no overall geographical trend in the results. Basins in which the 1 km model instance performs best are generally close to basins in which the 200 m model instance performs best. The 3 km model instance shows some clusters in the South and Pacific Northwest. The Rocky Mountains contain the best-performing 200 m model instances, and the Appalachia contain a mixture of 1 km and 200 m model instances.

3.3.3. THE EFFECT OF SPATIAL SCALE ON TERRAIN CHARACTERISTICS

We illustrate the effect of spatial scaling on the parameter sets of the 3 model instances by showing differences in topography and drainage density for 3 basins. To avoid presentation bias, the basins were sampled based on poor streamflow performance (ID:06878000), moderate performance (ID:02231342), and good performance (ID:06043500). Figure 3.11a shows the probability density functions (PDFs) of the height distributions of the model instances for each of the 3 basins, Figure 3.11b shows the slope distributions, and Figure 3.11c shows the profile-curvature distributions.

The height distribution of the DEM in Figure 3.11a shows, most clearly for basin ID 02231342, how the representation of the highest altitudes is underestimated by the 3 km model instance (orange) compared to the 200 m instance (green) and to a lesser extent the 1 km instance (red). Essentially, at coarser spatial resolutions, the terrain is flattened at high altitudes. An opposite effect is shown in this basin for the lower altitudes, where the finer resolution instances better capture gentle slopes that are flattened at coarse resolution. This effect is also detectable in the slope and profile curvature PDFs shown in Figures 3.11b and 3.11c. As can be expected from the height distributions, the 200 m instance has gentler and steeper slopes than the 1 km and 3 km instances. This is shown by the narrower slope distribution for the coarse spatial resolution that broadens with finer resolution. The differences in the mean slope of the basins between model instances are marginal, e.g., 0.00019 m m^{-1} for basin ID 02231342. The profile curvature in Figure

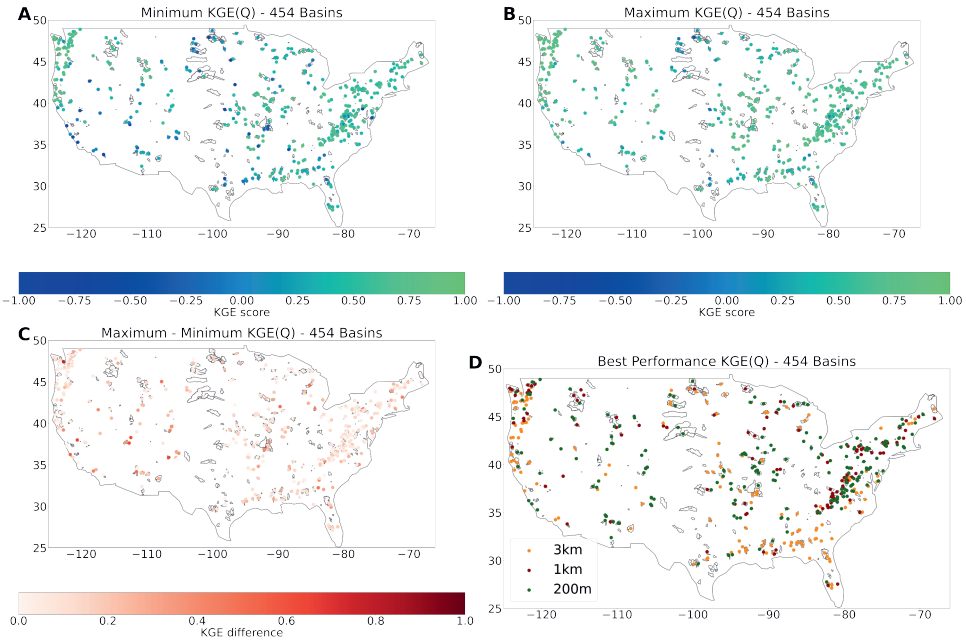


Figure 3.10: Evaluation-period results. (a) Minimum KGE score of the model instances. (b) Maximum KGE score of the model instances. (c) The difference between the minimum and maximum KGE scores. (d) Best-performing model instance based on KGE score for the evaluation period with 3 km in orange, 1 km in red, and 200 m in green.

Table 3.4: Drainage density defined as stream length divided by basin area for the three example basins.

Basin ID	3 km Drainage Density [m·m^{-2}]	1 km Drainage Density [m·m^{-2}]	200 m Drainage Density [m·m^{-2}]
06878000	0.0133	0.0141	0.0143
02231342	0.0056	0.0059	0.0063
06043500	0.0112	0.0123	0.0126

3.11c indicates whether a slope is linear (values close to 0), concave (values smaller than 0), or convex (values larger than 0). The 3 km and 1 km instances show similar slope geometries, with the 3 km instance having slightly more linear slopes. At the finest resolution (200 m), the slope geometry shifts from linear slopes to either convex or concave curvature profiles.

In addition to topography, we calculated the drainage density for each model instance, defined as total river length divided by basin area. The results in Table 3.4 show small differences between the model instances for each of the 3 basins.

3.4. DISCUSSION

3.4.1. BENCHMARKS

We applied an initial statistical benchmark based on streamflow observations for basin selection to identify basins whose streamflow estimates are deemed adequate for further analyses. This does not imply that excluded basins are less relevant. Instead, it implies that an in-depth model assessment is required to understand why the model is not able to simulate adequate streamflow estimates in these basins. The CDFs of the KGE score and its components based on the statistical benchmarks in Figure 3.6 show that the benchmark is relatively easy to beat given the KGE score distributions. We find, as one might expect, that the beta bias component has values close to 1 for the mean benchmark while this is not the case for the median benchmark. Bias values not close to 1 for the mean benchmark indicate that the flow regime changes from year to year. The hydrological model simulations were in most cases able to capture this change better than the benchmark. This shows that the model is able to capture yearly variability and is not overfitting due to extensive calibration. In addition to the benchmark, we added a layer of context by including results from the study by Knoben et al. (2020). This is an imperfect comparison due to differences in model inputs, numerical solvers, and simulation periods. However, the results do provide information on general model performance. The results show that the wflow_sbm streamflow estimates are in line with estimates of the mean of the 36 MARRMoT models. The spread of results is smaller for the 36 MARRMoT models which is due to averaging and likely the more extensive calibration routine of the conceptual models. It also implies that when only streamflow at the basin outlet is under consideration, users should carefully consider various model structures before model selection due to the small differences in results at the system-scale.

Other studies have conducted large-domain modelling efforts with the CONUS as a case study area (e.g. Mizukami et al., 2017; Rakovec et al., 2019). However, these are hard to compare with the results from this study as they did not use the same basins. To

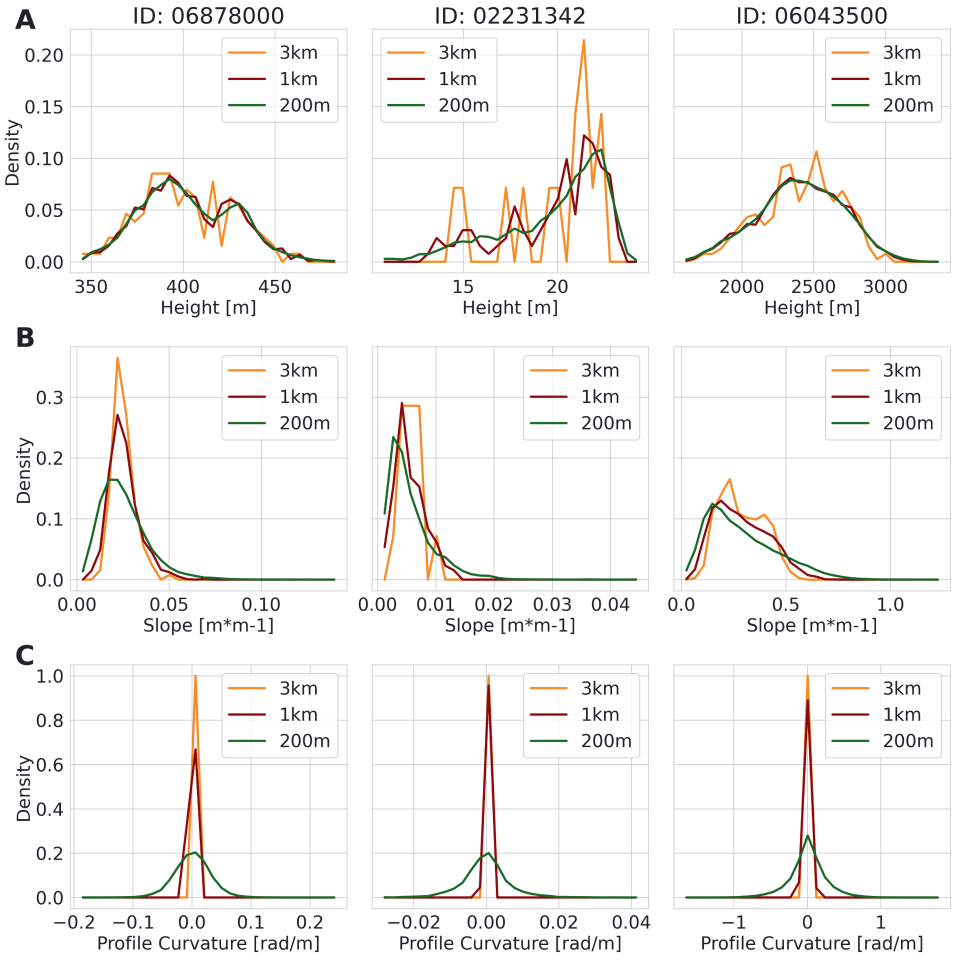


Figure 3.11: 3 example basins representing poor streamflow performance (ID:06878000), moderate performance (ID:02231342), and good performance (ID:06043500). (a) PDFs of height distributions for the 3 km (orange), 1 km (red), and 200 m (green) model instances. (b) PDFs of the slope distributions of the model instances per basin. (c) PDFs of profile-curvature distributions of the model instances per basin. Values equal to 0 indicate linear slope geometry, values smaller than 0 indicate concave slope geometry, and values larger than 0 indicate convex slope geometry.

improve future comparative work, we advocate for the creation of model output storage guidelines that use the CAMELS data set as case study area. These guidelines should encompass the differences between hydrological models, such as distributed and non-distributed modelling grids. A first step is the inclusion of distributed data sources in the CAMELS data set, e.g. meteorological data. This can be extended by including model evaluation products such as snow cover and soil moisture for further benchmarking. We further propose the use of a model experiment environment, such as eWaterCycle (Hut, Drost, van de Giesen, van Werkhoven, et al., 2021), to generate model results. This allows for similar pre-processing of inputs, standardization of outputs, and reproducible modelling studies. An example of how to apply these steps using the eWaterCycle platform is provided in the Jupyter Notebooks that supplement this publication (DOI:10.5281/zenodo.5724512). The ease of setting up a model experiment and storing its output is an incentive for users to store model results while conducting extensive modelling studies even when those results are deemed unsuitable for publication but might still benefit the community.

3.4.2. STREAMFLOW ESTIMATES AND UNCERTAINTY

At the start of the study, we hypothesized that differences between model instances would be small due to quasi-scale-invariant parameter sets and constant hydrological process descriptions within the model. The results of the calibration period in Figure 3.4 and the evaluation period in Figure 3.8 show that this is the case for the KGE score distributions based on streamflow estimates at the basin outlet. Although the differences are small, the crossing of the distribution lines is a strong indication that there is disagreement on KGE scores between model instances and that there is no single instance outperforming the rest consistently. The largest difference among the distributions for both periods is found in the beta bias component of the KGE score. The benchmark selection mainly excluded the basins that contained large bias values during the calibration period. For the bias component, the 3 km instance deviates from the 1 km and 200 m instances. This shows that the total volume of streamflow does differ for the 3 km instance. This difference was confirmed by testing the statistical significance of the differences between distributions of the model instances by applying the KS-statistic on the evaluation period results. Here, a statistically significant difference between distributions is found only for the 3 km and 200 m model-instance combination (Table 3.2).

This study applies a single objective function, the modified KGE (Kling et al., 2012), to determine simulated streamflow adequacy for the model calibration and evaluation time periods. To provide the reader with more context, we have included the KGE 2009 (Gupta et al., 2009), modified KGE, KGE non-parametric (Pool et al., 2018), and NSE objective function results in the data repository and in Figure 3.13 in the supplementary analyses at the end of this chapter. The conclusions based on hypothesis testing were not affected by the type of KGE objective function. We selected the modified KGE objective function as it is less influenced by extreme combinations of simulated and observed streamflow and less influenced by structural error in the meteorological input. The single objective function for a whole period approach is limited and can be improved by first determining the objective function for individual years and then averaging for the

whole period (Fowler et al., 2018). In addition, as stated by Clark et al. (2021), it is important to determine the sampling uncertainty of objective functions to avoid wrong conclusions at the system scale. Following their methodology, we investigated the sampling uncertainty by applying bootstrap and jackknife-after-bootstrap methods (Figure 3.9). Approximately half of the basins have a KGE uncertainty of 0.2 or lower based on the tolerance interval and half have an uncertainty higher than 0.2 with values surpassing 0.5 KGE uncertainty. The lower 25 % of the cumulative distribution function (Figure 3.8) contains the highest sample uncertainty and the upper 25 % the lowest (Table 3.3). Given these results, we argue that the differences found using the KS-statistic are easily within the sampling uncertainty margins and therefore not a valid basis for conclusions, especially when model and observational uncertainty are also taken into account. This demonstrates the drawbacks of conducting large-sample assessments and the sensitivity to sampling uncertainty of the results.

3.4.3. RELATIVE MODEL INSTANCE DIFFERENCES

We recognize that large-sample assessments obscure variations in simulations between instances due to the sample size. At the basin level, we find that local variations due to spatial resolution are in effect throughout the domain. This is depicted by the differences between the KGE scores of the model instances (Figure 3.10c). On average, there is a 0.22 KGE score difference with extremes of more than 0.5 KGE score difference at multiple basins. In this case we are interested in the relative differences between the instances to understand the effect of varying spatial resolution. According to Oreskes et al. (1994), this relative comparison is the only form of validation that is actually possible. Therefore, we argue that the differences are large locally even though they might be within the sampling uncertainty range.

We find that the 1 km instance performs best in basins where the difference between the minimum and maximum KGE scores of the 3 instances is small (Figure 3.10c). We attribute this partly to the calibration routine finding a better parameter value (KsatHorFrac) as its results are often close to those of the 200 m instance. For the best-performing 3 km or 200 m model instances (Figure 3.10d) there are only small geographical trends of best model performance in the South and Appalachia. This information is valuable for future research that does a more in-depth assessment of the internal states and fluxes as we now know where the effects of varying grid-cell size are small or large and which model is best performing.

We conducted a terrain analysis (Figure 3.11) to identify changes in terrain characteristics due to spatial resolution that might explain the differences between model-instance streamflow estimates. Minor effects are depicted by the profile curvatures present at 200 m resolution. Slopes are less linear at fine resolution than at coarse resolution. The effect on the hydrological response is, however, expected to be small as stated by Bogaart and Troch (2006). Similarly small changes are found for the differences in drainage density between model instances (Table 3.4). This confirms that the drainage network up-scaling method of Eilander et al. (2021) is (almost) consistent across spatial scales.

Larger differences between model instances are found for the height distribution of the DEM, which is flattened at coarse resolution compared to finer resolution. At high altitudes, this introduces changes in snow dynamics due to the use of the temperature

degree-day method by the hydrological model. The resulting effect on streamflow estimates depends on the relative contribution of snow melt. Although marginal at a basin level, the difference in slope between instances is expected to affect the partitioning of the lateral fluxes of the wflow_sbm model as the lateral connectivity between grid cells is slope-driven. An increase in slope would lead to larger lateral fluxes and vice versa. Increasing spatial resolution, and thus reducing DEM aggregation, results in a broader distribution of slopes that affects the volume and timing of streamflow estimates. The effect of terrain smoothing has been reported by Shrestha et al. (2015) and they found this to increase the lateral overland-flow flux. An in-depth assessment of the internal states and fluxes of the model instances is required to determine whether these components are the main causes of the differences in streamflow estimates.

We applied the same meteorological forcing products and pre-processing routine to each model instance. This ensured that the total volume of precipitation remained consistent across scales. A coarse grid cell contains a volume of precipitation that is equally redistributed over the equivalent number of finer grid cells. In reality, this redistribution of water might not be equal across the finer grid cells and therefore scaling behaviour is introduced due to the locality of precipitation. This has an effect on the streamflow estimates as the locality of precipitation directly influences hydrological processes that are dominant at different locations (e.g., hillslopes). In addition, due to the large difference between the native data resolution and the model-instance resolutions, it is likely that the effects of disaggregation of precipitation and temperature lapse-rate correction are main drivers of differences in streamflow estimates between model instances. However, Shuai et al. (2022) found by comparing the same forcing product with native implementations at various spatial resolutions that the effect on streamflow estimates was relatively small. This was not the case for distributed variables in the basin (e.g., snow water equivalent). It is of interest to investigate this effect on streamflow estimates and to determine the role of the native spatial resolution of the meteorological forcing.

COMPUTATIONAL COST

When we consider the increase in streamflow-based model performance as opposed to computational cost, we find that this does not scale linearly with the number of grid cells in the basin due to lateral connections in the hydrological model. The average nonparallel runtime of the 3 km instance is 157 seconds and that of the 200 m instance 12120 seconds with an average number of grid cell difference of 28872 cells. These results point toward the importance of conducting an initial spatial model resolution assessment at the start of a large sample assessment, as this can help avoid subpar or computationally expensive model runs. Note that this kind of information can stimulate scientific and/or computational developments, e.g. in the meantime the wflow code was rewritten in Julia (van Verseveld et al., 2021) roughly increasing performance by a factor of 3 while other improvements (threading, MPI) are being implemented. There are alternative approaches to the spatial discretization of basins that are computationally very efficient such as the vector-based configurations that have the added benefit of better capturing topographic details and are less influenced by native forcing resolution (Gharari et al., 2020).

3.4.4. OUTLOOK

The results from this study help model developers with model refinement by providing them with insight into where and under what circumstances differences due to spatial scaling occur. Based on the aggregated domain and basin-level results, we can conclude that increasing spatial resolution does not necessarily lead to better streamflow estimates at the basin outlet. The implications of the results for the user are that caution is advised when interpreting high-resolution model output as this does not directly translate into better model performance. Moreover, the computational cost of increasing model resolution is not always warranted compared to gains in streamflow-estimate-based model performance.

We conducted this study as an initial assessment to be followed by studies of scaling effects in distributed hydrological models. As the sampling-uncertainty results showed, it is very hard to draw conclusions from a large sample, and future research should therefore consider a smaller subset of basins to explore scaling effects in more detail. In this study, we did not investigate individual basins to avoid biased selection of case study areas. We suggest that future work should include basins that show large or small differences in model performance and lateral fluxes, as well as the effects of terrain aggregation. In addition, the evaluation should go beyond streamflow by using multiple evaluation data products (e.g., soil moisture, evaporation, and gravity anomalies); see Guse et al. (2021) for a recent overview. Conclusions might then be drawn about whether increasing spatial resolution leads to increased model fidelity or not. This should be tested using multiple forcing data sets to test the robustness of the conclusions. Including multiple timescales, as discussed in Melsen et al. (2016), could provide more information about the linearity of hydrological process descriptions in the model.

3.4.5. CONCLUSIONS

The aim of this study was to analyse the effects of varying spatial resolution on the streamflow estimates of the distributed wflow_sbm hydrological model. Distributions of model-instance KGE-score results were tested for significant differences, and their sampling uncertainty was quantified. A spatial distribution assessment was conducted to derive spatial trends from the results. The main findings of the study are the following:

- The distributions of streamflow estimates from the wflow_sbm model instances derived at multiple spatial grid resolutions (3 km, 1 km, and 200 m) differ significantly only between the 3 km and 200 m instances ($p < 0.05$). This confirms at an aggregated level the hypothesis that differences between model instances are small due to quasi-scale-invariant parameter sets and process descriptions that remain constant across scales in the hydrological model. However, the sampling uncertainty of the KGE score proved to be large throughout the domain. Therefore, we conclude that the minute differences found between model instances are too small to base conclusions on.

- Results show large differences in maximum and minimum KGE scores with an average difference of 0.22 between model instances throughout the CONUS. These results provide valuable information for follow-up research based on the locality of relative model scaling effects.

- There is no single best-performing model resolution across the domain. Finer spatial resolution does not always lead to better streamflow estimates at the outlet.
- Changes in terrain characteristics due to varying spatial resolution influence lateral-flux partitioning of the wflow_sbm model and might be an important cause of differences in streamflow estimates between model instances.

This study identified where results vary strongly because of changes in spatial resolution. Future research should conduct an in-depth assessment of basins in which streamflow estimates and lateral fluxes differ substantially among spatial scales. This will improve understanding of why spatial-scaling-related issues occur locally and the conditions under which they arise.

SUPPLEMENTARY ANALYSES

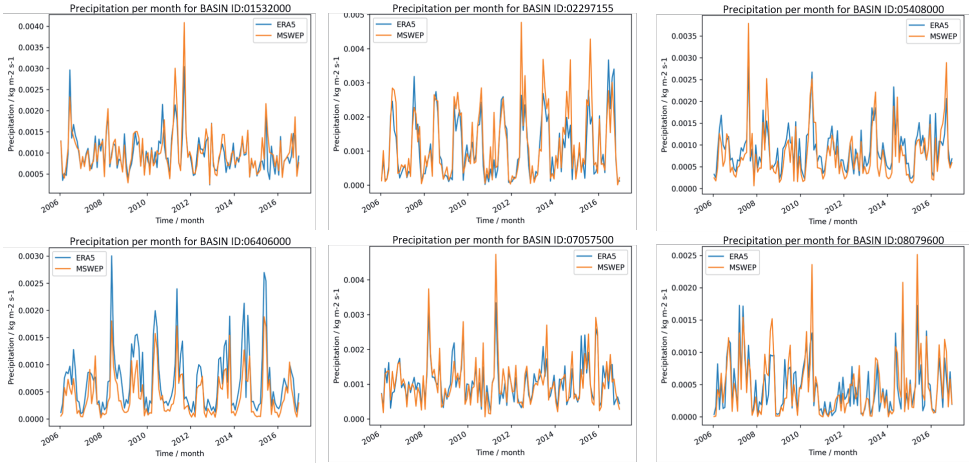


Figure 3.12: ERA5 and MSWEP forcing comparison for six basins in the CAMELS data set. Monthly precipitation values for the evaluation period are shown for ERA5 in blue and MSWEP in orange.

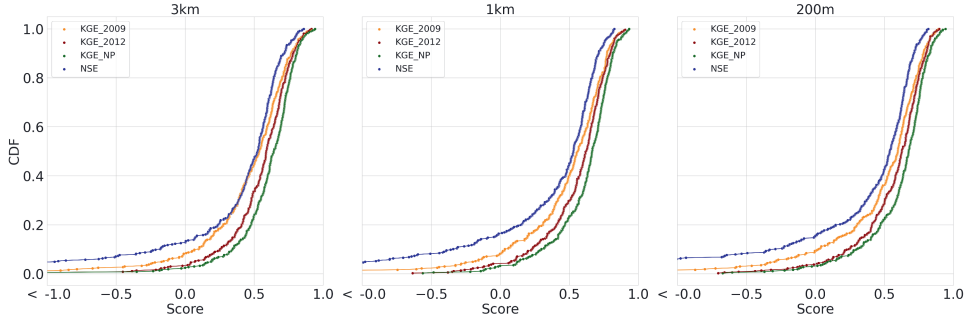


Figure 3.13: Evaluation-period cumulative distribution functions of multiple objective functions for the three model instances, with KGE 2009 in orange, KGE 2012 in red, non-parametric KGE in green, and NSE in blue.

4

ADVANCING HYDROLOGICAL MODEL DIAGNOSTICS: AN EXPLORATORY APPROACH USING RANDOM FOREST MODELS AND A LARGE-SAMPLE CATCHMENT DATA SET

Real stupidity beats artificial intelligence every time.

Terry Pratchett, *Hogfather*

This study introduces a diagnostic approach for understanding hydrological model performance by using random forest models (RFMs) to relate model outcomes to catchment attributes. Using the CAMELS-GB data set, we analyzed 93 catchment attributes from 617 catchments against performance metrics from the wflow_sbm and PCR-GLOBWB hydrological models. Our results reveal that the predictive skill of RFMs varies among hydrological targets, such as streamflow, actual evaporation, and soil moisture. Catchment area emerged as a primary predictor, challenging traditionally reported relationships with meteorological inputs and aridity. The key findings identify soil parameterization, process representation, and land-cover parameterization as promising areas for refining the hydrological models considered. This research emphasizes the value of data-driven methods for supporting hydrological model refinement. It offers a generalizable approach to exploratory model diagnostics and comparison, providing an avenue for actionable refinement across hydrological models and contexts.

Parts of this chapter are included verbatim in the manuscript *Advancing Hydrological Model Diagnostics: An Exploratory Approach Using Random Forest Models and a Large-Sample Catchment Data Set*, currently under review, by Aerts, J. P. M., Hut, R. W., van de Giesen, N. C., van der Heijden, T. J. T., Hoch, J. M., Weerts, A. H., and Sutanudjaja, E.

4.1. INTRODUCTION

As the hydrological cycle is expected to intensify, the frequency and intensity of droughts and extreme wet events are also expected to increase (Gründemann et al., 2022; Ingram, 2016; Lehner et al., 2006; Tabari, 2020). These shifts require hydrological models to accurately simulate the resulting impacts in support of water-resources assessments (e.g., Samaniego et al., 2018). This need emphasizes the importance of continuously refining the hydrological models at the core of such studies based on new insights. Model refinement typically occurs iteratively (e.g., Fenicia et al., 2008; Jakeman et al., 2006; Peters-Lidard et al., 2017) as new qualitative and quantitative data sources become available (e.g., Silberstein, 2006).

The iterative process of model refinement creates challenges when modifications to model processes and parameterizations must be verified and validated. This challenge is referred to as the equifinality problem (e.g., Beven, 2006b; Beven and Freer, 2001; Hornberger and Spear, 1981; Khatami et al., 2019; Kirchner, 2006) or, similarly, the affirmation fallacy (Oreskes et al., 1994). Models cannot be fully verified; instead, only certain properties can be partially confirmed. This limitation underscores the importance of model-comparison studies, which support the evaluation of model refinements by identifying and explaining discrepancies. In many cases, however, these studies lack actionable conclusions for model refinement (e.g., Aerts et al., 2022).

To address this limitation, model evaluations can incorporate approaches such as event modeling and multiple reference data sets. This extends the evaluation beyond considering a model's multidimensional state space based on a single flux, often streamflow. These methods enable a more comprehensive assessment of model fidelity (Beven, 2023; Clark et al., 2015; de Boer-Euser et al., 2017; Gupta et al., 2008; Vrugt and Sadegh, 2013). Such an assessment can provide valuable insights into how well the model represents underlying hydrological processes and can subsequently support the formulation and testing of hypotheses about alternative process and parameter representations (Clark, Kavetski, and Fenicia, 2011).

To test these hypotheses and evaluate the robustness of a model's process representation, researchers should consider a sufficiently large sample of catchments covering a wide range of hydroclimatic conditions and dominant hydrological processes (Andréassian et al., 2006; Gupta et al., 2014). Incorporating these considerations can help model-comparison studies produce more meaningful and reliable results for model refinement.

Researchers have begun to collect and use large-sample catchment data sets, such as CAMELS-US (Addor et al., 2017), that contain catchment boundaries, hydrometeorological time series, and catchment attributes, including topographic indices, land use, and soil properties. Studies use these data sets to investigate relationships among climatic attributes, catchment attributes, and hydrological model performance. For instance, Donnelly et al. (2016) assessed the suitability of the E-HYPE hydrological model in ungauged basins by examining relationships between catchment attributes and flow signatures. They found that flow signatures representing discharge regimes (low flow, long-term mean flow, and high flow) exhibited significant correlations (>50 %) with catchment attributes, including upstream area, percentage of agricultural land, elevation, mean precipitation, and mean temperature. By relating their findings to catchment attributes,

Massmann (2020) sought to determine the minimum required complexity of hydrological models. Aridity and baseflow emerged as the strongest negative predictors of model performance. The results indicated that arid regions require more complex modeling, in agreement with earlier research by Gudmundsson et al. (2012), Parajka et al. (2013), and van Esse et al. (2013).

More recently, McMillan et al. (2022) studied relationships between hydrological processes and signatures using multiple large-sample catchment data sets. They recommended signatures that they considered reliable for identifying hydrological processes. These findings complement the existing catchment attributes in large-sample data sets because they link to hydrological processes and, indirectly, to a model's ability to capture these processes.

This information is particularly valuable for distributed hydrological models because of their intricate spatiotemporal interactions and processes. This complexity makes it difficult to identify the causes of favorable or unfavorable model performance. For example, previous studies have noted that catchment responses and hydrological model performance cannot be attributed solely to a single catchment feature (Addor et al., 2018; Beck et al., 2016). Addressing this challenge therefore requires an exploratory approach that considers relationships extending beyond our current understanding of model behavior and beyond the assumptions underlying hydrological process descriptions (Gupta et al., 2008).

Understanding which catchment characteristics determine favorable or unfavorable model performance could help exploit differences among hydrological model structures and parameterizations to formulate new model-refinement hypotheses. However, limited research has explored such hypotheses based on how catchment characteristics influence hydrological model performance. In particular, few studies have comprehensively evaluated multiple states and fluxes to identify the key factors determining model fidelity.

In this chapter, we demonstrate how such a study can be conducted using data-driven techniques. These techniques can reveal nonlinear relationships between catchment attributes and hydrological model performance, providing insights that traditional evaluation approaches might not capture.

Random forest models (RFMs), a data-driven machine-learning technique introduced by Breiman (2001), have been applied in multiple studies combining large-sample catchment attributes with predictive modeling. For instance, Addor et al. (2018) combined catchment attributes and RFMs to examine the spatial predictability of hydrological signatures. Although this approach was promising, it produced results similar to those obtained through spatial autocorrelation. RFMs have also been successfully applied in other domains, such as predicting soil characteristics (Chaney et al., 2016; Hengl et al., 2017; Ramcharan et al., 2018) and evaluating model performance (Konapala et al., 2020). These studies highlight the potential value of RFMs as exploratory tools.

Using RFMs, we aim to answer several research questions focused on identifying potential avenues for model refinement through a novel model-comparison approach:

- How effective are RFMs in predicting hydrological model performance based on catchment attributes from a large-sample catchment data set?
- How does the resulting catchment attribute importance for prediction depend on the evaluated fluxes or states, and can this importance be linked to dominant hydrological processes?
- Can we leverage the importance of catchment attributes to identify avenues for model refinement?

To that end, we combine catchment attributes from the CAMELS-GB large-sample data set (Coxon et al., 2020b) and hydrological signatures from McMillan et al. (2022) with RFMs as an exploratory tool to identify model-refinement pathways through model comparison. We selected two hydrological models based on their opposing refinement strategies in relation to spatial scale. These models now meet on a 1 km × 1 km grid within the "hyper-resolution" domain (e.g., Beven and Cloke, 2012; Bierkens et al., 2015; Hoch et al., 2023; Wood, Roundy, Troy, van Beek, Bierkens, Blyth, Lettenmaier, et al., 2011).

4.2. METHODS AND DATA

The methodology builds on the hydrological models, model configurations, and model inputs previously used in Aerts et al. (2023). We broadened the scope of the investigation following the recommendation by Jain et al. (2022) to maximize the utility of existing model runs and increase the time available for novel research. In this study, we expanded the sample from 396 to all 617 catchments in the CAMELS-GB data set (Coxon et al., 2020b). We further extended the research by integrating actual evaporation flux and soil moisture state into the model-performance evaluation.

4.2.1. HYDROLOGICAL MODELS

DESCRIPTION OF CONCEPTUAL DIFFERENCES

Here, we focus on the main conceptual differences between the two physically based distributed hydrological models used in this study: wflow_sbm (van Verseveld et al., 2022) and PCR-GLOBWB (Hoch et al., 2023; Sutanudjaja et al., 2017). The wflow_sbm model concept was originally developed for shallow soils and steep headwater catchments based on the Topog_SBM model (Vertessy and Elsenbeer, 1999). PCR-GLOBWB was originally developed for large- to global-scale water-balance assessments and is now used at finer spatiotemporal scales. Both models can now be applied on a 1 km × 1 km grid at a daily temporal resolution.

The hydrological models were selected based on their complexity, conceptual similarities, and competing refinement strategies. Wflow_sbm is implemented at coarser resolutions than those for which it was originally developed, whereas PCR-GLOBWB is implemented at finer resolutions. The models differ in several respects, including their representations of groundwater, soil dynamics, lateral connectivity, and river routing:

- The PCR-GLOBWB model includes a separate groundwater store and assumes a linear relationship between storage and outflow (a linear reservoir). The wflow_sbm model contains unsaturated and saturated stores with a water-table approximation and accounts for exchange between the stores based on conductivity at the water table.

- The soil conceptualizations differ: PCR-GLOBWB uses two soil layers to simulate water storage, whereas wflow_sbm has three unsaturated soil layers. PCR-GLOBWB assumes no lateral connectivity within the groundwater or unsaturated zones, whereas wflow_sbm includes subsurface lateral connectivity through topographic-gradient-driven flow.

- The model setup uses the accumulated-travel-time method for river routing in PCR-GLOBWB and the kinematic-wave approximation in wflow_sbm. The kinematic-wave feature of PCR-GLOBWB is optimized to reduce runtimes on a 1 km × 1 km grid (Verkaik et al., 2022) but was not yet available at the start of this study.

Although the model-comparison approach should be compatible with any hydrological model, using models with similar properties facilitates more robust conclusions.

MODEL PARAMETERIZATION

Both hydrological models are used as intended by their developers. As model users, we retain the readily available parameters and conduct additional calibration only where prescribed by the model developers. PCR-GLOBWB therefore does not require additional calibration using streamflow observations, even though such calibration could improve model performance. Instead, best practice requires an extensive spin-up period before the model run. Wflow_sbm requires additional calibration of a single effective parameter (the horizontal conductivity fraction) based on streamflow observations. For more information and a demonstration of the effects of this calibration on streamflow simulations, see Aerts et al. (2022).

The parameterization of PCR-GLOBWB follows the method presented in Appendix A of Hoch et al. (2023). The wflow_sbm parameters are upscaled from the finest available data resolution to the hydrological model resolution. The flow-direction parameter is upscaled using the method presented by Eilander et al. (2021). Other parameters are upscaled based on the work of Imhoff et al. (2020), who used pedotransfer functions to derive parameters at the point scale before upscaling. An overview of the main parameter data sources is provided in Table 4.1.

4.2.2. HYDROLOGICAL MODEL RUNS

METEOROLOGICAL INPUT DATA

The meteorological data used to force the hydrological models are the same as those used to derive the catchment attributes in the CAMELS-GB data set (Coxon et al., 2020b). Consequently, differences in model results can be attributed to model processes and parameterization. The meteorological data sets have a daily temporal resolution and a spatial resolution of 1 km; they include temperature (CHESS-met; Robinson et al., 2020a),

Table 4.1: Overview of the main data sources used to derive the hydrological model parameter sets. Note the similarities between the two models in the topography, surface-water, and subsurface categories.

Model	Category	Datasets	References
wflow_sbm	Topography	MERIT Hydro DEM	Yamazaki et al. (2019)
	Surface water	GRaND, HydroLAKES	Lehner et al. (2011) and Messager et al. (2016)
	Meteorology	CHELSA, Köppen-Geiger	Karger et al. (2017) and Kottek et al. (2006)
	Land surface	VITO, MODIS LAI	Buchhorn et al. (2020) and Myneni et al. (2015)
	Subsurface	SoilGrids250m	Hengl et al. (2017)
PCR-GLOBWB	Topography	MERIT Hydro DEM	Yamazaki et al. (2019)
	Surface water	GRaND, Global Lakes and Wetlands Database	Lehner et al. (2011) and Messager et al. (2016)
	Groundwater	Global Hydrogeology MaPS	Gleeson et al. (2014)
	Land surface	Global Land Cover Characteristics v2.0, Global Food Security Support Analysis Data v.1.0, MIRCA2000, Global CropWater Model	Portmann et al. (2010), Siebert & Döll (2008), and Thenkabail et al. (2021)
	Subsurface	SoilGrids250m	Hengl et al. (2017)

precipitation (CEH-GEAR; Keller et al., 2015; Tanguy et al., 2022), and reference evaporation (CHESS-PE; Robinson et al., 2020b). Before use, these data sets are reprojected to align with the hydrological model grids using the ESMValTool climate-data postprocessor (Eyring et al., 2020).

MODEL RUNS

As with model parameterization, we conducted the model runs as intended by the model developers. PCR-GLOBWB was spun up through 30 consecutive repetitions of an average daily water year derived from the period from 1 October 2001 to 30 September 2007. The year 2008 was discarded before model performance was evaluated from 2009 to 2017. Wflow_sbm was calibrated from 2001 to 2007. The year 2008 was discarded, and the model was subsequently evaluated from 2009 to 2017. The model runs were conducted using the community-driven eWaterCycle platform for hydrological model experimentation (Hut et al., 2023).

4.2.3. HYDROLOGICAL MODEL PERFORMANCE

Evaluating hydrological model performance requires reference data. Here, we evaluate three output variables: discharge, actual evaporation, and soil moisture. We evaluate simulated discharge using the UK National River Flow Archive, which is included in the CAMELS-GB data set. The data set provides daily time series of streamflow measurements at the catchment outlet, expressed in cubic meters per second ($m^3\,s^{-1}$).

For actual evaporation, we used the FLUXCOM data set (Jung et al., 2019), which provides 8-day average values at a spatial resolution of 0.0833 degrees. Before use, the latent heat flux values were divided by the latent heat of vaporization to derive actual evaporation in millimeters per day ($mm\,d^{-1}$) (Allan et al., 1998).

As a soil-moisture reference, we selected the merged data set developed by Peng et al. (2021). They combined the ASCAT (H SAF 2020; Wagner et al., 2013) and SMAP (Entekhabi et al., 2010) remote-sensing products with output from the JULES-CHESS land-surface model (Best et al., 2011; Clark, Mercado, et al., 2011) using a triple-collocation merging technique (Stoffelen, 1998). The resulting weighted data set provides daily gridded estimates of volumetric water content at a spatial resolution of 1 km, expressed in cubic meters per cubic meter ($m^3\,m^{-3}$). This data set enables direct evaluation of the simulated soil-moisture states of the hydrological models.

OBJECTIVE FUNCTIONS

Model performance is assessed using objective functions that quantify the quality of the simulations over the evaluation period. Streamflow-based model performance is determined using the nonparametric Kling-Gupta efficiency (KGE-NP; Pool et al., 2018), which ranges from minus infinity to a perfect score of 1. Model performance for actual evaporation and soil moisture is measured using the relative root mean square error (RRMSE), calculated by averaging each time step across each catchment. Unlike conventional RMSE, this RRMSE variant is normalized by the standard deviation of the observations, ensuring comparability among catchments. A score of 0 denotes a perfect match between observations and model simulations (e.g., Hoch et al., 2023).

The evaluation periods for states and fluxes vary according to data availability. Streamflow is assessed using daily averages for the water years 2009–2015, while actual evaporation is averaged over 8-day periods for the same years and soil moisture is evaluated daily from 2015 to 2017. The objective functions are first determined for each water year and then averaged over the entire evaluation period.

4.2.4. RANDOM FOREST MODELS

Random forest models (RFMs; Breiman, 2001) are used as the primary method for evaluating relationships between catchment attributes and hydrological model performance. We use RFMs to estimate the importance of catchment attributes as features for predicting model performance. An RFM is an ensemble regression or classification approach that uses regression trees fitted to random subsets of features. Its output is the unweighted average of the individual tree outputs:

$$\bar{h}(x) = \frac{1}{K}\sum_{k=1}^{K} h\left(x, \theta_k\right),$$

where $\bar{h}$ is the average output of the regression trees h with input features x, feature set θ, and feature subset $\theta_k \subset \theta$.

Unlike more parsimonious methods, such as (multiple) linear regression, principal component analysis (e.g., Jolliffe and Cadima, 2016), and LASSO models (e.g., Ranstam and Cook, 2018), RFMs are less affected by strong collinearity among catchment attributes. An overview of the RFM setup is provided in Figure 4.1.

CATCHMENT ATTRIBUTES

The CAMELS-GB data set contains catchment and climate attributes represented by a single average value per catchment. These attributes cover meteorology, topography,

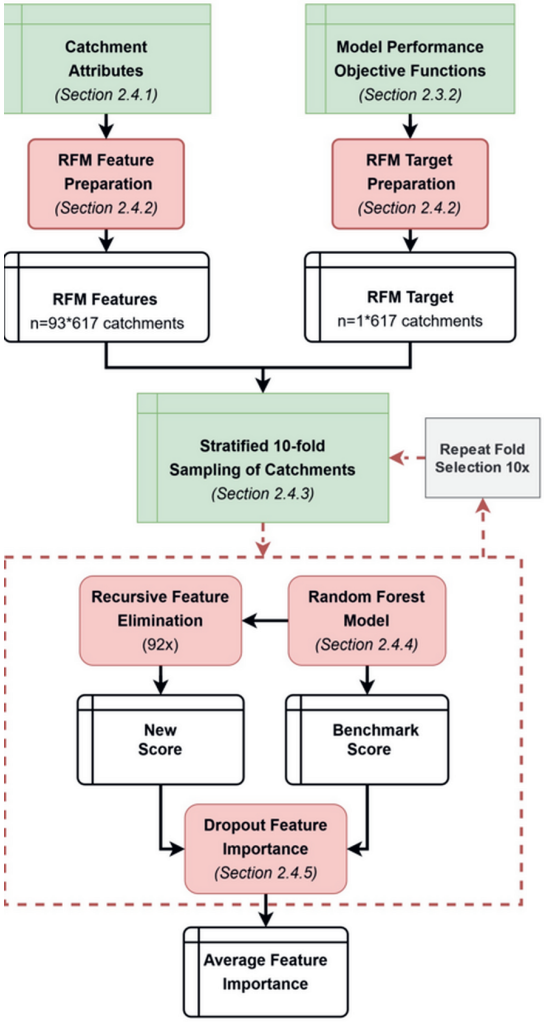


Figure 4.1: Flowchart of the random-forest-model (RFM)-based feature-importance method. (1) A set of model-performance results is selected as the RFM target. (2) Catchment attributes from the CAMELS-GB data set and McMillan et al. (2022) are extracted. (3) The catchment attributes are prepared as RFM features, and the model-performance results as RFM targets. (4) The features and targets for 617 catchments are divided into 10 folds. (5) The RFM is trained using 9 of the 10 folds (approximately 90 % of the catchments) and tested on the remaining fold (approximately 10 %). (6) The benchmark score is calculated using all RFM features. (7) One feature is removed from the complete feature set (93 features), and a new test score is calculated. (8) Dropout feature importance is calculated as the difference between the benchmark score and the new test score. (9) The removed feature is returned to the complete feature set and another feature is removed; this process is repeated until each feature has been removed once. (10) The feature-importance analysis is repeated for all 10 train-test fold combinations before the mean feature importance of each catchment attribute is calculated.

land cover, soil, hydrogeology, hydrometry, human influence, and hydrological signatures. We recreated the hydrological signatures for the study period (2008–2015) using the TOSSH toolbox (Gnann et al., 2021) and extended the feature set with the signatures recommended by McMillan et al. (2022). Together, the catchment characteristics and hydrological signatures are referred to as catchment attributes. Supplementary Table 4.5 provides an overview of the 93 selected attributes.

FEATURE AND TARGET PREPARATION

As with most statistical inference methods, an RFM optimizes its internal parameters to predict the target (dependent variable) from the features (independent variables). The RFMs used here are regressors and can therefore use continuous rather than categorical targets. In this study, the targets are the hydrological model-performance results calculated with the objective functions, and the features are the catchment attributes. To further investigate the relationship between catchment attributes and model performance, we also use the performance difference between the two hydrological models as an RFM target. This allows us to identify the catchment attributes that best explain discrepancies, or the lack thereof, between the model results. In total, this setup yields nine RFMs: one for each combination of target variable (streamflow, actual evaporation, or soil moisture) and model-performance result (wflow_sbm, PCR-GLOBWB, or the difference between the two models).

Features with more than 5 % missing data across all catchments are excluded. Missing data must be removed because RFMs might otherwise base feature importance on the presence or absence of data. In addition, the KGE-NP-based streamflow-performance target is balanced by setting extremely low values to -0.42. This value approximately corresponds to using the mean of the streamflow observations as a benchmark (Knoben, Freer, and Woods, 2019). The RRMSE results for actual evaporation- and soil-moisture-based model performance remain unchanged.

STRATIFIED 10-FOLD SAMPLING

Feature importance is evaluated using stratified 10-fold sampling of the features and target (e.g., Addor et al., 2018). Stratification aims to maintain the target-variable distribution when samples are randomly assigned to each fold. The RFM predictions are then evaluated through stratified 10-fold cross-validation, using 90 % of the catchments for training and 10 % for testing. This procedure is repeated for all 10 folds to cover the complete data set (Figure 4.2). It avoids dependence on a single division of the data set into training and test sets.

TRAINING AND TESTING

Like hydrological models, RFMs require extensive parameter calibration, referred to as a "hyperparameter search." This search optimizes several parameters, including the number of estimators (trees in the forest) and maximum tree depth. We conducted the search using the tree-structured Parzen estimator (TPE) method (Bergstra et al., 2011, 2013). This Bayesian optimization approach uses previous training results to construct a probabilistic model of the hyperparameters and estimate the probability associated with the



Figure 4.2: Schematic overview of the 10-fold cross-validation method. The test-set fraction (10 %) is shown in green and the training-set fraction (90 %) in blue. The process is repeated 10 times by iteratively selecting a new combination of training and test catchments.

loss function. RFM calibration was completed by selecting the parameter set that minimized the representative loss function—in this study, the mean absolute error—across all 10-fold combinations.

An important part of configuring RFMs is addressing overfitting by "pruning the trees," meaning reducing the width and depth of the decision trees. An overfit model performs well on the training data (the 90 % mentioned above) but generalizes poorly to independent test data (the remaining 10 %). When the hyperparameter search identifies multiple near-optimal solutions, we select the parameters that minimize tree width and depth, thereby reducing the risk of overfitting. Supplementary Table 4.6 presents the RFM target variables, benchmark scores, and standard deviations resulting from stratified 10-fold training and testing. The aim of this study is to determine feature importance rather than develop the best-predicting RFM. Benchmark scores can therefore be low because we use all catchment attributes as features, including those that reduce prediction accuracy.

DROPOUT FEATURE IMPORTANCE

Dropout feature importance—here, catchment-attribute importance—describes the contribution of each feature to predicting variation in the target (the hydrological model-performance results). First, the benchmark score is calculated using all features and the test data (10 % of the catchments). Second, one feature is removed from the complete feature set. Third, a new score is calculated using the reduced feature set. The difference between the benchmark and new scores is the dropout feature importance (e.g., Konapala et al., 2020). The removed feature is then returned, another feature is removed, and a new dropout score is calculated. This process continues until every feature has been removed once. Finally, the procedure is repeated for all 10 train-test fold combinations, after which the mean dropout feature importance is calculated.

We aggregated the dropout feature-importance results into the distinct classes used in the CAMELS-GB data set: meteorology, land surface, subsurface, and hydrology. Catchment attributes with negative feature importance were excluded, and the results were expressed as the fraction of positive feature importance. Supplementary Table 4.5 presents the catchment attributes and their respective classes.

4.3. RESULTS

This section presents our findings for three variables: streamflow, actual evaporation, and soil moisture. First, we present the hydrological model-performance results, followed by the positive fractions of dropout feature importance. Finally, we highlight the three highest-ranking catchment attributes based on their feature importance and relate them to model performance. Supplementary Figures 4.9–4.11 provide correlation plots of catchment attributes and model performance for these attributes.

4.3.1. STREAMFLOW

STREAMFLOW-BASED HYDROLOGICAL MODEL PERFORMANCE

Figure 4.3 shows the streamflow-based hydrological model-performance values. These results serve as the target variables for the subsequent RFM analysis.

Figure 4.3a shows a median KGE-NP value of 0.73 for wflow_sbm. Model performance is higher in the north and west and lower in the southeast. The PCR-GLOBWB results in Figure 4.3b have a median KGE-NP value of 0.42 and similarly show lower performance in the southeast. PCR-GLOBWB generally has slightly higher KGE-NP values than wflow_sbm in this region, as shown in Figure 4.3c. Based on the difference in KGE-NP between the models, wflow_sbm performs better than PCR-GLOBWB in 74 % of the catchments. This performance difference occurs mainly in the western and northern regions and is partly attributable to the additional streamflow-based calibration and the kinematic-wave approximation used for river routing at a daily time step in wflow_sbm (Section 2.2.1).

Both models show a distinct band of poor performance in the south, although the pattern is more pronounced for wflow_sbm. The karst geology of the southern region could contribute to this poor performance because it presents modeling challenges. This effect is particularly apparent when these results are compared with the better performance in catchments farther north that are dominated by Jurassic limestone.

FEATURE IMPORTANCE FOR PREDICTING STREAMFLOW-BASED MODEL PERFORMANCE

Next, we consider the RFM-based feature-importance results. These results are aggregated into feature classes and expressed as positive feature-importance fractions in Figure 4.4.

The results in Figure 4.4 show that the meteorology class has low overall feature importance and therefore plays a negligible role in predicting streamflow-based model performance for all three targets. The narrow boxplots show minimal variance, demonstrating that this finding is spatially consistent throughout the case-study area. The remaining hydrology, land-surface, and subsurface feature classes for wflow_sbm are well balanced, with similar mean feature-importance values of approximately 0.36.

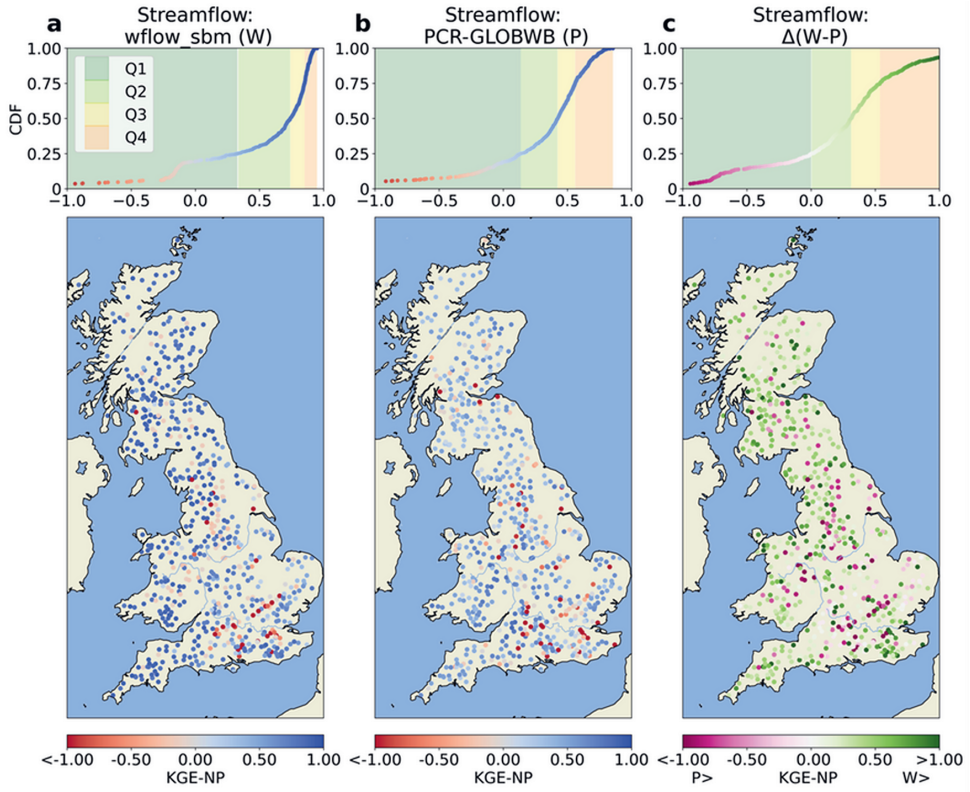


Figure 4.3: Comparison of streamflow-based model performance using the daily KGE-NP metric at the catchment outlet. The top panels show the cumulative distribution functions and distribution quantiles (Q1–Q4), and the bottom panels show the spatial distributions of model performance. (a) Wflow_sbm performance, with low KGE-NP values in red and high values in blue. (b) PCR-GLOBWB performance. (c) Performance difference between the two models, Δ(W-P), with better PCR-GLOBWB performance in purple and better wflow_sbm performance in green.

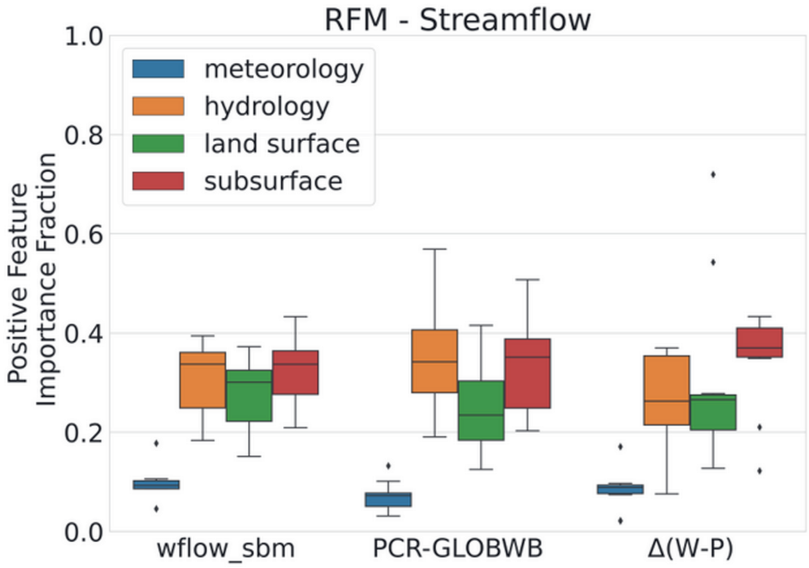


Figure 4.4: Boxplots of the positive feature-importance fraction for each feature class, using catchment attributes as RFM features and streamflow-based model-performance results as RFM targets. The targets are wflow_sbm (W), PCR-GLOBWB (P), and the difference between the two models, Δ(W-P). Results are grouped into four attribute classes: meteorology (blue), hydrology (orange), land surface (green), and subsurface (red). Each point in a boxplot represents one test fold (Figure 4.2); the variance indicates the degree of spatial variation in the results.

As in the wflow_sbm results, the remaining classes are balanced for PCR-GLOBWB. However, these classes have greater variance, indicating more spatial variation in the results. The land-surface class also has a relatively low mean feature-importance fraction of 0.25. For the model-performance difference target, Δ(W-P), the subsurface class has the highest positive feature-importance fraction. The magnitude of the discrepancy in streamflow-based model performance is therefore best predicted by subsurface attributes, making these attributes relevant candidates for further investigation and model refinement.

The feature importance of individual catchment attributes is shown in Table 4.2. The lower (Q1) and upper (Q4) quantiles of the model-performance distributions distinguish poor from good performance for the individual models (Figure 4.3). For the Δ(W-P) target, the lower quantile indicates better PCR-GLOBWB performance (purple), whereas the upper quantile indicates better wflow_sbm performance (green).

The RFM with the wflow_sbm target yields a benchmark score of 0.81 R^2. This score is considered high because the model includes catchment attributes with negative feature importance that reduce its predictive skill. Catchment area is the highest-ranking individual attribute; excluding it decreases the score by 0.13 R^2. The mean area of all 617 catchments is 417 km^2. The lower performance quantile for wflow_sbm has a mean area of 968 km^2, whereas the upper quantile has a mean of 278 km^2. Thus, wflow_sbm performs better, on average, in smaller catchments than in larger ones. The two lower-ranking catchment attributes each have a feature importance of 0.02 R^2. The RFM predicts better wflow_sbm performance in catchments with a relatively high runoff ratio and lower-than-average aridity.

The PCR-GLOBWB results show a more modest benchmark score of 0.46 R^2. Again, catchment area is the highest-ranking attribute. In contrast to the wflow_sbm findings, PCR-GLOBWB is predicted to perform better in larger catchments (904 km^2) than in smaller ones (179 km^2). The second-ranking attribute, the baseflow recession coefficient K, shows only minor deviations from the overall mean across the quantiles. Notably, catchments with shallower average root-zone depths, as indicated by lower total available water content, tend to show better PCR-GLOBWB streamflow-based performance.

For the model-performance difference target, Δ(W-P), the benchmark score is close to that of PCR-GLOBWB. Catchment area remains the highest-ranking attribute and confirms the previous finding of better wflow_sbm performance in smaller catchments (40 km^2) and better PCR-GLOBWB performance in larger catchments (1,292 km^2). Porosity is the only subsurface attribute among the three highest-ranking attributes. Thus, although the subsurface class is the best predictor (Figure 4.4), its individual attributes have positive but modest feature importance.

4.3.2. ACTUAL EVAPORATION

ACTUAL EVAPORATION-BASED HYDROLOGICAL MODEL PERFORMANCE

Hydrological model performance in simulating actual evaporation is assessed using the 8-day-average relative root mean square error (RRMSE) shown in Figure 4.5. The CDF of the wflow_sbm results in Figure 4.5a reveals two distinct performance clusters: most catchment simulations (80 %) represent actual evaporation well, with RRMSE values below 1 (Q1–Q3), whereas the remaining 20 % represent it poorly, with RRMSE values above

Table 4.2: Overview of the streamflow-based model-performance RFM targets. The RFM benchmark score (using all features) and the three highest-ranking catchment attributes are shown with their feature-importance scores and mean values. For wflow_sbm and PCR-GLOBWB, Q1 represents the 25 % of catchment simulations with the lowest performance, and Q4 represents the 25 % with the highest performance (Figure 4.3). For Δ(W-P), Q1 represents the 25 % of catchment simulations with better PCR-GLOBWB performance, and Q4 represents the 25 % with better wflow_sbm performance.

RFM Target	RFM Score R^2	Rank	Feature R^2	Catchment Attribute	Unit	Mean	Std. Dev.	Q1 Mean	Q4 Mean
wflow_sbm	0.81	1	0.13	Catchment Area	km^2	417	967	968	278
		2	0.02	Runoff Ratio	–	0.56	0.23	0.46	0.70
		3	0.02	Aridity	–	0.52	0.21	0.59	0.37
PCR-GLOBWB	0.46	1	0.06	Catchment Area	km^2	417	967	179	904
		2	0.02	Baseflow Recession K	–	0.13	0.07	0.14	0.10
		3	0.01	Total Available Water Content	mm	95	25	95	86
Δ(W – P)	0.45	1	0.17	Catchment Area	km^2	417	967	1292	40
		2	0.02	Porosity	–	48.0	6.2	49.0	46.7
		3	0.02	Low-flow Event Duration	d	7	8	6	8

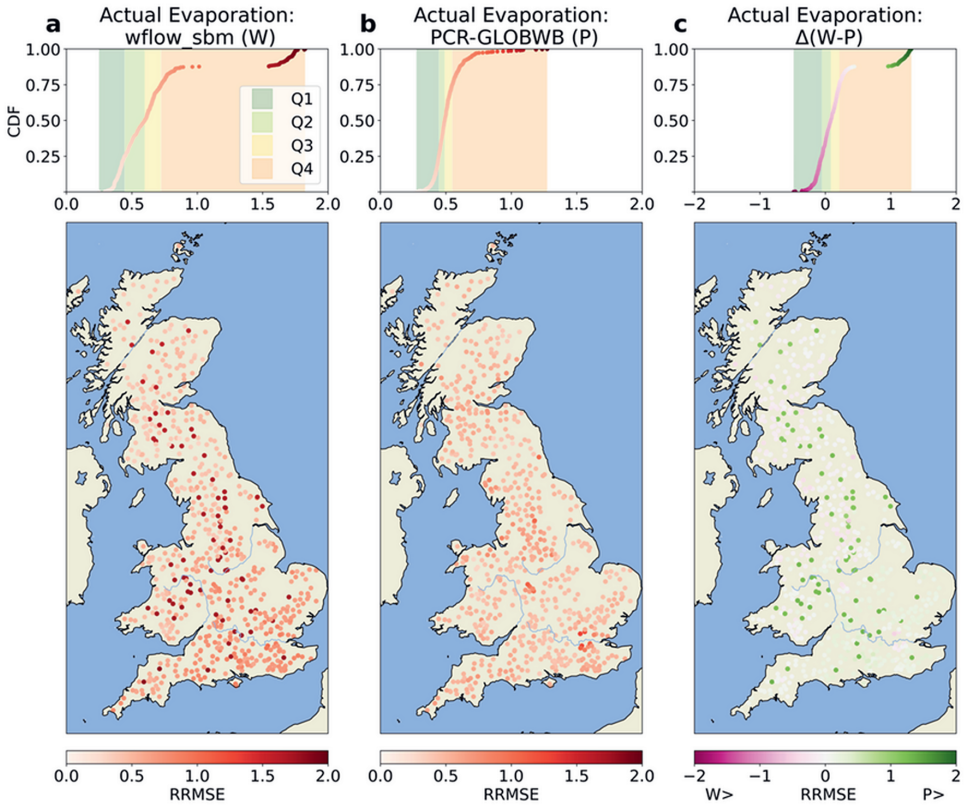


Figure 4.5: Comparison of catchment-averaged, actual evaporation-based model performance using the 8-day-average RRMSE metric. The top panels show the cumulative distribution functions and distribution quantiles (Q1–Q4), and the bottom panels show the spatial distributions of model performance. (a) Wflow_sbm performance, with low RRMSE values in white and high values in red. (b) PCR-GLOBWB performance. (c) Performance difference between the two models, Δ(W-P), with better wflow_sbm performance in purple and better PCR-GLOBWB performance in green. Note that this color order is the reverse of that in Figure 4.3.

1.5 (Q4). The PCR-GLOBWB results in Figure 4.5b are more consistent and show better overall performance, with only a small proportion of catchments having an RRMSE above 1. The Δ(W-P) results in Figure 4.5c emphasize this difference, with better PCR-GLOBWB performance shown in green. Notably, the poorer representation of actual evaporation by wflow_sbm does not necessarily coincide with poor streamflow-based performance (Figure 4.3a).

FEATURE IMPORTANCE FOR PREDICTING ACTUAL EVAPORATION-BASED MODEL PERFORMANCE

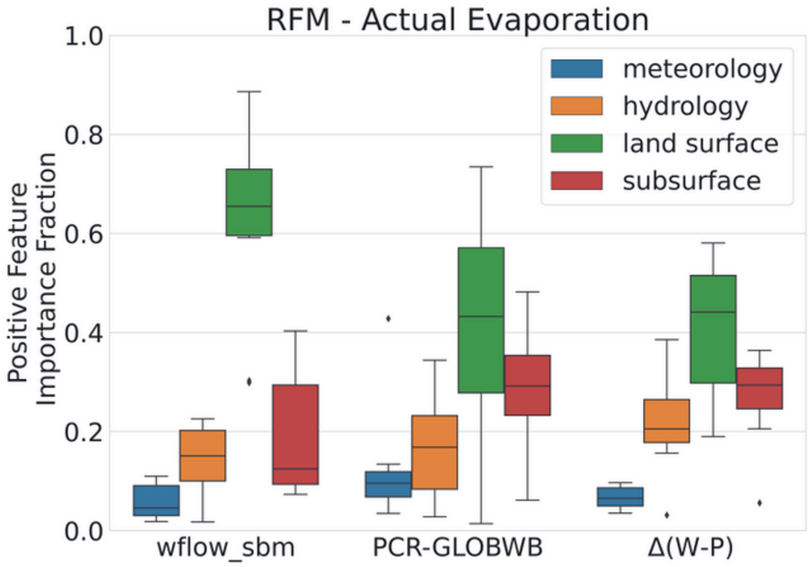


Figure 4.6: Boxplots of the positive feature-importance fraction for each feature class, using catchment attributes as RFM features and actual evaporation-based model-performance results as RFM targets. The targets are wflow_sbm (W), PCR-GLOBWB (P), and the performance difference, Δ(W-P). Results are grouped into four classes: meteorology (blue), hydrology (orange), land surface (green), and subsurface (red). Each point in a boxplot represents one test fold (Figure 4.2); the variance indicates the degree of spatial variation in the results.

The RFM results aggregated by attribute class are presented in Figure 4.6. The land-surface class consistently has the highest fraction of feature importance across all RFMs, particularly RFM-wflow_sbm. The high feature importance of land-surface attributes is consistent with their expected role in evaporation, especially for features associated with land cover. Meteorological features related to the timing and volume of precipitation, the magnitude of potential evapotranspiration, and indices such as aridity might also be expected to be important. However, their limited feature importance may suggest that both hydrological models are similarly proficient in simulating actual evaporation across diverse meteorological conditions.

As in the streamflow-based results, catchment area has the highest feature importance for RFM-wflow_sbm, with a score of 0.31 R^2 relative to a benchmark of 0.57 R^2. For

Table 4.3: Overview of the actual evaporation-based model-performance RFM targets. The RFM benchmark score (using all features) and the three highest-ranking catchment attributes are shown with their feature-importance scores and mean values. For wflow_sbm and PCR-GLOBWB, Q1 represents the 25 % of catchment simulations with the highest performance, and Q4 represents the 25 % with the lowest performance (Figure 4.5). For Δ(W-P), Q1 represents the 25 % of simulations with better wflow_sbm performance, and Q4 represents the 25 % with better PCR-GLOBWB performance. Note that the quantile order is the reverse of that in Table 4.2.

RFM Target	RFM Score R^2	Rank	Feature R^2	Catchment Attribute	Unit	Mean	Std. Dev.	Q1 Mean	Q4 Mean
wflow_sbm	0.57	1	0.31	Catchment area	km^2	417	967	230	987
		2	0.01	Mean PET	$mm\ d^{-1}$	1.34	0.11	1.22	1.38
		3	0.01	Shrubs Area	%	12	19	26	5
PCR-GLOBWB	0.52	1	0.11	Urban area	%	7.63	12.93	3.10	17.21
		2	0.01	Mean PET	$mm\ d^{-1}$	1.34	0.11	1.33	1.35
		3	0.01	Organic Content	%	4	5	5	3
Δ(W − P)	0.58	1	0.33	Catchment area	km^2	417	967	214	1006
		2	0.02	Bare soil area	%	0.6	2.0	0.9	0.4
		3	0.02	Frequency of high precipitation	$d\ yr^{-1}$	15.6	2.84	13.51	16.56

the RRMSE metric, the best-performing catchments belong to the lower quantile (Q1) and are characterized by smaller areas. Other identified attributes, such as mean potential evapotranspiration (PET) and the percentage of shrub cover, have relatively low feature importance. The results indicate that the best-performing catchments tend to have above-average shrub cover and lower mean PET.

In contrast to the streamflow-based results, catchment area is not among the three highest-ranking features for PCR-GLOBWB. Instead, the percentage of urban area is the highest-ranking feature, with a score of 0.11 R^2 relative to a benchmark of 0.52 R^2. Catchments with less urban area tend to outperform those with more extensive urban development.

The Δ(W-P) results show that catchment area remains the most important attribute for predicting discrepancies in actual evaporation-based performance between the models. The lower quantile (Q1) indicates better wflow_sbm performance in smaller catchments, whereas PCR-GLOBWB performs better in larger ones. Notably, catchment area consistently has the greatest explanatory importance in both the streamflow and actual evaporation evaluations.

4.3.3. SOIL MOISTURE

SOIL MOISTURE-BASED HYDROLOGICAL MODEL PERFORMANCE

In Figure 4.7, the soil moisture-based performance of wflow_sbm is similar to that of PCR-GLOBWB, with median RRMSE values of 2.76 and 2.92, respectively. The wflow_sbm (Figure 4.7a) and PCR-GLOBWB (Figure 4.7b) results show similar spatial patterns. These patterns distinguish areas of satisfactory and deficient soil-moisture representation relative to the reference product: performance is good in the southeast and poorer in the west and north of the case-study area. The spatial comparison in Figure 4.7c shows that the largest differences between the models occur in regions with generally poor model performance.

FEATURE IMPORTANCE FOR PREDICTING SOIL MOISTURE-BASED MODEL PERFORMANCE

The soil moisture-based RFM results in Figure 4.8 show that the subsurface class has the highest feature importance for RFM-wflow_sbm, followed by hydrology and land surface. The PCR-GLOBWB results do not show this hierarchy: median values are more similar among classes, but their large variance indicates spatial variation. For Δ(W-P), only the meteorology class has a lower median value (0.11), compared with positive feature-importance fractions of approximately 0.28 for the other classes.

The results in Table 4.4 show good benchmark scores for both RFM-wflow_sbm (0.68 R^2) and RFM-PCR-GLOBWB (0.75 R^2). Compared with the streamflow and actual evaporation findings, the highest-ranking feature-importance scores are modest to low, indicating a more even distribution of feature importance. For wflow_sbm, urban area is the most important catchment attribute, with a score of 0.02 R^2; the quantile means are close to the mean across all catchments. The PCR-GLOBWB results indicate greater sensitivity to climatic indices: the model tends to perform better in catchments characterized by extended periods of limited precipitation (frequency of dry days, Q4). The saturation-excess-effect attribute reinforces this observation, indicating that catchments with higher

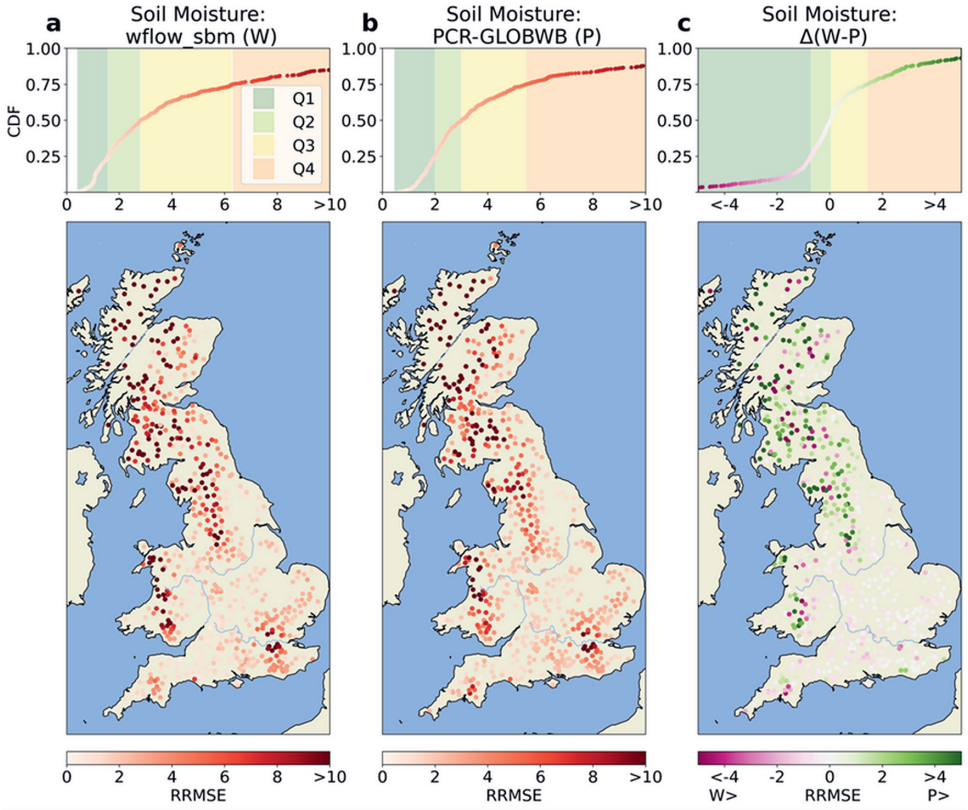


Figure 4.7: Comparison of catchment-averaged, soil moisture-based model performance using the daily-average RRMSE metric. The top panels show the cumulative distribution functions and distribution quantiles (Q1–Q4), and the bottom panels show the spatial distributions of model performance. (a) Wflow_sbm performance, with low RRMSE values in white and high values in red. (b) PCR-GLOBWB performance. (c) Performance difference between the two models, Δ(W-P), with better wflow_sbm performance in purple and better PCR-GLOBWB performance in green. Note that this color order is the reverse of that in Figure 4.3.

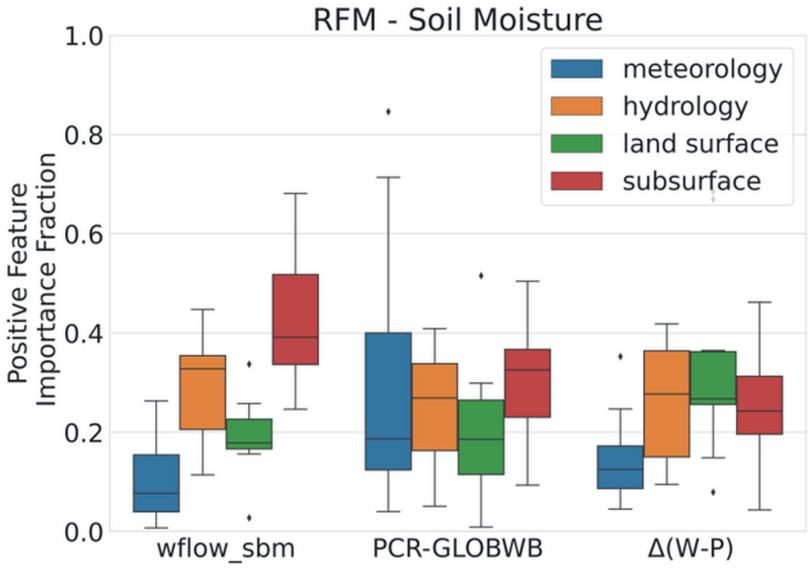


Figure 4.8: Boxplots of the positive feature-importance fraction for each feature class, using catchment attributes as RFM features and soil moisture-based model-performance results as RFM targets. The targets are wflow_sbm (W), PCR-GLOBWB (P), and the performance difference, Δ(W-P). Results are grouped into four classes: meteorology (blue), hydrology (orange), land surface (green), and subsurface (red). Each point in a boxplot represents one test fold (Figure 4.2); the variance indicates the degree of spatial variation in the results.

Table 4.4: Overview of the soil moisture-based model-performance RFM targets. The RFM benchmark score (using all features) and the three highest-ranking catchment attributes are shown with their feature-importance scores and mean values. For wflow_sbm and PCR-GLOBWB, Q1 represents the 25 % of catchment simulations with the highest performance, and Q4 represents the 25 % with the lowest performance (Figure 4.7). For Δ(W-P), Q1 represents the 25 % of simulations with better wflow_sbm performance, and Q4 represents the 25 % with better PCR-GLOBWB performance. Note that the quantile order is the reverse of that in Table 4.2.

RFM Target	RFM Score R^2	Rank	Feature R^2	Catchment Attribute	Unit	Mean	Std. Dev.	Q1 Mean	Q4 Mean
wflow_sbm	0.68	1	0.02	Urban area	%	8	13	6	5
		2	0.01	Recession parameters	–	0.23	0.62	0.14	0.24
		3	0.01	Porosity	–	48.0	6.2	45.6	53.3
PCR-GLOBWB	0.75	1	0.01	Frequency of dry days	$d\ yr^{-1}$	211	31	229	179
		2	0.008	Crop area	%	23	25	36	2
		3	0.003	Saturation excess effect	–	0.71	0.18	0.8	0.6
Δ(W – P)	0.35	1	0.06	Catchment area	km^2	417	968	908	243
		2	0.04	Mean drainage path slope	$m\ km^{-1}$	95	65	107	141
		3	0.03	Mean elevation	m.a.s.l.	200	126	228	296

saturation-excess thresholds generally perform better (Q1). In the Δ(W-P) analysis, catchment area is again the best predictor of the performance difference, with a value of 0.06 R^2 relative to a low benchmark of 0.35 R^2. In contrast to the previous findings, however, wflow_sbm performs better in larger catchments (Q1), whereas PCR-GLOBWB performs better in smaller catchments (Q4). Topographic indices are the leading predictors of the performance difference between the two models.

4.4. DISCUSSION

The objective of this study was to determine the effectiveness of random forest models (RFMs) in predicting hydrological model performance from catchment attributes, examine how attribute importance varies among model fluxes and states, and identify potential avenues for model refinement.

4.4.1. THE EFFECTIVENESS OF RANDOM FOREST MODELS

We first consider the effectiveness of RFMs in predicting model-performance targets from the attributes in a large-sample catchment data set. The RFM benchmark scores reveal substantial variation in predictive skill among model-performance targets (Tables 4.2, 4.3, and 4.4). Benchmark scores range from 0.57 to 0.81 R^2 for the wflow_sbm targets, from 0.46 to 0.75 R^2 for the PCR-GLOBWB targets, and from 0.35 to 0.58 R^2 for Δ(W-P). Supplementary Table 4.6 presents the variance resulting from the selection of train-test fold combinations. This variance indicates sensitivity to fold selection (subsamples) and spatial variation in RFM predictive skill. Given these nuances, we emphasize the value of using the complete data set across the training and testing folds.

To assess whether the individual folds differed significantly from the complete sample, we used the Kolmogorov-Smirnov test (Kolmogorov, 1933; Massey, 1951; Smirnov, 1933). The test found no significant differences between any fold and the complete sample at a significance level of 0.10. This result suggests that the sample size is sufficient for stratified 10-fold sampling, although a larger sample would allow assessment at a stricter significance level of 0.05. The large variance in RFM benchmark scores among folds provides another indication that the sample size is limited.

This study focuses on determining the importance of catchment attributes for predicting model performance rather than developing the best possible predictive model. The RFMs used here are effective for this purpose. If the goal were to obtain an RFM with the highest possible predictive capacity—a difficult task (Ley et al., 2015; Massmann, 2020; van Esse et al., 2013)—we would recommend applying a feature-selection method beforehand (e.g., Velthoen et al., 2022). Such a method would reduce the feature set to catchment attributes with a strong net positive effect on RFM predictive skill, thereby substantially improving predictive performance.

4.4.2. THE ADDED VALUE OF EXPLORATORY DIAGNOSTICS

Model performance is typically quantified using measures such as those shown in Figures 4.3, 4.5, and 4.7, but identifying the underlying causes of performance remains

challenging. RFMs are useful for this purpose compared with, for example, the correlation plots of catchment attributes and model-performance values provided in Supplementary Figures 4.9–4.11. These plots show the nonlinear distributions of catchment-attribute values across the catchments. RFMs add value by sorting these values through decision trees, enabling the evaluation of nonlinear relationships between catchment attributes and model performance.

Introducing an abstraction level through catchment-attribute classes helps evaluate how RFMs relate hydrological processes to the fluxes and states simulated by hydrological models. Figures 4.4 and 4.6 show that the meteorology class has a low feature-importance fraction for streamflow and actual evaporation. The wflow_sbm results in Figure 4.8 show a similar pattern for soil moisture. These findings contrast with previous research demonstrating the sensitivity of hydrological models to meteorological inputs (e.g., Gebrechorkos et al., 2023; Poncelet et al., 2017; Raimonet et al., 2017; Tarek et al., 2020). Our findings indicate that the two models have comparable proficiency in simulating streamflow and actual evaporation under varying meteorological conditions.

The streamflow-based results in Figure 4.4 show that the difference in model performance is best explained by the subsurface attribute class. Because both models use the SoilGrids data set for soil parameterization, an in-depth analysis of their subsurface structures and parameterizations might provide important insights for model refinement. Similar conclusions can be drawn from the influence of the land-surface class on actual evaporation performance (Figure 4.6) and the role of the subsurface class in the wflow_sbm soil-moisture results (Figure 4.8). Together, these findings highlight the nuanced interplay of specific catchment attributes in determining hydrological model results for a given flux or state.

THE INTERPLAY BETWEEN CATCHMENT ATTRIBUTES AND MODEL PERFORMANCE

In this study, we introduce a method for identifying conditions under which catchment-attribute values are associated with either good or poor model performance. We examine the individual catchment attributes in greater detail and relate our findings to previously established relationships between catchment attributes and model performance. One recognized relationship is the negative correlation between aridity and streamflow-based model performance documented in several studies (e.g., Gudmundsson et al., 2012; Massmann, 2020; Poncelet et al., 2017). Our results in Table 4.2 do not show this negative correlation across all hydrological model targets. A pattern of better performance in catchments with below-average aridity emerged only for the wflow_sbm target. Runoff ratio and small catchment area are more important RFM predictors of wflow_sbm performance.

Catchment area is particularly important for predicting all three variables (Tables 4.2, 4.3, and 4.4). These findings differ from previous studies: some found no discernible relationship between catchment area and model performance (Castellarin et al., 2013), whereas others observed better performance in larger catchments (Poncelet et al., 2017; van Esse et al., 2013). Massmann (2020) observed both positive and negative correlations and attributed this variation to aridity, which often coincided with larger catchments. In contrast, our data indicate better wflow_sbm performance in smaller catchments and better PCR-GLOBWB performance in larger catchments for both streamflow

and actual evaporation. The opposite relationship occurs for soil moisture: wflow_sbm performs better in larger catchments and PCR-GLOBWB in smaller catchments. These findings underscore that the relationship between model performance and catchment area is model-specific and depends on the predicted target.

The dependence of model performance on catchment area indicates that model-development history and refinement might play a role. Specifically, the originally coarser-scale PCR-GLOBWB model performs better in larger catchments, whereas the originally finer-scale wflow_sbm model performs better in smaller catchments. Reviewers of Aerts et al. (2022) hypothesized that a daily time step is too coarse for accurate simulations in small catchments. The wflow_sbm results provide no clear evidence for this hypothesis when the kinematic-wave approximation is used for river routing with subdaily calculations. PCR-GLOBWB does perform poorly in smaller catchments, which might be caused by its accumulated-travel-time routing approximation. This hypothesis could be tested by applying a setup similar to that of Cortés-Salazar et al. (2023) to PCR-GLOBWB and evaluating different routing schemes in relation to catchment area.

Alternatively, both models could be evaluated at multiple spatiotemporal resolutions to determine the effect of scale on model performance (e.g., Aerts et al., 2022; Melsen et al., 2016). RFMs using catchment attributes calculated at multiple spatial model scales could support this effort by identifying whether feature importance remains consistent across scales.

LIMITATIONS AND FUTURE RESEARCH

Despite the precautions taken in this study, the sample is limited in terms of both catchments and catchment attributes, as demonstrated by its sensitivity to fold selection. Small sample sizes can adversely affect random-forest prediction performance, although the required sample size depends on the characteristics of the data and prediction problem (Han et al., 2021; Luan et al., 2020). Our recommendation follows the well-known premise in data-driven research that "more high-quality data are better." For future research, we therefore advocate using the Caravan data set (Kratzert et al., 2022), which combines the available large-sample CAMELS data sets, to increase the sample size.

Further recommendations include expanding the data-set sample size to support more robust conclusions and exploring the potential of RFMs as model-selection tools. Such tools could use globally available catchment attributes to facilitate the creation of model ensembles that capitalize on the strengths of individual models.

4.5. CONCLUSION

At the outset of this study, we aimed to assess the effectiveness of random forest models (RFMs) in predicting hydrological model-performance targets using catchment attributes from a large-sample data set. We extended this investigation by evaluating how catchment-attribute feature importance depends on the targeted fluxes and states. We also aimed to use the importance of catchment attributes to identify potential pathways for model refinement and improvement. Based on our findings, we draw the following conclusions.

Our results indicate that RFM predictive skill and the importance of individual catchment attributes vary substantially among the selected hydrological targets (streamflow,

actual evaporation, and soil moisture). In addition, the RFM results varied considerably among the selected test subsamples (folds) of the 617 catchments, indicating strong spatial variation.

Our findings differ from relationships previously established in other studies. In particular, the recognized correlations among meteorological inputs, aridity, and streamflow performance were not consistently observed. Instead, catchment area emerged as a primary predictor, suggesting possible adverse scaling effects and model-development artifacts affecting performance.

This study identified several avenues for refining the selected hydrological models. Soil parameterization and process representation are key areas for further scrutiny and refinement to improve streamflow simulations with both wflow_sbm and PCR-GLOBWB. Improvements to PCR-GLOBWB land-cover parameterization, particularly the representation of urban areas, are expected to improve simulations of actual evaporation. In addition, a more advanced river-routing method might improve performance in smaller catchments.

The insights from this study underscore the potential of data-driven methods and comparative model analyses to refine hydrological models and deepen our understanding of hydrological processes. Because the presented approach is model-agnostic, it can also be applied to other hydrological models. This versatility improves our ability to identify how and why specific underlying processes produce different results among locations.

SUPPLEMENTARY ANALYSES

A.1 CATCHMENT ATTRIBUTE CLASSES

Table 4.5: Overview of the 93 catchment attributes used as random forest model features, with their respective attribute classes and data sources.

Catchment Attribute	Class	Source
AverageStorage	hydrology	TOSSH toolbox
BFI	hydrology	TOSSH toolbox
BaseflowRecessionK	hydrology	TOSSH toolbox
EventRR	hydrology	TOSSH toolbox
FDC_slope	hydrology	TOSSH toolbox
First_Recession_Slope	hydrology	TOSSH toolbox
HFD_mean	hydrology	TOSSH toolbox
IE_thresh	hydrology	TOSSH toolbox
Mid_Recession_Slope	hydrology	TOSSH toolbox
Q5	hydrology	TOSSH toolbox
Q95	hydrology	TOSSH toolbox
QP_elasticity	hydrology	TOSSH toolbox
Q_mean	hydrology	TOSSH toolbox
RR_Seasonality	hydrology	TOSSH toolbox
RecessionParameters_1	hydrology	TOSSH toolbox
RecessionParameters_2	hydrology	TOSSH toolbox
RecessionParameters_3	hydrology	TOSSH toolbox
Recession_a_Seasonality	hydrology	TOSSH toolbox
SE_effect	hydrology	TOSSH toolbox
SE_slope	hydrology	TOSSH toolbox
SE_thresh	hydrology	TOSSH toolbox
Spearmans_rho	hydrology	TOSSH toolbox
Storage_thresh	hydrology	TOSSH toolbox
TotalRR	hydrology	TOSSH toolbox
VariabilityIndex	hydrology	TOSSH toolbox
high_Q_dur	hydrology	TOSSH toolbox
high_Q_freq	hydrology	TOSSH toolbox
low_Q_dur	hydrology	TOSSH toolbox
low_Q_freq	hydrology	TOSSH toolbox
area	land surface	CAMELS-GB
bares_perc	land surface	CAMELS-GB
crop_perc	land surface	CAMELS-GB
dpsbar	land surface	CAMELS-GB
dwood_perc	land surface	CAMELS-GB
elev_10	land surface	CAMELS-GB
elev_50	land surface	CAMELS-GB
elev_90	land surface	CAMELS-GB
elev_max	land surface	CAMELS-GB
elev_mean	land surface	CAMELS-GB
elev_min	land surface	CAMELS-GB
ewood_perc	land surface	CAMELS-GB
gauge_elev	land surface	CAMELS-GB
grass_perc	land surface	CAMELS-GB
inwater_perc	land surface	CAMELS-GB
shrub_perc	land surface	CAMELS-GB

Continued on the next page

Table 4.5 continued from the previous page

Catchment Attribute	Class	Source
urban_perc	land surface	CAMELS-GB
aridity	meteorology	CAMELS-GB
frac_snow	meteorology	CAMELS-GB
high_prec_dur	meteorology	CAMELS-GB
high_prec_freq	meteorology	CAMELS-GB
low_prec_dur	meteorology	CAMELS-GB
low_prec_freq	meteorology	CAMELS-GB
p_mean	meteorology	CAMELS-GB
p_seasonality	meteorology	CAMELS-GB
pet_mean	meteorology	CAMELS-GB
bulkdens	subsurface	CAMELS-GB
bulkdens_5	subsurface	CAMELS-GB
bulkdens_50	subsurface	CAMELS-GB
bulkdens_95	subsurface	CAMELS-GB
clay_perc	subsurface	CAMELS-GB
conductivity_cosby	subsurface	CAMELS-GB
conductivity_cosby_5	subsurface	CAMELS-GB
conductivity_cosby_50	subsurface	CAMELS-GB
conductivity_cosby_95	subsurface	CAMELS-GB
conductivity_hypres	subsurface	CAMELS-GB
conductivity_hypres_5	subsurface	CAMELS-GB
conductivity_hypres_50	subsurface	CAMELS-GB
conductivity_hypres_95	subsurface	CAMELS-GB
inter_mod_perc	subsurface	CAMELS-GB
no_gw_perc	subsurface	CAMELS-GB
organic_perc	subsurface	CAMELS-GB
porosity_cosby	subsurface	CAMELS-GB
porosity_cosby_5	subsurface	CAMELS-GB
porosity_cosby_50	subsurface	CAMELS-GB
porosity_cosby_95	subsurface	CAMELS-GB
porosity_hypres	subsurface	CAMELS-GB
porosity_hypres_5	subsurface	CAMELS-GB
porosity_hypres_50	subsurface	CAMELS-GB
porosity_hypres_95	subsurface	CAMELS-GB
root_depth	subsurface	CAMELS-GB
root_depth_5	subsurface	CAMELS-GB
root_depth_50	subsurface	CAMELS-GB
root_depth_95	subsurface	CAMELS-GB
sand_perc	subsurface	CAMELS-GB
silt_perc	subsurface	CAMELS-GB
soil_depth_pelletier	subsurface	CAMELS-GB
soil_depth_pelletier_5	subsurface	CAMELS-GB
soil_depth_pelletier_50	subsurface	CAMELS-GB
soil_depth_pelletier_95	subsurface	CAMELS-GB
tawc	subsurface	CAMELS-GB
tawc_5	subsurface	CAMELS-GB
tawc_50	subsurface	CAMELS-GB
tawc_95	subsurface	CAMELS-GB

A.2 RANDOM FOREST MODEL BENCHMARK SCORES

Table 4.6: Overview of the random forest model target variables based on the hydrological model simulations. The table reports the mean coefficient of determination (R^2) and its standard deviation across the stratified 10-fold test combinations.

Target Variable	**Hydrological Model**	**RFM Target**	**Mean R^2**	**Std. Dev. R^2**
Streamflow	wflow_sbm (W)	Daily KGE-NP	0.81	0.10
	PCR-GLOBWB (P)	Daily KGE-NP	0.46	0.18
	Δ(W-P)	Daily KGE-NP	0.45	0.07
Actual evaporation	wflow_sbm (W)	8-day RRMSE	0.57	0.09
	PCR-GLOBWB (P)	8-day RRMSE	0.52	0.11
	Δ(W-P)	8-day RRMSE	0.58	0.16
Soil moisture	wflow_sbm (W)	Daily RRMSE	0.68	0.08
	PCR-GLOBWB (P)	Daily RRMSE	0.75	0.07
	Δ(W-P)	Daily RRMSE	0.35	0.09

A.3 MODEL PERFORMANCE AND CATCHMENT ATTRIBUTE CORRELATION PLOTS

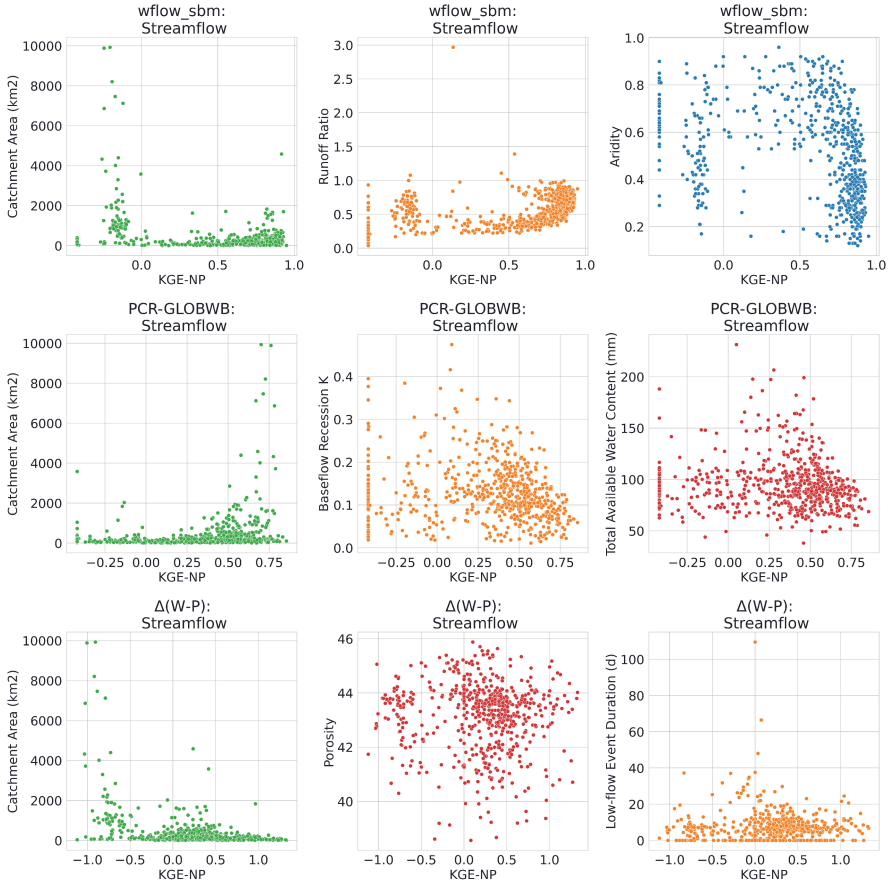


Figure 4.9: Correlations between catchment-attribute values and streamflow-based model performance for the three highest-ranking attributes in Table 4.2. Colors indicate attribute classes: land surface (green), hydrology (orange), meteorology (blue), and subsurface (red).

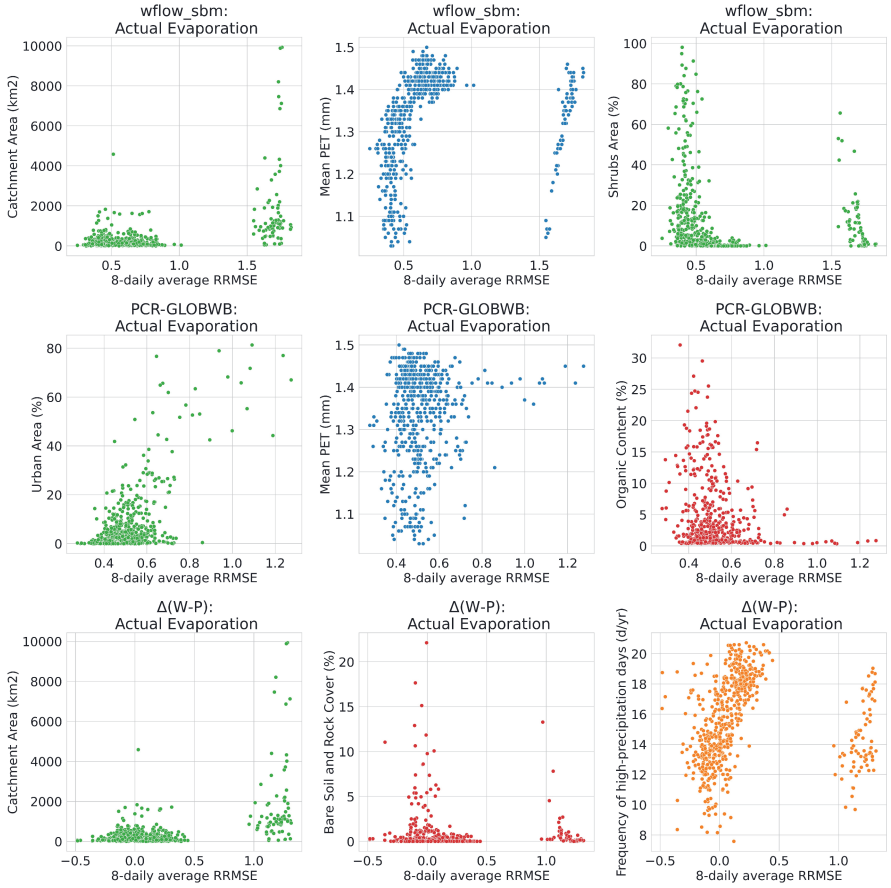


Figure 4.10: Correlations between catchment-attribute values and actual evaporation-based model performance for the three highest-ranking attributes in Table 4.3. Colors indicate attribute classes: land surface (green), hydrology (orange), meteorology (blue), and subsurface (red).

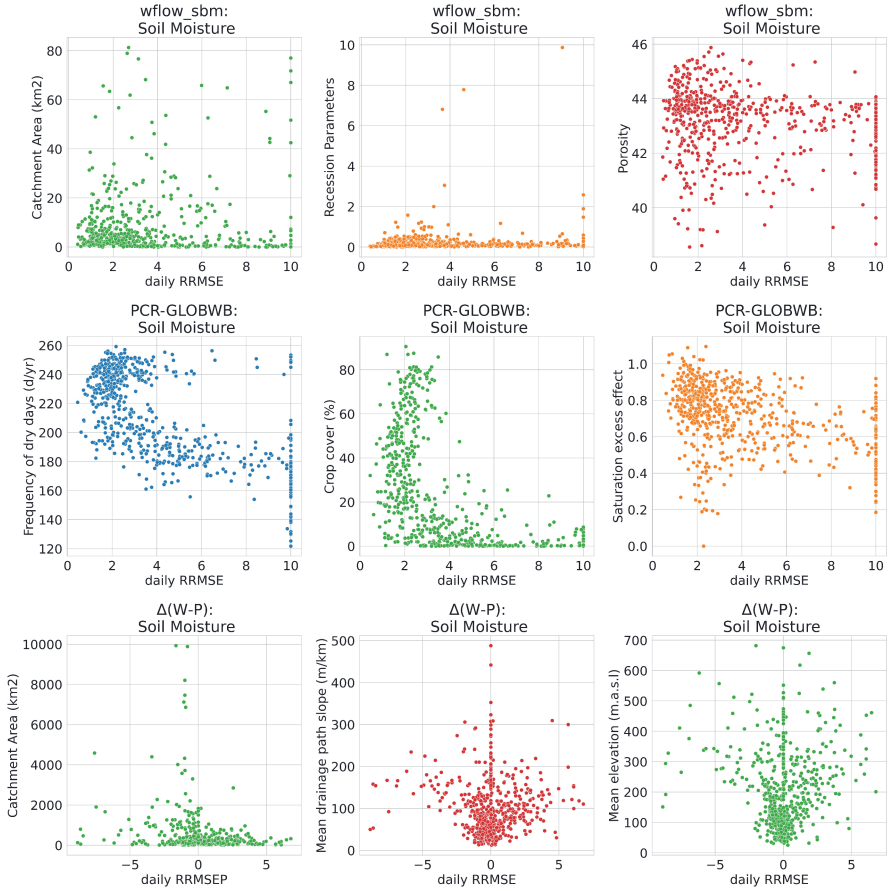


Figure 4.11: Correlations between catchment-attribute values and soil moisture-based model performance for the three highest-ranking attributes in Table 4.4. Colors indicate attribute classes: land surface (green), hydrology (orange), meteorology (blue), and subsurface (red).

5

ON THE IMPORTANCE OF DISCHARGE OBSERVATION UNCERTAINTY WHEN INTERPRETING HYDROLOGICAL MODEL PERFORMANCE

The first principle is that you must not fool yourself—and you are the easiest person to fool.

Richard Feynman

For users of hydrological models, the suitability of models can depend on how well their simulated outputs align with observed discharge. This study emphasizes the crucial role of factoring in discharge observation uncertainty when assessing the performance of hydrological models. We introduce an ad hoc approach, implemented through the eWaterCycle platform, to evaluate the significance of differences in model performance while considering the uncertainty associated with discharge observations. The analysis of the results encompasses 299 catchments from the CAMELS-GB large-sample catchment dataset and addresses 3 practical use cases for model users. These use cases involve assessing the impact of additional calibration on model performance using discharge observations, conducting conventional model comparisons, and examining how the variations in discharge simulations resulting from model structural differences compare with the uncertainties inherent in discharge observations.

Parts of this chapter have been published verbatim in Aerts, J. P. M., Hoch, J. M., Coxon, G., van de Giesen, N. C., and Hut, R. W.: On the importance of discharge observation uncertainty when interpreting hydrological model performance, Hydrol. Earth Syst. Sci., 28, 5011–5030, https://doi.org/10.5194/hess-28-5011-2024, 2024.

5.1. INTRODUCTION

Many fields in geoscience rely on uncertain data to accurately estimate states and fluxes that support decision-making. Uncertain hydrology data encompass multiple sources, including direct measurements, proxy-based measurements, interpolation techniques, scaling processes, and data management practices (McMillan, Westerberg, and Krueger, 2018). A large amount of literature has been devoted to discussing the effect of data quality limitations on hydrological modeling (e.g., Beven et al., 2011; Beven, 2016; Beven and Lane, 2022; Beven and Smith, 2015; Beven et al., 2022; Huang and Bardossy, 2020; Kauffeldt et al., 2013; Kirchner, 2006; Yew Gan et al., 1997). Data uncertainty can be distinguished into input data uncertainty (e.g., Kavetski et al., 2006a, 2006b) and evaluation data uncertainty (e.g., McMillan, Freer, et al., 2010).

Input data primarily comprises meteorological variables such as precipitation and temperature. Other input data sources include static data, such as soil and topographic properties used to estimate model parameters. The inherent uncertainties in input datasets influence the model's simulation of states and fluxes (e.g., Balin et al., 2010; Bárdossy and Das, 2008; Bárdossy and Anwar, 2023; Bárdossy et al., 2022; Beven, 2021; McMillan et al., 2011). The uncertainty propagation from input to model output is also closely influenced by the model structure (Butts et al., 2004; Liu and Gupta, 2007; Montanari and Di Baldassarre, 2013; Zhou et al., 2022). The effects of uncertainty propagation have been a focal point in the literature (e.g., Beven, 2006b; Gupta and Govindaraju, 2019; Montanari and Toth, 2007).

Evaluation of data uncertainty, the focus of this study, plays a pivotal role in determining hydrological models' potential accuracy and robustness. This is the case for model calibration, a process that involves fine-tuning model parameters to ensure that the model accurately and consistently reflects the observed historical behavior of the hydrologic system. Typically, this is based on discharge. When a model aims to replicate discharge values without including discharge observation uncertainty, the results are constrained to match a precise but potentially inaccurate representation of the hydrological response (Vrugt et al., 2005). Consequently, accurately calibrating the model becomes more challenging due to the need to incorporate evaluation data uncertainty into the calibration process to minimize bias in model parameters (McMillan, Freer, et al., 2010).

Multiple studies have demonstrated the importance of accounting for uncertainties in discharge observations. These mainly focus on hydrological model calibration (e.g., Beven and Binley, 1992; Beven and Freer, 2001; Beven and Lane, 2019; Beven and Smith, 2015; Blazkova and Beven, 2009; Coxon et al., 2015; Liu et al., 2009; McMillan, Freer, et al., 2010; McMillan, Westerberg, and Krueger, 2018; Westerberg et al., 2022). In these studies, multiple methodologies are used to quantify uncertainty estimates of discharge observations that are subsequently used for model calibration (overview in McMillan et al., 2012).

Combined, all uncertainty sources (input data, evaluation data, model structure, model parameters, initial conditions) add to a concept in hydrological modeling commonly referred to as the equifinality concept (Beven et al., 2011; Beven, 2006b; Beven and Freer, 2001; Clark, Slater, et al., 2008; Montanari and Grossi, 2008). The concept is

characterized by the circumstance of various model configurations yielding similar behavioral or acceptable results. Therefore, the recommendation is to account for all uncertainty sources simultaneously. An example of a method that includes all uncertainty sources during the parameter estimation process is the General Likelihood Uncertainty Estimation (GLUE; Beven and Freer, 2001) method. In practice, such methods are not always applied by model users, although the difficulty of implementation can be dispelled (Pappenberger and Beven, 2006).

Hydrological model evaluation by model users is often solely based on discharge observations. The inherent uncertainties in this single source of observational data might obscure the model's ability to simulate actual discharge. Therefore, omitting data uncertainty during model evaluation negatively affects the interpretation of relative model simulation differences, as these might fall within the uncertainty bounds of the observations.

Another challenging aspect of hydrological modeling is the hydrological system's large spatial and temporal variability. The large variety in landscape and hydrological heterogeneity can be captured when evaluating or comparing hydrological models using so-called large-sample catchment hydrology datasets. These large-sample catchment datasets contain hydro-meteorological time series, catchment boundaries, and catchment attributes for a large set of catchments. Each dataset is complemented by discharge observations at the catchment outlets and meteorological forcing datasets that include precipitation, temperature, and reference evaporation. The large-sample catchment datasets are collected using a consistent methodology across all catchments.

Recent large-sample datasets follow the structure introduced by Addor et al. (2017) in the form of the CAMELS(-US) dataset. More recently, Coxon et al. (2020b) released the CAMELS-GB, which includes estimates of quantified discharge observation uncertainty. This dataset describes 671 catchments in Great Britain. 503 catchments (gauging stations) are complemented with quantified discharge observation uncertainty estimates (Coxon et al., 2015). A recent effort by Kratzert et al. (2022) combined all available national CAMELS datasets in the overarching CARAVAN dataset for global consistency and to boost accessibility through data access via Google Earth Engine.

The access to large-sample catchment data has prompted a substantial body of research, as detailed in Addor et al. (2020). This includes applications in hydrological model testing and comparative analysis (e.g., Feng et al., 2022; Lane et al., 2019; Mizukami et al., 2017; Rakovec et al., 2019). One of the benefits of these datasets is that large samples of catchments allow for the evaluation of the robustness of model performance (Andréassian et al., 2006; Gupta et al., 2014). Identifying this robustness provides model users with valuable information on the presence or absence of consistency in the model results.

This study assesses the effect of omitting discharge observation uncertainty while interpreting model performance differences. Specifically, we focus on how this uncertainty influences model selection from the perspective of model users. We highlight the importance of incorporating discharge observation uncertainty during model calibration and model evaluation efforts. To achieve this, we developed a generic method applicable to any geoscience field where model results are compared to uncertain observations. This method determines, based on the 5th to 95th percentile range of flow,

whether model simulation differences are significant in the context of discharge observation uncertainties. In this study, we highlight 3 use cases based on 8 hydrological models that encompass model refinement efforts, conventional model comparisons, and the influence of model structure uncertainty in light of discharge observation uncertainty. Furthermore, we assess the spatial consistency of model performance results using a large-sample catchment dataset, and we evaluate the temporal consistency of model performance metrics by sub-sampling the observation and simulation pairs, as demonstrated by Clark et al. (2021). By doing so, more informed conclusions can be drawn about model performance based on individual or large samples of catchments.

5.2. METHODOLOGY

A generic tool is designed for assessing model simulations while considering the uncertainties inherent in evaluation data. First, the 3 use cases are presented. This is followed by the input data description, the evaluation data description, and the discharge observation uncertainty estimates used to conduct the analyses. Next, we describe the hydrological models and model runs employed for calibration and evaluation. The methodology concludes with an explanation of the uncertainty-based analyses.

In Figure 5.1, a graphical workflow provides an overview of the methodology. Figure 5.1a shows a typical model run with inputs and outputs, Figure 5.1b shows a conventional comparison of objective functions based on discharge observations and simulations, and Figure 5.1c describes the uncertainty analysis introduced in this study.

5.2.1. USE CASES

We devised 3 use cases based on 8 hydrological models that exemplify how users of models, who themselves are not the model developers, can interpret differences between model simulations in the context of discharge observation data uncertainty. The use cases are:

1. Model refinement in practice: This use case concerns additional model refinement by fine-tuning an effective model parameter based on discharge observations after initial calibration. It highlights the value of relative gains in model performance when not considering discharge observation uncertainty in the calibration process.

2. Model comparison for model selection: Here, two distributed hydrological models are compared against the backdrop of uncertainties in discharge observations. This analysis aims to pinpoint scenarios where the disparities between model results are within the margin of error of the discharge observations.

3. Model selection under model structural uncertainty: This use case involves contrasting the uncertainty inherent in the model's structure, as seen across various hydrological models, with the uncertainty in discharge observations.

An additional analysis quantifies uncertainty in the model performance objective functions due to temporal sampling of the discharge simulation and observation pairs. This temporal sampling uncertainty is detailed in Section 5.2.5.

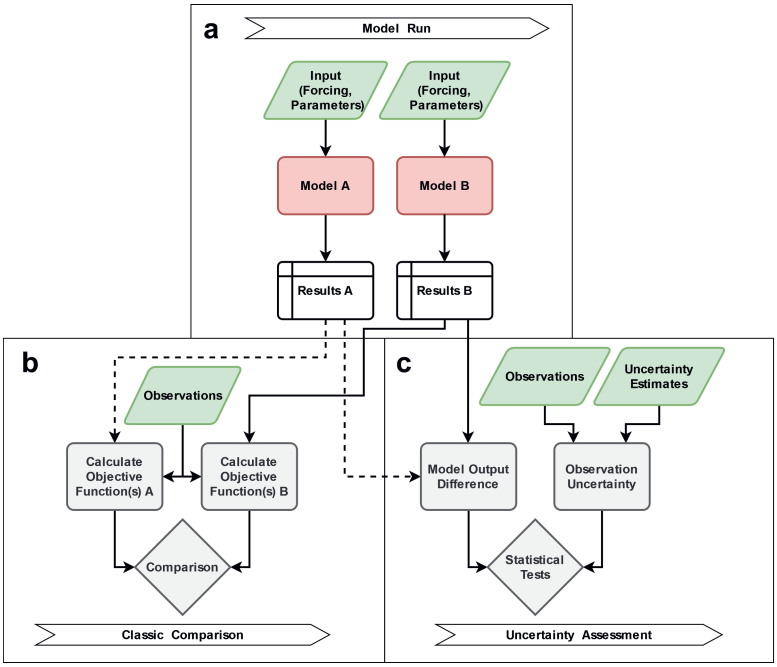


Figure 5.1: Graphical workflow of model experiments and analyses. The model experiment inputs are shown in green, the models in red, and the analysis components in grey. (a) the model runs of two models with inputs and outputs that result in simulation time series. (b) the conventional model comparison that compares objective functions based on simulation and observation time series. (c) the workflow of the proposed analysis that compares relative model simulation differences to discharge observation uncertainty estimates.

5.2.2. DATA

CASE STUDY AND CATCHMENT SELECTION PROCEDURE

The CAMELS-GB large-sample catchment dataset (Coxon et al., 2020a) serves as the case study area of the use cases and contains data (hydro-meteorological time series, catchment boundaries and catchment attributes) describing 671 catchments located across Great Britain. The underlying data used to create CAMELS-GB are publicly available and, therefore, suitable for evaluating hydrological models as the dataset can be easily extended. A unique feature of the dataset is the availability of quantified discharge observation uncertainty estimates for the flow percentiles of 503 catchments (see Coxon et al. (2015)).

The use cases in this study employ hydrological models with a daily time step. This can cause temporal discretization errors in small catchments due to peak precipitation and peak discharge occurring at the same time step. Therefore, these catchments are excluded through a selection procedure. This procedure calculates the cross-correlation between observed discharge and precipitation for a range of lag times. Catchments that predominantly show less than 1 day of lag between observed discharge and precipitation are excluded. Of the 503 catchments with uncertainty estimates, 299 are selected for the case study.

METEOROLOGICAL FORCING AND PRE-PROCESSING

In this study, we use the same meteorological forcing used to create the CAMELS-GB meteorological time series and climate indices as input to the hydrological models. This input consists of daily meteorological datasets on a 1 km × 1 km grid. The meteorological variables used in this study are precipitation (CEH-GEAR; Keller et al., 2015; Tanguy et al., 2021), reference evaporation (CHESS-PE; Robinson et al., 2020a), and temperature (CHESSmet; Robinson et al., 2020b). The distributed hydrological models use gridded inputs, and the conceptual hydrological models aggregate time series of meteorological variables readily available in the CAMELS-GB dataset.

DISCHARGE OBSERVATIONS AND QUANTIFIED UNCERTAINTY ESTIMATES

The discharge observations in the CAMELS-GB dataset were obtained from the UK National River Flow Archive and are daily values in cubic meters per second ($m^3\ s^{-1}$). As is common with large-sample catchment datasets, several catchments contain missing flow data in the time series. These missing values are not considered in this study's analyses.

A unique aspect of the CAMELS-GB dataset is the inclusion of quantified discharge observation uncertainty estimates created by Coxon et al. (2015). The uncertainty is quantified using a large dataset of quality-assessed rating curves and stage-discharge measurements. The mean and variance at each stage point are calculated and fitted in an iterative process using a LOWESS regression method that defines the rating curve and discharge uncertainty. Combining the LOWESS curves and variance in a Gaussian Mixture model based on a random draw from the measurement error distribution gives an estimate of discharge uncertainty; see Coxon et al. (2015) for more information.

5.2.3. HYDROLOGICAL MODELS

A mixture of distributed physical-process-based and lumped conceptual hydrological models is selected for the use cases, thereby showcasing the versatility of the analysis. The model refinement and model comparison use cases employ two physical-process-based hydrological models: wflow_sbm (van Verseveld et al., 2022) and PCR-GLOBWB (Hoch et al., 2023; Sutanudjaja et al., 2018). The rationale behind selecting these models lies in their differing approaches to conceptualizing hydrological processes and their respective optimization routines. Despite these differences, both models are suitable for comparison to a certain degree. This comparability stems from their shared classification as distributed hydrological models, similar complexity, parameterization, and applicability at a spatial resolution of 1 km.

For the model structure use case, 6 conceptual hydrological models are sourced from the Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT: Knoben, Freer, Fowler, et al., 2019; Trotter et al., 2022). These specific models are selected to encompass a wide array of model structures. The selection is based on the number of model stores, the quantity of parameters, and differing process representations.

DISTRIBUTED HYDROLOGICAL MODELS

The wflow_sbm physically based distributed hydrological model (van Verseveld et al., 2022) originated from the Topog_SBM model concept (Vertessy and Elsenbeer, 1999). This concept was developed for small-scale hydrologic simulations. The wflow_sbm model deviates from Topog_SBM by the addition of capillary rise, evapotranspiration and interception losses (Gash model; Gash, 1979), a root water uptake reduction function (Feddes and Zaradny, 1978), glacier and snow processes, and D8 river routing that uses the kinematic wave approximation in this study. The parameters (40 in total) are derived from open-source datasets and use pedo-transfer functions to estimate soil properties (see the HydroMT software package; Eilander and Boisgontier, 2022).

The model version with 1 km spatial resolution was aggregated from the finest available data scale (90 m). The hydraulic parameters related to the river network are upscaled using the method presented in Eilander et al. (2021). The parameter upscaling of the wflow_sbm model is based on the work by Imhoff et al. (2020) that uses point-scale (pedo)transfer-functions. This is similar to the multi-scale parameter regionalization method (Samaniego et al., 2010). Parameters are aggregated from the original data resolution with upscaling operators determined by a constant mean and standard deviation across various scales. Fluxes and states are checked for consistency during this process. See van Verseveld et al. (2022) for further information.

The PCR-GLOBWB physically based distributed hydrological model was initially developed for global hydrology and water resources assessments (Sutanudjaja et al., 2018). The PCR-GLOBWB model calculates the water storage in two soil layers, one groundwater layer, and the exchange between the top layer and the atmosphere. The model accounts for water use determined by water demand. We employ the version with 1 km spatial resolution introduced by Hoch et al. (2023). The model configuration in this study applies the accumulated travel time approximation for river routing.

Table 5.1: Overview of the six selected conceptual hydrological models showing the model name, number of stores, number of parameters, and key references (adapted from Knoben et al., 2020).

Model	Number of Stores	Number of Parameters	Key References
IHACRES	1	7	Croke and Jakeman, 2004; Ye et al., 1997
GR4J	2	4	Perrin et al., 2003; Santos et al., 2018
VIC	3	10	Liang et al., 1994
XINANJIANG	4	12	Jayawardena and Zhou, 2000
HBV-96	5	15	Lindström et al., 1997
SMAR	6	8	Tan and O'Connor, 1996

CONCEPTUAL HYDROLOGICAL MODELS

MARRMoT is a flexible modeling framework that houses an array of conceptual hydrological models (Knoben, Freer, Fowler, et al., 2019; Trotter et al., 2022). It is particularly valued in research for assessing model structure uncertainty, as highlighted in Knoben et al. (2020). One of the key advantages of MARRMoT is that the conceptual models share a uniform numerical implementation. To achieve this, alterations were made to the original model codes. These alterations ensure a consistent basis for model structure comparisons, allowing for a precise evaluation of differences in hydrological simulations due to varying model structures. The hydrological models IHACRES, GR4J, VIC, XINANJIANG, HBV-96, and SMAR are selected in this study. Table 5.1 provides an overview of the number of stores, parameters, and key references.

5.2.4. MODEL RUNS

The model runs that form the basis of the 3 use cases are performed as intended by the model developers. This means that this study employs calibration and/or optimization methodologies as recommended by the model developers for model users. The calibrated parameters for the distributed hydrological models were obtained from the model developers. Regarding the conceptual hydrological models, we follow the model run configuration of Knoben et al. (2020).

PCR-GLOBWB MODEL RUNS

The PCR-GLOBWB model does not require additional regional parameter optimization after deriving the parameter set, as this is typically not conducted by the model developers. However, this does not imply that the model would not benefit from additional optimization. The model does require a spin-up period at the start of the model run. The model is spun up 30 years back-to-back using a single water year climatology that is based on the average values of each calendar day between 1-10-2000 and 30-09-2007. The following water year, 2008, is discarded from analyses to avoid overfitting at the start of the evaluation period, and the model is evaluated for the water years 2009–2015.

DEFAULT AND OPTIMIZED WFLOW_SBM MODEL RUNS

The wflow_sbm model is spun-up using the water year 2000 and is additionally calibrated using discharge observations for the water years 2001-2007. Additional calibration is performed by optimizing a single parameter using the Kling-Gupta Efficiency Non-Parametric (KGE-NP) objective function (Pool et al., 2018) based on differences between discharge observations and simulations at the catchment outlet. This results in a single optimized parameter set per catchment. Imhoff et al. (2020) identified the horizontal conductivity fraction parameter (KsatHorFrac) as effective for single parameter value optimization. KsatHorFrac is an amplification factor of the vertical saturated conductivity that controls the lateral flow in the subsurface.

After calibration, the water year 2008 is discarded from analyses, and the model is evaluated for the water years 2009–2015. For more information on the effects of calibration, the reader is referred to Aerts et al. (2022), Section 3.1 and Figure 3. The default wflow_sbm model run sets the KsatHorFrac parameter value to the default value of 100. The calibration results of the wflow_sbm model are presented in Figure 5.8 in the supplementary analyses at the end of this chapter.

CONCEPTUAL HYDROLOGICAL MODEL RUNS

Similar to the other model runs, the conceptual hydrological model runs are spun up using the water year 2000 and calibrated using the water years 2001-2007. The calibration method uses the Covariance Matrix Adaptation Evolution Strategy (CMA-ES; Hansen, 2006; Hansen and Ostermeier, 2001; Hansen et al., 2003). This method optimizes a single-objective function to find global parameter optima based on non-separable data problems. A demonstration of the sensitivity of the calibration parameters is shown in Knoben et al. (2020). Following calibration based on the KGE-NP objective function, the water year 2008 is discarded. The models are evaluated based on the water years 2009-2015 using the same KGE-NP objective function.

EWATERCYCLE

This study is conducted using the eWaterCycle platform (Hut et al., 2022). eWaterCycle is a community-driven platform for running hydrological model experiments. All components that are required to run hydrological models are FAIR by design (Wilkinson et al., 2018). This is achieved by versioning models and datasets and creating reproducible workflows. Therefore, the platform is suitable for conducting model performance experiments. The model simulations were conducted on the Dutch supercomputer Snellius to ensure faster model run time. We created example notebooks that use the eWaterCycle platform on cloud computing infrastructure: https://doi.org/10.5281/zenodo.7956488.

5.2.5. ANALYSES

MODEL EVALUATION

The hydrological model runs (calibration and evaluation) are evaluated using the Kling-Gupta efficiency non-parametric (KGE-NP; Pool et al., 2018) objective function. This efficiency metric deviates from the more commonly used Kling-Gupta efficiency (KGE; Gupta et al., 2009) by calculating the Spearman rank correlation and the normalized-flow-duration curve instead of the Pearson correlation and variability bias. Values range

from $-\infty$ to 1 (perfect score). In addition to the KGE-NP metric, we consider the Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970) to demonstrate the sensitivity of the results to the selection of the objective function. We include the KGE-NP, KGE, modified KGE (Kling et al., 2012), and the NSE objective functions in the data repository for completeness and future reference.

DISCHARGE OBSERVATION UNCERTAINTY

The ad hoc discharge observation uncertainty-based analysis of model performance differences consists of 3 parts. The first part divides the observation and simulation pairs into 3 flow categories similar to Coxon et al. (2015): low flow, average flow, and high flow conditions. The low flow category is based on observed discharge values at the catchment outlet in the 5th to 25th percentile range, the average flow category is based on the 25th to 75th percentile range, and the high flow category is based on the 75th to 95th percentile range. Not all percentiles are included for the low and high flow categories due to limited data availability on quantified discharge observation uncertainty.

In the second part of the method, shown in Figure 5.1c, the first step is to calculate the absolute error between calibrated model simulations for each flow category and each catchment. This step is visualized as a hydrograph in Figure 5.2a (blue line) and is referred to as the model difference. Next, the discharge observations' upper and lower uncertainty bounds are taken from the CAMELS-GB dataset. For example, the upper uncertainty percentages of the flow category boundary percentiles correspond to 25 % and 15 %, respectively. These values are then averaged to obtain the upper uncertainty bound at 20 % discharge observation uncertainty, as shown by the orange line in Figure 5.2b. Similarly, the red line in Figure 5.2b is the calculated lower uncertainty bound. The observation uncertainty percentage is then calculated by taking the absolute average of both uncertainty bounds (17.5 %). This percentage is applied to the observations to quantify the portion of the observed discharge attributed to uncertainty (observation uncertainty) in cubic meters per second ($m^3\ s^{-1}$), as illustrated by the green line in Figures 5.2b and 5.2c.

The third part of the method applies a dependent t-test using the time series in Figure 5.2c with a 0.05 significance level to determine whether the observation uncertainty time series is greater than the model simulation difference time series.

This method is subject to certain limitations, particularly regarding the use of the discharge observation uncertainty estimates. Due to the absence of data, the upper and lower 5th percentiles of flow could not be included, although these data points may be most important for users to determine whether a model is fit for purpose. In addition, using the rating curve uncertainty rather than the uncertainty bounds of flow percentiles is preferred. We accept these limitations as we promote the use of existing datasets to ensure community participation in implementing the suggested evaluation procedure in other studies.

TEMPORAL SAMPLING UNCERTAINTY

Another aspect of model performance evaluations that might misinform model users is the sensitivity of objective functions to the temporal sampling of time series. Temporal sampling uncertainty determines whether the error distribution of simulation and observation pairs is heavily skewed. A few data pairs might have a disproportionate effect

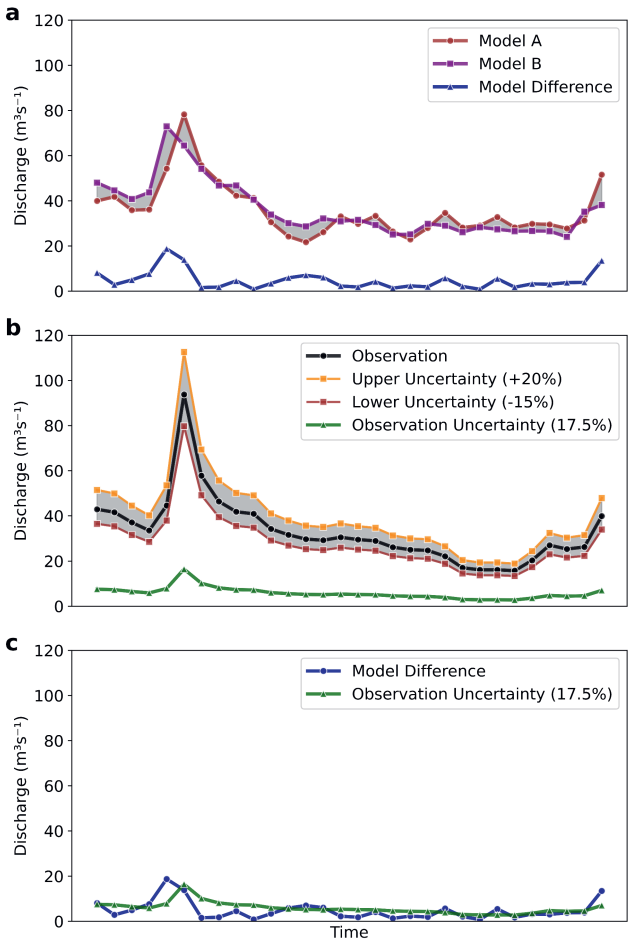


Figure 5.2: Example hydrographs of the discharge observation uncertainty analysis method. (a) Calculation of the absolute difference (blue) between calibrated model simulations (red and purple). (b) Calculation of streamflow observation uncertainty in $m^3\ s^{-1}$, shown in green. Orange and dark red lines indicate upper and lower bounds of observation uncertainty percentages that are averaged and multiplied by the observations in black. (c) Resulting time series, with, in blue, the absolute difference between calibrated model simulations and, in green, the averaged discharge observation uncertainty in $m^3\ s^{-1}$.

on the calculated objective functions that are used to determine model performance. The inclusion or exclusion of these data points due to the selection of the calibration and evaluation periods alters the consistency of model performance over time.

To quantify the temporal sampling uncertainty of the KGE-NP objective function, we applied the methodology of Clark et al. (2021). This method sub-samples the simulation and observation time series through bootstrapping (Efron, 1979) and jackknife-after-bootstrap (Efron and Tibshirani, 1986) methods. The change in objective function due to the shuffling of the sub-samples allows calculation of the standard error and its tolerance interval. The tolerance intervals corresponding to each model instance are averaged and referred to as the temporal sampling uncertainty. We extended the GUMBOOT software package (Clark et al., 2021) by adding the KGE-NP metric for this study.

5.3. RESULTS

In this section, we first present an overview of the discharge-based model performance results for each of the 3 use cases. Next, we detail the spatial distributions of the maximum model performance difference. This is followed by presenting uncertainty estimates for discharge observations categorized by flow. Next, the discharge observation uncertainty-based relative model performance analyses are presented. The section ends with the temporal sampling uncertainty analysis results.

The supplementary analyses at the end of this chapter contain the calibration results of the wflow_sbm model (Figure 5.8) and the Nash-Sutcliffe efficiency (NSE)-based model performance results of all considered models (Figure 5.9).

5.3.1. DISCHARGE-BASED MODEL PERFORMANCE

Model performance is assessed using discharge observations and simulations at 299 catchment outlets. The results are shown in Figure 5.3 as cumulative distribution functions (CDFs) of the KGE-NP objective function. These results offer insight into the model's accuracy in simulating observed discharge.

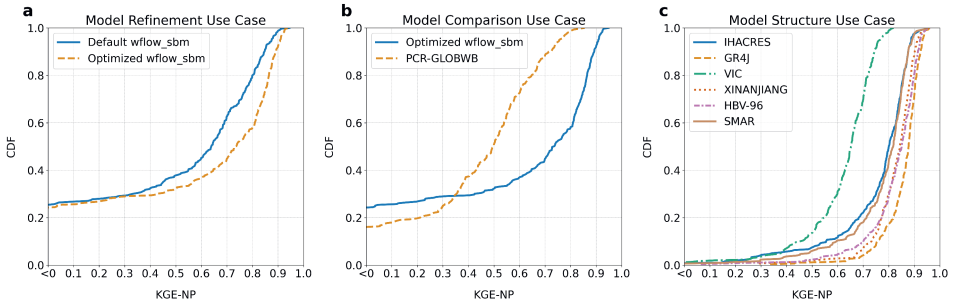


Figure 5.3: Cumulative distribution function (CDF) plots of the Kling-Gupta Efficiency non-parametric (KGE-NP) objective function, derived from discharge estimates and observations at 299 catchment outlets. (a) shows the CDF for the model refinement use case, optimizing the wflow_sbm hydrological model with a single parameter. (b) shows the CDF for the model comparison use case, comparing the optimized wflow_sbm and PCR-GLOBWB hydrological models. (c) demonstrates the CDF for the model structure use case, showcasing results from 6 conceptual hydrological models.

The CDF of the model refinement use case in Figure 5.3a establishes that optimizing a single effective parameter improves approximately 65 % of the catchment simulations. The improvements remain modest, as indicated by the median values of 0.64 KGE-NP for the default wflow_sbm model and 0.74 KGE-NP for the optimized wflow_sbm model. Larger model performance differences are found for the model comparison use case in Figure 5.3b. Here, the optimized wflow_sbm model performs better in 75 % of the catchments than the PCR-GLOBWB model. Both models demonstrate poor results for approximately 18 % to 24 % of the evaluated catchments (<0.40 KGE-NP).

The model structure use case results are based on 6 conceptual hydrological models that differ only in model structure (Figure 5.3c). The spread in model results shows that the VIC model lags behind in performance compared to the other models. The IHACRES and SMAR models yield very similar results despite large structural differences. The XINANJIANG and HBV-96 models produce comparable outcomes and share a more similar model structure. The GR4J model consistently outperforms the other models. The total model structure uncertainty, as expressed by the difference between the worst- and best-performing model CDFs, is substantial, while the differences between models can be subtle. Median KGE-NP values for the models are as follows: VIC at 0.65, IHACRES at 0.80, SMAR at 0.82, XINANJIANG at 0.84, HBV-96 at 0.85, and GR4J at 0.88 KGE-NP.

Next, we consider the spatial distribution of the results, as shown in Figure 5.4, which depicts the maximum KGE-NP difference between the models for each use case. Figure 5.4a indicates improvements after model refinement, with positive KGE-NP difference values in various parts of Great Britain. However, there are no discernible spatial trends in these improvements. Figure 5.4b compares the wflow_sbm and PCR-GLOBWB distributed hydrological models, revealing high spatial variability with no consistent patterns. Similarly, Figure 5.4c highlights differences for the model structure use case, where the largest differences are again observed in various regions without a clear spatial trend. While the spatial distribution is provided for completeness, no significant spatial trends are evident from the data.

5.3.2. DISCHARGE OBSERVATION UNCERTAINTY ESTIMATES

The discharge observation uncertainty estimates consider the 5th to 95th percentile range of flow. These estimates are categorized into 3 flow conditions and are presented in Figure 5.5. In the box plot for the low flow category, we observe a wide interquartile range, shown by the spread of the box. This indicates a higher variability in discharge observation uncertainty percentages. The median value, represented by the line within the box, is at the 20 % uncertainty mark. Many outliers above the box indicate occasional large deviations from the median value. The range of values for the average flow category is narrower than for the low flow category, with a median value of 15 %. The lowest median value is found for the high flow category at 12 %. It is important to mention that the uncertainty is expected to be considerably higher if the underlying data contain the upper 5th percentiles of flow for this category.

5.3.3. USE CASES

The discharge simulation difference time series of two models is expressed in cubic meters per second and compared to the discharge observation uncertainty time series in

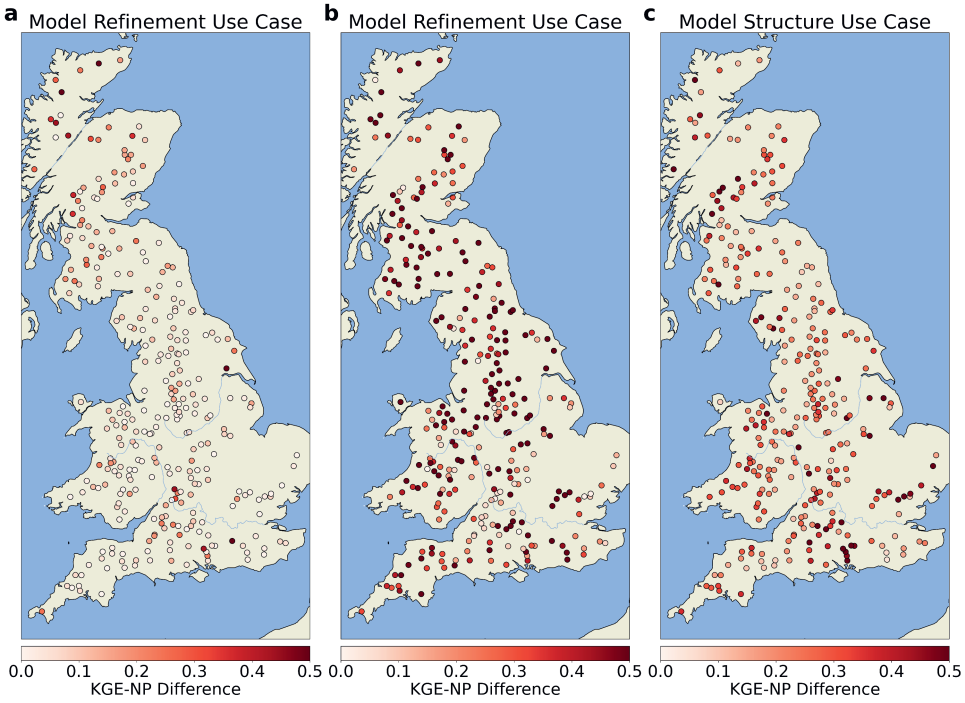


Figure 5.4: Spatial distribution of the absolute Kling-Gupta Efficiency non-parametric (KGE-NP) objective function difference between the worst and best model performances per catchment and use case. (a) The model refinement use case is based on the default and optimized wflow_sbm hydrological models. (b) The model comparison results are based on the optimized wflow_sbm and PCR-GLOBWB hydrological models. (c) The model structure use case results are based on the worst and best model performances of the 6 conceptual hydrological models.

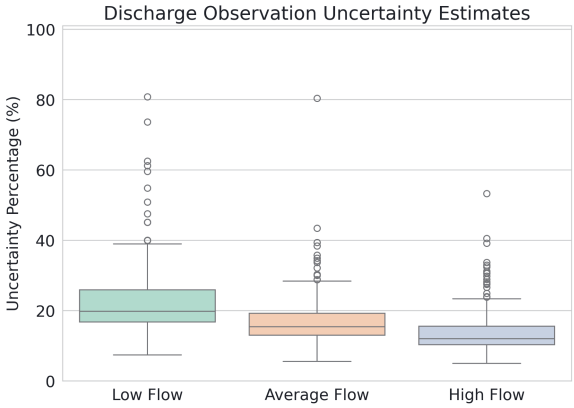


Figure 5.5: Discharge observation uncertainty estimates of 299 catchment outlets based on the work of Coxon et al., 2015 expressed as uncertainty percentages per flow category. (a) the low flow category uncertainty estimates are based on the 5th to 25th flow percentiles. (b) presents the 25th to 75th percentile average flow category. (c) the high flow discharge observation uncertainty estimates of the 75th to 95th flow percentiles are shown.

cubic meters per second. This is done using a t-test to determine whether the discharge observation uncertainty is larger than the simulation differences. The instances where this is the case are reported in Table 5.2 for the 3 use cases.

MODEL REFINEMENT

The model refinement use case results in Table 5.2 show that approximately one-third of the considered catchments contain instances of simulation differences between the wflow_sbm default and wflow_sbm optimized models that are statistically smaller than the discharge observation uncertainty estimates. This demonstrates the importance of incorporating (discharge) observation uncertainty when performing model refinement, especially based on a large-sample catchment dataset. This consideration should be part of the calibration and subsequent evaluation process. In addition, the results indicate that it is difficult to conclude whether the model performs better after refinement when discharge observation uncertainty is not considered. Overall, the results affirm the importance of incorporating discharge observation uncertainty in the optimization routine of the wflow_sbm model.

MODEL COMPARISON

For the model comparison use case (Table 5.2), there is a lower frequency of instances where discharge observation uncertainty surpasses differences in discharge simulations. The comparison between the optimized wflow_sbm model and the PCR-GLOBWB model reveals that discharge observation uncertainty exceeds simulation differences in 5 catchments for low flow, 4 for average flow, and 3 for high flow categories. These findings suggest that the interpretation of model performance is not significantly affected by the ad hoc addition of discharge observation uncertainty. However, catchments demonstrating the impact of observation uncertainty warrant careful examination.

Table 5.2: Overview of the number of instances per flow category where discharge observation uncertainty exceeds the simulation differences based on 299 catchments. Results are based on dependent t-tests with a significance level of 0.05.

Use Case	**Model Comparison**	**Flow Category**	**Obs. Uncertainty > Sim. Difference**
Model Refinement	wflow_sbm Default & Optimized	Low	98
		Average	98
		High	115
Model Comparison	wflow_sbm Optimized & PCR-GLOBWB	Low	5
		Average	4
		High	3
Model Structure	6 Conceptual Hydrological Models	Low	1
		Average	0
		High	0

MODEL STRUCTURE

The analysis of model structure uncertainty in the context of discharge observation uncertainty reveals that only a single instance of the low flow category contains discharge observation uncertainty that exceeds the simulation difference between all 6 conceptual hydrological models (Table 5.2). This shows that, based on the selected models, the model structure uncertainty, expressed as the difference in discharge simulations, is larger than this dataset's discharge observation uncertainty estimates. However, the investigation into the differences between the individual models yields several insights based on the results in Figure 5.6.

The VIC model results, characterized by their relatively lower performance, contain only a few instances where discharge observation uncertainty exceeds simulation differences, making it identifiable as the lesser-performing model. In contrast, the IHACRES and SMAR models exhibit a high level of simulation agreement, demonstrated by a large number of instances in Figure 5.6c. This is despite significant differences in their complexity and structural design. That is, IHACRES is a single-store hydrological model, and SMAR is a six-store hydrological model that accounts for soil moisture in a separate store. This alignment of simulation results between models with varying complexities highlights the nuanced influence of structural differences on simulation outcomes. The HBV-96 and XINANJIANG models most resemble each other regarding the number of stores and parameters; the parameters themselves and the process descriptions differ. Both models contain a low number of instances, allowing the identification of the better-performing model.

Next, we examine the results across the individual flow categories. The low flow category (Figure 5.6a) and the average flow category (Figure 5.6b) show similar trends, though with a lower number of instances for the average flow category. The high flow category (Figure 5.6c) is characterized by more frequent instances in which discharge

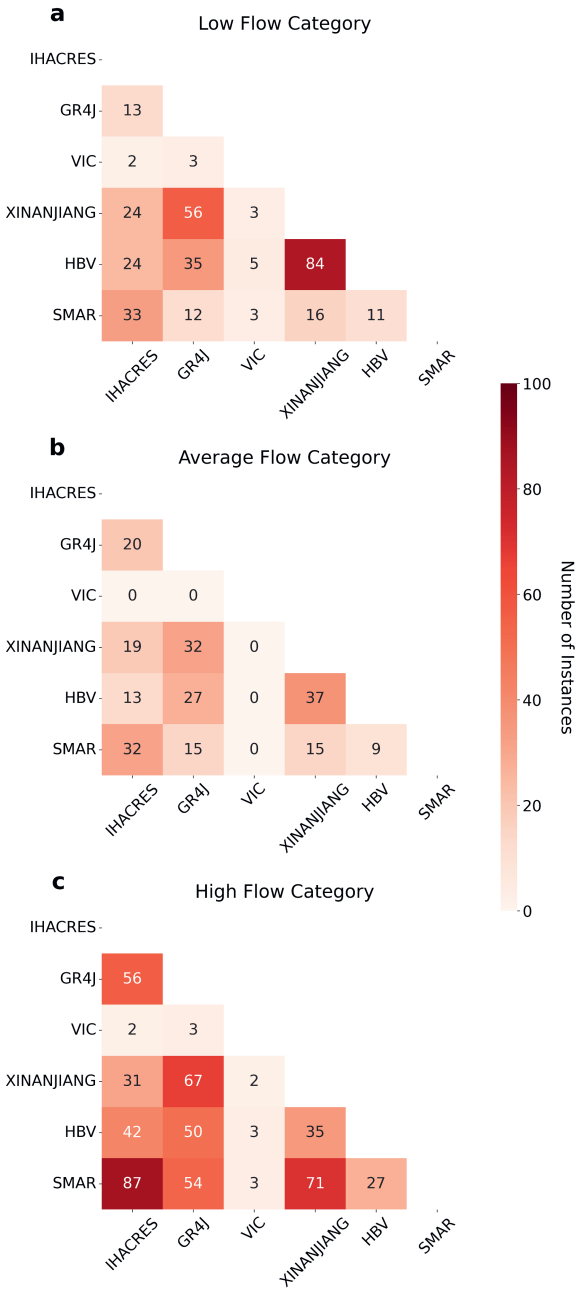


Figure 5.6: Heat map of the 6 conceptual hydrological models showing for each model combination the number of instances (n=299) that discharge observation uncertainty exceeds simulation differences. (a) number of instances for the low flow category, with low values shown in white and high values in red. (b) number of instances for the average flow category. (c) number of instances for the high flow category.

observation uncertainty surpasses simulation differences. This is especially evident between the IHACRES and SMAR models. The variability in structural design and parameterization among different hydrological models leads to notable output differences. This underscores the importance of selecting the appropriate model by including discharge observation uncertainty in the calibration and evaluation process.

5.3.4. TEMPORAL SAMPLING UNCERTAINTY

The temporal sampling uncertainty of the KGE-NP objective function is defined as the tolerance interval of the standard error of the objective function due to the sub-sampling of the simulation and observation pairs. This analysis provides insights into the temporal reliability and interpretability of hydrological model performance. Analysis of the results from the 6 conceptual hydrological models, as shown in Figure 5.7b, reveals a pattern consistent with the model performance depicted in Figure 5.3c. More precisely, the VIC model displays the highest KGE-NP uncertainty across all catchments, indicating its variability and the challenges in using this model's current setup for accurate predictions in different hydrological contexts.

The IHACRES and SMAR models and the GR4J, XINANJIANG, and HBV-96 model groups show similar KGE-NP-based temporal sampling uncertainty levels. This consistency across models with varying complexities suggests that KGE-NP uncertainty is influenced not only by the model design but also by hydrological conditions and data quality. Uncertainty values range widely, from about 0.1 KGE-NP to over 0.6, indicating significant variability in the temporal robustness of results (Figure 5.7b).

When comparing the average KGE-NP objective function uncertainty with the KGE-NP differences between individual models, it becomes clear that uncertainty often overshadows the differences between models. This is particularly the case in comparisons between GR4J - HBV-96, XINANJIANG - HBV-96, and SMAR - IHACRES. These findings imply that the inherent uncertainty in the objective functions may limit the ability to distinguish between model performances, complicating efforts to identify the most fit-for-purpose model based on this metric alone. This underscores the need for a more nuanced approach to model evaluation that considers objective function metrics, other contextual factors, and additional performance measures, ensuring more robust and reliable model selection processes.

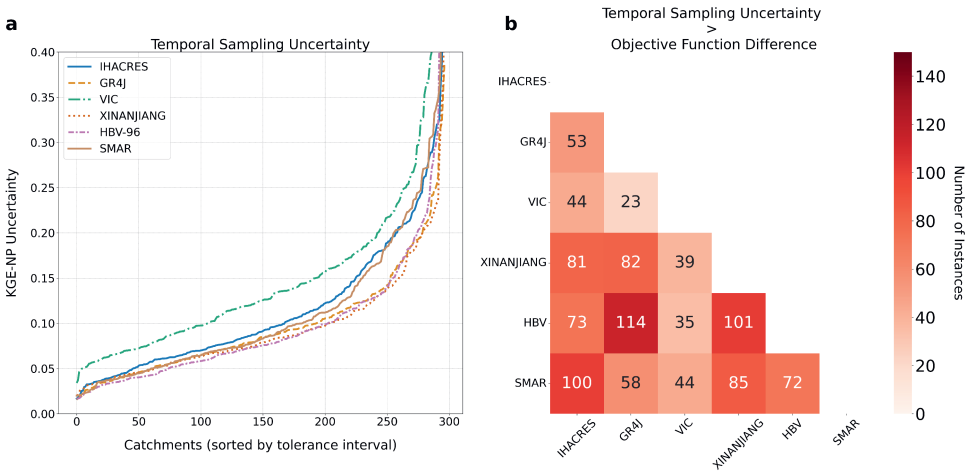


Figure 5.7: (a) Temporal objective function sampling uncertainty based on 6 conceptual hydrological models, expressed as the average tolerance interval of the standard error due to sub-sampling. The sorted values per catchment are shown on the horizontal axis, and the KGE-NP objective function uncertainty is shown on the vertical axis. (b) Heat map of the 6 conceptual hydrological models showing, for each model combination, the number of instances ($n = 299$) in which the average objective function uncertainty exceeds the objective function differences between model combinations. Low values in white and high values in red indicate the number of instances.

5.4. DISCUSSION

We introduced an ad hoc method highlighting the importance of including discharge observation uncertainty when evaluating hydrological models. Discharge observation uncertainty is frequently overlooked by model users, leading to potential misinterpretations of relative model performance. Our findings emphasize the significant impact of discharge observation uncertainty on model performance interpretation.

We acknowledge that observation uncertainty is not the only source of uncertainty, as there are uncertainties in model inputs, model structures, parameter sets, and initial or boundary conditions (e.g., Dobler et al., 2012; Hattermann et al., 2018; Moges et al., 2021; Renard et al., 2010). Therefore, the proposed generic tool does not replace a complete uncertainty analysis of modeling chains that also accounts for the impact of input uncertainties (Beven, 2006b; Beven and Freer, 2001; Pappenberger and Beven, 2006). Instead, it assists model users in interpreting relative model performance and highlights the importance of conducting a complete uncertainty analysis. Therefore, our study only constitutes a fraction of a broader challenge in which input uncertainty plays a substantial role, as demonstrated in Bárdossy and Anwar (2023).

5.4.1. PERFORMANCE INTERPRETATION UNDER DISCHARGE OBSERVATION UNCERTAINTY

Our analysis demonstrates that regionally optimizing the wflow_sbm hydrological model often results in only marginal improvements in model performance (Figure 5.3a). Although any improvement is beneficial, the findings suggest that discerning the superior

model variant becomes challenging without factoring in the uncertainty of discharge observations during calibration. This is evident in 98 instances in the low- and average-flow categories and 118 cases in the high-flow category (Table 5.2). The number of instances is expected to increase further when flows in the lower and upper 5th percentiles are included. Adopting an ad hoc measure, as introduced in this study, provides a practical but limited method for improving the interpretability of relative model results. Therefore, we recommend the integration of discharge observation uncertainty into both the model calibration and evaluation procedures, aligning with the consensus in the literature.

When comparing different hydrological models, we find that the uncertainty of discharge observations slightly masks the differences in relative model performance as shown by the 3 to 5 instances per flow category in Figure 5.3b and Table 5.2. Similarly to the model comparison use case, the model structure use case indicates that structural uncertainties overshadow the effects of discharge observation uncertainty. However, the comparison of individual models in Figure 5.6 shows many instances of discharge observation uncertainty exceeding model performance differences. For example, despite their structural differences, the IHACRES and SMAR models demonstrate a high level of simulation agreement (Figure 5.3c) and subsequent difficulty in discerning model performance differences in light of discharge observation uncertainty. In contrast, the VIC and XINANJIANG models, which have similar structures, exhibit two-thirds of the catchment simulation differences within the uncertainty bounds of the discharge observations. This underlines the complex interplay between model structures and subsequent performance, especially when contrasted with the uncertainty of discharge observation.

5.4.2. TEMPORAL ROBUSTNESS OF MODEL PERFORMANCE

Model performance can be heavily influenced by a few data points in the time series on which model performance is based (Clark et al., 2021). This can result in biased model performance interpretations depending on the selected time period for calibration and evaluation. When models are sensitive to specific data points, this can be due to inadequate process descriptions for the considered models. In addition, this might also indicate the presence of disinformative events and model invalidation sites where the runoff coefficient exceeds a value of 1 (Beven, 2023; Beven and Lane, 2022; Beven and Smith, 2015; Beven et al., 2022) or the presence of atypical data (e.g. Thébault et al., 2023).

Models ought to demonstrate adequate performance across the entire time series, which should be accurately represented in the performance outcomes. The assessment of temporal sampling uncertainty does not imply that this should not be the case; it instead points towards the model simulation and observation pairs that are worth investigating. These instances can serve as indicators that suggest areas where models may require further scrutiny and improvements. Knowing the temporal sampling uncertainty is relevant for model users as it provides information on the consistency of the model performance over time, which is necessary to determine whether a model is fit for purpose. Therefore, it is recommended to include alternative estimators better suited for skewed performance data in the reporting of model performance (e.g., Lamontagne et al., 2020; Shabestanipour et al., 2023; Towler et al., 2023).

5.4.3. PRACTICAL IMPLICATIONS FOR MODEL USERS

The method introduced in this study is purposely designed to be as generic and straightforward as possible to increase the potential for adoption in future studies. It can be applied to any hydrological state or flux where observation time series include uncertainty estimates. In addition, we recommend the routine reporting of evaluation data uncertainties and the temporal sampling uncertainty of objective functions. This would yield a clearer understanding of the relevance of differences between model outcomes and aid in identifying samples that require cautious interpretation. This reporting, however, does not replace model benchmarks that include complete uncertainty analyses (e.g., Lane et al., 2019) but enhances the interpretability of model performance in its absence.

For model users, this approach offers a pragmatic way to understand the implications of uncertainty in their model selection processes. While our method facilitates a clearer understanding of where and how uncertainties affect relative model performance differences, it should be viewed as a complementary step rather than a replacement for a thorough uncertainty analysis.

5.4.4. LIMITATIONS & OUTLOOK

The study presented faces several practical limitations. First, excluding the lower and upper 5th percentiles of flow from the analysis introduces a constraint on the uncertainty assessment, overlooking critical flow conditions often of significant interest in hydrological studies. This exclusion limits the ability to fully understand model performance under a complete range of hydrological conditions. Second, the reliance on uncertainty bounds rather than direct uncertainty estimates from rating curves, due to their absence in the CAMELS-GB dataset, poses another limitation. Using broad uncertainty bounds instead of precise estimates derived from rating curves, the analysis may not capture the variability and uncertainty inherent in the discharge observations. Additionally, displaying discharge observation uncertainty as relative values in the form of uncertainty percentages in Figure 5.5 might imply that the categories are affected at the same rate by the same physical phenomena. For instance, the largest changes in the rating curve for high flows are caused by changes in the river width, while low flow is most sensitive to sedimentation. Therefore, small changes in flow volume might have a larger effect on low flow conditions than high flow conditions. Expressing discharge observation uncertainty as absolute discharge values, as done in the methodology of this study, is not suitable for comparison across catchments because their flow volumes differ substantially, which is a common problem for large-sample catchment studies. Last, the study focuses solely on evaluating model performance primarily through discharge simulations, without delving into the reasons behind good or poor model performance, as this is outside the study's scope.

Looking forward, addressing discharge uncertainty on a global scale is of paramount importance. Accurate global assessments of discharge uncertainty are critical for informing water management strategies, policy decisions, and climate impact studies. Understanding and mitigating these uncertainties can lead to more reliable hydrological

models and enhance resource management worldwide. Although this study only provides a glimpse of what this might imply, it highlights the necessity for such global assessments by incorporating discharge observation and temporal sampling uncertainties into hydrological evaluations.

5.5. CONCLUSIONS

This study assesses the importance of including discharge observation uncertainty and temporal sampling uncertainty of objective functions in hydrological model performance evaluations based on a large-sample catchment dataset. This is done through statistical testing that determines whether the difference in discharge simulations between two hydrological models is larger or smaller than the discharge observation uncertainty estimates. Flow categories are created within the 5th to 95th percentile range of observed flow to support this analysis, and 3 use cases are devised.

In the model refinement use case, approximately one-third of the catchments showed discharge simulation differences between the default and optimized wflow_sbm models within the uncertainty bounds of discharge observations: 98 of 299 catchments for low and average flows and 118 of 299 catchments for high flows. This emphasizes the importance of integrating discharge observation uncertainty into the calibration process for model refinement. As a result, it is difficult to discern whether the optimization of the model leads to improved simulations of actual discharge. For the model comparison use case, we found that depending on the model combinations, a large fraction of catchments showed discharge observation uncertainty exceeding simulation differences. This suggests careful consideration of this uncertainty in model performance evaluations. The model structure uncertainty use case based on 6 conceptual hydrological models indicated only a few instances of discharge observation uncertainty exceeding simulation differences. This indicates that model structure uncertainty, expressed as discharge simulation differences, often exceeds discharge observation uncertainty. A comparison of the six individual hydrological models showed no clear relation between model complexity and model performance. Our study underscores the necessity of integrating discharge observation uncertainty and temporal sampling uncertainty into hydrological model evaluations to ensure accurate, reliable, and meaningful assessments of model performance. Implementing our proposed methodology in reporting practices is expected to improve the robustness of hydrological model result interpretation, aiding in more informed model selection and refinement decisions by model users.

Furthermore, addressing uncertainty on a global scale is of paramount importance. Accurate global assessments of discharge uncertainty are critical for informing water management strategies, policy decisions, and climate impact studies. Understanding and mitigating these uncertainties can lead to more reliable hydrological models and better resource management worldwide. This study provides a foundation for such global assessments by demonstrating the necessity of incorporating both discharge observation and temporal sampling uncertainties into hydrological evaluations.

SUPPLEMENTARY ANALYSES

A.1 WFLOW_SBM CALIBRATION

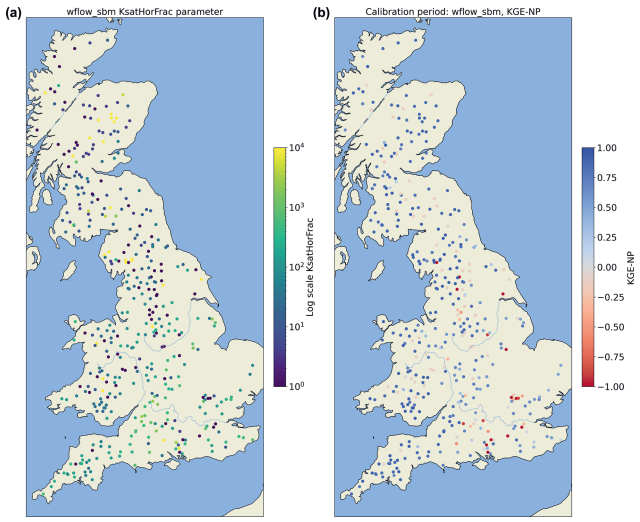


Figure 5.8: (a) Spatial distribution of the best-performing KsatHorFrac calibration parameter of the wflow_sbm model based on additional calibration using discharge observations. (b) Spatial distribution of the KGE-NP objective function based on the calibration period of the wflow_sbm model.

A.2 NSE-BASED MODEL PERFORMANCE RESULTS

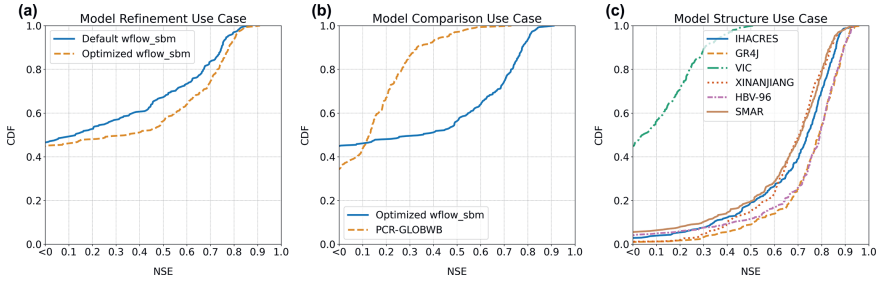


Figure 5.9: Cumulative distribution function plots of the Nash–Sutcliffe efficiency (NSE), derived from discharge estimates and observations at 299 catchment outlets. (a) The model refinement use case, optimizing the wflow_sbm hydrological model with a single parameter. (b) The model comparison use case, comparing the optimized wflow_sbm and PCR-GLOBWB hydrological models. (c) The model structure use case, showing results from 6 conceptual hydrological models.

6

SYNTHESIS

Remember kids, the only difference between screwing around and science is writing it down.

Adam Savage, *MythBusters*

6.1. INTRODUCTION

This dissertation is motivated by multiple open challenges and opportunities in hydrological modeling, specifically the need for reproducible workflows, large-sample catchment-based model experimentation, multi-variable diagnostics, and a pragmatic treatment of observational uncertainty. Chapter 1 introduces the broader scientific and societal rationale for reproducible hydrological modeling, while Chapter 2 develops a conceptual and technical framework for lowering reproducibility barriers through the eWaterCycle platform. Chapters 3 and 4 advance methodological innovations by testing how model complexity, spatial resolution, and multi-variable performance metrics interact in large-sample settings. Finally, Chapter 5 provides an empirical examination of observational uncertainty and its influence on model evaluation and benchmarking.

This synthesis brings together the core insights from each chapter, showing how these advances collectively contribute to a more rigorous, transparent, and robust hydrological modeling science.

The chapters of this dissertation can be grouped into three cross-cutting themes. The first theme, centered in Chapter 2, establishes the importance of reproducibility and open science in hydrology, leading up to the development of the eWaterCycle platform to facilitate multi-model experimentation. The second theme, spanning Chapters 3 and 4, focuses on large-sample catchment analyses and advanced diagnostic techniques, investigating how spatial resolution, scaling, and data-driven methods connect model performance to catchment attributes. The third theme, explored in Chapter 5, critically

examines observational uncertainty in discharge data and demonstrates how neglecting these uncertainties risks overstating or misinterpreting differences in model performance.

A unifying thread across these themes is the question of model selection. Each chapter contributes insights and practical guidance for selecting, refining, and applying hydrological models in ways that align with the research question at hand. By emphasizing reproducible workflows, the diagnostic power of large-sample catchment analyses, and the central role of observational uncertainty, this dissertation outlines key principles to inform both model choice and iterative model improvement across diverse hydrological contexts.

6.2. SUMMARY OF MAIN FINDINGS

REPRODUCIBILITY THROUGH THE EWATERCYCLE PLATFORM (CHAPTER 2)

A recurring challenge in hydrology is that model experiments are difficult to reproduce. Historically, fragmented code, proprietary data formats, and hidden dependencies made it nearly impossible to rerun or extend earlier studies. Chapter 2 addresses these technical barriers through the development of the eWaterCycle platform. The platform tackles these issues through a combination of the Basic Model Interface, containerization, automated preprocessing of meteorological forcing, and Jupyter notebook-based experimentation.

These are technical solutions to technical problems. The Basic Model Interface ensures that different models can be called in the same way, regardless of their original programming language. Containerization packages each model together with its libraries and dependencies, making the environment portable across machines. Automated preprocessing routines simplify the preparation of meteorological forcing datasets, allowing different gridded inputs, such as ERA5 or CMIP6, to be swapped in easily, thereby facilitating experimentation. Notebook-based experimentation ensures that code, parameters, intermediate outputs, and results are fully logged and documented.

By demonstrating five example experiments, from a simple "Hello World" setup to coupling models across languages, Chapter 2 shows how these components work together in practice. The platform promotes collaboration by enabling researchers to build on each other's code, datasets, and workflows, reducing the need to reinvent standard routines or processing steps. It facilitates scaling from single-catchment studies to large-sample experiments by automating model setup and standardizing workflows, thereby lowering the barrier for multi-model and multi-catchment analyses. It also ensures reproducibility by packaging each experiment's software environment, making it easier for others to rerun or adapt earlier work.

In short, Chapter 2 shows that reproducibility can be practiced without sacrificing flexibility. By removing technical hurdles, eWaterCycle creates space to focus on the hydrological questions themselves rather than the logistics of running the models. Chapter 2's case studies illustrate how these technical elements come together to support a wide range of research questions that range from routine catchment simulations to online coupled model experiments.

LARGE-SAMPLE ASSESSMENTS OF SPATIAL RESOLUTION (CHAPTER 3)

In Chapter 3, we tackled a major question in modern hydrology: the value of hyperresolution modeling. As computational power grows, there is a push toward finer spatial

grids under the assumption that higher resolution may yield more accurate or physically realistic outputs. Yet whether hyper-resolution truly improves catchment-scale streamflow predictions, and at what cost, remains unclear.

To investigate, we ran the wflow_sbm model on 454 CAMELS-US basins at three spatial resolutions: 3 km, 1 km, and 200 m. By calibrating each resolution against observed discharge, we compared the Kling-Gupta Efficiency (KGE) distributions to test if finer grids systematically improved performance at the catchment outlet. The key findings were:

- Modest overall performance gains. Across the full catchment sample, the performance difference between 1 km and 3 km, or between 200 m and 3 km, was typically small. This shows that "zooming in" does not automatically yield better hydrographs once aggregated to the outlet, and it suggests caution when interpreting high-resolution model outputs.
- Local improvements masked by large-sample averages. Some catchments did show substantial gains from finer resolution, particularly in areas with complex topography or specific land cover dynamics. However, these local improvements were averaged out across the sample, resulting in only limited overall gains in the KGE distribution.
- Uncertainty from forcing and sampling. When we incorporated temporal sampling uncertainty using bootstrap methods, the differences among resolutions often fell within confidence intervals. This finding reinforces that hyper-resolution does not consistently outperform coarser grids when observational data and parameter knowledge are uncertain, and it cautions against overinterpreting performance metrics based on single time series evaluations.

Chapter 3 concludes that finer grids do not inherently lead to more robust streamflow predictions, particularly when observational or forcing uncertainties are high or when parameter estimation is limited at small scales. These findings encourage researchers to balance computational costs, data availability, and the specific study objective, such as flood delineation versus annual water balance estimation, before adopting hyper-resolution approaches. In practical terms, investing heavily in computational resolution while the observational basis remains sparse offers limited returns. The gains from refining a grid from 3 km to 200 m are modest and often statistically indistinguishable, whereas a comparable investment in denser or more diverse observations could provide the constraints needed to actually improve process representation. Without supporting data, higher resolution does not translate into better models.

EXPLORATORY MODEL DIAGNOSTICS VIA LARGE-SAMPLE AND MACHINE LEARNING (CHAPTER 4)

Chapter 4 shifted focus from spatial scaling to a broader question. Why do some hydrological models perform better in certain catchments, and can we detect systematic links between catchment attributes and model performance?

While previous studies often relied on linear correlations, we applied a Random Forest approach to uncover potential non-linear or complex interactions. Using CAMELS-GB data from 617 catchments, we evaluated the wflow_sbm and PCR-GLOBWB models

across multiple variables, including streamflow, actual evaporation, and soil moisture. Regression Random Forest models were constructed with 93 catchment attributes covering topographic, climatic, land cover, subsurface, and hydrologic signatures. The key findings were:

- Catchment area as a dominant predictor. Catchment area consistently emerged as a key variable across several targets, such as streamflow KGE, model differences, or evaporation skill. For wflow_sbm, smaller catchments tended to show higher performance, while PCR-GLOBWB performed better in larger basins, reflecting the original design scales of each model.

- Subsurface and land cover effects. Variables such as porosity and urban fraction were among the most important predictors of differences in streamflow or evaporation performance, suggesting that improvements in soil parameterization or land cover representation could lead to measurable model gains.

- Flux-specific patterns. The importance of predictors shifted depending on whether we analyzed streamflow, evaporation, or soil moisture. This shows that there are no universal key attributes and that the dominant processes or parameters depend on the flux under investigation.

Chapter 4 demonstrated that exploratory diagnostics can point to concrete model improvements. For example, the link between poor soil moisture predictions and topsoil or slope attributes suggests a need for refined soil layering or slope-driven lateral processes. More broadly, this chapter showed how machine learning can complement physically based modeling by diagnosing structural weaknesses and systematically identifying targets for improvement.

INTEGRATING DISCHARGE OBSERVATION UNCERTAINTY (CHAPTER 5)

One of the most critical yet often overlooked factors in hydrological evaluation is the uncertainty in the observations themselves. Chapter 5 focused on discharge observation uncertainty, particularly errors arising from rating curve inaccuracies. Using 299 catchments from CAMELS-GB, the chapter examined how this uncertainty can complicate model evaluation:

- Model refinement. We tested how often a newly calibrated parameter (the horizontal conductivity fraction in wflow_sbm) produced improvements in streamflow KGE that fell within the 5th to 95th percentile bounds of observational error. In roughly 33% of cases, improvements that appeared meaningful were actually within the observational uncertainty margin, cautioning against overinterpreting calibration gains.

- Conventional model comparisons. When comparing wflow_sbm and PCR-GLOBWB, we found that in some basins, the difference in performance was effectively obscured by discharge observation uncertainty. In others, models diverged clearly beyond the error margin.

- Model structure uncertainty. Comparing six conceptual models (IHACRES, GR4J, VIC, XINANJIANG, HBV-96, SMAR), we found that large structural differences were often larger than observation error. Yet in some cases, structurally distinct models produced near-identical results, meaning that observational uncertainty masked the differences.

- Temporal sampling. Using a bootstrapping approach similar to Chapter 3, we discovered that a small number of events or time steps disproportionately influenced performance metrics like KGE, further complicating claims of significant difference.

Together, these findings led to a practical method for comparing absolute simulation differences (in m^3/s) with the estimated uncertainty margin of the discharge observations. The overall takeaway is that ignoring observational uncertainty risks overstating claims of improvement or difference between models, especially when differences are small, or models perform similarly. A particularly interesting result was the absence of a clear relationship between model complexity and model performance among the six conceptual models tested. This finding challenges the assumption that more complex models are inherently more capable. If adding process detail does not reliably improve performance, it is worth considering whether the effort invested in increasing structural complexity might, in some cases, be better directed toward improving the observations against which all models are judged.

6.2.1. CROSS-CUTTING THEMES

REPRODUCIBILITY MEETS LARGE-SAMPLE CATCHMENT HYDROLOGY

A recurring theme throughout this dissertation is the interplay between reproducibility (Chapter 2) and large-sample experimentation (Chapters 3 and 5). The two are mutually reinforcing: reproducibility makes large-sample experiments feasible, and large-sample studies demonstrate the value of reproducible workflows. The eWaterCycle platform supports this connection by offering a standardized model experimentation framework with systematic documentation, allowing:

- Standardizing input/output: By ensuring that each model instance uses near-identical forcing scripts and data preprocessing steps, the platform reduces hidden discrepancies and prevents configuration drift across experiments, eliminating repetitive manual setup. This lowers the effort required for hypothesis testing and model exploration. Under these workflows, large-sample catchment evaluations become more reliable because differences in outcomes are more likely to reflect real model differences rather than inconsistencies in data handling.

- Facilitating community contributions: eWaterCycle simplifies the process of adding new models and sharing results. This encourages collaboration by making it easier for researchers to build on each other's work and exchange reproducible workflows. This fosters a research environment where collective efforts from multiple groups drive both methodological and theoretical progress, benefiting the wider community.

At the same time, large-sample experimentation increases the value of reproducible workflows by enabling broader insights beyond single case studies. Chapter 3's analysis of spatial resolution across hundreds of basins would have been difficult to conduct without a reproducible way to configure and calibrate each model resolution. Chapter 4's machine learning diagnostics across 617 catchments show how reproducibility makes it possible to apply similar analyses in other regions or with other models. In this way, reproducibility not only supports replication but enables scaling, transferability, and continuous refinement of experiments across space, time, and models. In this sense, reproducibility is not just about repeating a study; it also means creating workflows that can be applied to different regions and models to address new hydrological questions.

BALANCING COMPLEXITY, DATA AVAILABILITY, AND PURPOSE

Chapters 3 to 5 highlight the importance of aligning model complexity with both the available data and the purpose of the modeling exercise. While high-resolution approaches (Chapter 3) offer finer spatial detail, their ability to improve streamflow predictions may be limited by uncertainties in forcing data or parameter estimates. Similarly, more complex models may introduce additional parameters that cannot be reliably constrained by the available observations, increasing the risk of overfitting. Likewise, Chapter 5 shows how observational discharge uncertainty can mask what appear to be improvements in model performance.

This leads to the principle of parsimony, also known as Occam's razor, that states that among competing explanations or models that achieve similar outcomes, the simplest one is preferred. In hydrological modeling, this translates to using the least complex model structure necessary to adequately answer the research question or achieve the required predictive performance. A simpler model is not inherently better, but complexity should only be added when supported by data and justified by purpose.

In practice, the optimal balance of complexity often involves:

- Catchment-specific complexity: Reserving detailed parameterizations for processes and regions where sufficient data exist to constrain and validate them. For example, a finely discretized groundwater module might be necessary in aquifer-dominated catchments but could introduce unnecessary computational overhead in regions with ephemeral (sub)surface flow. This selective use of complexity ensures that computational resources and data collection efforts are targeted where they matter most.

- Understanding input uncertainties: Recognizing that performance metrics like KGE or RMSE may fail to reflect real gains in physical accuracy if observational uncertainties or input errors dominate. As shown in Chapter 5, what appears as a performance improvement can lie within the bounds of observational uncertainty error. Therefore, adding complexity does not guarantee better realism if input or observational errors remain large.

- Purpose: Distinguishing between research-driven models aiming for detailed process understanding (e.g., high-resolution slope-runoff interactions) and operational

models oriented toward robust forecasts or water resource planning. The former may justify higher complexity, while the latter might prioritize computational speed and reliability. This distinction helps clarify whether the goal is explanation, prediction, or decision support, each requiring different trade-offs.

Thus, matching the model to the question becomes a central theme. Sometimes, simpler conceptual models can be more transparent and easier to calibrate. In other cases, fine-scale detail may be essential, such as understanding floodplain inundation or snowmelt processes. Ultimately, complexity should be added deliberately, guided by both the research objective and the diagnostic insights provided by large-sample, multi-objective evaluations. Selecting the appropriate level of complexity depends on the user's objectives, the available data, and the level of uncertainty that can be tolerated.

Achieving this balance requires looking beyond the models themselves. As argued in the Introduction (Section 1.4), the nonlinearity of hydrological processes means that effective parameters at any given scale require observations at that scale. A model can only be as complex as the available data allow it to be constrained. This implies that advancing hydrological modeling is not solely a matter of building better models or finer grids; it equally depends on investing in new observations.

To resolve processes at finer scales, the hydrological community needs observations that bridge the gap between point-scale field measurements and coarse satellite footprints. Technologies such as cosmic-ray neutron sensing, which estimates soil moisture at the intermediate hectare scale, and satellite-derived evapotranspiration products, which constrain water balance partitioning at catchment scales, show that this gap can be narrowed from both ends. Without such investments in data, increasing model complexity risks producing models that lack the observational support to distinguish between competing parameter sets, reinforcing rather than resolving the equifinality problem.

MULTI-FLUX, MULTI-OBJECTIVE EVALUATIONS

While streamflow remains the primary benchmark in most hydrological assessments, Chapter 4 shows that including additional fluxes and state variables, such as evaporation or soil moisture, can further support model refinement. For instance, a model that performs well in predicting daily runoff might still exhibit systematic biases in soil moisture, leading to errors in drought forecasting. Chapter 4 demonstrates how large-sample, multi-flux comparisons can highlight persistent weaknesses tied to specific land cover or subsurface attributes for the wflow_sbm and PCR-GLOBWB models.

Relying solely on streamflow can mask compensating errors in internal states, leading to solutions that reproduce hydrographs but misrepresent the underlying processes. By incorporating multiple fluxes into evaluation, it becomes harder for incorrect process assumptions to remain hidden behind a good streamflow fit, thereby strengthening the falsifiability of the model's hypotheses. Expanding the evaluation method, however, requires more than additional data. It demands robust methods for quantifying and integrating the uncertainty of each flux or state variable.

Chapter 5's exploration of discharge observation uncertainty reinforces this point by showing that even for a relatively well-monitored variable like streamflow in Great

Britain, rating-curve errors can mask meaningful differences in model performance. Extrapolating this finding to less frequently observed variables, such as groundwater level or actual evapotranspiration, emphasizes the need for improved observational networks, new remote sensing retrievals, and standardized protocols for estimating measurement errors. A further challenge is that some fluxes, like evaporation estimates, are themselves derived from models or reanalyses, complicating their role as fully independent benchmarks.

Ultimately, multi-objective evaluations offer a more comprehensive view of hydrological system dynamics but require consistent, high-quality datasets and transparent error characterization. By pairing eWaterCycle's reproducible environment with emerging data assimilation approaches, researchers can systematically test how well models capture a range of hydrological processes, from storm runoff peaks to seasonal soil moisture deficits. Moreover, the patterns of model mismatch identified across multiple fluxes and catchments can guide targeted field measurements, by pointing to the locations, times, and variables where observations would most help to falsify or refine specific process representations.

MODEL SELECTION

This brings us to the final theme of this synthesis: How should one choose (and refine) a model (architecture) in light of specific scientific or operational goals? The findings of all chapters answer this question by highlighting:

- Research objective and scale: In Chapter 3, hyper-resolution grids did not consistently outperform coarser ones for basin-scale outflows, reminding us that fine detail is not inherently better. Researchers must ask: is the focus on capturing localized processes (e.g., small-scale infiltration) or broader river basin dynamics (e.g., cumulative water balance)?

- Data requirements vs. data availability: More complex models often need correspondingly dense and accurate datasets. Chapters 4 and 5 show that parameter uncertainties and observational errors can determine the benefits of complexity, suggesting that simpler models may be more robust when data are sparse or uncertain.

- Process representation vs. accuracy: Chapter 4's machine learning approach demonstrates that these techniques can diagnose which catchment attributes drive performance differences. This insight can guide when to introduce more detailed process representations. The decision depends on whether the ultimate aim is purely predictive skill or also includes a deeper understanding of hydrological processes.

- Uncertainty integration: As underscored in Chapter 5, factoring in uncertainty from all inputs, meteorological forcings, and parameter sets is crucial when weighing model differences. A model might appear better performing but might be indistinguishable from its competitor once uncertainties are considered.

In this sense, model selection is a continuous, iterative process rather than a one-time choice. Researchers and practitioners can use reproducible frameworks (Chapter

2) to test multiple hypotheses, calibrate models under varying assumptions and target variables, and systematically compare results under uncertainty scenarios. As new observational datasets and measurement techniques emerge, and as computational capacities expand, the selected model can be further refined or even replaced by a structure better suited to the evolving research context. As a result, model selection should increasingly move from being an "art," guided by familiarity, tradition, or subjective preference, toward a more evidence-based practice. The tools developed and applied in this thesis, from reproducible multi-model frameworks to machine-learning diagnostics and uncertainty-aware evaluation, provide the means to make model selection more systematic and defensible. This does not eliminate the need for expert judgment, but it does mean that such judgment can now be informed by large-sample evidence and transparent diagnostics rather than relying on the pragmatic realities of data availability alone, all while maintaining a clear line of sight to the questions being asked and the uncertainties inherent in both measurements and predictions.

6.3. MAIN CONCLUSIONS

This thesis demonstrates how reproducible workflows, large-sample catchment hydrological experimentation, and careful treatment of observational uncertainty can collectively advance the discipline of hydrological modeling. While each chapter tackles different facets of these broad objectives, several overarching conclusions emerge:

1. **Reproducibility is fundamental to progress.** Without reproducible workflows, model experiments cannot be reliably extended, compared, or validated across studies. Chapter 2 underscores the necessity of transparent, open-science practices for enabling scientific collaboration. The eWaterCycle platform showcases how containerization, consistent data preprocessing, and clear documentation can greatly simplify multi-model comparisons and large-sample catchment studies. By embedding reproducibility into the experimental pipeline, the platform enables faster iteration, wider participation, and more durable scientific insights. It directly addresses longstanding critiques about the difficulty of reproducing hydrological experiments, setting the stage for broader collaboration and accelerated methodological innovation.

2. **Large-sample testing reveals generalizable patterns.** Chapters 3 to 5 highlight the value of examining model performance across hundreds of catchments rather than focusing on a single case study. This approach uncovers systematic trends, for example, how hyper-resolution sometimes yields only marginal discharge improvements (Chapter 3), or how catchment attributes (e.g., area, porosity) drive performance differences (Chapter 4). By expanding beyond individual basins, large-sample experiments expose which processes and parameters are consistently influential, and which are context-dependent. Large-sample catchment and multi-model experiments thus provide a more reliable basis for inferring which model structures or parameterizations are most universally effective.

3. **Observational uncertainty can obscure apparent model improvements.** A central finding from Chapter 5 is that even modest levels of measurement error, particularly in discharge, can obscure improvements from recalibrated models. Neglecting observation uncertainty risks overstating model skill or falsely favoring one structure over another. This emphasizes the need for explicit uncertainty quantification covering both forcing data (e.g., precipitation, temperature) and model outputs (e.g., streamflow, soil moisture) and including this in the calibration process to avoid overstating differences among models or concluding that a given model structure is "superior" when the difference may lie within the bounds of observational error. The notion that multiple model configurations are equally acceptable (equifinality) is better understood as a symptom of insufficient observational constraints. If multiple hypotheses survive testing, the response is not to accept all of them, but to seek the observations that can distinguish between them.

4. **Balanced complexity can outperform high complexity.** While high-resolution may seem appealing, Chapters 3 and 5 show that gains in predictive skill can be subtle and affected by data limitations. Chapter 4 further illustrates how machine learning approaches can expose where detailed process refinements might pay off and where simpler conceptual schemes already suffice. This reinforces the principle of parsimony: that increasing model complexity is only justified when supported by sufficient data and a clear added value for the modeling objective. The "best" model is not always the most complex; rather, it is the one that aligns process representation, available data, and research or management objectives in a parsimonious manner. At the same time, this also means that collecting more and better observations is the most direct way to justify and support greater model complexity. Where data are lacking, parsimony is a practical necessity. Where targeted observations can be obtained, more detailed process representations become defensible.

5. **Model selection is iterative.** Throughout the dissertation, model selection emerges as an iterative process, sensitive to the scale of study, data availability, and the nature of the posed hydrological questions. Rather than choosing a model once, selection and refinement must adapt as new data, questions, or constraints arise. By integrating reproducible frameworks (Chapter 2) with large-sample catchment diagnostic techniques (Chapters 3–5) and uncertainty assessments (Chapter 5), researchers can more systematically evaluate whether a given model is fit-for-purpose or whether an alternative, simpler, or more data-driven approach might be advantageous.

Combined, these conclusions suggest that future hydrological modeling efforts stand to benefit significantly from embedding reproducibility and large-sample comparisons into standard practice, explicitly integrating uncertainty in both calibration and evaluation, and tailoring model complexity to the quality and extent of available data. By embracing these principles, hydrological modeling can move toward a more transparent, adaptable, and decision-relevant science, one better equipped to inform water resource management, flood forecasting, and socio-environmental decision-making in an era of rapid global change.

7

DISCUSSION

Good food is very often, even most often, simple food.

Anthony Bourdain, *Kitchen Confidential*

7.1. TYPES OF MODEL USERS AND MODEL SELECTION

The main conclusion of this dissertation is that model selection must be guided by the study's purpose, the quality and availability of data, and the practical constraints faced by different user groups. In practice, these user groups can be broadly categorized into research-focused and practitioner-focused profiles. During the design of eWaterCycle, personas were used to ensure that different user needs were considered when building the platform. In this framework, professors and PhD candidates represent academic and research-oriented users who may need the flexibility to explore multiple model structures (Chapters 2–4), while consultants and water agencies exemplify more practitioner-focused users working under tight deadlines and often favoring robust and efficient workflows (Chapter 5).

In terms of model selection, Chapters 3 and 4 demonstrated that data availability and quality can and should determine the choice of spatial resolution or process complexity. For instance, in Chapter 3, high-resolution modeling did not consistently outperform coarser grids when forced by uncertain or coarse meteorological inputs. This finding should guide researchers who might otherwise assume that "finer is better." Chapter 4 uses a random forest approach to show that performance differences among models can be linked to basin characteristics (e.g., climate, topography, and soil type). Consequently, for a practitioner with limited resources, a moderately complex model might be more practical than a highly complex one. A PhD candidate interested in exploring hydrological process representations may still justify a more detailed model if sufficiently fine-scale, high-quality data are available.

In practice, operational constraints and model legacy frequently lead practitioners to continue using familiar models Addor and Melsen, 2019. Several barriers reinforce this status quo. Technically, proprietary data formats, restrictive IT rules, and limited computing resources make it difficult to install and run unfamiliar software. Organizationally, tight project deadlines further encourage staff to rely on established workflows. At the policy level, many agencies are restricted to pre-approved packages and must consider legal liability, which discourages experimentation. Educationally, staff training and university curricula remain anchored in older software, so even when better options emerge, many practitioners lack the capacity to adopt them.

This dissertation shows how eWaterCycle can help overcome some of these barriers by reducing the effort required for model experimentation (Chapter 2). Because the platform relies on containerization and the Basic Model Interface (BMI), practitioners can test new models without discarding their existing workflows. This promotes model selection based on suitability rather than legacy.

For model selection, large-sample catchment experimentation (Chapters 3–5) and random forest diagnostics (Chapter 4) provide systematic ways to evaluate how model performance varies across catchments and data conditions. These findings can guide both researchers and practitioners in selecting models that align with data availability and the hydroclimatic setting of their catchments. In this sense, the dissertation shows that model selection is not merely a theoretical decision but is also driven by pragmatic considerations. These considerations include operational needs, the reliability of input data, and the purpose of the study or intervention. By integrating reproducible frameworks such as eWaterCycle with robust diagnostic methods, hydrologists and water managers can better match the right tool to the right context instead of matching the context to the tool.

Education plays an important role in this transition. If the next generation of hydrology students is trained on open, multi-model platforms such as eWaterCycle from the start, evidence-based model selection will gradually replace legacy-driven practice. As discussed further in Section 7.3, Delft University of Technology already embeds eWaterCycle in its MSc curriculum; scaling this approach to hydrology programs worldwide could shift community norms within a decade.

7.2. MODEL PURPOSE: RESEARCH TOOL VERSUS PRACTICAL TOOL

Similar to the different types of model users, hydrological models can be classified according to their primary purpose: as research tools for expanding scientific understanding or as practical tools for reliable operational water management. Beneath these differences lies a philosophical distinction. From a research perspective, models function as means to test and refine our understanding of hydrological processes. In an operational setting, models are tools for immediate decision-making that depend on robustness and clarity. Recent AI and machine-learning models blur this distinction. These models excel

at rapid, data-driven prediction for operational tasks but, when used as black-box systems, contribute little to mechanistic understanding unless they are coupled with physical representations. Consequently, no single "workhorse" model fully satisfies the different needs of researchers and practitioners. A model used for day-to-day operations does not necessarily need to include all the processes relevant to understanding long-term catchment dynamics.

The notion of a universal model is also in tension with the "reality of locality" argued in this thesis (Section 1.4). Every catchment presents a different combination of processes, scales, and data constraints, making it unlikely that a single model can serve all purposes and locations. Rather than pursuing a universal workhorse model, the community would benefit from investing in flexible frameworks that allow models to be adapted, compared, and selected based on the specific context. Modular platforms (e.g., RAVEN and SUMMA) already offer steps in this direction. Researchers can toggle process descriptions (hypotheses) on or off to test competing hypotheses. Over time, this flexible approach may help diminish the distinction between "research" and "operational" models by allowing users to adjust the level of complexity based on the research question or operational goal. Large-sample catchment experiments further support this approach because they enable us to examine how models of varying complexity perform across diverse catchments, highlighting where detailed processes add value and where simplified parameterizations are sufficient.

There is also a risk that the hydrological community is overinvesting in machine learning for prediction at the expense of using it for diagnosis. As Chapter 4 demonstrates, methods such as random forests can reveal which catchment attributes drive model failure and where process representations need refinement; these are insights that predictive skill alone cannot provide. Yet the majority of machine-learning applications in hydrology focus on replacing process-based models rather than interrogating them. A rebalancing toward diagnostic applications of machine learning would yield more lasting scientific value.

7.3. REPRODUCIBILITY IN HYDROLOGICAL SCIENCES

Progress in hydrological modeling depends on workflows that can be reliably reproduced. The eWaterCycle platform facilitates this by allowing users to replicate, adapt, and extend earlier studies while addressing key technical barriers such as fragmented code, incompatible data formats, and hidden software dependencies. By removing these obstacles, the platform allows users to rerun and adjust existing experiments with minimal configuration effort, transforming what was once a time-consuming task into a routine part of the research process.

Nevertheless, adoption remains the platform's greatest challenge, and it would be dishonest not to acknowledge this. Concepts such as software containers and the Basic Model Interface require technical knowledge that many hydrologists lack, and the initial effort required to install and configure the environment can deter potential users before they experience its benefits. Including new models requires adapting them to conform to the BMI standard, which, while conceptually straightforward, demands developer time that model authors may not be willing or able to invest.

These barriers are real, but they are also likely to diminish. The broader trend in scientific computing favors cloud-based environments over local installations, with containerization increasingly becoming the default rather than the exception. Moreover, rapidly developing assistants based on large language models are making high-level programming interfaces, such as the one eWaterCycle offers through BMI, increasingly accessible. Whereas users currently need to understand Docker and Python to interact with the platform, natural-language interfaces may soon allow researchers to configure, run, and analyze model experiments through conversational interaction. In this sense, eWaterCycle's architecture is well positioned for a future in which AI-assisted workflows substantially lower the technical barriers that currently limit adoption.

In the meantime, education offers the most direct path to overcoming these hurdles. By embedding eWaterCycle in coursework, future hydrologists acquire open-science habits from the outset and come to regard reproducible, shareable workflows as the norm rather than the exception. Delft University of Technology already follows this approach, using eWaterCycle in two MSc courses so that students gain hands-on experience in running benchmark experiments, exchanging notebooks, and extending existing studies.

In the long run, eWaterCycle's longevity will depend on widespread community uptake, stable funding, and a shared commitment to open data. The platform offers a feasible blueprint for reproducible and collaborative modeling. Its modular software architecture can accommodate new data sources, new models, and evolving user needs. By lowering technical barriers to reproducibility and demonstrating the benefits presented in this thesis, eWaterCycle and similar tools have the potential to transform hydrology over the long term.

7.4. TESTING MODEL HYPOTHESES WITH FIELD MEASUREMENTS

From a Popperian perspective, hydrological models can be viewed as sets of falsifiable hypotheses about how catchments function. These hypotheses may concern whether a model emphasizes infiltration-excess runoff, saturation-excess runoff, or complex subsurface flow. If two models with different representations of soil-water movement show comparable performance but different internal states under limited data, they may contain untested assumptions about how water moves through soils. Large-sample catchment comparisons, such as those presented in Chapters 3 and 4, help identify more precisely where and when model performance diverges. These comparisons can therefore suggest which field measurements are needed to distinguish which model is closer to reality in a given region or climate.

A targeted measurement strategy is fundamental to the scientific method. A hypothesis that cannot be tested, or one for which the necessary data will forever remain out of reach, cannot be falsified. For instance, a model may hypothesize that infiltration processes dominate storm runoff in certain soils, yet the soil properties used to parameterize these processes may never be measured in situ and may instead be estimated using pedotransfer functions. The hypothesis then remains untested. Therefore, focused field campaigns are necessary to confirm or falsify specific process representations. Designing these campaigns based on large-sample catchment insights not only optimizes

resource allocation but also ensures that the “small data” from field observations address the most important uncertainties identified by “big data” modeling results.

The financial and logistical challenges of fieldwork are real, but they should not serve as an excuse for complacency. Other measurement-intensive disciplines have shown that data scarcity is not a permanent condition but an obstacle that can be overcome through sustained technological investment. In meteorology, systematic investment in radiosonde networks, weather radar, and satellite constellations over the past century transformed weather forecasting from an empirical exercise into one of the most successful applications of physics-based modeling. The cost of many sensing technologies has dropped dramatically over the past decades, and hydrology should pursue a comparable ambition rather than resigning itself to sparse observations. Techniques such as distributed temperature sensing (DTS) can resolve subsurface flow variability at meter-scale resolution along fiber-optic cables, while stable-isotope tracers provide independent constraints on water transit times and flow pathways that discharge measurements alone cannot offer. These tools exist and have proven their value in individual studies, yet they remain far from routine practice.

Equally important, and often overlooked, is the need to rehabilitate and maintain existing gauging-station networks. Many countries have seen a decline in the number of operational discharge and meteorological stations over recent decades, precisely when large-sample hydrological studies, such as those in Chapters 3–5, depend on long, continuous, quality-controlled records. Restoring and expanding these networks is essential. Without reliable observations, no amount of model improvement can produce trustworthy results.

Moreover, large-sample comparisons can miss outlier processes, such as karst drainage or wetland dynamics, that require localized data collection. Failing to measure these processes in a way that tests model predictions prevents effective model falsification. Therefore, any hypothesis that cannot be tested using direct field instrumentation or remote-sensing measurements is of limited value to hydrology. A truly robust hydrological science requires an iterative loop in which models guide measurement strategies and field data refine or reject model assumptions.

7.5. CAN MODELING IMPROVE HYDROLOGICAL UNDERSTANDING?

A continuing debate in hydrology is whether understanding can be gained through modeling efforts or only by improving perceptual models based on measurements. This dissertation argues that model experiments can, in some cases, improve understanding.

A more critical question is why different model structures (hypotheses) can achieve similar performance metrics. Data and computational limitations make it inevitable that processes such as preferential flow pathways are omitted, while calibration may simply tune parameters to match discharge without adding physical realism. The result can be an illusion of understanding in which apparent skill is driven more by parameter optimization than by correct process representation.

Yet the work in this dissertation shows that model-based experimentation can still advance understanding, provided that the hypotheses embedded in each model are confronted with sufficiently rich and independent observations. Innovative data sources such as commercial microwave-link rainfall estimates, stable-isotope tracers, video-derived discharge estimates, and discharge estimates derived from satellite altimetry should therefore play a more prominent role. These measurements provide additional, partly independent constraints on states and fluxes that were previously unobservable at scale, reducing the freedom of competing hypotheses to remain indistinguishable when evaluated against discharge alone.

A model that deviates substantially from sufficiently reliable observations cannot be considered an adequate representation of the system; at least one of its underlying hypotheses must therefore be rejected. In single-catchment, single-model studies, disentangling all the hypotheses embedded in a model is difficult, which hinders progress in understanding. Large-sample datasets and multi-model, multi-catchment studies conducted with identical, reproducible workflows (Chapters 3 and 4) make it possible to identify which hypotheses fail systematically and under which hydroclimatic conditions. This, in turn, guides both future field campaigns targeting the variables and sites that matter most and iterative model improvement.

When multi-year streamflow records are evaluated alongside remote-sensing products such as soil moisture retrievals, evapotranspiration estimates, or discharge estimates derived from satellite data, a model that reproduces these signatures across diverse catchments earns more credible support for its underlying process descriptions, even though the true system remains uncertain. The systematic mismatches exposed by large-sample, multi-model analyses isolate the process descriptions that require revision, steering hydrological science toward more clearly falsifiable hypotheses.

Realizing this potential, however, requires bridging a persistent cultural divide between modelers and field scientists. In practice, many modelers rarely engage in fieldwork, while experimentalists are often disconnected from the models their data are intended to constrain. Closing this gap demands deliberate institutional effort. Joint field and modeling campaigns, in which modelers help design measurement strategies and experimentalists participate in model evaluation, could foster mutual understanding of what each group needs. Graduate programs that require students both to collect field data and to run models would cultivate researchers who can work across this divide. Platforms such as eWaterCycle can support this integration by making it straightforward to confront model predictions with new field observations as they become available, turning the model-measurement feedback loop from an aspiration into a routine part of hydrological research.

8

OUTLOOK

Not all those who wander are lost.

J. R. R. Tolkien, *The Fellowship of the Ring*

As hydrological modeling advances, one of the most promising paths forward lies in building stronger collaborative networks, sharing data and results more openly, and creating reproducible workflows that accelerate collective learning. This thesis has shown that reproducibility (Chapter 2), large-sample experimentation (Chapters 3–5), and the explicit treatment of observational uncertainty (Chapter 5) are central to making hydrology more robust and transparent. The next step is to ensure that these practices scale beyond individual projects and become embedded in community-wide infrastructure.

A first priority is the establishment of open-access repositories that share not only raw datasets but also model results, parameter sets, metadata, and evaluation metrics in standardized formats. Such repositories would allow researchers to compare models across regions, systematically assess which model structures generalize well, and identify where region-specific refinements are needed. Platforms such as eWaterCycle already provide the computational framework for reproducible experiments. A natural extension is to connect these platforms to public repositories of model experiments so that outputs, scripts, and calibrated models can be reused, modified, and built on across research groups. By doing so, the community could gradually assemble a mosaic of regional models covering diverse hydroclimatic zones while acknowledging that no single model fits all conditions.

Beyond sharing results, collaborative networks must also focus on harmonizing data protocols, metadata standards, and evaluation procedures. Only then can large-sample catchment studies be reliably compared across countries and datasets, thereby avoiding hidden inconsistencies in data processing or model implementation. Strengthening these collaborative and technical foundations is critical for scaling insights from individual studies to global hydrological challenges.

A second avenue for progress is expanding the types of observations used to constrain and evaluate models. This dissertation demonstrated that uncertainty in discharge measurements (Chapter 5) can mask differences between models, limiting confidence in performance claims. A wider set of hydrological observations, such as soil moisture, evapotranspiration, groundwater levels, and snow cover, can provide additional, partly independent checks on model states and fluxes. Multi-objective evaluations that integrate multiple fluxes (Chapter 4) can better isolate which model components need refinement. Investments in observational networks, standardized uncertainty quantification, and shared datasets will therefore directly support more effective model falsification and improvement.

Satellite-based observations, in particular, offer untapped potential for hydrology. Missions such as GRACE-FO provide terrestrial water storage anomalies at basin scales, while SWOT now provides observations of surface water elevation and data products for river discharge at high spatial resolution. Yet routine assimilation of these products into hydrological models remains rare. More strikingly, satellite-derived surface temperature, available at high temporal frequency from MODIS and at high spatial resolution from Landsat, is widely accessible but is not commonly assimilated as a state variable in hydrological models, despite its clear potential to constrain energy-balance partitioning and evapotranspiration estimates. Beyond dedicated missions, opportunistic data sources such as commercial microwave links in telecommunications networks already provide area-averaged rainfall estimates at scales relevant to urban and small-catchment hydrology. Pursuing these opportunities more actively would directly strengthen the multi-objective evaluation framework advocated in this thesis.

Third, better hydrology will depend on a closer feedback loop between models and targeted field measurements. Large-sample diagnostic methods, as used in Chapters 3 and 4, can highlight where models systematically fail or underperform. These patterns can indicate where new field campaigns would be most valuable. Such campaigns should target the processes, locations, and variables that most constrain model uncertainty or differentiate competing hypotheses. In this way, modeling and measurement become iterative partners: models reveal where data are most needed, and new data challenge or validate model assumptions.

To make this feedback loop operational rather than aspirational, it must be embedded in the organization of hydrological research. This means designing field campaigns explicitly around competing model hypotheses, for example by deploying fiber-optic sensing in catchments where models disagree on subsurface flow paths or by installing soil moisture sensor networks where the random forest diagnostics of Chapter 4 identified soil parameterization as the dominant source of model divergence. It also means training a generation of hydrologists who engage in both observation collection and modeling, treating data collection and model experimentation as two sides of the same scientific process.

Finally, progress will depend on maintaining a balance between technological innovation and conceptual clarity. While advances in machine learning, remote sensing, and computational power offer exciting tools, they should strengthen rather than bypass our

mechanistic understanding of hydrological systems. The principle of parsimony, highlighted in Chapters 3–5, reminds us that complexity must be justified by data and purpose. More data, more variables, or finer grids do not automatically yield better models unless they are meaningfully linked to process understanding and decision-making needs.

Taken together, four priorities form the foundation for better hydrological science: fostering collaborative networks and data sharing; integrating diverse observations with explicit uncertainty estimates; creating stronger feedback loops between models and field measurements; and promoting a balanced, question-driven approach to model complexity.

In conclusion, hydrology stands at a pivotal moment—a sentence many PhD candidates have written before me. The tools for large-sample analysis, reproducible experimentation, and multi-objective evaluation are here. The challenge is to embed these tools in an open, collaborative culture that integrates field data, remote sensing, and model diagnostics into an iterative cycle of improvement. By doing so, hydrologists can move toward a more transparent, testable, and decision-relevant science capable of guiding sustainable water management across diverse regions and scales.

Hydrological models, like the data that feed them, are imperfect approximations of a system that we observe only partially. This thesis is no exception. But imperfection is not a reason for complacency. It is a reason to measure more, model more transparently, and hold our tools to higher standards. The path forward is not to accept the limitations we have inherited but to systematically reduce them.

BIBLIOGRAPHY

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'Connell, P. E., & Rasmussen, J. (1986). An introduction to the European Hydrological System — Systeme Hydrologique Europeen, "SHE", 2: Structure of a physically-based, distributed modelling system. *Journal of Hydrology*, *87*(1), 61–77. https://doi.org/10.1016/0022-1694(86)90115-0

Addor, N., & Melsen, L. A. (2019). Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models. *Water Resources Research*, *55*(1), 378–390. https://doi.org/10.1029/2018WR022958

Addor, N., 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. https://doi.org/10.1029/2018WR022606

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. https://doi.org/10.1080/02626667.2019.1683182

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Aerts, J. P. M., Hoch, J. M., Coxon, G., van de Giesen, N. C., & Hut, R. W. (2023). On the importance of observation uncertainty when evaluating and comparing models: A hydrological example. *EGUsphere*, 1–26. https://doi.org/10.5194/egusphere-2023-1156

Aerts, J. P. M., Hut, R. W., van de Giesen, N. C., Drost, N., van Verseveld, W. J., Weerts, A. H., & Hazenberg, P. (2022). Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model. *Hydrology and Earth System Sciences*, *26*(16), 4407–4430. https://doi.org/10.5194/hess-26-4407-2022

Aerts, J. P. M., Weerts, A., van Verseveld, W., Hazenberg, P., Drost, N., Hut, R. W., & van de Giesen, N. C. (2021). Large-sample based evaluation of the spatial resolution discretization of the wflow_sbm model for the CAMELS dataset. *EGU General Assembly Conference Abstracts*, EGU21–10680.

Albergel, C., Dutra, E., Munier, S., Calvet, J.-C., Munoz-Sabater, J., de Rosnay, P., & Balsamo, G. (2018). ERA-5 and ERA-Interim driven ISBA land surface model simulations: Which one performs better? *Hydrology and Earth System Sciences*, *22*(6), 3515–3532. https://doi.org/10.5194/hess-22-3515-2018

Albers, T. (2020). *Hydrologisch model PCR-GLOBWB forceren met zowel regenval als verdamping* [BSc]. Delft University of Technology.

Allan, R., Pereira, L., & Smith, M. (1998, January). *Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56* (Vol. 56).

Andréassian, V., Hall, A., Chahinian, N., & Schaake, J. (2006). Introduction and synthesis: Why should hydrologists work on a large number of basin data sets? In *Large sample basin experiments for hydrological parametrization : Results of the models parameter experiment- MOPEX. IAHS Red Books Series n° 307* (pp. 1–5). AISH. https://hal.inrae.fr/hal-02588687

Arnold, J. G., Srinivasan, R., Muttiah, R. S., & Williams, J. R. (1998). Large Area Hydrologic Modeling and Assessment Part I: Model Development. *JAWRA Journal of the American Water Resources Association*, *34*(1), 73–89. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x

Balin, D., Lee, H., & Rode, M. (2010). Is point uncertain rainfall likely to have a great impact on distributed complex hydrological modeling? *Water Resources Research*, *46*(11). https://doi.org/10.1029/2009WR007848

Bárdossy, A. (2007). Calibration of hydrological model parameters for ungauged catchments. *Hydrology and Earth System Sciences*, *11*(2), 703–710. https://doi.org/10.5194/hess-11-703-2007

Bárdossy, A., & Das, T. (2008). Influence of rainfall observation network on model calibration and application. *Hydrology and Earth System Sciences*, *12*(1), 77–89. https://doi.org/10.5194/hess-12-77-2008

Bárdossy, A., & Singh, S. K. (2008). Robust estimation of hydrological model parameters. *Hydrology and Earth System Sciences*, *12*(6), 1273–1283. https://doi.org/10.5194/hess-12-1273-2008

Bárdossy, A., & Anwar, F. (2023). Why do our rainfall–runoff models keep underestimating the peak flows? *Hydrology and Earth System Sciences*, *27*(10), 1987–2000. https://doi.org/10.5194/hess-27-1987-2023

Bárdossy, A., Kilsby, C., Birkinshaw, S., Wang, N., & Anwar, F. (2022). Is Precipitation Responsible for the Most Hydrological Model Uncertainty? *Frontiers in Water*, *4*. https://www.frontiersin.org/journals/water/articles/10.3389/frwa.2022.836554

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., & Bruijnzeel, L. A. (2016). Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, *52*(5), 3599–3622. https://doi.org/10.1002/2015WR018247

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Dijk, A. I. J. M. v., McVicar, T. R., & Adler, R. F. (2019). MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment. *Bulletin of the American Meteorological Society*, *100*(3), 473–500. https://doi.org/10.1175/BAMS-D-17-0138.1

Bell, V. A., Kay, A. L., Jones, R. G., & Moore, R. J. (2007). Development of a high resolution grid-based river flow model for use with regional climate model output. *Hydrology and Earth System Sciences*, *11*(1), 532–549. https://doi.org/10.5194/hess-11-532-2007

Benning, R. (1995). Towards a new lumped parameterization at catchment scale. *Master Thesis, University of Wageningen*.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 24). Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning* (pp. 115–123, Vol. 28). PMLR. https://proceedings.mlr.press/v28/bergstra13.html

Bergström, S., & Forsman, A. (1973). Development of a conceptual deterministic rainfall-runoff model. *Hydrology Research*, *4*(3), 147–170. https://doi.org/10.2166/nh.1973.0012

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Ménard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B., & Harding, R. J. (2011). The joint uk land environment simulator (jules), model description – part 1: Energy and water fluxes. *Geoscientific Model Development*, *4*(3), 677–699. https://doi.org/10.5194/gmd-4-677-2011

Beven, K. (2001a). How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences*, *5*(1), 1–12. https://doi.org/10.5194/hess-5-1-2001

Beven, K. (2006a). Searching for the Holy Grail of scientific hydrology: Q_t=(*S, R, Δt*)*A* as closure. *Hydrology and Earth System Sciences*, *10*(5), 609–618. https://doi.org/10.5194/hess-10-609-2006

Beven, K., Smith, P. J., & Wood, A. (2011). On the colour and spin of epistemic error (and what we might do about it). *Hydrology and Earth System Sciences*, *15*(10), 3123–3133. https://doi.org/10.5194/hess-15-3123-2011

Beven, K. J. (2001b). *Rainfall-Runoff Modelling: The Primer*. Wiley. https://doi.org/10.1002/9781119951001

Beven, K. J. (2019). Models as multiple working hypotheses: Hydrology in the post-normal science era. *Water Resources Research*. https://doi.org/10.1029/2019WR025370

Beven, K. J., Lamb, R., Quinn, P. F., Romanowicz, R., & Freer, J. (1988). Process heterogeneity in HYMOD conceptual model. In *Conceptual Problems in Hydrology*. IAHS.

Beven, K. (2006b). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K. (2016). Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrological Sciences Journal*, *61*(9), 1652–1665. https://doi.org/10.1080/02626667.2015.1031761

Beven, K. (2021). An epistemically uncertain walk through the rather fuzzy subject of observation and model uncertainties1. *Hydrological Processes*, *35*(1), e14012. https://doi.org/10.1002/hyp.14012

Beven, K. (2023). Benchmarking hydrological models for an uncertain future. *Hydrological Processes*, *37*(5), e14882. https://doi.org/10.1002/hyp.14882

Beven, K., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298. https://doi.org/10.1002/hyp.3360060305

Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*, *249*(1), 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

Beven, K., & Lane, S. (2019). Invalidation of Models and Fitness-for-Purpose: A Rejectionist Approach. In C. Beisbart & N. J. Saam (Eds.), *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives* (pp. 145–171). Springer International Publishing. https://doi.org/10.1007/978-3-319-70766-2_6

Beven, K., & Lane, S. (2022). On (in)validating environmental models. 1. Principles for formulating a Turing-like Test for determining when a model is fit-for purpose. *Hydrological Processes*, *36*(10), e14704. https://doi.org/10.1002/hyp.14704

Beven, K., Lane, S., Page, T., Kretzschmar, A., Hankin, B., Smith, P., & Chappell, N. (2022). On (in)validating environmental models. 2. Implementation of a Turing-like test to modelling hydrological processes. *Hydrological Processes*, *36*(10), e14703. https://doi.org/10.1002/hyp.14703

Beven, K., & Smith, P. (2015). Concepts of Information Content and Likelihood in Parameter Calibration for Hydrological Simulation Models. *Journal of Hydrologic Engineering*, *20*(1), A4014010. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000991

Beven, K. J., & Chappell, N. A. (2021). Perceptual perplexity and parameter parsimony. *WIREs Water*, *8*(4), e1530. https://doi.org/10.1002/wat2.1530

Beven, K. J., & Cloke, H. L. (2012). Comment on "Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water" by Eric F. Wood et al. *Water Resources Research*, *48*(1). https://doi.org/10.1029/2011WR010982

Bierkens, M. F. P. (2015a). The geography of water resources research: Facing the curse of locality in a data-scarce world. *Nature Geoscience*, *8*, 455–456. https://doi.org/10.1038/ngeo2530

Bierkens, M. F. P. (2015b). Global hydrology 2015: State, trends, and directions. *Water Resources Research*, *51*(7), 4923–4947. https://doi.org/10.1002/2015WR017173

Bierkens, M. F. P., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., de Roo, A., Döll, P., Drost, N., Famiglietti, J. S., Flörke, M., Gochis, D. J., Houser, P., Hut, R. W., Keune, J., Kollet, S., Maxwell, R. M., Reager, J. T., Samaniego, L., … Wood, E. F. (2015). Hyper-resolution global hydrological modelling: What is next? *Hydrological Processes*, *29*(2), 310–320. https://doi.org/10.1002/hyp.10391

Blazkova, S., & Beven, K. (2009). A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. *Water Resources Research*, *45*(12). https://doi.org/10.1029/2007WR006726

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes*, *9*(3-4), 251–290. https://doi.org/10.1002/hyp.3360090305

Bogaart, P. W., & Troch, P. A. (2006). Curvature distribution within hillslopes and catchments and its effect on the hydrological response. *Hydrology and Earth System Sciences*, *10*(6), 925–936. https://doi.org/10.5194/hess-10-925-2006

Booij, M. (2005). Impact of climate change on river flooding assessed with different spatial model resolutions. *Journal of Hydrology*, *303*(1-4), 176–198. https://doi.org/10.1016/j.jhydrol.2004.07.013

Bouaziz, L., Duan, Q., Delannoy, G., Bouaziz, O., & Jaffrain, J. (2018). Net intercatchment groundwater fluxes in a Budyko framework. *Hydrology and Earth System Sciences*, *22*, 6415–6431. https://doi.org/10.5194/hess-22-6415-2018

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. https://hess.copernicus.org/articles/22/6415/2018/

Brakensiek, D. L., Rawls, W. J., & Stephenson, G. R. (1984). Modifying SCS hydrologic soil groups and curve numbers for rangeland soils. *ASAE Paper No. PNR-84-203. American Society of Agricultural Engineers, St. Joseph, MI, USA.*

Bras, R. L. (2015). Complexity and organization in hydrology: A personal view. *Water Resources Research*, *51*(8), 6532–6548. https://doi.org/10.1002/2015WR016958

Brauer, C. C., Teuling, A. J., Torfs, P. J. J. F., & Uijlenhoet, R. (2014). The Wageningen Lowland Runoff Simulator (WALRUS): A lumped rainfall–runoff model for catchments with shallow groundwater. *Geoscientific Model Development*, *7*(5), 2313–2332. https://doi.org/10.5194/gmd-7-2313-2014

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brooks, R. H., & Corey, A. T. (1964). Hydraulic Properties of Porous Media. *Hydrology Papers, Colorado State University, Fort Collins, Colorado*, 37.

Bruin, H. A. R. d., Trigo, I. F., Bosveld, F. C., & Meirink, J. F. (2016). A Thermodynamically Based Model for Actual Evapotranspiration of an Extensive Grass Field Close to FAO Reference, Suitable for Remote Sensing Application. *Journal of Hydrometeorology*, *17*(5), 1373–1382. https://doi.org/10.1175/JHM-D-15-0006.1

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., & Smets, B. (2020). Copernicus Global Land Cover Layers—Collection 2. *Remote Sensing*, *12*(6), 1044. https://doi.org/10.3390/rs12061044

Burnash, R. J. C. (1973). *A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers.* U. S. Department of Commerce, National Weather Service, State of California, Department of Water Resources.

Butts, M. B., Payne, J. T., Kristensen, M., & Madsen, H. (2004). An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. *Journal of Hydrology*, *298*(1), 242–266. https://doi.org/10.1016/j.jhydrol.2004.03.042

Castellarin, A., Botter, G., Hughes, D. A., Liu, S., Ouarda, T. B. M. J., Parajka, J., Post, D. A., Sivapalan, M., Spence, C., Viglione, A., & Vogel, R. M. (2013). Prediction of flow duration curves in ungauged basins. In A. Viglione, G. Blöschl, H. Savenije, M.

Sivapalan, & T. Wagener (Eds.), *Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales* (pp. 135–162). Cambridge University Press. https://doi.org/10.1017/CBO9781139235761.010

Chaney, N. W., Wood, E. F., McBratney, A. B., Hempel, J. W., Nauman, T. W., Brungard, C. W., & Odgers, N. P. (2016). Polaris: A 30-meter probabilistic soil series map of the contiguous united states. *Geoderma*, *274*, 54–67. https://doi.org/10.1016/j.geoderma.2016.03.025

Chevuru, S., van Beek, R. L. P. H., van Vliet, M. T. H., Aerts, J. P. M., & Bierkens, M. F. P. (2024). Relevance of feedbacks between water availability and crop systems using a coupled hydrology – crop growth model. *EGUsphere*, 1–27. https://doi.org/10.5194/egusphere-2024-465

Ciarapica, L., & Todini, E. (2002). TOPKAPI: A model for the representation of the rainfall-runoff process at different scales. *Hydrological Processes*, *16*(2), 207–229. https://doi.org/10.1002/hyp.342

Clark, D. B., Mercado, L. M., Sitch, S., Jones, C. D., Gedney, N., Best, M. J., Pryor, M., Rooney, G. G., Essery, R. L. H., Blyth, E., Boucher, O., Harding, R. J., Huntingford, C., & Cox, P. M. (2011). The joint uk land environment simulator (jules), model description – part 2: Carbon fluxes and vegetation dynamics. *Geoscientific Model Development*, *4*(3), 701–722. https://doi.org/10.5194/gmd-4-701-2011

Clark, M. P., Kavetski, D., & Fenicia, F. (2008). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*. https://doi.org/10.1029/2007WR006089

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., & Rasmussen, R. M. (2015). A unified approach for process-based hydrologic modeling: 1. modeling concept. *Water Resources Research*, *51*(4), 2498–2514. https://doi.org/10.1002/2015WR017198

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A., Uddstrom, M., & Vrugt, J. A. (2005). Hydrological data assimilation with the particle filter: Use of streamflow observations to update states in a distributed hydrological model. *Water Resources Research*. https://doi.org/10.1029/2004WR003169

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N., Cai, X., Wood, A. W., & Peters-Lidard, C. D. (2017). The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. *Hydrology and Earth System Sciences*, *21*(7), 3427–3440. https://doi.org/10.5194/hess-21-3427-2017

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*(9). https://doi.org/10.1029/2010WR009827

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., & Hay, L. E. (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2007WR006735

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R., & Papalexiou, S. M. (2021). The Abuse of Popular Performance Metrics in Hydrologic Modeling. *Water Resources Research*, *57*(9), e2020WR029001. https://doi.org/10.1029/2020WR029001

Cortés-Salazar, N., Vásquez, N., Mizukami, N., Mendoza, P. A., & Vargas, X. (2023). To what extent does river routing matter in hydrological modeling? *Hydrology and Earth System Sciences*, *27*(19), 3505–3524. https://doi.org/10.5194/hess-27-3505-2023

Cosby, B. J., Hornberger, G. M., Clapp, R. B., & Ginn, T. R. (1984). A Statistical Exploration of the Relationships of Soil Moisture Characteristics to the Physical Properties of Soils. *Water Resources Research*, *20*(6), 682–690. https://doi.org/10.1029/WR020i006p00682

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). https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9

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. https://doi.org/10.1002/2014WR016532

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. (2020b). CAMELS-GB: Hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. *Earth System Science Data*, *12*(4), 2459–2483. https://doi.org/10.5194/essd-12-2459-2020

Craig, J. R., & Horritt, D. G. (2020). *Raven HYDRO: A flexible modeling framework*. University of Saskatchewan.

Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., Mai, J., Serrer, M., Sgro, N., Shafii, M., Snowdon, A. P., & Tolson, B. A. (2020). Flexible watershed simulation with the Raven hydrological modelling framework. *Environmental Modelling & Software*, *129*, 104728. https://doi.org/10.1016/j.envsoft.2020.104728

Crawford, N., & Linsley, R. (1966). DIGITAL SIMULATION IN HYDROLOGY' STANFORD WATERSHED MODEL 4. https://www.semanticscholar.org/paper/DIGITAL-SIMULATION-IN-HYDROLOGY%27-STANFORD-WATERSHED-Crawford-Linsley/708df5dca480e1d418abf0078e2447cc1e0ea6ba

Croke, B. F. W., & Jakeman, A. J. (2004). A catchment moisture deficit module for the IHACRES rainfall-runoff model. *Environmental Modelling & Software*, *19*(1), 1–5. https://doi.org/10.1016/j.envsoft.2003.09.001

Cunderlik, J. M., & Burn, D. H. (2013). Hydrological model intercomparison: What have we learned? *Journal of Hydrology*. https://doi.org/10.1016/j.jhydrol.2012.08.037

de Boer-Euser, T., Bouaziz, L., De Niel, J., Brauer, C., Dewals, B., Drogue, G., Fenicia, F., Grelier, B., Nossent, J., Pereira, F., Savenije, H., Thirel, G., & Willems, P. (2017).

Looking beyond general metrics for model comparison – lessons from an international model intercomparison study. *Hydrology and Earth System Sciences*, *21*(1), 423–440. https://doi.org/10.5194/hess-21-423-2017

de Marsily, G., Delay, F., Teles, V., Barbecot, F., Stoppok, E., Bouchard, J., Durand, M., & Harvey, C. F. (2005). Dealing with spatial heterogeneity. *Advances in Water Resources*. https://doi.org/10.1016/j.advwatres.2005.01.013

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., … Vitart, F. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Dobler, C., Hagemann, S., Wilby, R. L., & Stötter, J. (2012). Quantifying different sources of uncertainty in hydrological projections in an Alpine watershed. *Hydrology and Earth System Sciences*, *16*(11), 4343–4360. https://doi.org/10.5194/hess-16-4343-2012

Donnelly, C., Andersson, J. C., & Arheimer, B. (2016). Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe. *Hydrological Sciences Journal*, *61*(2), 255–273. https://doi.org/10.1080/02626667.2015.1027710

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Efron, B., & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, *1*(1), 54–75. https://doi.org/10.1214/ss/1177013815

Efron, B. (1992). Jackknife-After-Bootstrap Standard Errors and Influence Functions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*(1), 83–111. https://doi.org/10.1111/j.2517-6161.1992.tb01866.x

Eilander, D., & Boisgontier, H. (2021, April). HydroMT. https://github.com/Deltares/hydromt

Eilander, D., & Boisgontier, H. (2022, February). hydroMT. https://doi.org/10.5281/zenodo.6107669

Eilander, D., van Verseveld, W., Yamazaki, D., Weerts, A., Winsemius, H. C., & Ward, P. J. (2021). A hydrography upscaling method for scale-invariant parametrization of distributed hydrological models. *Hydrology and Earth System Sciences*, *25*(9), 5287–5313. https://doi.org/10.5194/hess-25-5287-2021

Elshafei, Y., Coletti, J. Z., Sivapalan, M., & Hipsey, M. R. (2015). A model of the socio-hydrologic dynamics in a semiarid catchment: Isolating feedbacks in the coupled human-hydrology system. *Water Resources Research*, *51*(8), 6442–6471. https://doi.org/10.1002/2015WR017048

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J. K., Goodman, S. D., Jackson, T. J., Johnson, J., Kimball, J., Piepmeier, J. R., Koster, R. D., Martin, N., McDonald, K. C., Moghaddam, M., Moran, S., Reichle, R., Shi, J. C., … Van Zyl, J. (2010). The Soil Moisture Active Passive (SMAP)

Mission. *Proceedings of the IEEE*, *98*(5), 704–716. https://doi.org/10.1109/JPROC.2010.2043918

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Brötz, B., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., Davin, E. L., Davini, P., Debeire, K., de Mora, L., Deser, C., & … Zimmermann, K. (2020). Earth system model evaluation tool (esmvaltool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of earth system models in cmip. *Geoscientific Model Development*, *13*(7), 3383–3438. https://doi.org/10.5194/gmd-13-3383-2020

Fan, Y., Clark, M., Lawrence, D. M., Swenson, S., Band, L. E., Brantley, S. L., Brooks, P. D., Dietrich, W. E., Flores, A., Grant, G., Kirchner, J. W., Mackay, D. S., McDonnell, J. J., Milly, P. C. D., Sullivan, P. L., Tague, C., Ajami, H., Chaney, N., Hartmann, A., … Yamazaki, D. (2019). Hillslope Hydrology in Global Change Research and Earth System Modeling. *Water Resources Research*, *55*(2), 1737–1772. https://doi.org/10.1029/2018WR023903

Feddes, R. A., Kowalik, P. J., & Zaradny, H. (1978). *Simulation of Field Water Use and Crop Yield*. Wiley.

Feddes, R. A., & Zaradny, H. (1978). Model for simulating soil-water content considering evapotranspiration — Comments. *Journal of Hydrology*, *37*(3), 393–397. https://doi.org/10.1016/0022-1694(78)90030-6

Feng, D., Beck, H., Lawson, K., & Shen, C. (2022, August). *The suitability of differentiable, learnable hydrologic models for ungauged regions and climate change impact assessment* (preprint). Catchment hydrology/Modelling approaches. https://doi.org/10.5194/hess-2022-245

Fenicia, F., Kavetski, D., & Clark, M. P. (2011). A comparison of diagnostic approaches for hydrological model evaluation: Examining the role of internal states. *Water Resources Research*. https://doi.org/10.1029/2011WR010612

Fenicia, F., McDonnell, J. J., & Savenije, H. H. G. (2008). Learning from model improvement: On the contribution of complementary data to process understanding. *Water Resources Research*, *44*(6). https://doi.org/10.1029/2007WR006386

Finnerty, B. D., Smith, M. B., Seo, D.-J., Koren, V., & Moglen, G. E. (1997). Space-time scale sensitivity of the Sacramento model to radar-gage precipitation inputs. *Journal of Hydrology*, *203*(1), 21–38. https://doi.org/10.1016/S0022-1694(97)00083-8

Fowler, K., Peel, M., Western, A., & Zhang, L. (2018). Improved Rainfall-Runoff Calibration for Drying Climate: Choice of Objective Function. *Water Resources Research*, *54*(5), 3392–3408. https://doi.org/10.1029/2017WR022466

Freeze, R. A., & Harlan, R. L. (1969). Blueprint for a physically-based, digitally-simulated hydrologic response model. *Water Resources Research*. https://doi.org/10.1029/WR005i001p00001

Gan, T., Tarboton, D. G., Dash, P., Gichamo, T. Z., & Horsburgh, J. S. (2020). Integrating hydrologic modeling web services with online data sharing to prepare, store, and execute hydrologic models. *Environmental Modelling & Software*, *130*, 104731. https://doi.org/10.1016/j.envsoft.2020.104731

Gao, H., Hrachowitz, M., Fenicia, F., Gharari, S., & Savenije, H. H. G. (2014). Testing the realism of a topography-driven model (FLEX-Topo) in the nested catchments of

the Upper Heihe, China. *Hydrology and Earth System Sciences*, *18*(5), 1895–1915. https://doi.org/10.5194/hess-18-1895-2014

Gao, H., Sabo, J. L., Chen, X., Liu, Z., Yang, Z., Ren, Z., & Liu, M. (2018). Landscape heterogeneity and hydrological processes: A review of landscape-based hydrological models. *Landscape Ecology*, *33*(9), 1461–1480. https://doi.org/10.1007/s10980-018-0690-4

Garrick, M., Cunnane, C., & Nash, J. E. (1978). A criterion of efficiency for rainfall-runoff models. *Journal of Hydrology*, *36*(3), 375–381. https://doi.org/10.1016/0022-1694(78)90155-5

Gash, J. H. C. (1979). An analytical model of rainfall interception by forests. *Quarterly Journal of the Royal Meteorological Society*, *105*(443), 43–55. https://doi.org/10.1002/qj.49710544304

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrology and Earth System Sciences*, *25*(4), 2045–2062. https://doi.org/10.5194/hess-25-2045-2021

Gebrechorkos, S. H., Leyland, J., Dadson, S. J., Cohen, S., Slater, L., Wortmann, M., Ashworth, P. J., Bennett, G. L., Boothroyd, R., Cloke, H., Delorme, P., Griffith, H., Hardy, R., Hawker, L., McLelland, S., Neal, J., Nicholas, A., Tatem, A. J., Vahidi, E., … Darby, S. E. (2023). Global scale evaluation of precipitation datasets for hydrological modelling. *Hydrology and Earth System Sciences Discussions*, 1–33. https://doi.org/10.5194/hess-2023-251

Gharari, S., Fenicia, F., Hrachowitz, M., Kavetski, D., & Clark, M. P. (2014). Hydrological model conditioning at the regional scale: Overcoming data scarcity with a parsimonious approach. *Water Resources Research*. https://doi.org/10.1002/2013WR014812

Gharari, S., Clark, M. P., Mizukami, N., Knoben, W. J. M., Wong, J. S., & Pietroniro, A. (2020). Flexible vector-based spatial configurations in land models. *Hydrology and Earth System Sciences*, *24*(12), 5953–5971. https://doi.org/10.5194/hess-24-5953-2020

Gichamo, T. Z., Sazib, N. S., Tarboton, D. G., & Dash, P. (2020). HydroDS: Data services in support of physically based, distributed hydrological models. *Environmental Modelling & Software*, *125*, 104623. https://doi.org/10.1016/j.envsoft.2020.104623

Gleeson, T., Moosdorf, N., Hartmann, J., & van Beek, L. P. H. (2014). A glimpse beneath earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity. *Geophysical Research Letters*, *41*(11), 3891–3898. https://doi.org/10.1002/2014GL059856

Gnann, S. J., Coxon, G., Woods, R. A., Howden, N. J. K., & McMillan, H. K. (2021). TOSSH: A Toolbox for Streamflow Signatures in Hydrology. *Environmental Modelling & Software*, *138*, 104983. https://doi.org/10.1016/j.envsoft.2021.104983

Grayson, R. B., Brookshaw, W. R., & Blöschl, G. (1997). Variability in catchment responses to rainfall-runoff processes. *Hydrological Processes*. https://doi.org/10.1002/(SICI)1099-1085(19970430)11:5<523::AID-HYP534>3.0.CO;2-A

Grayson, R. B., Moore, I. D., & McMahon, T. A. (1992). Physically based hydrologic modeling: 1. A terrain-based model for investigative purposes. *Water Resources Research*, *28*(10), 2639–2658. https://doi.org/10.1029/92WR01258

Gründemann, G. J., van de Giesen, N. C., Brunner, L., & van der Ent, R. (2022). Rarest rainfall events will see the greatest relative increase in magnitude under future climate change. *Communications Earth & Environment*, *3*(1), 1–9. https://doi.org/10.1038/s43247-022-00558-8

Gudmundsson, L., Wagener, T., Tallaksen, L. M., & Engeland, K. (2012). Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in europe. *Water Resources Research*, *48*(11). https://doi.org/10.1029/2011WR010911

Gupta, A., & Govindaraju, R. S. (2019). Propagation of structural uncertainty in watershed hydrologic models. *Journal of Hydrology*, *575*, 66–81. https://doi.org/10.1016/j.jhydrol.2019.05.026

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*(2), 463–477. https://doi.org/10.5194/hess-18-463-2014

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrological Processes*, *22*(18), 3802–3813. https://doi.org/10.1002/hyp.6989

Gupta, H. V., Wagener, T., & Liu, Y. (2012). Decomposition of sources of uncertainty in hydrological modeling using a chaos-theoretic approach. *Water Resources Research*. https://doi.org/10.1029/2011WR010489

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1-2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gupta, V. K., Rodríguez-Iturbe, I., & Wood, E. F. (1986). *Scale Problems in Hydrology*. Springer Dordrecht. https://doi.org/10.1007/978-94-009-4678-1

Guse, B., Fatichi, S., Gharari, S., & Melsen, L. A. (2021). Advancing Process Representation in Hydrological Models: Integrating New Concepts, Knowledge, and Data. *Water Resources Research*, *57*(11), e2021WR030661. https://doi.org/10.1029/2021WR030661

Guse, B., Kiesel, J., Pfannerstill, M., & Fohrer, N. (2020). Assessing parameter identifiability for multiple performance criteria to constrain model parameters. *Hydrological Sciences Journal*, *65*(7), 1158–1172. https://doi.org/10.1080/02626667.2020.1734204

Haddeland, I., Lettenmaier, D. P., & Skaugen, T. (2006). Reconciling Simulated Moisture Fluxes Resulting from Alternate Hydrologic Model Time Steps and Energy Budget Closure Assumptions. *Journal of Hydrometeorology*, *7*(3), 355–370. https://doi.org/10.1175/JHM496.1

Hall, C. A., Saia, S. M., Popp, A. L., Dogulu, N., Schymanski, S. J., Drost, N., van Emmerik, T., & Hut, R. W. (2021). A Hydrologist's Guide to Open Science. *Hydrology and*

Earth System Sciences Discussions, 1–23. https://doi.org/10.5194/hess-2021-392

Han, S., Williamson, B. D., & Fong, Y. (2021). Improving Random Forest predictions in small datasets from two-phase sampling designs. *BMC Medical Informatics and Decision Making*, *21*(1), 322. https://doi.org/10.1186/s12911-021-01688-3

Hancock, G., & Willgoose, G. (2001). The interaction between hydrology and geomorphology in a landscape simulator experiment. *Hydrological Processes*, *15*(1), 115–133. https://doi.org/10.1002/hyp.143

Hansen, N. (2006). The CMA Evolution Strategy: A Comparing Review. In J. A. Lozano, P. Larrañaga, I. Inza, & E. Bengoetxea (Eds.), *Towards a New Evolutionary Computation: Advances in the Estimation of Distribution Algorithms* (pp. 75–102). Springer. https://doi.org/10.1007/3-540-32494-1_4

Hansen, N., Müller, S. D., & Koumoutsakos, P. (2003). Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation*, *11*(1), 1–18. https://doi.org/10.1162/106365603321828970

Hansen, N., & Ostermeier, A. (2001). Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, *9*(2), 159–195. https://doi.org/10.1162/106365601750190398

Hansen, N., yoshihikoueno, ARF1, Nozawa, K., Chan, M., Akimoto, Y., & Brockhoff, D. (2021, June). CMA-ES/pycma: R3.1.0. https://doi.org/10.5281/zenodo.5002422

Haren, R., Camphuijsen, J., Dzigan, Y., Drost, N., Alidoost, F., Andela, B., Aerts, J. P. M., Weel, B., Hut, R. W., Zimmermann, K., Kalverla, P., Vreede, B., Paçal, A., Smeets, S., Verhoeven, S., Malinina, E., & Schilperoort, B. (2023, April). Era5cli. https://doi.org/10.5281/zenodo.7851975

Hattermann, F. F., Vetter, T., Breuer, L., Su, B., Daggupati, P., Donnelly, C., Fekete, B., Flörke, F., Gosling, S. N., Hoffmann, P., Liersch, S., Masaki, Y., Motovilov, Y., Müller, C., Samaniego, L., Stacke, T., Wada, Y., Yang, T., & Krysnaova, V. (2018). Sources of uncertainty in hydrological climate impact assessment: A cross-scale study. *Environmental Research Letters*, *13*(1), 015006. https://doi.org/10.1088/1748-9326/aa9938

Hengl, T., Jesus, J. M. d., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hoch, J. M., Sutanudjaja, E. H., Wanders, N., van Beek, R. L. P. H., & Bierkens, M. F. P. (2023). Hyper-resolution PCR-GLOBWB: Opportunities and challenges from refining model spatial resolution to 1 km over the European continent.

Hydrology and Earth System Sciences, *27*(6), 1383–1401. https://doi.org/10.5194/hess-27-1383-2023

Hoogeveen, J., Faurès, J.-M., Peiser, L., Burke, J., & van de Giesen, N. C. (2015). GlobWat – a global water balance model to assess water use in irrigated agriculture. *Hydrology and Earth System Sciences*, *19*(9), 3829–3844. https://doi.org/10.5194/hess-19-3829-2015

Hornberger, G. M., & Spear, R. C. (1981). Approach to the preliminary analysis of environmental systems. *J. Environ. Manage.*, *12*(1). https://doi.org/10.1016/0301-4797(81)90005-9

Horritt, M. S., & Bates, P. D. (2001). Effects of spatial resolution on a raster based model of flood flow. *Journal of Hydrology*, *253*(1), 239–249. https://doi.org/10.1016/S0022-1694(01)00490-5

Horsburgh, J., Morsy, M., Castronova, A., Goodall, J., Gan, T., Yi, H., Stealey, M., & Tarboton, D. (2015). HydroShare: Sharing Diverse Environmental Data Types and Models as Social Objects with Application to the Hydrology Domain. *Journal of the American Water Resources Association*, 873–889. https://doi.org/10.1111/1752-1688.12363

Horton, R. E. (1932). Drainage-basin characteristics. *Transactions, American Geophysical Union*, *13*, 350–361.

Houze Jr., R. A. (2012). Orographic effects on precipitating clouds. *Reviews of Geophysics*, *50*(1). https://doi.org/10.1029/2011RG000365

Hrachowitz, M., & Clark, M. P. (2017). HESS Opinions: The complementary merits of competing modelling philosophies in hydrology. *Hydrology and Earth System Sciences*, *21*(8), 3953–3973. https://doi.org/10.5194/hess-21-3953-2017

Huang, Y., & Bardossy, A. (2020). Impacts of Data Quantity and Quality on Model Calibration: Implications for Model Parameterization in Data-Scarce Catchments. *Water*, *12*(9), 2352. https://doi.org/10.3390/w12092352

Hut, R. W., Aerts, J. P. M., Wiersma, P., Hoogelander, V., van de Giesen, N. C., Drost, N., Kalverla, P., van Werkhoven, B., Verhoeven, S., Alidoost, F. S., et al. (2023). The eWaterCycle platform for open and FAIR hydrological collaboration.

Hut, R. W., Drost, N., Alidoost, F., Verhoeven, S., Smeets, S., Kalverla, P., Vreede, B., Aerts, J. P. M., van Werkhoven, B., & van de Giesen, N. C. (2021, October). eWaterCycle tech paper example notebooks. https://doi.org/10.5281/zenodo.5543899

Hut, R. W., Drost, N., van de Giesen, N. C., van Werkhoven, B., Abdollahi, B., Aerts, J. P. M., Albers, T., Alidoost, F., Andela, B., Camphuijsen, J., Dzigan, Y., van Haren, R., Hutton, E. W. H., Kalverla, P., van Meersbergen, M., van den Oord, G., Pelupessy, I., Smeets, S., Verhoeven, S., … Weel, B. (2022). The eWaterCycle platform for open and FAIR hydrological collaboration. *Geoscientific Model Development*, *15*(13), 5371–5390. https://doi.org/10.5194/gmd-15-5371-2022

Hut, R. W., Drost, N., van de Giesen, N. C., van Werkhoven, B., Aerts, J. P. M., Alidoost, F., Kalverla, P., Verhoeven, S., Andela, B., Camphuijsen, J., et al. (2021). The eWaterCycle platform for open and FAIR computational hydrological research. *AGU Fall Meeting Abstracts*, *2021*, H35K–1157.

Hut, R. W., van de Giesen, N. C., & Drost, N. (2017). Comment on "Most computational hydrology is not reproducible, so is it really science?" by Christopher Hutton et

al.: Let hydrologists learn the latest computer science by working with Research Software Engineers (RSEs) and not reinvent the waterwheel ourselves. *Water Resources Research*, *53*(5), 4524–4526. https://doi.org/10.1002/2017WR020665

Hutton, C., Wagener, T., Freer, J., Han, D., Duffy, C., & Arheimer, B. (2016). Most computational hydrology is not reproducible, so is it really science? *Water Resources Research*, *52*(10), 7548–7555. https://doi.org/10.1002/2016WR019285

Hutton, E. W. H., Piper, M., & Tucker, G. (2020). The Basic Model Interface 2.0: A standard interface for coupling numerical models in the geosciences. *Journal of Open Source Software*, *5*(51), 2317. https://doi.org/10.21105/joss.02317

Hutton, E. W. H., Piper, M. D., & Tucker, G. E. (2016). The Basic Model Interface: A standard for coupling Earth system models. *Environmental Modelling Software*. https://doi.org/10.1016/j.envsoft.2015.08.004

Imhoff, R. O., van Verseveld, W. J., van Osnabrugge, B., & Weerts, A. H. (2020). Scaling Point-Scale (Pedo)transfer Functions to Seamless Large-Domain Parameter Estimates for High-Resolution Distributed Hydrologic Modeling: An Example for the Rhine River. *Water Resources Research*, *56*(4), e2019WR026807. https://doi.org/10.1029/2019WR026807

Ingram, W. (2016). Increases all round. *Nature Climate Change*, *6*(5), Article 5. https://doi.org/10.1038/nclimate2966

Jain, S., Mindlin, J., Koren, G., Gulizia, C., Steadman, C., Langendijk, G. S., Osman, M., Abid, M. A., Rao, Y., & Rabanal, V. (2022). Are We at Risk of Losing the Current Generation of Climate Researchers to Data Science? *AGU Advances*, *3*(4), e2022AV000676. https://doi.org/10.1029/2022AV000676

Jakeman, A. J., Letcher, R. A., & Norton, J. P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software*, *21*(5), 602–614. https://doi.org/10.1016/j.envsoft.2006.01.004

Jayawardena, A. W., & Zhou, M. C. (2000). A modified spatial soil moisture storage capacity distribution curve for the Xinanjiang model. *Journal of Hydrology*, *227*(1), 93–113. https://doi.org/10.1016/S0022-1694(99)00173-0

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., & Reichstein, M. (2019). The fluxcom ensemble of global land-atmosphere energy fluxes. *Scientific Data*, *6*(1), Article 1. https://doi.org/10.1038/s41597-019-0076-8

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, *4*(1), 170122. https://doi.org/10.1038/sdata.2017.122

Karssenberg, D. J., Schmitz, O., Salamon, P., de Jong, K., & Bierkens, M. F. P. (2010). A software framework for construction of process-based stochastic spatio-temporal models and data assimilation. *Environmental Modelling and Software*, *25*, 489–502. https://doi.org/10.1016/j.envsoft.2009.10.004

Kauffeldt, A., Clark, M. P., McMillan, H. K., Woods, R. A., Kavetski, D., & Fenicia, F. (2016). Benchmarking hydrological models at the global scale: eWaterCycle intercomparison experiments. *Hydrology and Earth System Sciences*, *20*, 801–819. https://doi.org/10.5194/hess-20-801-2016

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C.-Y., & Westerberg, I. K. (2013). Disinformative data in large-scale hydrological modelling. *Hydrology and Earth System Sciences*, *17*(7), 2845–2857. https://doi.org/10.5194/hess-17-2845-2013

Kavetski, D., Kuczera, G., & Franks, S. W. (2006a). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2005WR004368

Kavetski, D., Kuczera, G., & Franks, S. W. (2006b). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2005WR004376

Keller, V. D. J., Tanguy, M., Prosdocimi, I., Terry, J. A., Hitt, O., Cole, S. J., Fry, M., Morris, D. G., & Dixon, H. (2015). CEH-GEAR: 1 km resolution daily and monthly areal rainfall estimates for the UK for hydrological and other applications. *Earth System Science Data*, *7*(1), 143–155. https://doi.org/10.5194/essd-7-143-2015

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*(11), 8922–8941. https://doi.org/10.1029/2018WR023750

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). https://doi.org/10.1029/2005WR004362

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, *424-425*, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., & Woods, R. A. (2020). A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments. *Water Resources Research*, *56*(9), e2019WR025975. https://doi.org/10.1029/2019WR025975

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

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Kollet, S. J., & Maxwell, R. M. (2008). Integrated surface–groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model. *Water Resources Research*. https://doi.org/10.1029/2006WR005142

Kolmogorov, A. N. (1933). *Foundations of the theory of probability*. Chelsea Publishing Co.

Konapala, G., Kao, S.-C., Painter, S. L., & Lu, D. (2020). Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. *Environmental Research Letters*, *15*(10), 104022. https://doi.org/10.1088/1748-9326/aba927

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the köppen-geiger climate classification updated. *Meteorologische Zeitschrift*, *15*(3), 259–263. https://doi.org/10.1127/0941-2948/2006/0130

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., & Matias, Y. (2022). Caravan - A global community dataset for large-sample hydrology. https://eartharxiv.org/repository/view/3345/

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. https://doi.org/10.1029/2020WR027101

Lane, R. A., Coxon, G., Freer, J. E., Wagener, T., Johnes, P. J., Bloomfield, J. P., Greene, S., Macleod, C. J. A., & 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. https://doi.org/10.5194/hess-23-4011-2019

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*(3), 273–299. https://doi.org/10.1007/s10584-006-6338-4

Lehner, B., Liermann, C. R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J. C., Rödel, R., Sindorf, N., & Wisser, D. (2011). High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. *Frontiers in Ecology and the Environment*, *9*(9), 494–502. https://doi.org/10.1890/100125

Ley, R., Hellebrand, H., Casper, M. C., & Fenicia, F. (2015). Comparing classical performance measures with signature indices derived from flow duration curves to assess model structures as tools for catchment classification. *Hydrology Research*, *47*(1), 1–14. https://doi.org/10.2166/nh.2015.221

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. https://doi.org/10.1029/94JD00483

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, *201*(1), 272–288. https://doi.org/10.1016/S0022-1694(97)00041-3

Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., & Arheimer, B. (2010). Development and testing of the HYPE (Hydrological Predictions for the Environment) water

quality model for different spatial scales. *Hydrology Research*, *41*(3), 295–319. https://doi.org/10.2166/nh.2010.007

Liu, Y., Freer, J., Beven, K., & Matgen, P. (2009). Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. *Journal of Hydrology*, *367*(1), 93–103. https://doi.org/10.1016/j.jhydrol.2009.01.016

Liu, Y., & Gupta, H. V. (2007). Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, *43*(7). https://doi.org/10.1029/2006WR005756

Luan, J., Zhang, C., Xu, B., Xue, Y., & Ren, Y. (2020). The predictive performances of Random Forest models with limited sample size and different species traits. *Fisheries Research*, *227*, 105534. https://doi.org/10.1016/j.fishres.2020.105534

Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, *46*(253), 68–78. https://doi.org/10.2307/2280095

Massmann, C. (2020). Identification of factors influencing hydrologic model performance using a top-down approach in a large number of U.S. catchments. *Hydrological Processes*, *34*(1), 4–20. https://doi.org/10.1002/hyp.13566

McDonnell, J. J., Sivapalan, M., & Woods, R. A. (2007). Moving beyond heterogeneity and process complexity: A new vision for hydrology. *Water Resources Research*. https://doi.org/10.1029/2006WR004867

McMillan, H. K., Westerberg, I. K., & Rahman, A. (2018). Observational uncertainty for hydrological model evaluation: Multiple data sets and methods. *Water Resources Research*. https://doi.org/10.1002/2017WR020908

McMillan, H. K., Westerberg, I. K., & Trosset, M. W. (2010). A framework for describing and deriving catchment characteristic response functions. *Water Resources Research*.

McMillan, H., Freer, J., Pappenberger, F., Krueger, T., & Clark, M. (2010). Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrological Processes*, *24*(10), 1270–1284. https://doi.org/10.1002/hyp.7587

McMillan, H., Jackson, B., Clark, M., Kavetski, D., & Woods, R. (2011). Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models. *Journal of Hydrology*, *400*(1-2), 83–94. https://doi.org/10.1016/j.jhydrol.2011.01.026

McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes*, *26*(26), 4078–4111. https://doi.org/10.1002/hyp.9384

McMillan, H., Montanari, A., Cudennec, C., Savenije, H., Kreibich, H., Krueger, T., Liu, J., Mejia, A., Loon, A. V., Aksoy, H., Baldassarre, G. D., Huang, Y., Mazvimavi, D., Rogger, M., Sivakumar, B., Bibikova, T., Castellarin, A., Chen, Y., Finger, D., … Xia, J. (2020). Panta Rhei 2013–2015: Global perspectives on hydrology, society and change. *Hydrological Sciences Journal*, *65*(S1), 1174–1191. https://doi.org/10.1080/02626667.2016.1159308

McMillan, H. K., Gnann, S. J., & Araki, R. (2022). Large Scale Evaluation of Relationships Between Hydrologic Signatures and Processes. *Water Resources Research*, *58*(6), e2021WR031751. https://doi.org/10.1029/2021WR031751

McMillan, H. K., Westerberg, I. K., & Krueger, T. (2018). Hydrological data uncertainty and its implications. *WIREs Water*, *5*(6), e1319. https://doi.org/10.1002/wat2.1319

Melsen, L. A., Teuling, A. J., Torfs, P. J. J. F., Uijlenhoet, R., Mizukami, N., & Clark, M. P. (2016). HESS Opinions: The need for process-based evaluation of large-domain hyper-resolution models. *Hydrology and Earth System Sciences*, *20*(3), 1069–1079. https://doi.org/10.5194/hess-20-1069-2016

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., & Schmitt, O. (2016). Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, *7*(1), 13603. https://doi.org/10.1038/ncomms13603

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., Rakovec, O., & Samaniego, L. (2017). Towards seamless large-domain parameter estimation for hydrologic models. *Water Resources Research*, *53*(9), 8020–8040. https://doi.org/10.1002/2017WR020401

Moges, E., Demissie, Y., Larsen, L., & Yassin, F. (2021). Review: Sources of Hydrological Model Uncertainties and Advances in Their Analysis. *Water*, *13*(1), 28. https://doi.org/10.3390/w13010028

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. https://doi.org/10.1016/j.advwatres.2012.09.007

Montanari, A., & Grossi, G. (2008). Estimating the uncertainty of hydrological forecasts: A statistical approach. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2008WR006897

Montanari, A., & Toth, E. (2007). Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins? *Water Resources Research*, *43*(5). https://doi.org/10.1029/2006WR005184

Mott, R., Vionnet, V., & Grünewald, T. (2018). The Seasonal Snow Cover Dynamics: Review on Wind-Driven Coupling Processes. *Frontiers in Earth Science*, *6*, 197. https://doi.org/10.3389/feart.2018.00197

Myneni, R., Knyazikhin, Y., & Park, T. (2015). MCD15A3H MODIS/Terra+Aqua Leaf Area Index/FPAR 4-day L4 Global 500m SIN Grid V006. https://doi.org/10.5067/MODIS/MCD15A3H.006

Nash, J., & Sutcliffe, J. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., & 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. https://doi.org/10.5194/hess-19-209-2015

Or, D. (2019). Scaling issues in hydrological modeling: The struggle between fine-scale processes and catchment-scale predictions. *Water Resources Research*. https://doi.org/10.1029/2018WR024147

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science*, *263*(5147), 641–646. https://doi.org/10.1126/science.263.5147.641

Pan, M., Wood, E. F., & Zaitchik, B. F. (2020). Global integration of satellite-derived evapotranspiration with hydrological modeling. *Water Resources Research*. https://doi.org/10.1029/2019WR025528

Pappenberger, F., & Beven, K. J. (2006). Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resources Research*, *42*(5). https://doi.org/10.1029/2005WR004820

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., & Salamon, P. (2015). How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology*, *522*, 697–713. https://doi.org/10.1016/j.jhydrol.2015.01.024

Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins – part 1: Runoff-hydrograph studies. *Hydrology and Earth System Sciences*, *17*(5), 1783–1795. https://doi.org/10.5194/hess-17-1783-2013

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., Zhang, L., Amiro, B., Ammann, C., … Papale, D. (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, *7*(1), 225. https://doi.org/10.1038/s41597-020-0534-3

Peng, J., Tanguy, M., Robinson, E. L., Pinnington, E., Evans, J., Ellis, R., Cooper, E., Hannaford, J., Blyth, E., & Dadson, S. (2021). Estimation and evaluation of high-resolution soil moisture from merged model and earth observation data in the great britain. *Remote Sensing of Environment*, *264*, 112610. https://doi.org/10.1016/j.rse.2021.112610

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*(1), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Peter-Lidard, C. D., Pan, M., Wood, E. F., & Zaitchik, B. F. (2017). Hyper-resolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*. https://doi.org/10.1002/2016WR019259

Peters-Lidard, C. D., Clark, M., Samaniego, L., Verhoest, N. E. C., van Emmerik, T., Uijlenhoet, R., Achieng, K., Franz, T. E., & Woods, R. (2017). Scaling, similarity, and the fourth paradigm for hydrology. *Hydrology and Earth System Sciences*, *21*(7), 3701–3713. https://doi.org/10.5194/hess-21-3701-2017

Pfeffer, W. T., Arendt, A. A., Bliss, A., Bolch, T., Cogley, J. G., Gardner, A. S., Hagen, J.-O., Hock, R., Kaser, G., Kienholz, C., Miles, E. S., Moholdt, G., Mölg, N., Paul, F., Radić, V., Rastner, P., Raup, B. H., Rich, J., Sharp, M. J., & Consortium, T. R. (2014). The Randolph Glacier Inventory: A globally complete inventory of glaciers. *Journal of Glaciology*, *60*(221), 537–552. https://doi.org/10.3189/2014JoG13J176

Poncelet, C., Merz, R., Merz, B., Parajka, J., Oudin, L., Andréassian, V., & Perrin, C. (2017). Process-based interpretation of conceptual hydrological model performance

using a multinational catchment set. *Water Resources Research*, *53*(8), 7247–7268. https://doi.org/10.1002/2016WR019991

Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: Towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*, *63*(13-14), 1941–1953. https://doi.org/10.1080/02626667.2018.1552002

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Routledge. https://philotextes.info/spip/IMG/pdf/popper-logic-scientific-discovery.pdf

Portmann, F. T., Siebert, S., & Döll, P. (2010). MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochemical Cycles*, *24*(1). https://doi.org/10.1029/2008GB003435

Raimonet, M., Oudin, L., Thieu, V., Silvestre, M., Vautard, R., Rabouille, C., & Moigne, P. L. (2017). Evaluation of Gridded Meteorological Datasets for Hydrological Modeling. *Journal of Hydrometeorology*, *18*(11), 3027–3041. https://doi.org/10.1175/JHM-D-17-0018.1

Rakovec, O., Mizukami, N., Kumar, R., Newman, A. J., Thober, S., Wood, A. W., Clark, M. P., & 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. https://doi.org/10.1029/2019JD030767

Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., & Thompson, J. (2018). Soil property and class maps of the conterminous united states at 100-meter spatial resolution. *Soil Science Society of America Journal*, *82*(1), 186–201. https://doi.org/10.2136/sssaj2017.04.0122

Ranstam, J., & Cook, J. A. (2018). Lasso regression. *BJS (British Journal of Surgery)*, *105*(10), 1348–1348. https://doi.org/10.1002/bjs.10895

Refsgaard, J. C., & Storm, B. (1995). MIKE SHE: Systems description and user manual. *Journal of Hydrology*. https://doi.org/10.1002/(SICI)1099-1085(199504)9:3<313::AID-HYP438>3.0.CO;2-8

Refsgaard, J. C., & Storm, B. (1996). Distributed engineering hydrology models—development and applications. *Advances in Water Resources*. https://doi.org/10.1016/0262-8856(95)00063-7

Reggiani, P., Hassanizadeh, S. M., Sivapalan, M., & Gray, W. G. (1999). A unifying framework for watershed thermodynamics: Constitutive relationships. *Advances in Water Resources*, *23*(1), 15–39. https://doi.org/10.1016/S0309-1708(99)00005-6

Reggiani, P., & Schellekens, J. (2003). Modelling of hydrological responses: The representative elementary watershed approach as an alternative blueprint for watershed modelling. *Hydrological Processes*, *17*(18), 3785–3789. https://doi.org/10.1002/hyp.5167

Reggiani, P., Sivapalan, M., & Majid Hassanizadeh, S. (1998). A unifying framework for watershed thermodynamics: Balance equations for mass, momentum, energy and entropy, and the second law of thermodynamics. *Advances in Water Resources*, *22*(4), 367–398. https://doi.org/10.1016/S0309-1708(98)00012-8

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*(5). https://doi.org/10.1029/2009WR008328

Richards, L. A. (1931). Capillary conduction of liquids through porous mediums. *Physics*, *1*, 318–333.

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., de Mora, L., Diblen, F., Dreyer, L., Drost, N., Earnshaw, P., Hassler, B., Koldunov, N., Little, B., Loosveldt Tomas, S., & Zimmermann, K. (2020). Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview. *Geoscientific Model Development*, *13*(3), 1179–1199. https://doi.org/10.5194/gmd-13-1179-2020

Roberts, W., Williams, G. P., Jackson, E., Nelson, E. J., & Ames, D. P. (2018). Hydrostats: A Python Package for Characterizing Errors between Observed and Predicted Time Series. *Hydrology*, *5*(4), 66. https://doi.org/10.3390/hydrology5040066

Robinson, E. L., Blyth, E. M., Clark, D. B., Comyn-Platt, E., & Rudd, A. C. (2020a). Climate hydrology and ecology research support system meteorology dataset for Great Britain (1961-2017) [CHESS-met]. https://doi.org/10.5285/2ab15bf0-ad08-415c-ba64-831168be7293

Robinson, E. L., Blyth, E. M., Clark, D. B., Comyn-Platt, E., & Rudd, A. C. (2020b). Climate hydrology and ecology research support system potential evapotranspiration dataset for Great Britain (1961-2017) [CHESS-PE]. https://doi.org/10.5285/9116e565-2c0a-455b-9c68-558fdd9179ad

Rouholahnejad Freund, E., Zappa, M., & Kirchner, J. W. (2020). Averaging over spatiotemporal heterogeneity substantially biases evapotranspiration rates in a mechanistic large-scale land evaporation model. *Hydrology and Earth System Sciences*, *24*(10), 5015–5025. https://doi.org/10.5194/hess-24-5015-2020

Rutter, A. J., Kershaw, K. A., Robins, P. C., & Morton, A. J. (1971). A predictive model of rainfall interception in forests, 1. Derivation of the model from observations in a plantation of Corsican pine. *Agricultural Meteorology*, *9*, 367–384. https://doi.org/10.1016/0002-1571(71)90034-3

Rutter, A. J., Morton, A. J., & Robins, P. C. (1975). A Predictive Model of Rainfall Interception in Forests. II. Generalization of the Model and Comparison with Observations in Some Coniferous and Hardwood Stands. *Journal of Applied Ecology*, *12*(1), 367–380. https://doi.org/10.2307/2401739

Samaniego, L., Thober, S., Kumar, R., Wanders, N., Rakovec, O., Pan, M., Zink, M., Sheffield, J., Wood, E. F., & Marx, A. (2018). Anthropogenic warming exacerbates european soil moisture droughts. *Nature Climate Change*, *8*(5), Article 5. https://doi.org/10.1038/s41558-018-0138-0138

Samaniego, L., Brenner, J., Craven, J., Cuntz, M., Dalmasso, G., Demirel, C. M., Jing, M., Kaluza, M., Kumar, R., Langenberg, B., Mai, J., Müller, S., Musuuza, J., Prykhodko, V., Rakovec, O., Schäfer, D., Schneider, C., Schrön, M., Schüler, L., … Attinger, S. (2021, March). Mesoscale Hydrologic Model - mHM v5.11.1. https://doi.org/10.5281/zenodo.4575390

Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*(5). https://doi.org/10.1029/2008WR007327

Santos, L., Thirel, G., & Perrin, C. (2018). Continuous state-space representation of a bucket-type rainfall-runoff model: A case study with the GR4 model using state-space GR4 (version 1.0). *Geoscientific Model Development*, *11*(4), 1591–1605. https://doi.org/10.5194/gmd-11-1591-2018

Savenije, H. H. G. (2009). HESS opinions "The art of hydrology". *Hydrology and Earth System Sciences*, *13*, 339–344. https://doi.org/10.5194/hess-13-339-2009

Schaake, J. C., Arsenault, K. R., Duan, Q., Georgakakos, K. P., Hall, A., Nash, J. E., Patterson, T. N., & Pitman, A. J. (1996). Simple water balance models for estimating monthly runoff: Evaluation of MOPEX results. *Journal of Geophysical Research*. https://doi.org/10.1029/95JD02092

Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes*, *21*(15), 2075–2080. https://doi.org/10.1002/hyp.6825

Schellekens, J., Beven, K. J., Klanke, T., Panzeri, M., Peel, M., & Salamon, P. (2017). Benchmarking hydrological models at the global scale: eWaterCycle intercomparison experiments. *Hydrology and Earth System Sciences*, *21*, 1595–1610. https://doi.org/10.5194/hess-21-1595-2017

Schellekens, J., Drost, N., Mao, L., Tudorache, T., Deo, R., Berghuijs, W., Bronstert, A., & Clark, M. P. (2020a). *Wflow: A software framework for hydrological modeling* (tech. rep.). Zenodo.

Schellekens, J., et al. (2020b, May). Openstreams/wflow: Bug fix release for 2020.1. https://doi.org/10.5281/ZENODO.593510

Seibert, J., & McDonnell, J. J. (2002). On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria calibration. *Water Resources Research*. https://doi.org/10.1029/2001WR000978

Seibert, J. (2001). On the need for benchmarks in hydrological modelling. *Hydrological Processes*, *15*(6), 1063–1064. https://doi.org/10.1002/hyp.446

Seibert, J., Vis, M. J. P., Lewis, E., & van Meerveld, H. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, *32*(8), 1120–1125. https://doi.org/10.1002/hyp.11476

Shabestanipour, G., Brodeur, Z., Farmer, W. H., Steinschneider, S., Vogel, R. M., & Lamontagne, J. R. (2023). Stochastic Watershed Model Ensembles for Long-Range Planning: Verification and Validation. *Water Resources Research*, *59*(2), e2022WR032201. https://doi.org/10.1029/2022WR032201

Sherman, L. K. (1932). Streamflow from rainfall by unit-graph method. *Engineering News-Record*.

Shrestha, P., Sulis, M., Simmer, C., & Kollet, S. (2015). Impacts of grid resolution on surface energy fluxes simulated with an integrated surface-groundwater flow model. *Hydrology and Earth System Sciences*, *19*(10), 4317–4326. https://doi.org/10.5194/hess-19-4317-2015

Shuai, P., Chen, X., Mital, U., Coon, E. T., & Dwivedi, D. (2022). The effects of spatial and temporal resolution of gridded meteorological forcing on watershed hydrological responses. *Hydrology and Earth System Sciences*, *26*(8), 2245–2276. https://doi.org/10.5194/hess-26-2245-2022

Siebert, S., & Döll, P. (2008). *The global crop water model (gcwm): Documentation and first results for irrigated crops* (Frankfurt Hydrology Paper 07). Institute of Physical Geography, University of Frankfurt. Frankfurt am Main, Germany.

Silberstein, R. P. (2006). Hydrological models are so good, do we still need data? *Environmental Modelling & Software*, *21*(9), 1340–1352. https://doi.org/10.1016/j.envsoft.2005.04.019

Smirnov, N. V. (1933). Estimate of Deviation between Empirical Distribution Functions in Two Independent Samples. *Bulletin of Moscow University*, *2*, 3–16.

Sorooshian, S., & Gupta, V. K. (1983). Automatic calibration of conceptual rainfall-runoff models: The question of parameter observability and uniqueness. *Water Resources Research*, *19*(1), 260–268. https://doi.org/10.1029/WR019i001p00260

Stoffelen, A. (1998). Toward the true near-surface wind speed: Error modeling and calibration using triple collocation. *Journal of Geophysical Research: Oceans*, *103*(C4), 7755–7766. https://doi.org/10.1029/97JC03180

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., & Bierkens, M. F. P. (2017). PCR-GLOBWB 2: A 5 arc-minute global hydrological and water resources model. *Geosci. Model Dev. Discuss.*, *2017*, 1–41. https://doi.org/10.5194/gmd-2017-288

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., & Bierkens, M. F. P. (2018). PCR-GLOBWB 2: A 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, *11*(6), 2429–2453. https://doi.org/10.5194/gmd-11-2429-2018

Tabari, H. (2020). Climate change impact on flood and extreme precipitation increases with water availability. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-70816-2

Tan, B. Q., & O'Connor, K. M. (1996). Application of an empirical infiltration equation in the SMAR conceptual model. *Journal of Hydrology*, *185*(1), 275–295. https://doi.org/10.1016/0022-1694(95)02993-1

Tanaka, T., & Tachikawa, Y. (2015). Testing the applicability of a kinematic wave-based distributed hydrological model in two climatically contrasting catchments. *Hydrological Sciences Journal*, *60*(7-8), 1361–1373. https://doi.org/10.1080/02626667.2014.967693

Tanguy, M., Dixon, H., Prosdocimi, I., Morris, D. G., & Keller, V. D. J. (2021). Gridded estimates of daily and monthly areal rainfall for the United Kingdom (1890-2019) [CEH-GEAR]. https://doi.org/10.5285/dbf13dd5-90cd-457a-a986-f2f9dd97e93c

Tanguy, M., Peng, J., Robinson, E., Pinnington, E., Evans, J., Ellis, R., Cooper, E., Hannaford, J., Blyth, E., & Dadson, S. (2022). Soil moisture product merged from satellite and modelled data for Great Britain, April 2015-December 2017. https://doi.org/10.5285/26b8ddd4-09fd-4e40-a556-6a8f3a7481ea

Tarboton, D., Idaszak, R., Horsburgh, J., Heard, J., Ames, D., Goodall, J., Band, L., Merwade, V., Couch, A., Arrigo, J., Hooper, R., Valentine, D., & Maidment, D. (2014). HydroShare: Advancing Collaboration through Hydrologic Data and Model Sharing. *International Congress on Environmental Modelling and Software*. https://scholarsarchive.byu.edu/iemssconference/2014/Stream-A/7

Tarek, M., Brissette, F. P., & Arsenault, R. (2020). Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. *Hydrology and Earth System Sciences*, *24*(5), 2527–2544. https://doi.org/10.5194/hess-24-2527-2020

Tetzlaff, D., Uhlenbrook, S., & McDonnell, J. J. (2008). Conceptualizing spatial variability of soil and subsurface hydraulic properties within a catchment. *Hydrological Processes*. https://doi.org/10.1002/hyp.6991

Thébault, C., Perrin, C., Andréassian, V., Thirel, G., Legrand, S., & Delaigue, O. (2023). Impact of suspicious streamflow data on the efficiency and parameter estimates of rainfall–runoff models. *Hydrological Sciences Journal*, *68*(12), 1627–1647. https://doi.org/10.1080/02626667.2023.2234893

Thenkabail, P., Teluguntla, P. G., Xiong, J., Oliphant, A., Congalton, R. G., Ozdogan, M., Gumma, M. K., Tilton, J. C., Giri, C., Milesi, C., Phalke, A., Massey, R., Yadav, K., Sankey, T. T., Zhong, Y., Aneece, I., & Foley, D. (2021). *Global cropland-extent product at 30-m resolution (GCEP30) derived from Landsat satellite time-series data for the year 2015 using multiple machine-learning algorithms on Google Earth Engine cloud* (tech. rep. No. 1868). U.S. Geological Survey. https://doi.org/10.3133/pp1868

Tonkin, M. J., & Doherty, J. (2005). A hybrid regularized inversion methodology for highly parameterized environmental models. *Water Resources Research*, *41*(10). https://doi.org/10.1029/2005WR003995

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

Troch, P. A., Carrillo, G., & Camerlynck, C. (2009). The role of hillslope-riparian connectivity in catchment hydrology. *Water Resources Research*. https://doi.org/10.1029/2008WR007637

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006). Threshold relations in subsurface stormflow: 1. A 147-storm analysis of the Panola hillslope. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003778

Trotter, L., Knoben, W. J. M., Fowler, K. J. A., Saft, M., & Peel, M. C. (2022). Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v2.1: An object-oriented implementation of 47 established hydrological models for improved speed and

readability. *Geoscientific Model Development*, *15*(16), 6359–6369. https://doi.org/10.5194/gmd-15-6359-2022

Tucker, G. E., Hutton, E. W. H., Piper, M. D., Campforts, B., Gan, T., Barnhart, K. R., Kettner, A., Overeem, I., Peckham, S. D., McCready, L., & Syvitski, J. (2021). CSDMS: A community platform for numerical modeling of Earth-surface processes. *Geoscientific Model Development Discussions*, 1–40. https://doi.org/10.5194/gmd-2021-223

van Beek, L. P. H., & Bierkens, M. F. P. (2008). The PCR-GLOBWB 1 global hydrological model for streamflow and groundwater simulations. *Hydrology and Earth System Sciences*, *12*, 2695–2711. https://doi.org/10.5194/hess-12-2695-2008

van Dijk, A. I. J. M., Renzullo, L. J., Werners, S. E., & Su, C. H. (2011). Integrating observations of soil moisture from satellite and in situ sensors for improved hydrological modeling. *Water Resources Research*. https://doi.org/10.1029/2011WR010702

van Esse, W. R., Perrin, C., Booij, M. J., Augustijn, D. C. M., 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. https://doi.org/10.5194/hess-17-4227-2013

van Verseveld, W. J., Weerts, A. H., Visser, M., Buitink, J., Imhoff, R. O., Boisgontier, H., Bouaziz, L., Eilander, D., Hegnauer, M., ten Velden, C., & Russell, B. (2022). Wflow_sbm v0.6.1, a spatially distributed hydrologic model: From global data to local applications. *Geoscientific Model Development Discussions*, 1–52. https://doi.org/10.5194/gmd-2022-182

van Verseveld, W., Visser, M., Bootsma, H., Boisgontier, H., & Bouaziz, L. (2021, September). Wflow.jl. https://doi.org/10.5281/zenodo.5384924

Van Der Knijff, J. M., Younis, J., & De Roo, A. P. J. (2010). LISFLOOD: A GIS-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, *24*(2), 189–212. https://doi.org/10.1080/13658810802549154

Velthoen, J., Cai, J.-J., & Jongbloed, G. (2022). Forward variable selection for random forest models. *Journal of Applied Statistics*, *0*(0), 1–21. https://doi.org/10.1080/02664763.2022.2095362

Venhuizen, G. J., Hut, R. W., Albers, C., Stoof, C. R., & Smeets, I. (2019). Flooded by jargon: How the interpretation of water-related terms differs between hydrology experts and the general audience. *Hydrology and Earth System Sciences*, *23*(1), 393–403. https://doi.org/10.5194/hess-23-393-2019

Verhoeven, S., Drost, N., Weel, B., Kalverla, P., Alidoost, F., & Andela, B. (2021, September). eWaterCycle infra. https://doi.org/10.5281/zenodo.5356689

Verkaik, J., Sutanudjaja, E. H., Oude Essink, G. H. P., Lin, H. X., & Bierkens, M. F. P. (2022). GLOBGM v1.0: A parallel implementation of a 30 arcsec PCR-GLOBWB-MODFLOW global-scale groundwater model. *Geoscientific Model Development Discussions*, 1–27. https://doi.org/10.5194/gmd-2022-226

Vertessy, R. A., & Elsenbeer, H. (1999). Distributed modeling of storm flow generation in an Amazonian rain forest catchment: Effects of model parameterization. *Water Resources Research*, *35*(7), 2173–2187. https://doi.org/10.1029/1999WR900051

Vionnet, V., Marsh, C. B., Menounos, B., Gascoin, S., Wayand, N. E., Shea, J., Mukherjee, K., & Pomeroy, J. W. (2021). Multi-scale snowdrift-permitting modelling of mountain snowpack. *The Cryosphere*, *15*(2), 743–769. https://doi.org/10.5194/tc-15-743-2021

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

Vrugt, J. A., & ter Braak, C. J. F. (2013). Multi-objective calibration and uncertainty analysis of a watershed model using DREAM. *Water Resources Research*. https://doi.org/10.1002/2013WR014673

Vrugt, J. A., ter Braak, C. J. F., Diks, C. G., Robinson, B. A., Hyman, J. M., & Higdon, D. (2003). Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *Water Resources Research*. https://doi.org/10.1029/2002WR001746

Vrugt, J. A., Bouten, W., Gupta, H. V., & Sorooshian, S. (2002). Toward improved identifiability of hydrologic model parameters: The information content of experimental data. *Water Resources Research*, *38*(12), 48–1–48–13. https://doi.org/10.1029/2001WR001118

Vrugt, J. A., Diks, C. G. H., Gupta, H. V., Bouten, W., & Verstraten, J. M. (2005). Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water Resources Research*, *41*(1). https://doi.org/10.1029/2004WR003059

Wagener, T. (2004). Catchment classification and model improvement using data-driven methods. *Hydrological Processes*. https://doi.org/10.1002/hyp.5622

Wagener, T., McIntyre, N., Gupta, H. V., & Wheater, H. S. (2001). Application of Topmodel in a semi-arid environment. *Water Resources Research*. https://doi.org/10.1029/2000WR900328

Wagener, T., & Wheater, H. S. (2006). Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty. *Journal of Hydrology*, *320*(1), 132–154. https://doi.org/10.1016/j.jhydrol.2005.07.015

Wagner, W., Hahn, S., Kidd, R., Melzer, T., Bartalis, Z., Hasenauer, S., Figa, J., De Rosnay, P., Jann, A., & Schneider, S. (2013). The ASCAT soil moisture product: A review of its. *Meteorologische Zeitschrift*, *22*(1), 1–29. https://www.researchgate.net/profile/Yong-Wang-209/publication/235553667_The_ASCAT_Soil_Moisture_Product_A_Review_of_its_Specifications_Validation_Results_and_Emerging_Applications/links/5b5092150f7e9b240fee72d5/The-ASCAT-Soil-Moisture-Product-A-Review-of-its-Specifications-Validation-Results-and-Emerging-Applications.pdf

Weigel, K., Bock, L., Gier, B. K., Lauer, A., Righi, M., Schlund, M., Adeniyi, K., Andela, B., Arnone, E., Berg, P., Caron, L.-P., Cionni, I., Corti, S., Drost, N., Hunter, A., Lledó, L., Mohr, C. W., Paçal, A., Pérez-Zanón, N., … Eyring, V. (2021). Earth System

Model Evaluation Tool (ESMValTool) v2.0 – diagnostics for extreme events, regional and impact evaluation, and analysis of Earth system models in CMIP. *Geoscientific Model Development*, *14*(6), 3159–3184. https://doi.org/10.5194/gmd-14-3159-2021

Westerberg, I. K., McMillan, H. K., & McIntyre, N. (2020). Data uncertainty in hydrological modeling: Effects of rainfall and discharge errors on model calibration. *Water Resources Research*. https://doi.org/10.1029/2019WR024865

Westerberg, I. K., Sikorska-Senoner, A. E., Viviroli, D., Vis, M., & Seibert, J. (2022). Hydrological model calibration with uncertain discharge data. *Hydrological Sciences Journal*, *67*(16), 2441–2456. https://doi.org/10.1080/02626667.2020.1735638

Wiersma, P., Aerts, J. P. M., Zekollari, H., Hrachowitz, M., Drost, N., Huss, M., Sutanudjaja, E. H., & Hut, R. W. (2022). Coupling a global glacier model to a global hydrological model prevents underestimation of glacier runoff. *Hydrology and Earth System Sciences*, *26*(23), 5971–5986. https://doi.org/10.5194/hess-26-5971-2022

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., & Dumontier, M. (2018). A design framework and exemplar metrics for FAIRness. *Scientific Data*, *5*(1), 180118. https://doi.org/10.1038/sdata.2018.118

Wood, E. F., Roundy, J. K., Troy, T. J., van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., Lettenmaier, D. P., Pitman, A., Sivapalan, M., & Vörösmarty, C. (2011). Hyper-resolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water cycle. *Water Resources Research*. https://doi.org/10.1029/2011WR011222

Wood, E. F., Roundy, J. K., Troy, T. J., van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J., Gochis, D., van de Giesen, N. C., Houser, P., Jaffé, P. R., Kollet, S., Lehner, B., Lettenmaier, D. P., Peters-Lidard, C., Sivapalan, M., … Whitehead, P. (2011). Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*, *47*(5). https://doi.org/10.1029/2010WR010090

Wood, E. F., Sivapalan, M., Beven, K., & Band, L. (1988). Effects of spatial variability and scale with implications to hydrologic modeling. *Journal of Hydrology*, *102*(1), 29–47. https://doi.org/10.1016/0022-1694(88)90090-X

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset. *Water Resources Research*, *55*(6), 5053–5073. https://doi.org/10.1029/2019WR024873

Ye, W., Bates, B., Viney, N., Sivapalan, M., & Jakeman, A. (1997). Performance of Conceptual Rainfall-Runoff Models in Low-Yielding Ephemeral Catchments. *Water Resources Research - WATER RESOUR RES*, *33*, 153–166. https://doi.org/10.1029/96WR02840

Yew Gan, T., Dlamini, E. M., & Biftu, G. F. (1997). Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. *Journal of Hydrology*, *192*(1-4), 81–103. https://doi.org/10.1016/S0022-1694(96)03114-9

Zhang, Y., Gao, H., & Sivapalan, M. (2019). Exploring evapotranspiration partitioning using satellite data: Insights for model evaluation. *Water Resources Research*. https://doi.org/10.1029/2018WR024597

Zhou, L., Liu, P., Gui, Z., Zhang, X., Liu, W., Cheng, L., & Xia, J. (2022). Diagnosing structural deficiencies of a hydrological model by time-varying parameters. *Journal of Hydrology*, *605*, 127305. https://doi.org/10.1016/j.jhydrol.2021.127305

ACKNOWLEDGEMENTS

Completing a PhD is often presented as an individual achievement, while in reality, it is anything but. This thesis was shaped not only by research questions, models, data, and many lines of code, but also by the people who offered ideas, encouragement, distraction, and, when necessary, a gentle push in the right direction. Looking back, I feel fortunate that these years brought me much more than a collection of publications.

Rolf, you gave me the opportunity to start this PhD, and it is difficult to know where to begin thanking you. From the start, you welcomed me not only into the research, but into PhD life in the broadest sense: from pressing cider at Bart's and having drinks in Haarlem to birthdays and finding our way around New Orleans. Your supervision was energetic, optimistic, and typical Rolf. I will always remember you moving rapidly back and forth in front of a blackboard, filling it with diagrams while somehow making a complex problem appear simple. Your enthusiasm for teaching and sharing knowledge was infectious and helped awaken the same passion in me. Thank you for the opportunity and for showing me that good science and generosity can go hand in hand.

Nick, thank you for continually encouraging me to look beyond the immediate problem and consider the larger challenges facing hydrology. I am especially grateful for your guidance on the parts of a PhD that never appear explicitly in its publications: learning to ask meaningful questions, position research within a wider debate and become an independent scientist. You also provided advice of a more practical nature. I will not forget your recommendation to stand in front of an impressive bookcase when having your picture taken, because people will automatically assume you are more intelligent. Nor will I forget arriving in the nature near Nijmegen to find you carrying a spade and challenging me, as the resident Earth scientist, to demonstrate how one should dig and interpret the soil. Rolf and Nick, your approaches were different but complementary, and I have been fortunate to learn from you both.

I would also like to thank my office mates over the years: Sha, Dengxiao, Juan Carlos, the Davids, Monica, Bart, Laurene, Gaby, Bahereh and Vincent. My first office taught me that finding a temperature acceptable to every culture could be very difficult: one person would be wearing a winter coat while another was visibly sweating. David S and Sergio, thank you in particular for helping me find my feet, both technically and socially; Thursday drinks and kapsalon proved to be very important. Sharing an office with Laurene, Bart and Gaby offered a lot of gezelligheid. Gaby, I greatly valued our coffees and conversations about life; Laurene, it was a joy having you around; and Bart, you became a true friend. Vincent, it was rewarding to pass on some of the lessons that others had taught me.

Beyond the office, I was lucky to share PhD life with Chelsea, Bas T, Paul, Judith, Ibrahim, Boran, Roberto, Mario, Fransje, Gijs, Coco, Banafsheh, Bahereh, Luuk, Bas W, Thijs and many others. Thank you for the laughter, coffees, walks and the reassuring

knowledge that someone else always understood how a seemingly minor technical problem could consume an entire week.

This thesis would not have been possible without the eWaterCycle team. Niels, Peter, Yifat, Gijs, Bart, Stefan and everyone else involved: thank you for building both the technical framework and the community behind it. I am equally grateful to the researchers from other institutes who strengthened this work. Jannis, since taking me under your wing at EGU when I attended as an MSc student, I have valued your advice, support, and our beers in Utrecht. Albrecht, thank you for your consistently valuable feedback and continued interest in my work.

The PhD also took me far beyond Delft. Lauren, thank you for the adventures during my time in the United States. Now that you live in Luxembourg, I look forward to our European adventures. I also thank AGU H3S, particularly Tom and Caithlyn, for making my time in the United States so memorable. Closer to home, Ruben, Janneke, Raquel, Fransje, Jorge and everyone in YHS-NL: it was great to give the network a fresh start, and naturally, fun to join many borrels.

Two people deserve a special mention as my paranymphs. Ties, I met you on my first day at TU Delft and immediately recognised someone both exceptionally sociable and deeply driven. We became friends through late nights, scientific discussions over beer, shared travels. No scientific outcome can compete with that, so choosing you as my paranymph was easy.

Bart, somewhere between sharing an office, working on eWaterCycle, pressing cider, brewing beer and travelling together, you became a great friend. Our road trip through the American Midwest in a white convertible Ford Mustang remains unforgettable: long roads, good music, podcasts and plenty of adventures. Being the officiant at your wedding was a great honour, which makes it especially meaningful that you will now stand beside me as my paranymph.

Speaking of friends, Monica, Juan, Mario, thank you for your continued support and belief in me and my academic abilities, even when my own confidence occasionally faltered. Your friendship has meant a great deal throughout this journey. Sergio, my Mexican buddy: ride the wave.

I am also happy to have been part of the wider Thursday PSOR community: Robert, Jessica, Boran, David, Oswaldo, Antonella, Carina, Patricia, Su, Davide, Job, Roberto, Akshit, Mario, Dirk. Thank you for the drinks, conversations, laughter and Friday mornings that were occasionally harder than strictly necessary.

My thanks also go to the wider department and its staff, including Thom, Erik, Ruud, Markus, Edo, Mark, Miriam, Maurits, Saket and many others. Remko, I owe you particular thanks for our enjoyable collaboration, for giving me the opportunity to continue at TU Delft as a researcher and now for serving as an opponent during my defence. Your confidence and support came at an important point in my career and have meant a great deal to me.

I would also like to thank my colleagues at HKV, with whom I began the next chapter of my career. Although you were not directly involved in the research presented in this thesis, working with you has allowed me to put many of its lessons into practice in the Netherlands and beyond.

Outside the university, Jorik, Timothy, Huib and Joost, my former housemates, thank you for your friendship and support. To the Chillings group—Huib, Jessse, Jan, Joost, Jelle, Julius, Justin, Max and Diederik—thank you for the many good times and for the distractions outside of work.

To my parents, my brothers and the rest of my family, including those who sadly cannot be here: thank you for always supporting and believing in me, even when the details of what I was working on were not entirely clear. Who would have thought that I would one day defend a PhD? Certainly not me.

Finally, Mar. Meeting you in Mexico was by far the best outcome of any journey. Thank you for your love, support and warmth, for the many, many laughs and for making every place feel like home. Te amo.

With this thesis, a long chapter comes to an end, but the lessons, friendships and curiosity it brought will remain. I look forward to putting what I have learned into practice, taking on new questions and challenges, and seeing where the next chapter takes us.

Curriculum Vitæ

Jerom Pierre Marie Aerts

About the Author

Jerom Aerts was born on 12 December 1991 in Amsterdam the Netherlands. He studied Earth Sciences at Vrije Universiteit Amsterdam and obtained an MSc degree in Hydrological Sciences there in 2018. Later that year, he joined the Water Resources section at Delft University of Technology as a PhD candidate. His research focused on making hydrological model experiments more reproducible and informative by combining the eWaterCycle platform with large-sample experimentation, model diagnostics, and the explicit consideration of observational uncertainty.

After completing his doctoral research, Jerom continued at Delft University of Technology as a researcher. In 2024, he joined HKV as a Water and Climate consultant. His work covers hydrological modelling, precipitation nowcasting, flood forecasting, and early-warning systems in the Netherlands and internationally. In this role, he continues to connect scientific research with its practical application in operational water management.

Education

2016	BSc Earth Sciences Vrije Universiteit Amsterdam
2018	MSc Hydrological Sciences Vrije Universiteit Amsterdam
2018–2023	Doctoral research in Water Resources Delft University of Technology *Thesis:* *eWaterCycle in Practice: Reproducible and Diagnostic Approaches for Large-Sample Catchment Based Hydrological Model Evaluation* *Promotor:* Prof. dr. ir. N. C. van de Giesen *Copromotor:* Dr. ir. R. W. Hut

PROFESSIONAL EXPERIENCE

2018–2023	PhD candidate, Water Resources Delft University of Technology; research on reproducible and diagnostic approaches for large-sample hydrological model evaluation.
2023–2024	Researcher Delft University of Technology; application of hydrological research in operational forecasting and early-warning systems.
2024–present	Consultant, Water and Climate HKV; hydrological and hydraulic modelling, precipitation nowcasting, flood forecasting, and operational early-warning systems.

OTHER RELEVANT TRAINING

2021	Catchment Hydrology and Modelling, Canada
2019	ParFlow Hydrological Modelling, United States of America
2019	Ensemble Forecasting and Data Assimilation, United States of America

List of Publications

Peer-reviewed Scientific Publications

8. S. Chevuru, R. L. P. H. van Beek, M. T. H. van Vliet, **J. P. M. Aerts**, and M. F. P. Bierkens, *Relevance of feedbacks between water availability and crop systems using a coupled hydrological–crop growth model*, Hydrology and Earth System Sciences **29**, 4219–4239 (2025).

7. W. Hazeleger, **J. P. M. Aerts**, P. Bauer, et al., *Digital twins of the Earth with and for humans*, Communications Earth & Environment **5**, 463 (2024).

6. **J. P. M. Aerts**, J. M. Hoch, G. Coxon, N. C. van de Giesen, and R. W. Hut, *On the importance of discharge observation uncertainty when interpreting hydrological model performance*, Hydrology and Earth System Sciences **28**, 5011–5030 (2024).

5. T. C. van Hateren, H. J. Jongen, H. Al-Zawaidah, et al., *Where should hydrology go? An early-career perspective on the next IAHS Scientific Decade: 2023–2032*, Hydrological Sciences Journal **68**, 529–541 (2023).

4. P. G. Wiersma, **J. P. M. Aerts**, H. Zekollari, M. Hrachowitz, N. Drost, M. Huss, E. H. Sutanudjaja, and R. W. Hut, *Coupling a global glacier model to a global hydrological model prevents underestimation of glacier runoff*, Hydrology and Earth System Sciences **26**, 5971–5986 (2022).

3. **J. P. M. Aerts**, R. W. Hut, N. C. van de Giesen, N. Drost, W. J. van Verseveld, A. H. Weerts, and P. Hazenberg, *Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model*, Hydrology and Earth System Sciences **26**, 4407–4430 (2022).

2. R. W. Hut, N. Drost, N. C. van de Giesen, B. van Werkhoven, B. Abdollahi, **J. P. M. Aerts**, et al., *The eWaterCycle platform for open and FAIR hydrological collaboration*, Geoscientific Model Development **15**, 5371–5390 (2022).

1. **J. P. M. Aerts**, S. Uhlemann-Elmer, D. Eilander, and P. J. Ward, *Comparison of estimates of global flood models for flood hazard and exposed gross domestic product: a China case study*, Natural Hazards and Earth System Sciences **20**, 3245–3260 (2020).

Manuscript Included in This Thesis

1. **J. P. M. Aerts**, R. W. Hut, N. C. van de Giesen, T. J. T. van der Heijden, J. M. Hoch, A. H. Weerts, and E. H. Sutanudjaja, *Advancing hydrological model diagnostics: an exploratory approach using random forest models and a large-sample catchment data set*, manuscript under review and included in this thesis.

ORCID: 0000-0003-0157-4818

