



**Politecnico
di Torino**

ScuDo

Scuola di Dottorato ~ Doctoral School

WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Civil and Environmental Engineering (38th Cycle)

An exploration of post-processing uncertainty estimation in conceptual rainfall-runoff modelling

Luca Lombardo

Supervisor

Prof. Alberto Viglione

Co-Supervisor

Prof. Simon M. Papalexiou

Politecnico di Torino
April 30, 2026

This thesis is licensed under a Creative Commons License, Attribution - Noncommercial - NoDerivative Works 4.0 International: see www.creativecommons.org. The text may be reproduced for non-commercial purposes, provided that credit is given to the original author.

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

A handwritten signature in black ink, appearing to read 'Luca Lombardo', written in a cursive style.

Luca Lombardo
Turin, April 30, 2026

Summary

Conceptual rainfall–runoff models are widely used in hydrology and represent a primary tool for generating predictions, particularly in data-scarce contexts. However, in many practical applications, these models are still employed in a deterministic framework, without adequately accounting for the multiple sources of uncertainty affecting their outputs. This thesis provides a comprehensive investigation of uncertainty in conceptual rainfall–runoff modelling, with the dual objective of clarifying its main sources and developing practical tools for its quantification in operational settings.

Chapter 2 begins with the development of a large-scale regional rainfall–runoff model over the Greater Alpine Region, with a focus on reconstructing flood events. This application serves as an example to highlight key modelling limitations, showing how uncertainties increase significantly near extreme events and when models are applied to ungauged catchments. These findings demonstrate that neglecting uncertainty can lead to misleading results, particularly in operational contexts.

Building on this evidence, Chapter 3 investigates the statistical properties of rainfall–runoff model residuals through a large multi-catchment, multi-model study across the continental United States. The analysis focuses on key features such as marginal distributions, dependence structures, and temporal correlations. Both general patterns and catchment-specific behaviors are identified and related to climatic and morphological characteristics. In addition, the impact of residual transformations aimed at stabilizing statistical properties is assessed, highlighting which features can be effectively controlled and which remain inherently difficult to model within standard error modelling frameworks.

Based on these insights, Chapter 4 then explores practical post-processing approaches for aggregated uncertainty estimation. The focus is placed on parametric methods, due to their potential improvements in out-of-sample applications and regionalization. A first approach consists of a copula-based method (C-method) designed to stochastically generate residual and discharge time series. Its performance is evaluated against established non-parametric resampling techniques, showing comparable performance but also revealing limitations that would require increased model complexity, potentially hindering regional applicability.

To address this issue, in Chapter 5, a second approach is developed, focusing on the direct estimation of confidence intervals without generating full time series. This method, formulated as a parametric copula-based reinterpretation of the BLUECAT (Koutsoyiannis and Montanari, 2022) framework, is implemented in two versions targeting different flow regimes. The introduction of parametric tail modeling allows extrapolation beyond observed extremes. Results indicate performance comparable to non-parametric methods, with improved estimation of uncertainty for high-flow conditions, as confirmed by peak-over-threshold analyses.

Chapter 6 of the thesis addresses uncertainty estimation for predictions in ungauged catchments through a regionalization of the previously proposed parametric BLUECAT framework. Model components, including deterministic predictions, copula dependence structures, and discharge distributions, are transferred to ungauged sites to estimate confidence intervals. The results are again comparable to those obtained with non-parametric approaches, while highlighting areas for methodological refinement.

Overall, the findings do not support, within the tested conditions, the hypothesis that parametric post-processing methods systematically outperform simpler non-parametric alternatives. At the same time, the results identify clear directions for future research, suggesting that further methodological developments may enable parametric approaches to achieve improved performance, especially in ungauged catchments, in uncertainty estimation for rainfall–runoff modelling.

Table of Contents

1.	Introduction.....	1
1.1	Rainfall-Runoff models.....	1
1.1.1	Modelling approaches in rainfall–runoff hydrology	1
1.1.2	Study area and modelling framework.....	4
1.1.3	Beyond local predictions: the challenge of prediction in ungauged basins	5
1.2	Uncertainty in conceptual Rainfall-Runoff models	7
1.2.1	Multiple sources of uncertainty	7
1.2.2	Generalized Likelihood Uncertainty Estimation - GLUE	9
1.2.3	Formal Bayesian Approaches	11
1.2.4	Post-processing (PP) approaches	13
2.	Regional Rainfall-runoff model for the Greater Alpine Region.....	16
2.1	A model fit to purpose: why and how we model floods.....	16
2.2	Study Area: the scale as a challenge.....	18
2.2.1	Constraints and implications in modelling the GAR.....	18
2.2.2	Climate Input Data	20
2.2.3	Discharge observations	21
2.2.4	Hydrological Units.....	22
2.3	Model and Methods.....	23
2.3.1	Model structure	23
2.3.2	HUs attribution.....	26
2.3.3	Local calibration strategies	26
2.3.4	Regionalization procedure	29
2.3.5	Validation metrics	31
2.4	Results.....	34
2.4.1	Does a tailored dataset improve HUs attribution?	34
2.4.2	A necessary premise about input uncertainty.....	36
2.4.3	Local (gauged) calibrations efficiency	37
2.4.4	Regional (ungauged) efficiencies	40

2.4.5	Regression trees: what drives the regional model?.....	45
2.5	Discussion	49
2.5.1	Usefulness of the results: are they fit-to-purpose?.....	49
3.	General properties of rainfall-runoff model residuals	52
3.1	From a model to its uncertainty	52
3.1.1	About the generality of rainfall-runoff residuals.....	52
3.1.2	Desired (ideal) properties of model residuals	54
3.1.3	Study Area and multi-model ensemble.....	57
3.1.4	Residuals definition and catchments' clustering	59
3.2	Statistical Characteristics of Residuals	60
3.2.1	Do residuals display seasonality?	60
3.2.2	What is the probability distribution of model residuals?	61
3.2.3	How much does model selection matter in shaping distribution marginal properties?	66
3.2.4	Is heteroskedasticity reduced by a data transformation?.....	67
3.3	Temporal Correlation Characteristics of the Residuals	69
3.3.1	How persistent are the model errors?	69
3.3.2	Temporal Correlation's upper and lower tail dependency	73
3.4	What can we learn: from properties to error models	76
3.4.1	Difference between regimes and general effect of transformations ...	76
3.4.2	General insights for modelling	77
3.4.3	Note about deseasonalization.....	78
3.4.4	Note about intermittent catchments	78
3.4.5	Limits in our approach	79
4.	Post-processing methods.....	81
4.1	Deciding a direction	81
4.1.1	Methods and operativity	81
4.1.2	Parametric vs non-parametric: why are we testing parametric approaches?	83
4.2	A new error model.....	84

4.2.1	Looking backwards to decide how to move forward	84
4.2.2	General notions about copulas.....	89
4.2.3	Study Area and RR-model.....	90
4.2.4	Benchmark non-parametric methods.....	91
4.2.5	Note on the calibration and validation periods selection	94
4.2.6	Validation metrics	95
4.3	Comparative results.....	97
4.3.1	Is the independence assumption between D-model predictions and transformed residuals reasonable?.....	97
4.3.2	Marginal and temporal correlation properties of the stochastically generated residuals	98
4.3.3	Marginal and temporal correlation properties of the stochastically generated discharge series.....	101
4.3.4	S-model reliability	106
4.3.5	Back Correction Factor (BCF)	109
4.4	Results assessment: how to move forward?.....	112
4.4.1	Is the assumption of independence really reasonable for the C-model?	112
5.	Beyond residual modelling: an alternative approach to stochastic rainfall-runoff modelling.....	115
5.1	A different philosophy.....	115
5.1.1	Bias correction and direct confidence bound estimation: BLUECAT and its parametric evolution	115
5.1.2	Main operational differences.....	117
5.1.3	Parametric copula model tailored for high flows	122
5.1.4	Validation metrics for the S-model	123
5.2	Results analysis	125
5.2.1	Parametric tail and mixed cumulative distribution functions	125
5.2.2	Which parametric copula can model the data structure?	129
5.2.3	Predictive Probability Plots and Interval Score	132
5.2.4	Focus on the peak-over-threshold.....	136

5.2.5	Properties of the ensemble median.....	139
5.3	Did the new paradigm pay off?.....	141
5.3.1	Not better, not worst (but for floods).....	141
5.3.2	About the generality of the copula structures.....	142
5.3.3	Computational trade-offs.....	143
5.3.4	Toward regionalization	144
6.	Regional estimation of conceptual RR-model uncertainty.....	146
6.1	PP-methods regionalization perspectives.....	146
6.1.1	A short history of uncertainty estimation in ungauged catchments..	146
6.2	Regional P-BLUECAT model	148
6.2.1	Model components and challenges.....	148
6.2.2	Intertwining effects on S-model location and width	149
6.2.3	Leave-One-Out (LOO) regional model.....	151
6.2.4	Regionalization of the copula parameters	153
6.2.5	Conceptual dilemma of regional O-CDF.....	154
6.2.6	Regional correction: which CDF?	155
6.2.7	BB7 copula model: regionalization of the period-of-record CDF	155
6.2.8	Joe copula model: regionalization of the “period-of-record” CDF...	159
6.3	Benchmark and validation metrics.....	160
6.3.1	Benchmark non-parametric PP regionalization method.....	160
6.4	Validation Metrics for regional S-model	162
6.5	Results and Comparative assessment.....	162
6.5.1	Copula parameters regression	162
6.5.2	Predictive Probability Plots.....	163
6.5.3	Interval Skill Score (ISS)	166
6.6	About the usefulness of regional uncertainty estimation.....	167
6.6.1	Where does most of the uncertainty lay?.....	167
6.6.2	Regional uncertainty for high-flow conditions: potential avenues ...	169
7.	Discussion and conclusion.....	170

7.1	Fitting the pieces together	170
7.2	Extended approaches comparison	172
7.2.1	Is there a “best” model based on the Interval Score?	172
7.2.2	Is there an optimal error model for the flood extremes?	177
7.2.3	Parametric vs non-parametric: are there real advantages?	179
7.3	Other operative setups: forecast.....	181
7.4	About the urgency of synthesis.....	182
7.4.1	Generality and synthesis projects efforts.....	182
7.5	Rainfall-runoff model and uncertainty: but for who?.....	184
	Reference.....	186
	APPENDIX.....	200
	R SCRIPTS.....	202

ACKNOWLEDGMENTS

Before getting into the thesis itself, I would like to acknowledge a few people who, in different ways, contributed to this work and supported me throughout my Ph.D.

First of all, I would like to thank my supervisor, Professor Alberto Viglione. Over the past three years, he has guided and supported me as much as possible, always with great care and availability. At the same time, he managed to do so without ever being intrusive or adding unnecessary pressure. This balance is far from obvious, and it is a big reason why, despite the usual challenges of a Ph.D., this period never truly felt overwhelming.

A special acknowledgment also goes to Professor Simon M. Papalexiou. He first supervised me during my visiting period at the University of Calgary and later continued to support me and my research even after I returned to Italy. While Alberto was my main supervisor, Simon has been a key guide, especially when I approached research topics outside my original area of expertise. Without his help, I would likely have gotten stuck or lost at some point. Instead, thanks to him, I was able to stay both productive and rigorous in a field where that is often harder than it seems.

I would also like to thank all the co-authors of the papers produced during these years, those already published, those currently under review, and those still to come. I was always the least experienced person in the group, and I learned a lot from your professionalism. Your example has set a standard I hope to carry forward. Among them, I would like to especially thank Anna Basso, who collaborated closely with me in Turin, and Richard M. Vogel, whose enthusiasm, despite being “officially” retired, is truly contagious.

Finally, a big thank you to everyone who supported me more indirectly: my colleagues at the university, as well as my friends and family.

PREFACE

It is always difficult to find the right way to begin a paper, or in this case, a Ph.D. thesis. One tends to search for the perfect opening sentence, something that can immediately set the stage and capture the spirit of what follows. Unfortunately, this has never been one of my strengths. I have always been more comfortable doing things than describing them in writing. For this reason, I hope the reader will allow me, in this opening paragraph (and only in this one), a slightly more informal and less pragmatic tone, while I briefly introduce the motivations behind the research questions explored in this work. Like most students embarking on a Ph.D., I began this journey with a general topic and a few core ideas already in mind. Still, it would have been impossible to foresee all the challenges, detours, and unexpected turns that would appear along the way, as well as the achievements that would eventually emerge from them. In fact, I should probably consider myself fortunate that, despite the different directions explored over the past three years, some more successfully than others, the topics I worked on always retained a common thread. From my perspective, this has allowed the final result to remain coherent, rather than becoming a simple collection of loosely connected pieces.

Three main themes have guided the work presented in this thesis: uncertainty estimation, prediction in ungauged basins, and river floods. None of these are new problems in hydrology; many researchers before me have studied one or more of them in great depth. How this work attempts to contribute, however modestly, to these topics will form the core of the following chapters. Another element that deserves to be mentioned from the start is the main tool used throughout the research. Among the many possible approaches available to tackle these problems, we chose to rely on conceptual rainfall–runoff models.

Finally, I would like to briefly highlight the general philosophy that guided the research here presented. Although a Ph.D. is often primarily expected to focus on fundamental research, we tried to keep in mind what this work might eventually enable, if not immediately, then in future developments. In other words, the aim was to avoid research carried out purely for its own sake, and instead to pursue ideas that might eventually prove useful in practical or operational contexts. This perspective influenced many of the choices made along the way. Often, when several alternatives were available, the option that appeared simpler, but potentially more attractive for

real-world applications, was preferred. The idea of producing research that would remain confined to the pages of an article or this thesis, without ever potentially contributing to practical use, simply to say “it was done,” was not particularly appealing to us. This is not meant as criticism of purely theoretical research; rather, it reflects my own background and inclination as an engineer. Whether this goal has actually been achieved is, of course, not for me to decide.

But now that the ice has been broken, I will leave the reader to the more formal introduction that follows, hoping that the rest of this thesis will prove, at least to some extent, both interesting and enjoyable to read.

Chapter 1

1. Introduction

1.1 Rainfall-Runoff models

1.1.1 Modelling approaches in rainfall–runoff hydrology

As anticipated in the preface, the main methodological tool used throughout this thesis is the application of rainfall-runoff (RR) models. Given the specific focus on floods, and more generally on high-discharge quantiles, this was not the only possible methodological choice. Conventional purely statistical approaches could also have been adopted for both prediction and the quantification of predictive uncertainty. The choice to rely on RR models is motivated by their main advantage over purely statistical methods: they explicitly represent the temporal evolution of hydrological processes, rather than focusing only on specific statistics. In other words, RR models attempt to reproduce the entire hydrograph at a river cross-section, providing a more comprehensive representation of catchment dynamics. In this work, mainly due to data limitations, the analysis is conducted at the daily time scale. However, most of the methodological framework presented here could in principle be extended to sub-daily (e.g., hourly) resolutions, or more generally to any time scale of interest.

Rainfall–runoff modelling has a long history, dating back more than a century (and even earlier for the simplest formulations). These models were initially developed to address practical operational needs arising from the limitations of hydrological measurements. Hydrological variables cannot be observed continuously in space and time at the desired resolution, and modelling therefore provides a way to infer the behavior of hydrological systems beyond the limited available observations. RR models thus allow extrapolation in both space and time based on incomplete measurements. One of the earliest examples of a rainfall–runoff model can be traced back more than 150 years to the work of Thomas James Mulvaney, who proposed a

simple formulation for estimating peak discharge (Mulvaney, 1851, reproduced by Loague, 2010). This approach later became known as the Rational Method, after being further developed and published by multiple authors using different empirical datasets (see Dooge, 1957).

$$Q_p = CA\bar{R} \quad (\text{Eq. 1.1})$$

where A is the catchment area, $\bar{R}$ is the average rainfall intensity, and C is a runoff coefficient describing the catchment response. Although extremely simple and limited to peak discharge estimation (rather than the full hydrograph), this early model already highlights several challenges that remain relevant in modern RR modelling. As discussed by Keith J. Beven (Beven, 2012), hydrologists still face difficulties in accurately representing the nonlinear response of hydrological systems, particularly during extreme events. In addition, model parameters are generally easier to estimate when observations are available for calibration than in out-of-sample conditions, such as model validation or prediction in ungauged basins. Another persistent challenge is the difficulty of clearly separating the effects of runoff generation and flow routing within model parameterizations. Nevertheless, despite these limitations, RR models can still provide useful predictions. For instance, the Rational Method itself was widely used as a practical engineering design tool throughout much of the twentieth century (e.g., Plate et al., 1988).

The Rational Method would today be classified as an empirical model, also referred to as a black-box model. Empirical (or data-driven) models describe the relationship between hydrological inputs and outputs through statistical or mathematical relationships derived directly from observed data, without explicitly representing the underlying physical processes of the catchment. Their structure is therefore primarily determined by predictive performance rather than hydrological interpretability, and model parameters typically lack a direct physical meaning (Singh, 1995).

Alternative modelling philosophies emerged during the second half of the twentieth century, particularly from the 1960s onward, as the increasing availability of computers allowed the implementation of more computationally demanding models. At one end of the modelling spectrum lie physically based models, which attempt to represent hydrological processes through equations derived from fundamental physical principles, such as the conservation of mass, momentum, and energy. These models explicitly describe processes such as infiltration, evapotranspiration, and subsurface

flow using parameters that correspond to measurable physical properties of the catchment (e.g., soil hydraulic conductivity or porosity). As a consequence, they generally require detailed spatial information and substantial computational resources (Simonovic, 2012). Between empirical and physically based approaches lies a third class: conceptual models. These models represent the dominant hydrological processes through simplified conceptual elements, typically reservoirs and transfer functions designed to reproduce the overall behavior of a catchment. Although inspired by physical understanding, their parameters are usually not directly measurable and must be estimated through calibration against observed discharge data. Conceptual models therefore provide a compromise between physical realism and parsimony, and have become widely used in catchment-scale hydrological studies (Wagener et al., 2004).

While physically based models may in principle allow a more detailed representation of hydrological processes, their practical application at the catchment scale faces significant challenges. Solving the governing differential equations requires detailed information on initial and boundary conditions, which is rarely available in practice. Furthermore, computational demands increase rapidly with the spatial extent of the system, making such models more suitable for localized studies or small experimental basins. Conceptual models, on the other hand, have been successfully applied in hundreds of catchment-scale studies (Perrin et al., 2003; Seibert & Bergström, 2022), making them a practical and widely adopted modelling approach.

Within conceptual RR models, a further distinction is commonly made according to the spatial representation of hydrological processes, leading to three main categories: lumped, semi-distributed, and distributed models. (I) Lumped models (e.g., HBV, Bergström, S., 1995) treat the entire catchment as a single homogeneous unit, assuming that precipitation, evapotranspiration, and other hydrological processes are spatially uniform across the basin. Consequently, the model is forced using spatially averaged inputs and produces discharge at the basin outlet using a single set of parameters representing the integrated catchment behavior. This approach greatly simplifies model structure and calibration, although it neglects internal spatial variability. (II) Semi-distributed models (e.g., TOPMODEL, Beven & Kirkby, 1979) introduce a partial representation of spatial heterogeneity by dividing the catchment into several interconnected sub-units, such as sub-basins or hydrological response units. Each unit is assumed to be internally homogeneous but may differ from others in terms of parameters or forcing inputs. Runoff is simulated separately for each unit and then routed through the river network toward the catchment outlet. This structure allows

spatial variability to be considered while keeping data and computational requirements relatively moderate. (III) Finally, distributed models (e.g., SHE, Abbott et al., 1986a, 1986b) explicitly represent hydrological processes across the spatial domain of the catchment by discretizing the basin into grid cells or small computational elements. Each element has its own state variables and parameters, allowing spatial variability in catchment properties and forcing data to be represented explicitly. Water and energy fluxes are computed locally and subsequently routed through the drainage network. Although this approach provides the most detailed spatial representation of hydrological processes, it generally requires large volumes of input data and significant computational resources. Starting from these considerations, Section 1.1.2 will introduce the specific modelling framework adopted in this work and discuss the reasons behind this selection.

1.1.2 Study area and modelling framework

The applications presented in Chapter 2 to Chapter 6 (with the exception of Chapter 3, although similar considerations still apply), are conducted over the Greater Alpine Region (GAR). A detailed description of the study area, together with the specific information relevant for each modelling application, will be provided in the respective chapters. However, the study area is briefly anticipated here in order to motivate the selection of the class of RR models adopted throughout the thesis for the generation of hydrographs. From a hydrological perspective, the GAR represents a very large and complex region. Applying a fully physically based hydrological model over such an extended domain would not only be impractical but effectively unfeasible. Even under the unrealistic assumption that the required initial and boundary conditions could be accurately characterized, which is not the case at this spatial scale, the computational requirements would be unfeasibly demanding. For these reasons, the use of conceptual RR models represents the only realistic modelling strategy for the type of regional analyses conducted in this work. Nevertheless, this choice should not be interpreted merely as a constraint imposed by the scale of the study area. Conceptual models offer a well-established compromise between process representation, model flexibility, and computational efficiency, which has contributed to their widespread adoption in both hydrological research and operational applications. (Jakeman & Hornberger, 1993; Perrin et al., 2003).

1.1.3 Beyond local predictions: the challenge of prediction in ungauged basins

Conceptual RR models are commonly applied in gauged catchments, where streamflow observations at a measurement station allow for model calibration and validation. However, a large proportion of catchments worldwide are ungauged, meaning that no discharge observations are available. Despite this limitation, the ability to make predictions in ungauged basins is of fundamental importance in applied hydrology. Predicting hydrological behavior in ungauged catchments is still widely regarded as one of the major unresolved challenges in the field (Blöschl, 2006; Vogel, 2005; Wagener et al., 2004). One of the key difficulties is that no universal physical laws exist that allow runoff to be directly predicted at the catchment scale. Much of the theoretical understanding of RR processes derives from observations and experiments conducted at local or laboratory scales, where processes can be more easily controlled and measured (Blöschl, 2005; Dooge, 1987). A possible strategy would be to subdivide a catchment into sufficiently small units so that known physical equations could be applied locally. This approach, often referred to as the Newtonian or reductionist approach, leads to modelling frameworks similar to fully distributed models. As discussed previously, such models are generally unsuitable for large-scale regional studies because they require detailed information on boundary and initial conditions that is rarely available in practice, and they often lead to biased predictions when applied at the catchment scale (Blöschl et al., 2013). For these reasons, most regionalization approaches rely instead on catchment-scale statistical methods. These approaches typically relate hydrological model parameters or hydrological signatures to observable catchment descriptors using regression techniques or similarity-based methods (Richard M. Vogel, 2005). The general idea is that catchments that are similar from a morpho-climatic perspective are likely to exhibit similar hydrological responses. In practice, parameters are first estimated through calibration in gauged catchments and are subsequently related to catchment descriptors through regression models. In this way, the regionalization problem can be formulated as a statistical regression problem in which model parameters represent the dependent variables and catchment descriptors act as explanatory variables. Numerous examples of such approaches have been reported in the hydrological literature across different climatic and physiographic regions (e.g., Boughton & Chiew, 2007; Fernandez et al., 2000; Hundecha et al., 2008; Merz & Blöschl, 2004; Oudin et al., 2008; Xie et al., 2007). Despite their widespread application, regression-based regionalization methods present several conceptual limitations. Ideally, the statistical relationships linking model

parameters to catchment descriptors should be consistent with hydrological understanding. However, although hydrological reasoning can guide the selection of descriptors, the resulting relationships often remain purely statistical and their physical interpretation may not always be straightforward (Beven, 2012). Furthermore, the problem of equifinality (see Section 1.2.1) implies that multiple parameter sets may produce similarly acceptable model performances, making it difficult to determine which calibrated parameters should be used when deriving regional regression relationships. In addition to regression approaches, regionalization methods often rely on catchment clustering or pooling-group techniques. In these cases, similarity measures are defined in order to quantify the degree of similarity between catchments, and information from gauged basins is transferred to ungauged locations based on this similarity. Typically, similarity is evaluated as a distance in a multidimensional space defined by normalized catchment descriptors. However, this strategy also raises methodological questions, such as the choice of the similarity metric and the selection of the descriptors to be included. Various similarity measures have been proposed in the literature (see Wagener et al., 2007), but no universally accepted criteria exist for their selection. Moreover, it has been argued that physical similarity does not necessarily imply hydrological similarity, thereby challenging the underlying assumption of descriptor-based similarity approaches (Oudin et al., 2010). In response to the need for a more coordinated effort to address these issues, the Prediction in Ungauged Basins (PUB) initiative was launched by the International Association of Hydrological Sciences with the aim of consolidating existing knowledge and improving methodologies for hydrological prediction in ungauged catchments. The initiative brought together a large international research community and concluded in 2013 with the publication of a synthesis volume edited by Günter Blöschl et al. (Blöschl et al., 2013). Within the broader context of research on prediction in ungauged basins, another important initiative was the Model Parameterization Experiment (MOPEX), which evaluated the performance of several rainfall–runoff models across a large set of catchments distributed worldwide (see Q. Duan et al., 2006). Results from this experiment showed that when model parameters were specified a priori, without calibration using local observations, none of the tested models was able to reproduce observed discharge satisfactorily across all catchments. The MOPEX dataset has since been widely used to test alternative modelling approaches and regionalization strategies aimed at improving predictions in ungauged basins (e.g., Ao et al., 2006; Huang & Liang, 2006). These studies highlight that the regionalization of hydrological

models remains a challenging and partially unresolved problem. More generally, hydrological predictions are affected by multiple sources of uncertainty that arise even in well-instrumented gauged catchments and become even more pronounced when predictions are extended to ungauged basins. The following section therefore discusses the main sources of uncertainty in RR modelling.

1.2 Uncertainty in conceptual Rainfall-Runoff models

1.2.1 Multiple sources of uncertainty

The challenges associated with regionalization and prediction in ungauged basins highlight a broader issue that affects hydrological modelling more generally, namely the presence of multiple sources of uncertainty in RR predictions. Uncertainty is inherent to hydrological modelling and arises even when models are applied in gauged catchments where calibration and validation data are available. Hydrological systems are highly complex, nonlinear, and spatially heterogeneous, while the available observations are typically sparse and affected by measurement errors. As a consequence, any modelling framework inevitably involves a degree of approximation, and model predictions must always be interpreted in light of the uncertainties affecting both the data and the modelling process.

Several sources of uncertainty can be identified in RR modelling. A first important component is related to input data, particularly meteorological forcings such as precipitation and temperature. These variables are often measured at a limited number of stations and must be interpolated in space, which may introduce errors and representativeness issues, especially in regions characterized by complex topography or strong spatial variability of precipitation. Additional uncertainty arises from errors in the observational data used for model calibration and validation, such as inaccuracies in streamflow measurements or rating curves.

Another major source of uncertainty concerns the parameters of the model. In conceptual RR models, parameters typically represent aggregated characteristics of the catchment and are rarely directly measurable. Instead, they must be estimated through calibration procedures based on observed discharge data. The resulting parameter estimates depend not only on the available observations but also on the chosen calibration strategy, the objective function used to evaluate model performance, and the specific period used for calibration. Consequently, different calibration choices

may lead to different parameter sets, each providing an acceptable representation of the observed hydrograph.

A further important component of uncertainty is associated with the structure of the model itself. All hydrological models represent simplified descriptions of the real system and rely on assumptions about the dominant processes controlling the RR transformation. These simplifications inevitably introduce structural uncertainty, as some processes may be neglected or represented in an approximate way. Structural uncertainty is particularly relevant in conceptual models, where complex physical processes are represented through simplified reservoirs and transfer functions designed to reproduce the overall behavior of the catchment rather than its detailed internal dynamics. Additional uncertainties may also arise from imperfect knowledge of the initial conditions of the model states at the beginning of the simulation period.

These different components are often broadly classified into two main categories: aleatory uncertainty, which reflects the inherent variability and randomness of natural processes, and epistemic uncertainty, which arises from incomplete knowledge of the system being modelled and from limitations in data, model structure, or parameter estimation (Beven, 2012). While aleatory uncertainty is generally considered irreducible, epistemic uncertainty may be partially reduced through improved observations, better model structures, or more robust estimation techniques.

An important conceptual framework for understanding uncertainty in hydrological modelling is the concept of equifinality, introduced and extensively discussed by Keith J. Beven. The equifinality hypothesis states that, in complex environmental systems, multiple different model structures or parameter sets may produce similarly acceptable simulations of the observed data. In other words, the available observations may not contain sufficient information to uniquely identify a single “optimal” representation of the system. As a consequence, calibration procedures often identify not a single optimal parameter set but rather a range of parameter combinations that yield comparable model performance. This situation reflects both the limited information content of the available data and the simplifying assumptions inherent in model structures. Equifinality has important implications for RR modelling. It suggests that the goal of model calibration should not necessarily be the identification of a unique optimal parameter set, but rather the characterization of the range of plausible model representations that are consistent with the available observations. Different parameter sets within this range may lead to different predictions, particularly under conditions that differ from those used for calibration, such as extreme events or simulations in

ungauged basins. For this reason, explicitly accounting for uncertainty has become a central aspect of modern hydrological modelling. A variety of methodologies have therefore been developed to quantify and propagate uncertainty in RR predictions, including approaches based on Monte Carlo sampling, likelihood-based inference, and Bayesian frameworks. These approaches will be introduced and discussed in the following sections.

1.2.2 Generalized Likelihood Uncertainty Estimation - GLUE

The first approach to RR model uncertainty estimation here presented is the Generalized Likelihood Uncertainty Estimation (GLUE - Beven & Binley, 1992). The GLUE approach is based on the same concept of equifinality already described in Section 1.2.1, accepting that multiple set of different (often substantially) calibrated parameter sets can produce similar runoff prediction in terms of the selected efficiency measure selected in calibration. If no “optimal” parameter set can be identified, it seems natural to start considering multiple parameter sets, each with associated a degree of belief. The optimal parameter set will have the highest belief, but the multiple other parameter sets with similar performance will now still be accounted for with a lower, yet close, degree of belief. At the same time, parameter sets that generate predictions that do not reproduce the observed runoff process can be rejected by associating a degree of belief equal to 0. The GLUE approach then simply account for uncertainty by a weighted average, with weights proportional to the degrees of belief, of the parameter sets with degree of belief greater than 0. The generation of multiple parameter sets follow a similar procedure to the Bayesian approaches later described in Section 1.2.3: a prior distribution is defined for each RR model parameter; at each iteration (in Monte-Carlo style) for each parameter a realization is sampled from its prior distribution and the model is forced at each iteration with the same input. The resulting runoff is then associated with its degree of belief using a likelihood measure. The GLUE method can also be conceptually easily extended for multi-model and/or multiple calibration period scenarios.

Despite being one, if not the most used RR model uncertainty estimation method, several critiques (Beven et al., 2007; M. P. Clark et al., 2011, 2012; Montanari, 2005; Stedinger et al., 2008) have underscored the limitations of the GLUE approach, opening a large discussion and calling for more rigorous treatments of residual distributions. A first and frequently raised criticism concerns the statistical interpretation of the likelihood measure adopted within GLUE. In many applications,

the likelihood function used to assign degrees of belief to the sampled parameter sets does not correspond to a true likelihood in the strict statistical sense, but rather to a performance metric or an informal transformation of an efficiency measure. As a consequence, the resulting weights cannot always be interpreted as probabilities in a fully coherent Bayesian framework. This aspect has led some authors to classify GLUE as an “informal Bayesian” approach, highlighting that it does not strictly follow the requirements of Bayesian inference. Another commonly discussed issue relates to the definition of the so-called behavioral threshold. In the GLUE framework, parameter sets are typically classified as behavioral or non-behavioral depending on whether their likelihood measure exceeds a predefined threshold. However, the choice of this threshold is often subjective and may significantly influence the resulting uncertainty bounds. Different thresholds may lead to different sets of accepted parameter combinations and therefore to different predictive uncertainty estimates. Similarly, the selection of the likelihood measure itself and the definition of the prior parameter distributions introduce additional degrees of subjectivity that can affect the final results.

Nevertheless, despite these criticisms, the GLUE approach presents several advantages that have contributed to its widespread adoption in hydrological applications. One of its most important strengths lies in its conceptual flexibility and in its explicit acknowledgment of equifinality in environmental modelling. By allowing multiple parameter sets to contribute to the predictive distribution, GLUE naturally accounts for the fact that complex hydrological systems are often poorly constrained by the available observations. In this sense, the method emphasizes the exploratory nature of model calibration and uncertainty analysis, rather than focusing exclusively on the identification of a single optimal parameter set. Another important advantage of the GLUE framework is that it does not require strong assumptions regarding the statistical properties of the model residuals. Many formal statistical approaches rely on specific assumptions about the distribution of residual errors, such as normality, independence, and homoscedasticity. However, these assumptions are rarely satisfied in hydrological applications, where residuals often exhibit heteroscedasticity, temporal correlation, and systematic biases due to structural deficiencies in the model. The GLUE methodology avoids imposing a rigid statistical structure on the residuals and instead evaluates model performance through likelihood measures that can be tailored to the specific objectives of the analysis. This flexibility allows the modeler to incorporate alternative performance criteria, for example emphasizing peak flows, low

flows, or other hydrological signatures of interest. Furthermore, the GLUE approach is relatively simple to implement and computationally straightforward, as it relies primarily on Monte Carlo sampling of the parameter space. This simplicity has made it particularly attractive in situations where the structure of the residual errors is difficult to characterize or when the goal is to obtain a first-order assessment of predictive uncertainty.

While GLUE remains a flexible and widely used framework for uncertainty analysis, the methodological debate surrounding its statistical interpretation has stimulated the development of alternative approaches based on formal Bayesian inference. These methods aim to provide a more rigorous probabilistic framework for parameter estimation and uncertainty quantification in RR modelling. The main principles of Bayesian uncertainty estimation are therefore introduced in the following section.

1.2.3 Formal Bayesian Approaches

In conceptual RR modeling, Bayesian inference provides a formal framework to quantify predictive uncertainty by combining prior knowledge about model parameters with observed data. Formally, the posterior distribution of model parameters θ given observed discharges Q^0 is expressed by Bayes' theorem:

$$p(\theta | Q^0) = \frac{p(Q^0 | \theta) p(\theta)}{p(Q^0)} \quad (\text{Eq. 1.2})$$

where $p(\theta)$ is the prior distribution of parameters, $p(Q^0 | \theta)$ is the likelihood function representing the statistical behavior of model residuals, and $p(Q^0)$ is the marginal likelihood ensuring normalization. In RR models, residuals can be defined as

$$\epsilon_t = Q_t^0 - Q_t^S(\theta, I) \quad (\text{Eq. 1.3})$$

with I representing input data such as precipitation or evaporation. For simple cases, residuals may be assumed independent and Gaussian, leading to a likelihood of the form

$$L(\theta \mid Q^0) = \prod_{t=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon_t^2}{2\sigma^2}\right) \quad (\text{Eq. 1.4})$$

while more complex likelihoods can incorporate autocorrelation, heteroscedasticity, or multiplicative errors to better reflect hydrological processes (Evin et al., 2013, 2014; Schoups & Vrugt, 2010; Tajiki et al., 2020). Within this framework, joint inference approaches estimate model parameters, error parameters, and sometimes input or observation uncertainties simultaneously, providing a statistically coherent representation of predictive uncertainty and capturing interactions between model and error components (Kuczera et al., 2006; Renard et al., 2011; Schaeffli et al., 2007). Joint inference also allows, in principle, for disaggregation of uncertainty sources, separating contributions from parameters, inputs, structural deficiencies, and observation errors (Renard et al., 2010).

Despite these advantages, the practical implementation of Bayesian inference in RR models faces several substantial challenges. First, the residuals themselves are complex and often violate common statistical assumptions (see Chapter 3). They can exhibit strong temporal autocorrelation and can be heteroscedastic, with variance increasing for higher flows. They may also be non-Gaussian or influenced by structural model errors and uncertain inputs, making it difficult to define a likelihood function that fully captures their behavior (Draper, 1997; Feyen et al., 2007). Applying standard iid Gaussian likelihoods in such cases can lead to over-conditioning, producing highly peaked posterior distributions where small differences in error variance are exaggerated, resulting in misleading confidence in specific parameter sets. Second, equifinality presents a persistent problem: multiple combinations of parameter values can produce similar model outputs, yielding multimodal posterior distributions that are difficult to sample and interpret (Beven et al., 2008). This not only complicates numerical convergence in MCMC (Markov Chain Monte Carlo) simulations but also challenges the identifiability of model parameters, particularly in high-dimensional spaces. Different realizations of the calibration dataset can shift posterior peaks, highlighting the sensitivity of parameter estimates to the particular data used for calibration. Third, methods that attempt to “normalize” residuals, such as the Box–Cox transformation (Box & Cox, 1964) or the normal quantile transform (Kelly & Krzysztofowicz, 1997; Montanari & Brath, 2004), can mitigate some statistical issues but introduce approximations that may distort the underlying information content of

the data. For example, extreme events may be under- or over-weighted, affecting uncertainty estimates for high flows, which are often the primary hydrological interest. Finally, the computational burden of joint Bayesian inference for complex conceptual RR models is considerable. Each iteration of a high-dimensional MCMC requires a full model simulation, and the presence of correlated parameters and complex residual structures further slows convergence. While these computational challenges can be partially mitigated through advanced sampling techniques or surrogate models, they remain a practical limitation for large-scale or high-resolution applications.

In summary, while Bayesian inference provides a theoretically rigorous framework for probabilistic uncertainty quantification in RR models, its practical application is constrained by the complexity of residual behavior, equifinality, sensitivity to calibration data, and computational cost. These challenges highlight the importance of carefully specifying likelihood functions, verifying residual assumptions, and, when necessary, considering complementary approaches such as post-processing methods.

1.2.4 Post-processing (PP) approaches

Building on the previous discussion of GLUE and Bayesian joint inference, an alternative strategy that has gained increasing attention in hydrological modeling is represented by post-processing (PP) approaches for uncertainty estimation. While both GLUE and Bayesian inference attempt to estimate model parameters and predictive uncertainty within a unified calibration framework, post-processing methods adopt a different philosophy: the rainfall–runoff model is first calibrated using a chosen objective function, and the uncertainty of the resulting predictions is then estimated in a second step by statistically analyzing and correcting the residual errors between observed and simulated discharge (Evin et al., 2014). In other words, rather than embedding the uncertainty estimation directly within the parameter estimation procedure, PP approaches treat the hydrological model as a deterministic predictor and subsequently model the statistical properties of its errors. This separation between calibration and uncertainty estimation offers several practical advantages. First, it significantly reduces the computational burden relative to full Bayesian joint inference, where the evaluation of the posterior distribution typically requires thousands or millions of model simulations within a MCMC scheme. In contrast, PP methods rely on a single calibrated model (or a limited ensemble of calibrated models) and focus on the residual structure of the resulting predictions. Second, the separation alleviates some of the identifiability issues that arise in joint inference frameworks when model

parameters and error model parameters are estimated simultaneously. When both sets of parameters are inferred together, interactions between them can lead to equifinality and poorly identifiable posterior distributions, particularly in conceptual RR models with relatively simple structures but complex error behavior (Evin et al., 2013; Renard et al., 2011). By estimating the hydrological model parameters first, PP methods reduce the dimensionality of the statistical inference problem and often lead to more stable uncertainty estimates in practice.

Within the broad class of PP approaches, different methodological variants can be identified. Some methods remain likelihood-based, in the sense that they assume a statistical model for the residual errors and use it to derive predictive distributions for future simulations. Others adopt likelihood-free or semi-parametric strategies, relying on resampling or empirical distributions of residuals to generate predictive uncertainty bounds (Montanari & Koutsoyiannis, 2012; Quilty & Adamowski, 2020). In this context, bootstrap-based procedures have been widely used to estimate prediction intervals while relaxing strict assumptions on the residual distribution (Sikorska et al., 2015; Koutsoyiannis & Montanari, 2022a; Shabestanipour et al., 2023). These techniques allow the empirical distribution of residuals to be propagated into future simulations, providing uncertainty estimates even when the residuals are heteroscedastic, skewed, or non-Gaussian. Another important distinction within PP frameworks concerns whether the different sources of uncertainty are explicitly disaggregated or instead aggregated into a single residual term. In principle, uncertainty in RR modeling originates from multiple sources, including errors in forcing data, parameter uncertainty, structural deficiencies of the model, and observation errors in discharge measurements. Some approaches attempt to explicitly represent these components separately (Renard et al., 2010), providing valuable diagnostic insight into the origin of predictive uncertainty. However, in practice this disaggregation can introduce additional identifiability problems, since different uncertainty sources may produce similar effects on model outputs. For this reason, several studies have shown that aggregated residual approaches, in which all sources of uncertainty are treated as a single error term, can produce more robust and reliable prediction intervals (e.g., Valdez et al., 2022). Despite their operational simplicity, PP methods remain grounded in a formal statistical framework. They still require assumptions about the statistical behavior of model residuals and therefore inherit some of the challenges discussed in the context of Bayesian inference. Residuals in RR modeling often exhibit temporal autocorrelation, heteroscedasticity, and non-stationary behavior, particularly during

extreme hydrological events. Even non-parametric approaches such as bootstrap resampling rely on implicit assumptions about the representativeness and independence of the residual samples used to generate prediction intervals. As a result, careful diagnostic analysis of residual properties remains an essential step in the application of PP techniques. Nevertheless, when compared with the two approaches discussed previously, PP methods offer an appealing compromise between theoretical rigor and practical feasibility. Unlike GLUE, which has been criticized for relying on subjective likelihood measures and lacking a clear probabilistic interpretation, PP methods maintain a closer connection with statistical inference and the probabilistic characterization of residual errors. At the same time, they avoid some of the computational complexity and identifiability challenges associated with fully Bayesian joint inference frameworks. For these reasons, the research presented in this thesis focuses exclusively on post-processing approaches for uncertainty estimation. This choice reflects the general philosophy that guided the work, namely the prioritization of methods that remain theoretically sound while also being sufficiently simple and robust to be applicable in practical or operational contexts. Within the wide range of available uncertainty estimation techniques, PP methods appear to offer a particularly suitable balance between methodological rigor, computational efficiency, and practical applicability. To conclude this introductory chapter, the structure of the remainder of the thesis is briefly outlined. Chapter 2 presents an application of a deterministic regional RR model for prediction in ungauged basins across the Greater Alpine Region, with particular emphasis on large flood events. This case study also serves to illustrate some of the uncertainties inherently associated with RR modelling and prediction in ungauged catchments. Chapter 3 provides a detailed analysis of the statistical properties of RR model residuals, with the aim of identifying general patterns in their behavior across different catchments and models. This analysis lays the foundation for the uncertainty estimation methods explored in the subsequent chapters. Chapter 4 and Chapter 5 then investigate and evaluate different post-processing strategies for quantifying predictive uncertainty in conceptual RR models at gauged locations. Finally, Chapter 6 addresses the less explored problem of estimating predictive uncertainty for ungauged basins, providing initial insights into how the approaches discussed earlier may be extended to this more challenging context.

Chapter 2

2. Regional Rainfall-runoff model for the Greater Alpine Region

2.1 A model fit to purpose: why and how we model floods

In the preface and Chapter 1, a general overview of RR models and the challenges associated with regionalization for prediction in ungauged basins has been presented. These concepts are now reintroduced through the practical application described in this chapter. Rather than repeating the material already discussed, the aim here is to highlight an additional point that is essential for properly framing the work that follows. RR models can be developed for a wide variety of purposes, and their calibration, validation, and, more generally, the choices involved in their construction strongly depend on the specific objective for which the model is intended. Even when the same model architecture is adopted, different modelling choices may lead to substantially different outcomes depending on the aspect of the hydrological response that is of interest. The idea that hydrological models should be developed as “fit-to-purpose” tools is well established in the hydrological literature (Beven, 2012; M. Clark et al., 2011). A natural question therefore arises: why develop a regional model, and for what purpose should it be considered fit-to-purpose? In this work, among the many possible characteristics that can be derived from a hydrograph, particular attention is given to flood events and, more generally, to the upper part of the discharge distribution (i.e., high-flow quantiles). Floods are among the most impactful hydrological hazards, causing significant economic losses and threats to human safety worldwide. Their analysis and prediction therefore represent a central topic in hydrology and water resources management (Blöschl et al., 2013; B. Merz et al., 2010). From a scientific perspective, extreme flows are also particularly challenging to model, as they often arise from complex interactions between meteorological forcings, catchment

processes, and river network dynamics. As a consequence, accurately reproducing flood magnitudes and frequencies remains a major objective in hydrological modelling and regional analysis.

Given this focus, a further question arises: why rely on a regional RR model rather than on more traditional statistical approaches for flood analysis? Classical flood frequency analysis is typically performed at individual gauging stations and focuses on estimating the probability distribution of annual maxima or peak flows at a single location. While such approaches are extremely valuable, they are inherently point-based. In contrast, regional flood events are distributed not only in time but also across space, often affecting large portions of a river network simultaneously. To investigate these spatial patterns, it is useful to rely on modelling frameworks capable of reproducing the hydrological response across an entire river network rather than at isolated points. For this reason, the approach adopted here relies on a regional RR model able to simulate discharge across multiple catchments simultaneously. This consideration also motivated one of the modelling choices introduced later in Section 2.3.1, where a modification to an existing RR model was implemented to explicitly represent the routing of water through the river network. Such a representation allows the model to reproduce not only the magnitude of flood events at individual locations but also their spatial propagation along the river network

It should be noted, however, that while the focus on floods remains a central theme throughout this thesis, the explicit representation of spatial flood dynamics was partly motivated by research objectives extending beyond the scope of the present work. The model described in this chapter was originally developed with subsequent studies in mind, and some of the modelling choices were therefore influenced by requirements that may appear less directly relevant for the specific objectives of this thesis. For the purposes of the present work, the model should therefore be regarded primarily as an illustrative example of regional RR model. In particular, it provides a useful framework for highlighting some of the limitations, biases, and sources of uncertainty that inevitably arise in regional hydrological modelling, especially within the discharge ranges that are most critical for risk-related and operational applications. These aspects will naturally motivate the focus on uncertainty estimation that forms the core of the following Chapters 3 to 6.

2.2 Study Area: the scale as a challenge

2.2.1 Constrains and implications in modelling the GAR

The following chapter will describe in depth the construction and validation of a regional RR model for the entire Greater Alpine Region (GAR). The GAR is a particularly complicated area to model due to its climatic and physiographic characteristics. Most of the catchments are of small or medium size, already very well known to be harder to represent due to lower flows and reduced data availability, and are mostly 100% mountainous alpine catchments, dominated by snow related processes (accumulation and melting). The characteristics of the region makes it of fundamental importance for the countries that relies on the mountains as a reservoir of water, and make it particularly vulnerable to the effects of climate change, upon all the rise in temperatures. Another additional challenge in modeling the GAR is due to the transboundary nature of the region (it extents over Italy, France, Austria, Switzerland, and Slovenia) and most of its catchments: this turned out to be particularly relevant in the process of data collection of climate and discharge data, used to force and calibrate the model, posing some challenging constrains in modelling decisions.

It is important to clarify since the very beginning some terminology that will be used throughout the following sections that may not be otherwise clear by someone non-familiar with this field of research: in first place, when referring to a “local model”, it will be intended the process of fine tuning the parameters of the selected RR model architecture for a specific river cross section at which observed discharge measurements are available (gauged site). The parameter set that optimizes an arbitrary (there can be more than a single possible choice) efficiency function, is said to define the “best” parameterization of the model. The same concept of “best” is somewhat arbitrary, and depends on the Model Efficiency (ME) function applied in calibration as well as on the specific hydrologic signatures of interest the model is aimed at reproducing. Whit “regional model” is instead intended the process of defining a Regional Functional Relationship (RFR) valid for all the catchments in the region, connecting some characteristics of the catchments (usually called catchment descriptors or CDs for brevity) to the model parameters. The process of inferring such RFR starts from the local model calibrated parameters at multiple locations (only gauged), and once defined can be applied to all catchments (also ungauged) because the CDs, differently from the discharge observations, are continuously available everywhere. There are several ways to define these connections between CDs and

model parameters, but all are generally based on the assumption of similarity: similar catchments (so with similar CDs) will tend to exhibit a similar hydrologic response to climate forcings (and so will have similar model parameters). Later in Section 2.3.4 it will be presented the specific algorithm used for the GAR. Another fundamental concept is that a regional model, trying to define a single “rule” to define model parameters for a region, is expected to return lower values of the ME at the gauge location when compared to the local models there constructed. The goal of a regional model is to well represent, on average, all the catchments, and to do so some trade off at individual locations are expected and necessary. At the same time, some catchments may exhibit runoff generation mechanisms that differ substantially from those of the other basins in the region, despite having similar catchment descriptors (CDs). In such cases, the estimated RFR, which essentially captures average relationships across the region, may fail to represent these sites adequately, leading to very low regional predictive efficiencies. These additional challenges, compared with locally calibrated models, are compensated by the main advantage of a regional approach: once validated, it enables the generation of continuous discharge time series at ungauged locations. This, together with an estimate of the reliability of the predictions, allows hydrological analyses to extend well beyond the limited number of gauged catchments. It is worth recalling that, worldwide, the majority of river catchments are either ungauged or poorly gauged (Hrachowitz et al., 2013; Sivapalan et al., 2003).

When discussing model validation in regional modelling studies, it is necessary to distinguish between two types of validation. First, temporal validation, which refers to evaluating discharge predictions over a time period different from the one used to locally calibrate model parameters. Second, spatial validation, which concerns the assessment of model performance in ungauged basins, typically achieved using cross-validation techniques such as Leave-One-Out, as described in later sections. These two approaches can also be combined to perform spatial-temporal validation, which represents the most challenging case, where predictions in ungauged basins are evaluated over a time period different from the calibration period. A final note concerns the terminology of Hydrological Units (HUs) used in the model. As described in Section 2.2.4, catchments are discretized into HUs, which can be interpreted as land areas connected through water flow pathways. Each HU may receive inflow from multiple upstream HUs but has only one downstream outflow connection, resulting in a single HU directly connected downstream. Each catchment is therefore composed of multiple HUs. In cases where one catchment is nested within another, the smaller

catchment shares all its HUs with the larger one, while the larger catchment includes additional downstream-connected units. HUs and their connections define the river network and water flow structure of the region. Since HUs are the fundamental modelling blocks, their spatial size should be comparable with the resolution of the climate forcing data. Very small HUs may not be adequately represented when driven by coarse-resolution inputs. In general, the input data resolution should be comparable to the average characteristic size of the HUs used in the model.

2.2.2 Climate Input Data

The climate dataset used in this study was selected based on two primary criteria. First, its spatial resolution needed to be compatible with the average size of the model's HUs. Coarse-resolution datasets risk failing to capture critical spatial variability. Second, the dataset had to provide full coverage of the study area to ensure consistency in climate forcings across all HUs. Based on these constraints, two regional reanalysis products were initially considered: UERRA (Uncertainties in Ensembles of Regional ReAnalyses - spatial resolution ~5.5 km, daily time resolution) and ERA5-Land (ECMWF Reanalysis v5 - spatial resolution ~9 km, hourly time resolution). To assess their suitability, two complete local simulations of the model were performed over multiple test catchments in Piedmont, Italy (a western sub-region of the GAR), following the same methodology later described in Sections 2.3.3 and 2.3.4.

In order to evaluate the impact of this selection compared to the use of local/national datasets (such as those from Meteo-France or the Italian ARPA network), Section 2.4.2 also presents, for same smaller sub-region, a local calibration using the Optimal Interpolation (OI) dataset produced by ARPA Piemonte (NWIOI, 2011). Although using higher-quality local datasets may improve model performance in specific sub-regions, doing so requires patching different products and could introduce spatially inconsistent biases due to differing methodologies in producing the datasets, undermining comparability and compromising the study's objectives.

The hydrological model used operates at a daily time step and requires three primary climate inputs: total daily precipitation (P), mean daily temperature (T), and daily potential evapotranspiration (PET). Precipitation and temperature were directly obtained from the global reanalysis product. PET was calculated using the temperature-based formula proposed by Oudin et al., 2005,

$$\begin{cases} \text{PET} = \frac{R_e}{\lambda \rho} * \frac{T_a + 5}{100} & \text{if } T_a + 5 > 0 \\ \text{PET} = 0 & \text{otherwise} \end{cases} \quad (\text{Eq. 2.1})$$

Where R_e is the amount of extraterrestrial radiation (function only of latitude and day of the year), λ is the latent heat flux, ρ is the density of water, and T_a is the mean daily temperature. Latitude is derived from the centroid of each grid cell. Although this method provides less accurate estimates than physically based approaches such as Priestley–Taylor or Penman–Monteith, its minimal data requirements and computational efficiency make it appropriate for application at the regional scale of this study. Moreover, being the focus of the current study on floods, evapotranspiration fluxes are often negligible when compared to precipitation fluxes during the intense precipitation events that triggered those events.

2.2.3 Discharge observations

The discharge dataset used for model calibration, regionalization, and validation consists of daily runoff time series from gauge stations distributed across the entire Greater Alpine Region (GAR). A comprehensive data collection effort was conducted to collect all available raw data from multiple sources. The main contributors included the Global Runoff Data Centre (GRDC) and national or regional environmental agencies in Italy, France, Switzerland, Germany, Austria, and Slovenia. Although some agencies provide discharge measurements at hourly resolution, others offer only daily records. To ensure consistency across the dataset, daily time steps were adopted throughout the analysis. The raw dataset was subsequently filtered using two main criteria: temporal coverage and data quality. For temporal consistency, a 20-year time window (2000–2020) was selected. This period was divided into a 10-year calibration period (2010–2019) and a 10-year validation period (2000–2009), preceded by a two-years model warm-up period (1998–1999). A station, to be considered suitable, needed at least 7 years of data in both sub-periods. In terms of quality control, the primary objective was to exclude stations exhibiting strong anthropogenic influences inconsistent with the assumptions of the hydrological model, which is designed to simulate near-natural catchment responses. While moderate level of human modifications are common and tolerated, basins dominated by infrastructure such as large dams, hydropower facilities, or major water abstractions were excluded due to their impact on runoff regimes.

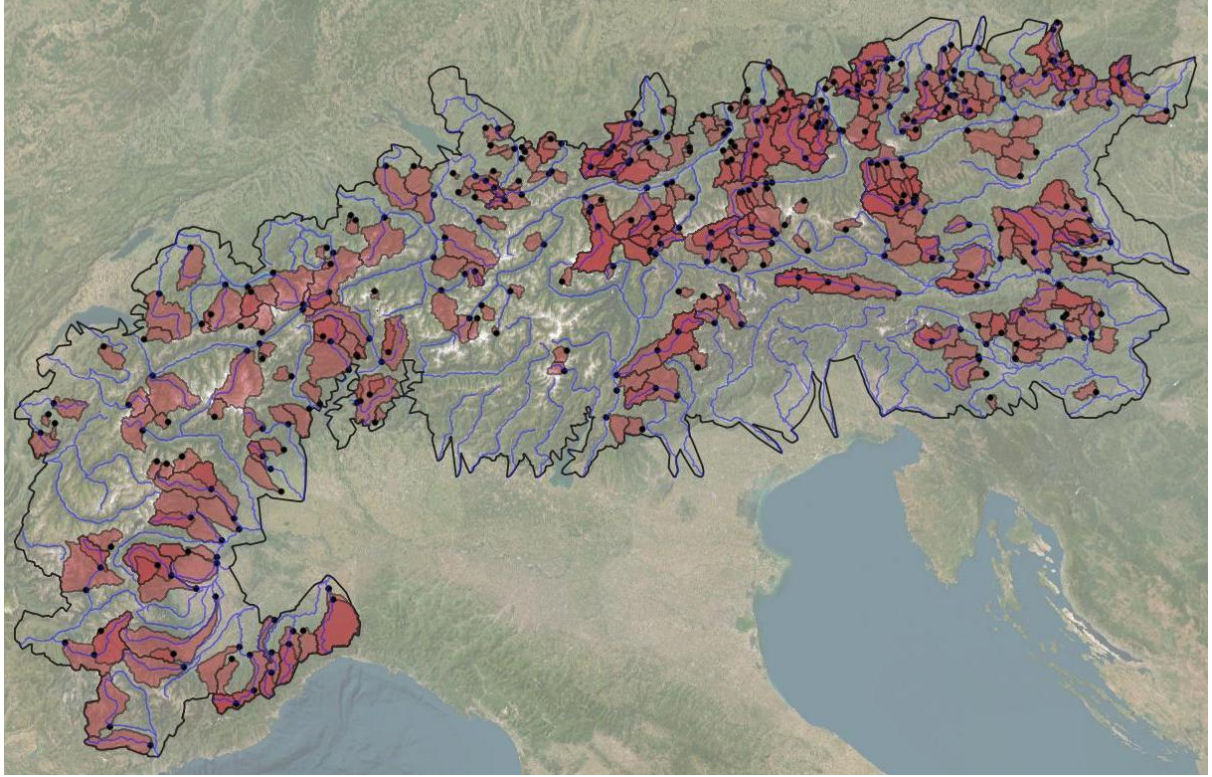


Figure 1. Location of the discharge gauges used in calibration throughout the study (points) and delimited catchments. Darker shades for areas shared across multiple nested catchments.

Moreover, only the catchments with an area lower than 1000 km^2 have been considered as required for the proper application of the regionalization approach (see Section 2.3.4). Additional filtering steps were applied, as detailed in Sections 2.3.3, to ensure proper attribution of gauge stations to HUs and to support different regionalization configurations. The selection resulted in a final total of 369 sites usable for model calibration and regionalization across the different setups. A full overview of the gauge stations and the corresponding delimited catchments included in this study is provided in Figure 1.

2.2.4 Hydrological Units

HUs in the model are defined by the spatial boundaries of sub-basins. Combined with their topological structure, that is, how each unit connects to upstream and

downstream neighbors, these HUs constitute the fundamental “skeleton” of the hydrological model. Two strategies were considered for defining the HUs: (1) using a pre-existing dataset covering the entire Greater Alpine Region, or (2) generating a new dataset from scratch. In the first case, predefined sub-basins had to be validated to ensure that each discharge gauge corresponds to one of the designated catchments outlet points. This required attributing gauge locations to the closest available HU outlet. While this method is more time-efficient, it may introduce significant spatial mismatches in contributing areas to the outlets, which can degrade model performance. The alternative approach involves deriving HUs directly from a Digital Elevation Model (DEM), allowing outlet points to be more accurately aligned with gauge station locations using widely available GIS tools (e.g., GRASS). Although more time-consuming, this method offers greater flexibility and can improve the spatial accuracy of station-to-sub-basin assignments.

The ECRINS (European Catchments and Rivers Network System, EEA, 2012) dataset was initially evaluated as a candidate, given its complete regional coverage, internal consistency, and sub-basins with an average area of approximately 62 km². As a comparison, a custom HUs dataset was also developed to assess whether a DEM-derived configuration could reduce spatial attribution errors. The validation methodology, evaluation results, and the final selection of the HUs dataset are detailed in Section 2.3.2 and Section 2.4.1.

2.3 Model and Methods

2.3.1 Model structure

The hydrological model used in this study is a modified version of the TUWmodel (Parajka et al., 2007). The original TUWmodel consists of a set of routines that simulate the hydrological behavior of a catchment: a snow routine (for accumulation and melt), a soil moisture routine, and a response and transfer function representing routing on hillslopes. The standard model can be applied in lumped or semi-distributed form but is limited to simulations of entire catchments, as it does not allow for the explicit transfer of water between interconnected sub-basins in a fully distributed configuration. Given the spatial scale of this study and the nature of the simulated events of interest, often large-scale and regional, it was essential to incorporate flows between HUs. To address this, the original TUWmodel (referred to hereafter as the

“generative” component, responsible for converting climate inputs into runoff volumes) was extended by coupling it with a Nash Cascade routing scheme (Nash, 1957). The Nash Cascade is a series of identical linear reservoirs with an analytical solution for its Instantaneous Unit Hydrograph (IUH), making it computationally efficient when applied via convolution with the discharge input signal:

$$IUH[t] = \frac{1}{k^N(N-1)!} * t^{N-1} * \exp^{-\left(\frac{t}{k}\right)} \quad (\text{Eq. 2.2})$$

Where N are the number of identical linear reservoirs constituting the Nash-Cascade, t is the timestep of the simulation and k is the response constant of the reservoirs. The input to the routing module consists of the sum of the generative component for a HU and the inflow from all upstream HUs. If multiple upstream sub-basins are present, their contributions are added up. If no upstream HU exists, only the generative runoff is routed. A schematic representation of the modified model structure is provided in Figure 2.

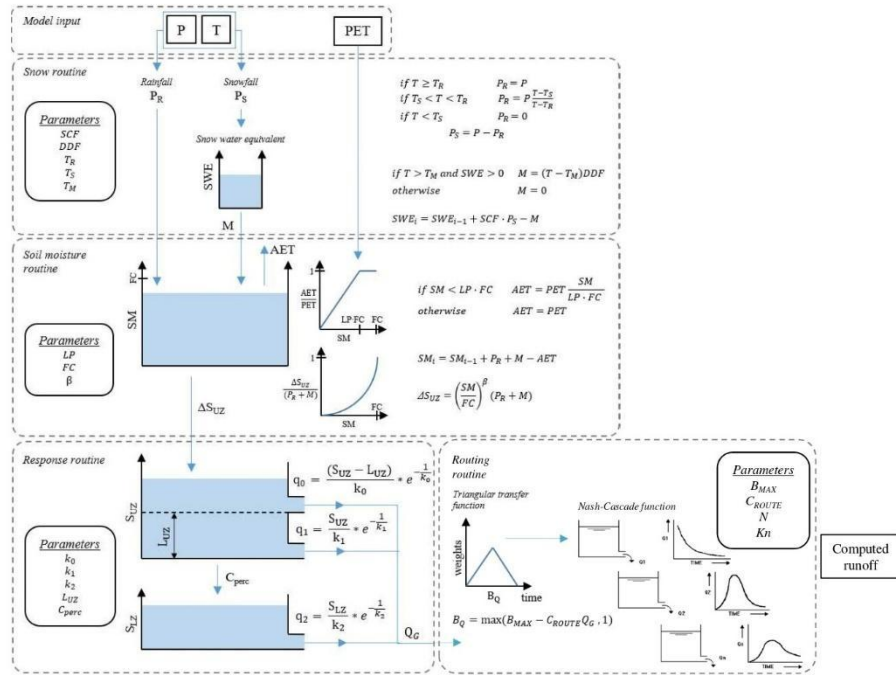


Figure 2. Modified structure of the TUWmodel, accounting for the addition of the Nash-Cascade module.

The model has 17 parameters: the original 15 from TUWmodel and two additional parameters for the routing component (parameter ranges shown in Table 1). To simplify calibration and regionalization, an additional assumption was introduced: given the relatively uniform size of HUs, the number of reservoirs in the Nash Cascade is fixed at 1 for all individual units. This means that the cases of catchments constituted by n HUs in series, have a routing behavior corresponding to a Nash-Cascade with n reservoirs, but with potentially varying k . This simplification reduces the total number of parameters to be calibrated to 16 and leads to a simplified IUH (i.e., linear reservoir):

$$IUH[t] = \frac{1}{k} * \exp^{-\left(\frac{t}{k}\right)} \quad (\text{Eq. 2.3})$$

While this assumption limits model flexibility by neglecting potential effects induced by sub-basin shape variability, it significantly enhances computational efficiency, as further discussed in Section 2.3.3.

Table 1. Modified TUWmodel parameters ranges and units.

Parameter	Lower Boundary	Upper Boundary	Description
SCF[-]	0.9	1.5	Snow Correction Factor
DDF [mm/°C/day]	0	5	Degree-day factor
Tr [°C]	1	3	Rain Temperature
Ts [°C]	-3	1	Snow Temperature
Tm [°C]	-2	2	Melting Temperature
LP [-]	0	1	Parameter related to the limit of evaporation
FC [mm]	0	600	Filed Capacity
β [-]	0	20	Non-linear parameter for runoff production
K_0 [days]	0	2	Storage coefficient for very fast response
K_1 [days]	2	30	Storage Coefficient for fast response
K_2 [days]	30	250	Storage Coefficient for slow response
lsuz [mm]	1	100	Threshold storage state
C_p [mm/day]	0	8	Constant percolation rate
Bmax [day]	0	30	Maximum base at low flows
Croute [day ² /mm]	0	50	Free scaling parameter
N [-]	1	1	Number of linear reservoir (Nash)
K_n [days]	0.1	2	Storage coefficient of the linear reservoir (Nash)

2.3.2 HUs attribution

In Section 2.2.4, the necessity of test and validation of two candidate HUs datasets has been briefly introduced. To do so, following the technical notes of Godet (Godet et al., 2024) we relied on the application of a contour-based method: such method is based on calculating a Critical Success Index (CSI) between two spatial polygons: polygon A, representing the boundary of a candidate catchment in the test dataset, and polygon B, representing the boundary from the tested catchment in the benchmark dataset instead. The CSI is consequently defined as:

$$CSI = \frac{a}{a + b + c} \quad (\text{Eq. 2.4})$$

where a is the intersection area between polygon A and polygon B, b is the area of polygon A not included in polygon B, and c is the area of polygon B not included in polygon A. When polygons A and B coincide perfectly, both b and c are zero, resulting in a CSI of 1. Conversely, if the two polygons do not overlap at all, a is zero and CSI drops to 0. Intermediate levels of overlap yield CSI values between 0 and 1. It is important to note that, except for these two extremes, CSI values should not be interpreted as percentages. While more data-intensive than simpler alternatives (e.g., area-based comparison), this method offers higher precision and therefore a more rigorous validation. If a dataset fails under this test, it is unlikely to perform better with a less robust approach. The results of this validation procedure are presented in Section 2.4.1.

2.3.3 Local calibration strategies

Considering the model structure presented in Section 2.3.1, two different local model calibration strategies were possible, from now on called respectively “independent” and “network”. In the first case, each catchment (as identified by the HU attributed to the discharge gauge station and all the HUs connected upstream of it) can be considered on its own, looking for a unique set of model parameters that assigned to all the sub-basins optimize a performance metric, similarly to a standard lumped calibration procedure. In the second case, instead, if multiple gauging stations are collocated along connected river branches, partial zones of the catchment can be calibrated in series, starting from upstream and moving downstream. Here the set of parameters optimized for the upstream areas already previously calibrated are kept fixed when calibrating downstream regions, and the simulated discharge from those

areas is used as fixed input for the now calibrating downstream zones. The regionalization procedure (Section 2.3.4) can be used starting from both local calibration approaches. The choice between the two approaches has to be based on several considerations. First of all, the computational time: in general, the “network” approach is substantially faster than the “independent” approach. The reason is simply due to the fact that in the first case, each sub-basin is calibrated only once, while in the second case, if any HU is part of multiple calibratable catchments, it must be re-calibrated the same amount of times. The larger the region, and the more numerous the number of stations, and their number on connected river branches, the larger the difference between the two approaches becomes. Secondly, the flexibility: the “independent” approach coincides with the “network” approach only for the headwater sub-basins (those that do not have any other gauge upstream of it connected to the same river branch); in any other case, the “network” approach has the advantage to calibrate the same area with more parameters. Intuitively, better calibrations, in an ideal scenario, have to be expected from the “network” approach as a direct consequence. At the same time, the very same potential strength of the “network” approach is also its greatest vulnerability: if the calibrated sub-areas depend on each other, in case of poorly calibrated upstream zones, they will inevitably condition the quality of all the calibrated areas connected downstream too. For example, if poor observations, or “area errors” (see Section 2.4.1) are introduced in the model to characterize upstream areas, they would now have repercussions to multiple catchments instead of just one. The “independent” approach, despite being slower and less flexible, is not affected by potential poor upstream calibrations: if an upstream catchment is, for whatever reason, badly calibrated, any downstream catchment also including the same area is now calibrated again from scratch (and so eventual inconsistencies, if limited to single basins, are not influential for all the others). For the current study, considering the size of the region, and the amount of stations allocated on connected river branches in the network, the “independent” approach has been preferred despite being in theory slower, to avoid the necessity of stopping the calibration, removing the “bad-behaving” stations after having checked that the reason for the poor calibration where external and not a inherent limit of the model, and re-starting it in case any is found (to assess if downstream gauges could be well-calibrated without them), because overall it would have been more time efficient. For applications on smaller areas with fewer stations, the “network” approach is instead strongly encouraged. Calibration code for the

“independent” approach used in this study, developed in R software environment, is available in the appendix.

Two model efficiency functions have been used throughout the current study. The first, the weighted Nash-Sutcliffe efficiency (wNSE - Hundecha & Bárdossy, 2004), is a variation of the commonly used Nash-Sutcliffe efficiency (NSE - Nash & Sutcliffe, 1970), that gives additional weight on high flows:

$$wNSE = 1 - \frac{\sum_{t=1}^T Q_o^t * (Q_o^t - Q_m^t)^2}{\sum_{t=1}^T Q_o^t * (Q_o^t - \bar{Q}_o)^2} \quad (\text{Eq. 2.5})$$

where t is the simulation timestep, T is the total number of timesteps, Q_o^t is the observed discharge at time t , Q_m^t is the simulated discharge at time t , and $\bar{Q}_o$ is the average discharge of the observed timeseries. Additionally the Kling-Gupta Efficiency (KGE - Gupta et al., 2009) has also been used:

$$KGE = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (\text{Eq. 2.6})$$

Where r is the Pearson correlation coefficient, α represents the variability of prediction errors, and β is the bias term.

Two preliminary local calibrations were conducted for all gauged catchments using the “independent” calibration approach, each separately optimizing one of the two performance metrics over the 10-year calibration period, then validated for the 10-year validation period. All optimizations were carried out in the R software environment using the “DEoptim” package (Mullen et al., 2011), which implements the Differential Evolution (DE) algorithm as introduced by Storn & Price, 1997.

Before proceeding to regionalization, a critical decision was required: whether to include all stations in the regional calibration, or to exclude those with poor local performance. Including all stations may appear comprehensive, but it poses significant risks. For instance, if 10% of the locally calibrated catchments are poorly represented, it is improbable that their performance will improve through regionalization, given that local calibration represents (in theory) the maximum possible performance for a given model structure. At the same time, the inclusion of such low-quality parameter sets may degrade the performance of otherwise well-behaving catchments, as the applied regionalization procedure, based on transferring information about parameter–descriptor relationships, does not incorporate any weighting based on local model skill. An alternative is to apply a performance threshold and retain only those catchments

that locally surpass it. While this mitigates the influence of poorly performing catchments, it introduces a degree of arbitrariness and the threshold could be applied to one metric alone or based on a composite criterion. In this study, the following approach was adopted, resulting in three different setups:

- Setup I: given the primary focus on accurately reproducing high-flow events, the wNSE was used to define a threshold that ensured both acceptable performance and spatial coverage across the GAR. The regionalization was then performed using the locally calibrated parameters obtained through wNSE optimization, with the regional model likewise maximizing wNSE. This resulted in a selection of 328 stations.
- Setup II: Using the same 328 stations as in Setup I, a second regionalization was performed, this time starting from the local calibrations optimized for KGE and maximizing KGE in the regional model. While this ensured direct comparability with Setup I, it introduced some locally suboptimal catchments (in terms of KGE) into the regionalization procedure.
- Setup III: To assess the effect of including poorly performing (KGE) stations in Setup II, a third regionalization was performed by selecting a new subset of 328 stations based on a KGE threshold, independently of Setups I and II. This allowed for a direct comparison between Setup II and Setup III on the shared stations between the two (87.5%), isolating the impact of suboptimal local calibration on regional performance.

2.3.4 Regionalization procedure

The primary objective of the proposed model is to enable discharge simulations at ungauged locations. To achieve this, a regionalization procedure is performed based on the information derived from the locally calibrated catchments. Most regionalization techniques in the literature rely on proximity and similarity criteria (e.g., He et al., 2011; Oudin et al., 2008), built on the assumption that hydrologically similar catchments will exhibit comparable responses and, by extension, similar model parameters. In this study, we adopt the PASS algorithm (PArAmeter Set Shuffling, Merz et al., 2020). The PASS approach is a data-driven method designed to derive regionally consistent parameter sets for distributed hydrological models without imposing a priori assumptions on parameter–descriptor relationships. PASS builds on

the idea that spatial patterns of model parameters can be inferred from multiple locally calibrated parameter sets. First, several well-performing parameter sets are obtained through independent local calibrations for each catchment. At each iteration, one parameter set per catchment is randomly selected from this pool and used to infer regional functional relationships between model parameters and catchment descriptors using a simple machine-learning technique (random forests). These relationships are then applied to predict spatially distributed parameters, which are evaluated by running the distributed model and assessing regional model performance. The shuffling and regionalization steps are repeated, progressively refining the selection of parameter sets, until regional performance no longer improves. In this way, PASS explicitly accounts for parameter equifinality and identifies parameter combinations that are both locally acceptable and regionally consistent. It must be also noted how at each iteration, new local parameters are also inferred through the defined RFR: these parameters sets are used to run local simulations, evaluating also in this case model performance. If one of this new parameter sets returns ME greater than a threshold ($0.9 * \max(\text{ME}_{\text{local}})$) it is added to the pool of possible sets for that catchments for the random selection in the first step of the regionalization procedure (this way incorporating a learning component in the regionalization methodology). For each setups, 10 parallel “groups” are defined: each group in an independent regional model, and multiple of them are used to account for variability in the random selection of parameter sets, that will, in general, lead to different final models. In the following sections, for all setups, results will be reported only for the group with the highest median performance across the entire GAR. The required starting multiple parameter sets (20) needed for PASS, have been obtained for each catchment following the procedure detailed in Section 2.3.3. Also note that the inclusion or not of a catchment in the regionalization routine for the different setups based on thresholds, was based on the best performing parameters set obtained for a catchment. While the machine learning model operates as a black box and does not provide explicit analytical equations, interpretability is still partially possible as the relative importance of individual descriptors can be extracted from the regression tree structure (see Section 2.4.5).

The catchment descriptors used in this study can be classified into five main categories. (I) Climatic descriptors: include climatologic characteristics, like mean annual precipitation, mean temperature aridity index, and extreme climate indices; (II) Topographic descriptors: encompassing mean elevation, slope, and distance from the

sea, among other attributes defining the hydrological unit's (HU) position and topography; (III) Land-use descriptors: representing the percentage of the HU covered by different land cover types, including forests, small vegetation, urban areas, croplands, and wetlands; (IV) Soil properties: including soil texture, granulometry, and drainage density; (V) others: accounting for physiographic factors not explicitly modeled, such as the density of karst areas and the presence of permanent glaciers, which strongly influence hydrological behavior in mountainous or geologically complex terrain. All descriptors were derived from open-access European datasets, including HWSO (HWSO; FAO et al., 2012) for soil data, Corine Land Cover (CLC; EEA, 2018) for land use, the Mediterranean Karst Aquifer Map (MEDKAM – Stevanović et al., 2015) for karstic characteristics, and the Randolph Glacier Inventory (RGI v6.0; RGI Consortium, 2017, Pfeffer et al., 2014) for the presence of glaciers. Climatic variables were computed from ERA5-Land reanalysis data, and topographic attributes were extracted from a Digital Elevation Model (DEM – GLO-90; ESA, 2021). A complete list of the used CDs is available in the appendix.

2.3.5 Validation metrics

Both the local and regional models were subjected to validation procedures. For local models, performance was assessed based on model efficiency during the validation period. In contrast, regional models underwent both temporal and spatial validation. To evaluate regional performance, the gauged stations used for constructing the regional models were randomly split into two equally sized sub-groups. For each of the three setups described in Section 2.3.3, the regionalization procedure was applied twice, that is, in each setup, one sub-group served as the training dataset while the other acted as the testing dataset. This procedure allowed model performance to be assessed across four key scenarios: (A) calibration period on the training catchments; (B) validation period on the training catchments; (C) calibration period on the test catchments, and; (D) validation period on the test catchments, the most stringent test of generalization capability. All results from this cross-validation process are presented in Section 2.4.3. It is important to note that the same sub-group divisions were used for Setups I and II, as they relied on the same initial set of catchments. However, the sub-groups for Setup III differ, as the set of included catchments is not identical to those used in the first two setups.

Synthetic model efficiency metrics, such the ones introduced in Section 2.3.3, are valuable for model calibration. However, they are not sufficient on their own to ensure

that the model reproduces the targeted hydrologic signatures and it is fit to purpose, and often display considerable limitations (M. P. Clark et al., 2021; Wouter J. M. Knoben et al., 2019). To further assess model performance, particularly in capturing flow peaks, two additional metrics will be employed in Section 2.4. The first of these is the percentage bias of event peaks' error, defined as:

$$\%BIAS_C^P = \frac{\sum_{i=1}^n (P_{i,C}^S - P_{i,C}^O)}{\sum_{i=1}^n P_{i,C}^O} \quad (\text{Eq. 2.7})$$

Here, C represents a specific catchment, $\%BIAS_C^P$ denotes the percentage bias on the event peaks for catchment C, $P_{i,C}^S$ is the i-th event peak discharge of the simulated discharge for catchment C, and $P_{i,C}^O$ is the i-th event peak discharge in the observed time series. Additionally, a second metric focusing on discharge volumes is also applied:

$$\%BIAS_C^V = \frac{\sum_{i=1}^n (V_{i,C}^S - V_{i,C}^O)}{\sum_{i=1}^n V_{i,C}^O} \quad (\text{Eq. 2.8})$$

Where now V denotes the volume corresponding to a specific event, and superscripts and subscripts have the same meaning as in Eq. 2.7.

Both evaluation metrics are event-based and rely on the identification of the N most significant high-flow events within each catchment and simulation period (i.e., calibration and validation). Specifically, events are first selected from the observed time series, and corresponding events are then identified within the simulated time series. In this study, an average of four events per year was targeted, resulting in N = 40 events over each ten-year periods considered. It is important to note that the selected events do not necessarily correspond to the N highest daily discharge values in the observed records. This is because individual high-flow events often span multiple days, and thus may include several high daily discharges. Consequently, care must be taken to distinguish discrete events from consecutive daily peaks that are part of the same hydrologic event. To systematically extract hydrologically relevant events, the following multi-step procedure was applied:

1. Flow peaks were identified from the observed discharge time series as local maxima, where the discharge value exceeds that of both the preceding and following days. Only peaks above a threshold value, such as the Q90 of the full

hydrograph, were retained to ensure relevance while avoiding the exclusion of moderate but potentially significant events;

2. A smoothed baseline discharge curve (T_0) was computed for the observed time series, derived by applying a 91-day moving average to the observed record;

3. For each day of the year, median (50th percentile) and Q75 (75th percentile) discharges were calculated across all years, producing two seasonal reference curves for the observed discharge, T_{050} and T_{075} ;

4. A scaling coefficient was defined as $KK_0 = \text{mean}(T_{075}/T_{050})$. It was used to scale the baseline curve, yielding an adjusted reference series $T_0' = T_0 * KK_0$;

5. For each peak identified in Step 1, an event window was defined as the period between the first day before the peak where discharge falls below T_0' , and the first following day where it again falls below T_0' . If multiple peaks fell within the same window, only the highest was retained. This process yielded a ranked list of 40 observed events, each characterized by its peak value, start date, and end date.

Once the events were identified in the observed time series, corresponding events were then systematically extracted from the simulated time series:

6. The simulated event peak was defined as the day with maximum discharge within a ± 1 -day window centered on the observed peak date, to allow for minor timing discrepancies in precipitation input.

7. To ensure comparability, an event time window equal to the observed event (same number of days before and after the peak) was considered also for the simulated event, allowing for consistent volume integration.

For both observed and simulated events, the event volume was computed as the sum of daily discharge values within the defined event window. A schematic overview of this entire procedure is presented in Figure 3, illustrating each step for greater clarity.

This event-based analysis was applied to both locally calibrated models and the final regional models (for both training and test catchments) across both calibration and validation periods, enabling robust comparative evaluation of events peak reproduction and volumetric accuracy.

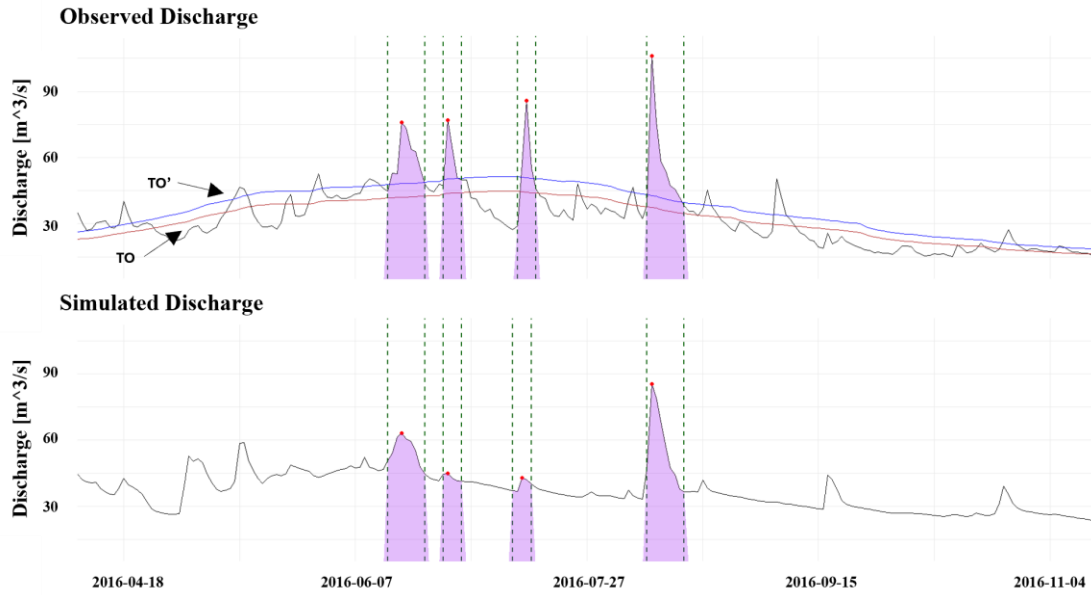


Figure 3. Schematic representation of event identification. Four events identified in the observed time series (peaks in red, event windows delimited by the green dashed lines), and the corresponding event windows in the simulated time series. In purple shades, the total events volumes.

2.4 Results

2.4.1 Does a tailored dataset improve HUs attribution?

To evaluate the most suitable sub-basin delineation for the GAR, two hydrological units datasets were compared: the existing ECRIN dataset and a newly constructed HUs dataset. The latter was generated using the same DEM already used to derive the topographic CDs, targeting an average HU area of 100 km² and explicitly aligning HUs' outlets with known discharge gauge locations. Both datasets were tested using a benchmark set of 197 catchments in Piedmont, Italy, for which high-quality catchment

outlines were available. For each benchmark catchment, the Critical Success Index (CSI), introduced in Section 2.3.2, was computed across all possible HUs combinations in the test datasets. The HU with the highest CSI value was retained as the best match. Results are reported in Figure 4.

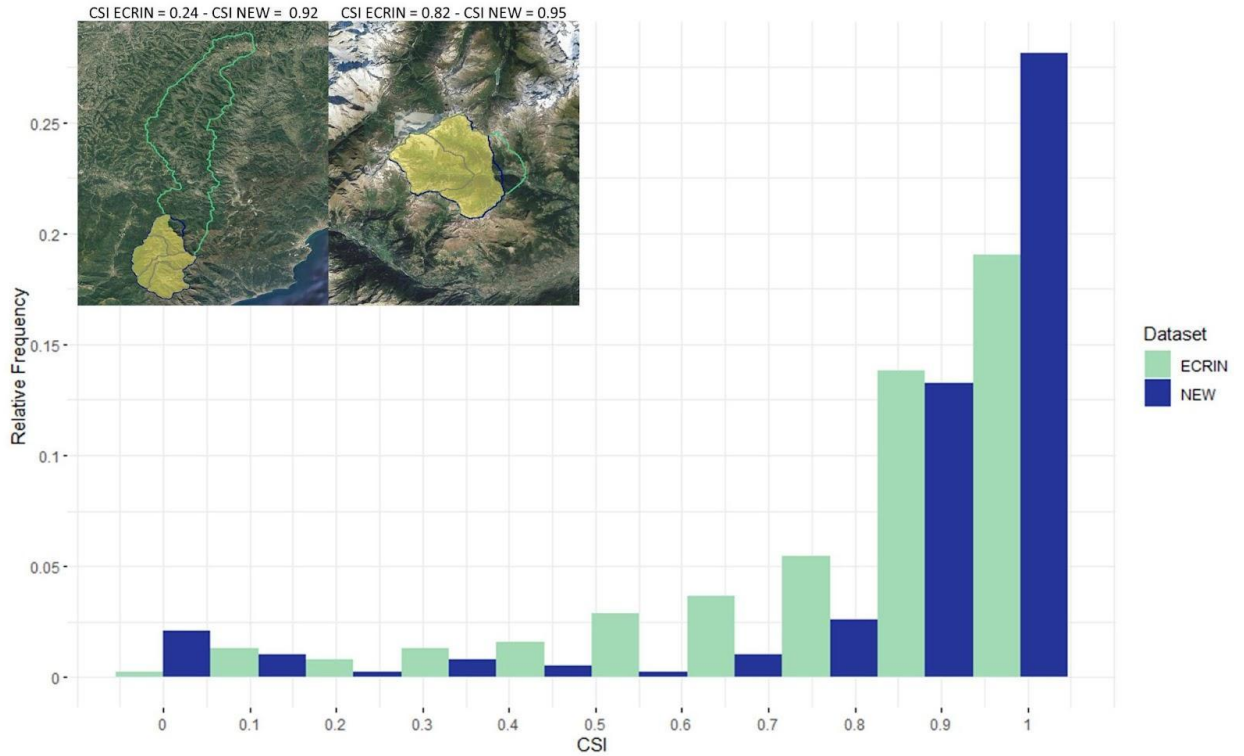


Figure 4. CSI distribution comparison between the ECRIN dataset (light-blue) and the newly created dataset (blue) for the 197 catchments in the Piedmont region. In the sub-panel, two examples of catchment delimitation improvements.

The ECRIN dataset yielded generally high CSI values, often close to 1, but a substantial number of catchments showed lower agreement (CSI between 0.7 and 0.9). These discrepancies reflect notable mismatches in catchment shape and area, often exceeding 10%, and were primarily due to misalignments between the gauge locations and HUs' outlets. In contrast, the newly constructed HUs dataset showed clear improvements: a greater number of catchments achieved near-perfect CSI scores. However, some catchments also exhibited very low CSI values (~ 0). These cases were traced to misplaced gauge stations during dataset construction, specifically in areas near river

confluences, where station locations were assigned with the help of basemaps. Erroneous placement upstream or downstream of a confluence could drastically alter the delineated catchment area, resulting in near-zero CSI. Crucially, these outliers reflect gauge positioning errors, rather than inherent flaws in the delineation method. Given its overall superior performance and better alignment with gauge locations, the new HU dataset was selected for further use. In regions of the GAR where no benchmark catchment outlines were readily available, dataset quality was validated by comparing the nominal contributing area of each station with its modeled catchment area. Stations with substantial discrepancies were excluded. Notably, all gauge stations shown in Figure 1 have verified contributing area errors lower than 5%.

2.4.2 A necessary premise about input uncertainty

As discussed in Section 2.2.2, selecting a climate forcing for the Greater Alpine Region (GAR) requires balancing data quality against spatial consistency. To quantify how this choice affects model skill, we first carried out a targeted experiment on 37 catchments in Piedmont, within the GAR, delineated with the new HU dataset (Section 2.4.1). Each catchment was calibrated twice, once to maximize wNSE and once to maximize KGE, using three alternative climate inputs: (I) a high-quality dataset covering the test area (Optimal Interpolation), (II) ERA5-land, (III) UERRA. Summary outcomes are shown in Figure 5.

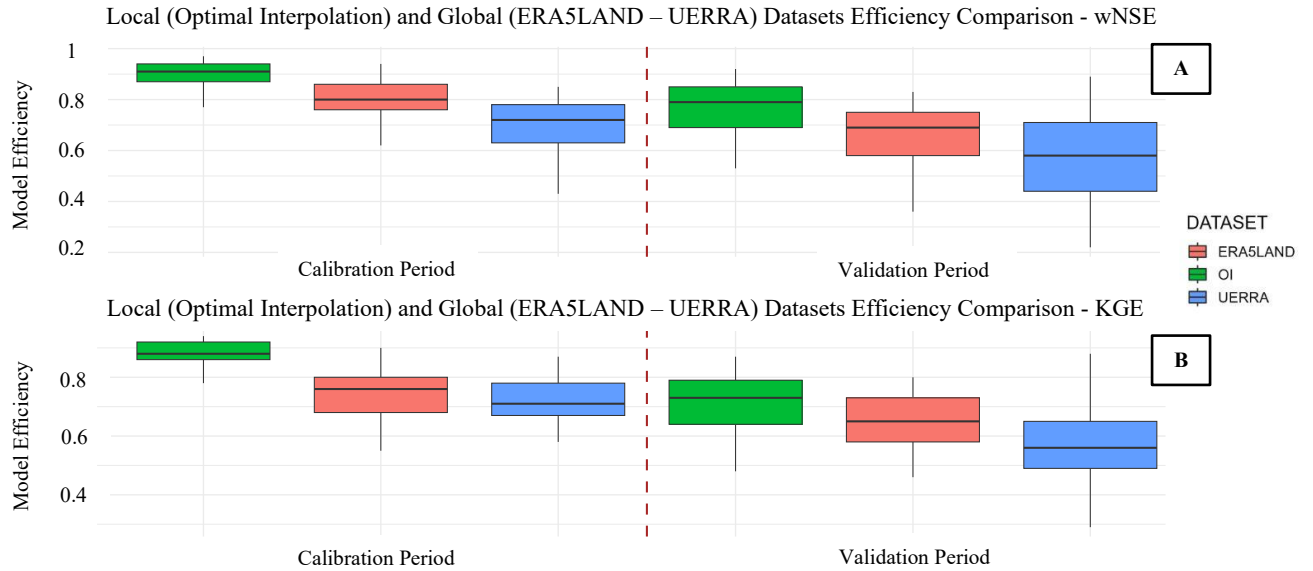


Figure 5. Comparison in model efficiency between a local (Optimum Interpolation – OI – dataset, in green) climate input for the Piedmont region, ERA5-Land climate inputs (in red), and UERRA climate inputs (in blue). (A) Efficiencies optimizing the weighted Nash Sutcliffe (wNSE); (B) Efficiencies optimizing the Kling-Gupta Efficiency (KGE).

In Figure 5A it can be seen how both ERA5-Land and UERRA forcing lead to a systematic decline in wNSE relative to OI forcing across all catchments. In contrast, the OI dataset yields higher wNSE. In Figure 5B a similar, though slightly less pronounced, pattern emerges for the KGE. Because all other model components were identical, performance differences stem solely from the climate input. This experiment highlights the added uncertainty inherent in globally and regionally consistent products and should inform the interpretation of GAR-wide results in Section 2.4.3 and Section 2.4.4, which represent the best performance achievable with a global dataset, yet still carry greater uncertainty than would be expected with higher-quality local inputs. The results here also justify, among the global climate forcings, the use of ERA5-land over UERRA.

2.4.3 Local (gauged) calibrations efficiency

As described in Section 2.3.3, three different setups were considered for the regionalization procedure, all based on local calibration results. This section focuses on setups I and II (setup III is addressed later in Section 2.4.4). The model efficiency

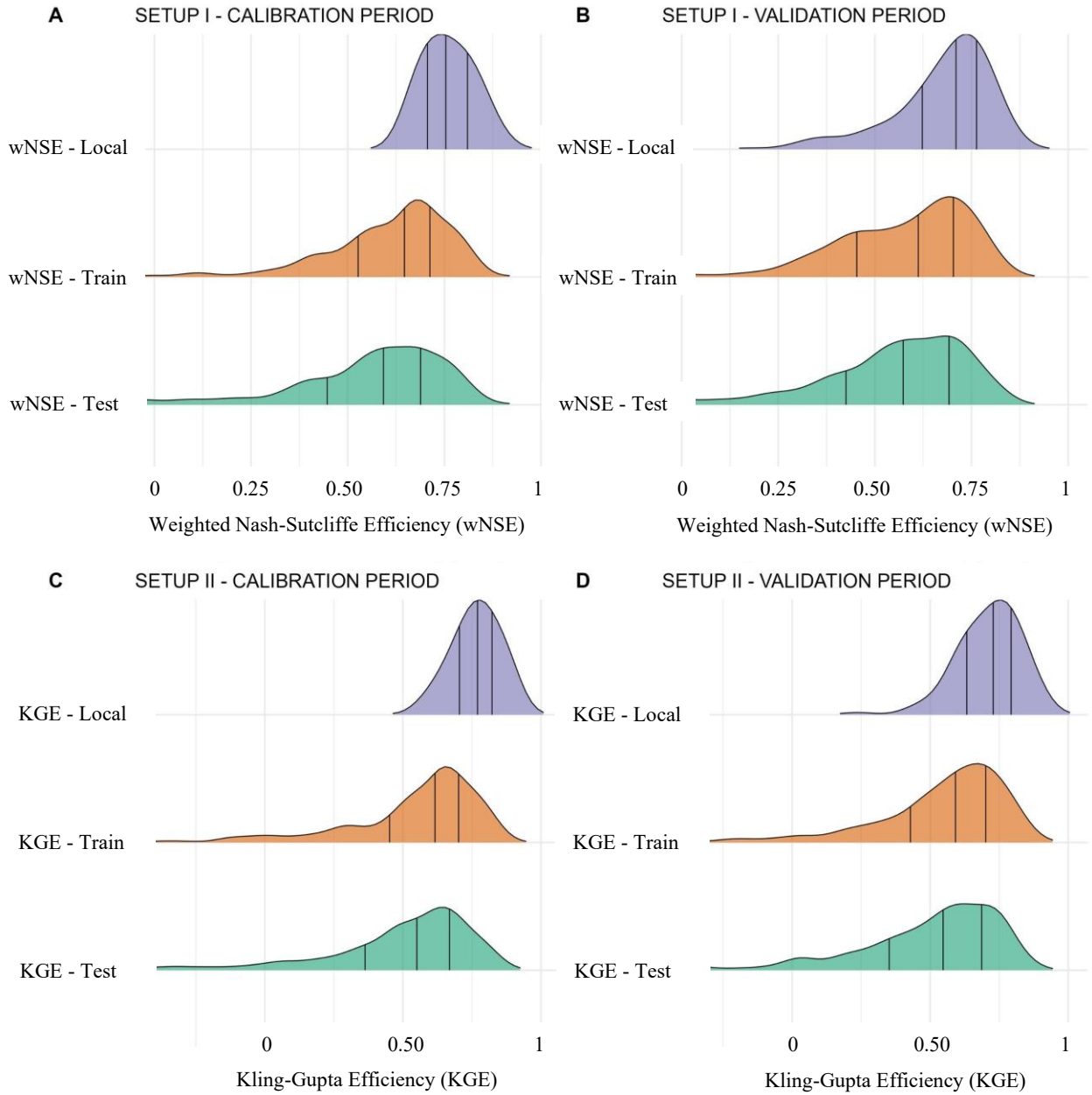
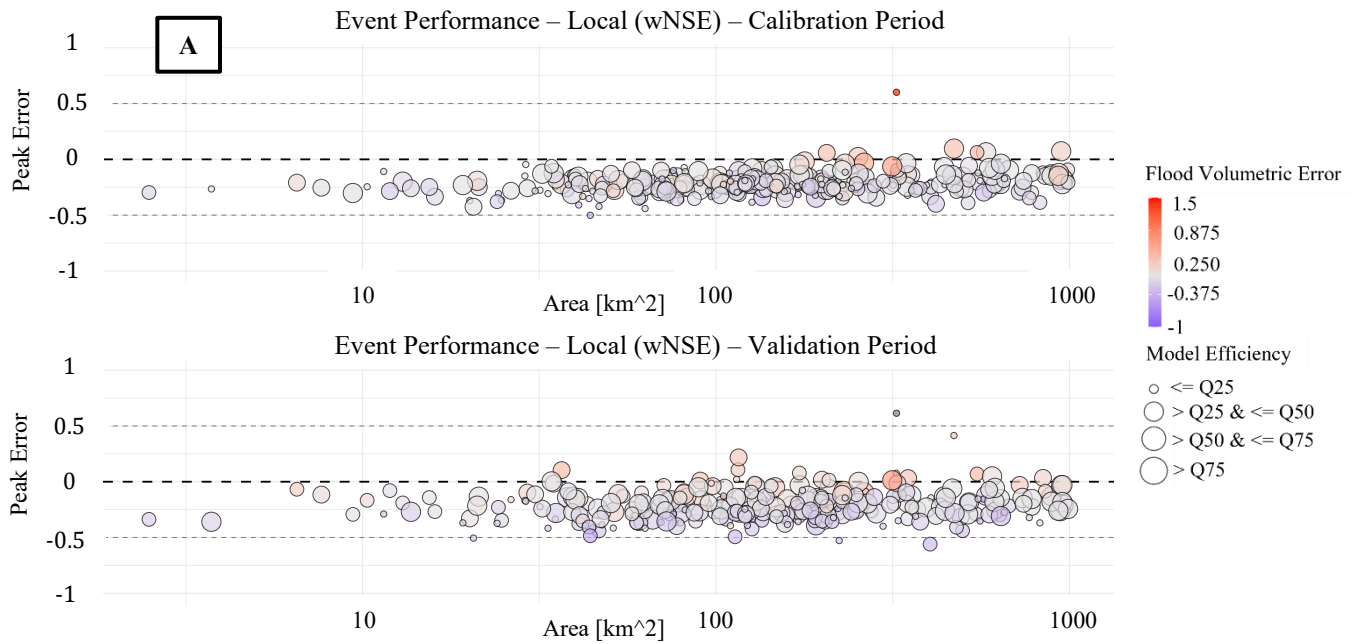


Figure 6. Empirical distribution function of model efficiencies for the calibration period (A and C) and validation period (B and D), for setups I (wNSE) and II (KGE). In the plots are reported local (blue), regional in train (orange) and regional in test (green) efficiencies.

empirical probability distributions are shown in Figure 6, with panels A and C corresponding to the calibration period, and panels B and D to the validation period. The x-axis in the figures is lower bounded to 0 for wNSE and to -0.41 for KGE following their respective benchmark values (Wouter J. M. Knoben et al., 2019). In all cases, less than 2.5% of the catchments were lower than those benchmarks. The information about the local calibration are already displaying only the catchments after the application of the filtering threshold (wNSE = 0.60) introduced in Section 2.3.3.

The local calibrations yielded median efficiencies of 0.754 (wNSE) for setup I and 0.770 (KGE) for setup II during calibration, and 0.709 (wNSE) and 0.729 (KGE), respectively, during validation. These results indicate that model performance remains generally robust also in temporal validation. It should be noted that wNSE and KGE are not directly comparable, and a lower numerical value of one metric does not necessarily imply poorer performance. To complement this analysis, the second validation component introduced in Section 2.3.5, which focuses on the reconstruction of high-flow events, was used to assess model behavior in more detail.



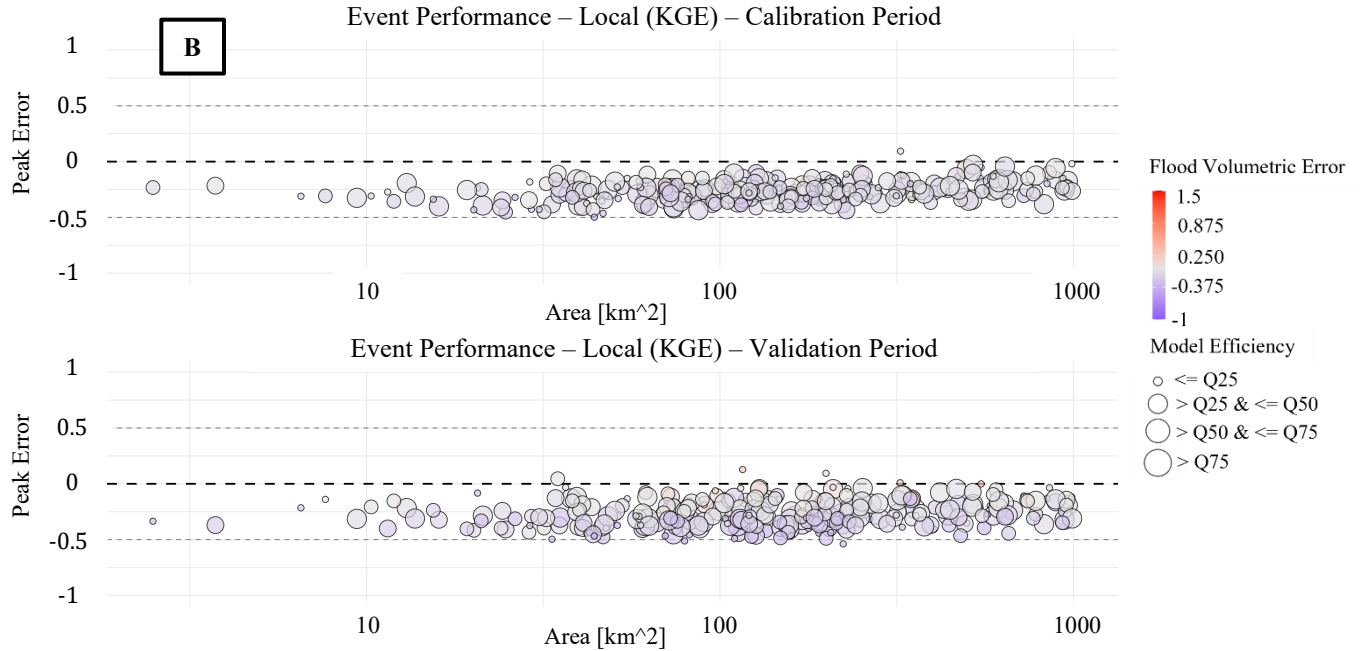


Figure 7. Local top-40 peak events validation. Y-axis %BIAS of the peak; color %BIAS of the volumes; bubble size proportional to the model efficiency quantiles (very small = worst 25%, small = between 25% and 50%, medium = between 50% and 75%, large = best 75%). Panel A for setup (I), panel B for setup (II).

As already detailed in Section 2.3.4, multiple parameter sets were generated for each catchment regardless of the setup. The results presented in Figure 7 are based on the simulated discharges obtained using the parameter set that yielded the highest model efficiency.

2.4.4 Regional (ungauged) efficiencies

The results of the regionalization procedure using the PASS algorithm are presented hereafter. For clarity, the results are discussed separately for each of the three calibration setups introduced in Section 2.3.3. For setup I, the complete set of 328 catchments was randomly divided into two mutually exclusive clusters, referred to as C1 and C2. To evaluate the transferability of regional relationships, the entire regionalization procedure was carried out twice: once using C1 for model training and C2 for model testing, and once with the roles reversed. The results are shown in Figure 6A (calibration period) and Figure 6B (validation period), where cumulative

distribution functions (CDFs) of wNSE efficiency scores are reported for both training and test catchments and compared to local efficiencies. As anticipated, regional model performance is consistently lower than that of the local calibrations, which are optimized for individual catchments. The regional approach, in contrast, attempts to generalize a single predictive relationship across diverse hydrological and physiographic conditions, inevitably leading to performance tradeoffs. Nonetheless, the regional model demonstrates acceptable levels of skill: for the training catchments, the median wNSE values were 0.647 for the calibration period and 0.612 for the validation period. For the test catchments, these values declined further to 0.592 and 0.581, respectively. Interestingly, the reduction in performance between training and test catchments was more pronounced in the calibration period than in the validation period, suggesting that the regional model retains some robustness under the more challenging condition of simultaneous spatial and temporal extrapolation.

To better understand spatial variations in model skill, efficiency maps were generated and are shown in Figure 8. These spatial plots indicate that regional performance is not confined to any particular sub-area within the Greater Alpine Region (GAR), and that there are no clear spatial clusters of consistent underperformance. This is a desirable feature for any regional model, as it increases confidence in its applicability across the entire domain. In addition to standard efficiency metrics, the evaluation also included the validation of flood-event reconstruction, following the methodology introduced in Section 2.3.5. The results for setup I for the train catchments in the calibration period, and for the test catchments in the validation period are presented in Figure 9. As expected, the regional model is generally less accurate in reproducing high-flow events, with a more substantial decrease in accuracy for test catchments.

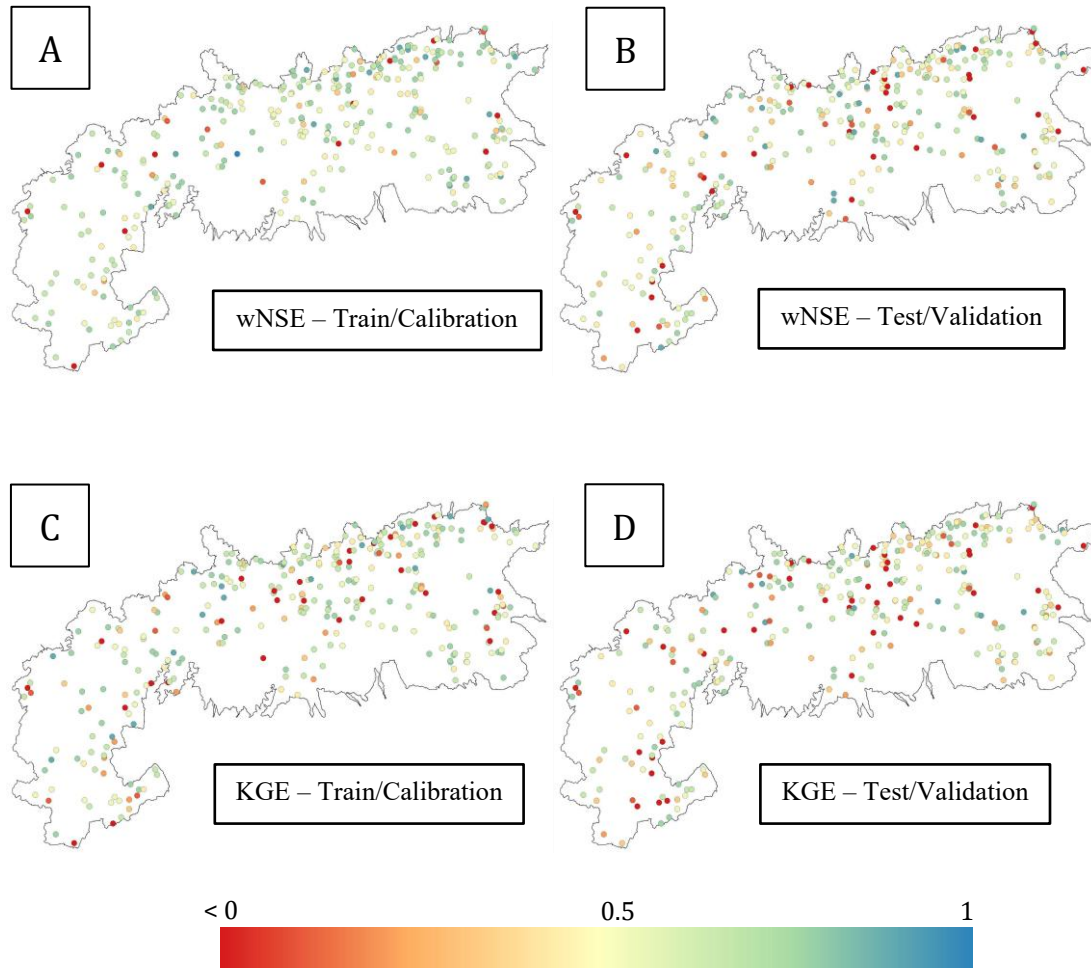


Figure 8. Spatial distribution of model efficiency for setup I. (A) wNSE train catchments for the calibration period and (B) test catchments for the validation period. (C) KGE train catchments for the calibration period and (D) test catchments for the validation period.

The decline in performance is mostly concentrated among basins with already low wNSE scores in the regional model (represented by smaller bubbles). In contrast, catchments with relatively high regional wNSE still show underestimation patterns for flood peaks and volumes similar to those observed in the locally calibrated simulations. For the group of catchments with lower regional efficiency, errors in peak flow reproduction were fairly balanced between underestimation and overestimation. Importantly, although nearly all catchments with poor flood-event performance also had low regional wNSE, the reverse is not true: not all low-efficiency catchments

performed poorly in reconstructing flood peaks, indicating that event-based errors can be partially decoupled from overall model fit.

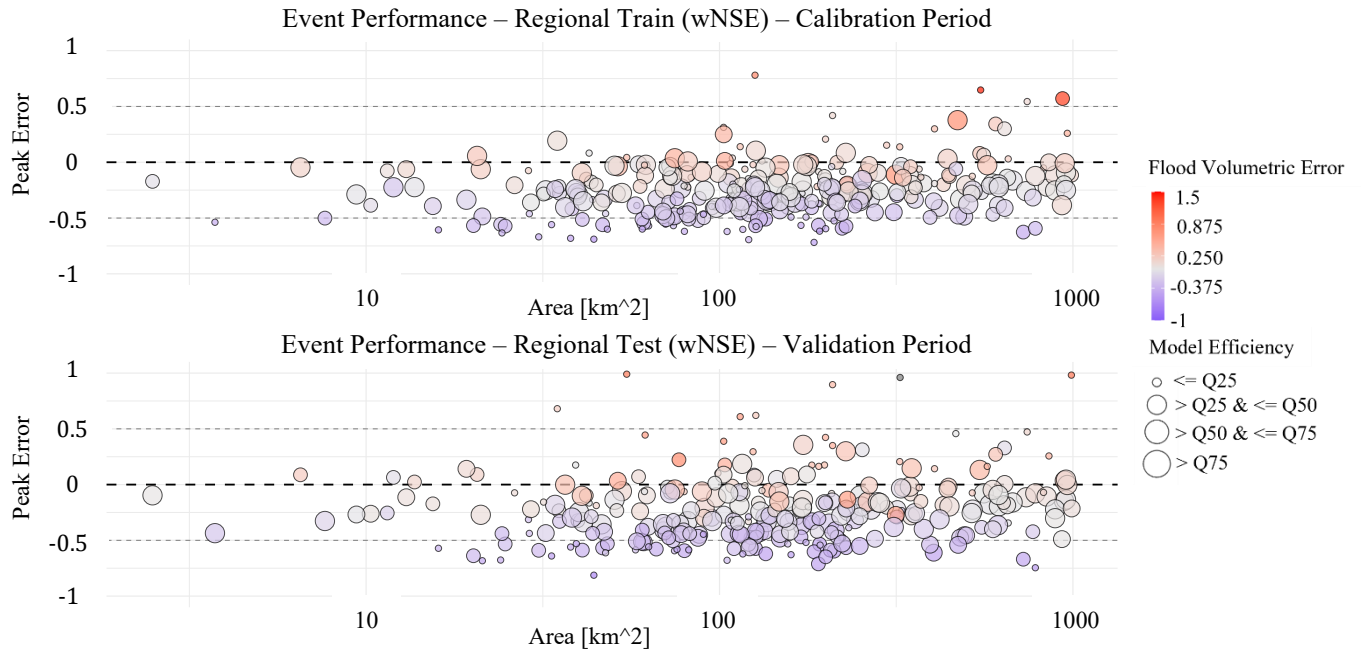


Figure 9. Regional top-40 peak events for setup I (wNSE). Y-axis %BIAS of the peaks; color %BIAS of the volumes; bubble size proportional to the model efficiency quantiles (very small = worst 25%, small = between 25% and 50%, medium = between 50% and 75%, large = best 75%). Top for train catchments during the calibration period, bottom for test catchments during the validation period.

For setup II, the same 328 catchments were used, but this time the regional model was optimized using the KGE. The results are presented in Figure 6C and Figure 6D. As with setup I, a reduction in model skill is observed when moving from local to regional performance. The median KGE for training catchments was 0.616 during the calibration period and 0.592 during the validation period. For the test catchments, the median efficiency values were lower, at 0.551 and 0.547, respectively. Although wNSE and KGE are not directly comparable due to their differing formulations and sensitivity profiles, a stronger degradation in performance is evident for the KGE-based setup when transitioning from local calibration to the regional model.

Despite this, the spatial distribution of model skill remains relatively uniform across the GAR, and no evident geographical bias can be identified from the efficiency maps. This again confirms the spatial robustness of the regional relationships produced by the PASS algorithm. Flood-event validation for setup II is summarized in Figure 10. A general tendency toward stronger underestimation of peak flows is again observed. Interestingly, in contrast to setup I, the basins with low regional efficiency in setup II exhibit a more consistent pattern of overestimation in peak flows, rather than the more evenly balanced distribution between under and overestimation observed previously.

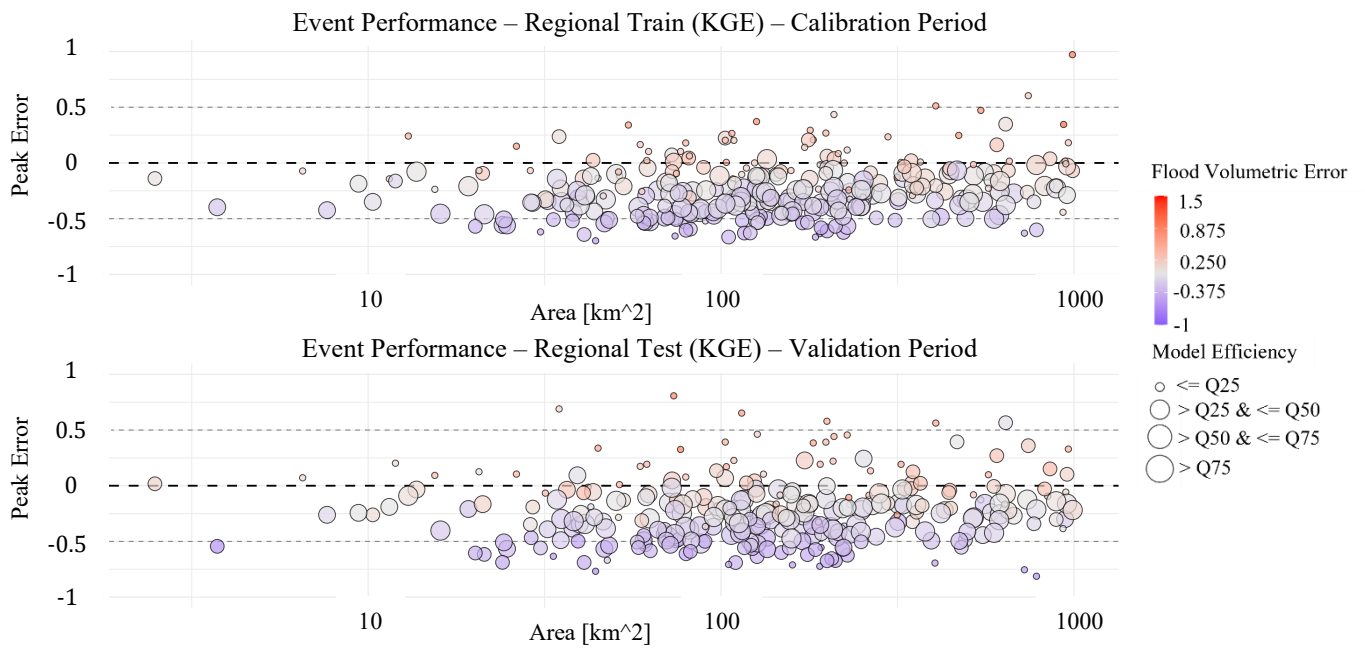


Figure 10. Regional top-40 peak events for setup II (KGE). Y-axis %BIAS of the peaks; color %BIAS of the volumes; bubble size proportional to the model efficiency quantiles (very small = worst 25%, small = between 25% and 50%, medium = between 50% and 75%, large = best 75%). Top for train catchments during the calibration period, bottom for test catchments during the validation period.

Finally, setup III was designed to test the impact of including sub-optimally performing catchments in the regional calibration process. Here, the relevant comparison is the one considering only the 287 catchments shared between the two setups.

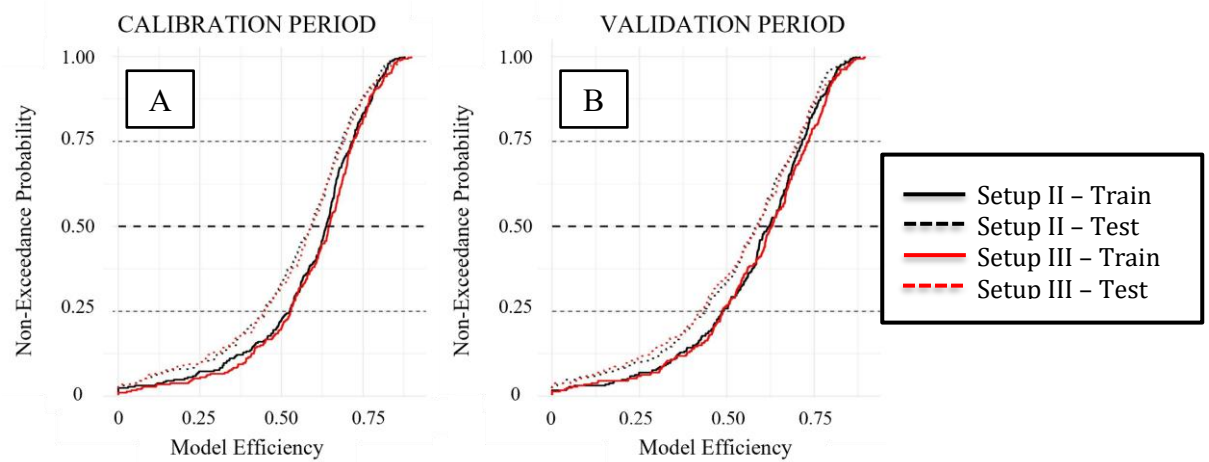


Figure 11. Regional model efficiency (KGE) comparison between setup II (in black) and setup III (in red) only considering the catchments shared by both setups. Train catchments continuous line, test catchments dashed line. (A) Calibration period; (B) validation period.

The results, reported in Figure 11A (calibration period) and Figure 11B (validation period), show that performance differences between setup II and setup III are negligible. This holds for both the calibration and validation periods, and for both training and test catchments. The similarity in performance indicates that the inclusion of the sub-optimal 12.5% of catchments did not result in any meaningful degradation (or improvement) of regional model performance within the shared subset. This suggests that the PASS regionalization is relatively robust to the presence of moderately underperforming locally calibrated basins, at least within the conditions here tested.

Flood-event validation plots for setup III are not shown; nonetheless, results from that analysis are consistent with the findings from setups I and II.

2.4.5 Regression trees: what drives the regional model?

Although the random forest algorithm functions as a black box and does not provide an explicit regional functional relationship linking CDs to model parameters, a partial interpretation of its behavior is still possible. Specifically, the relative importance of each descriptor in determining individual model parameters can be assessed. This is illustrated in Figure 12 and Figure 13, which report results for the regionalization based on wNSE. Here are reported statistics using all 10 regionalization

groups (see section 2.3.4), and not exclusively the best performing one. Remember how each group is an independent regional model, with his unique RFR: looking at CDs relevance in each of them, allow for consideration about the generality of RFR between model parameters and CDs.

For each parameter, the figures show the mean relevance (bar height), as well as the range (minimum to maximum) across the ten independently generated PASS regional groups. Consistent with the findings of R. Merz et al., 2020, no single set of descriptors consistently explains the regional variability of model parameters. For example, in the case of parameter T_m (Threshold temperature above which melt starts – Figure 12), certain descriptors are highly influential in some regional models but show negligible importance in others, as indicated by the wide variability in importance values (black vertical lines extending from 0 to a positive maximum). This suggests that comparable regional performances can be achieved using different combinations of descriptors, likely due to correlations among them. In practice, this flexibility is advantageous, as it implies that reliable regional models can still be developed even in regions where only a limited set of catchment descriptors is available. However, from a process understanding perspective, this variability also highlights a limitation in our ability to clearly identify the dominant physical controls on hydrological behavior. There are, however, a few notable exceptions to this general pattern. Field capacity (FC, Figure 12) consistently shows strong associations with the area covered by small vegetation, the extent of glacier coverage, and long-term climate variables such as mean annual temperature and potential evapotranspiration. Similarly, the very fast and fast storage coefficients (k_0 and k_1 , Figure 13) exhibit nearly identical patterns, with consistently high relevance attributed to the same climatic and physiographic variables: glacier coverage, mean annual temperature, potential evapotranspiration, and mean elevation. For k_1 , an additional influence from small vegetation cover and the aridity index is also evident. Lastly, the constant percolation rate (c_{perc} , Figure 13) frequently correlates with mean elevation and climate-related descriptors including temperature, potential evapotranspiration, and the aridity index.

Comparable patterns were observed for the alternative regionalization setups based on the KGE objective function, with the most robust relationships for these four parameters holding consistently.

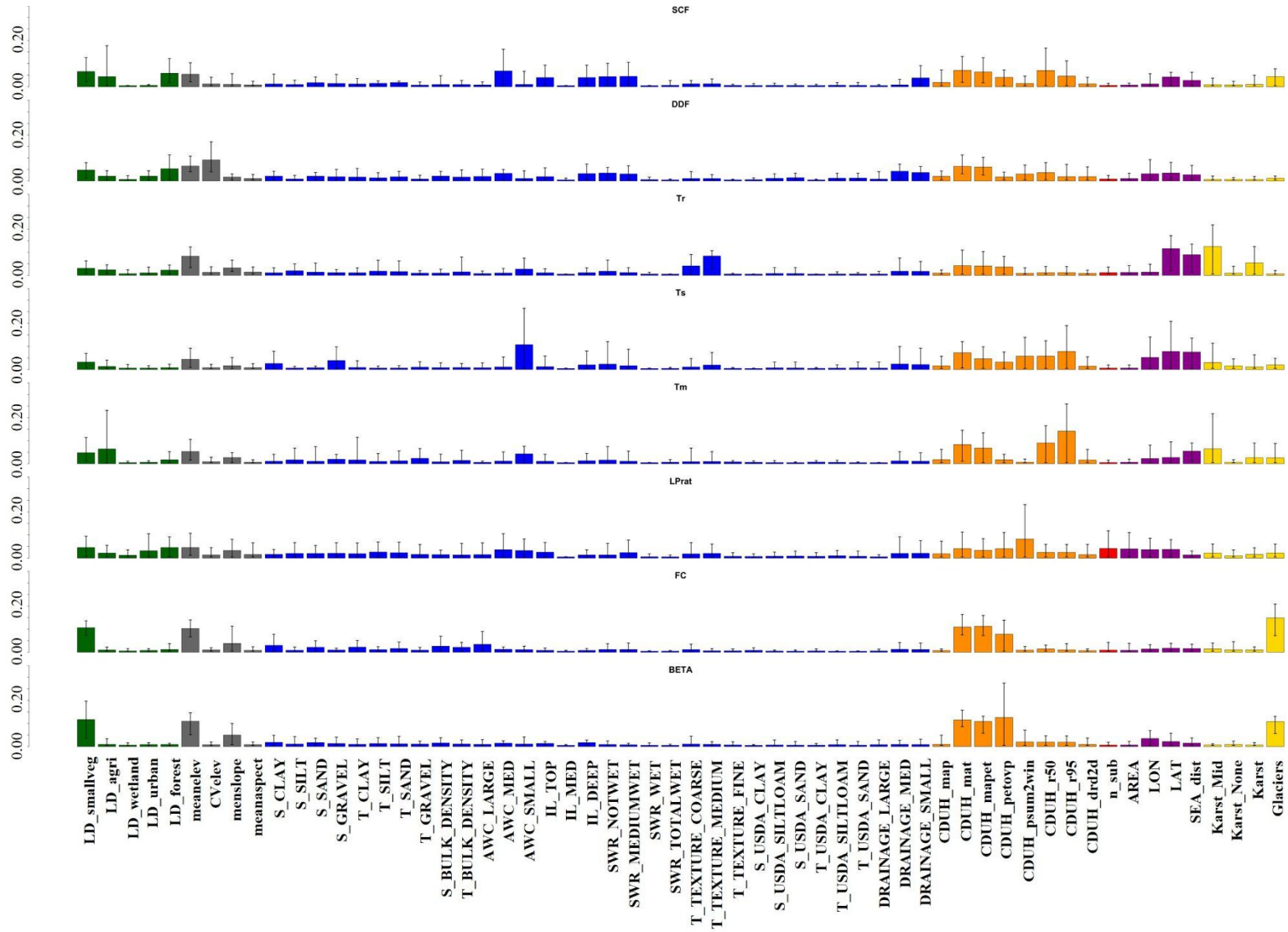


Figure 12. Percentual catchment descriptors relevance in random tree regression (SCF, DDF, Tr, Ts, Tm, LPrat, FC, BETA) for setup I. Descriptors on the x-axis; colors indicates macro descriptor classes (in green land-use, in gray topography, in blue soil, in orange climate, in red connectivity, in purple position and in yellow karst and presence of glaciers).

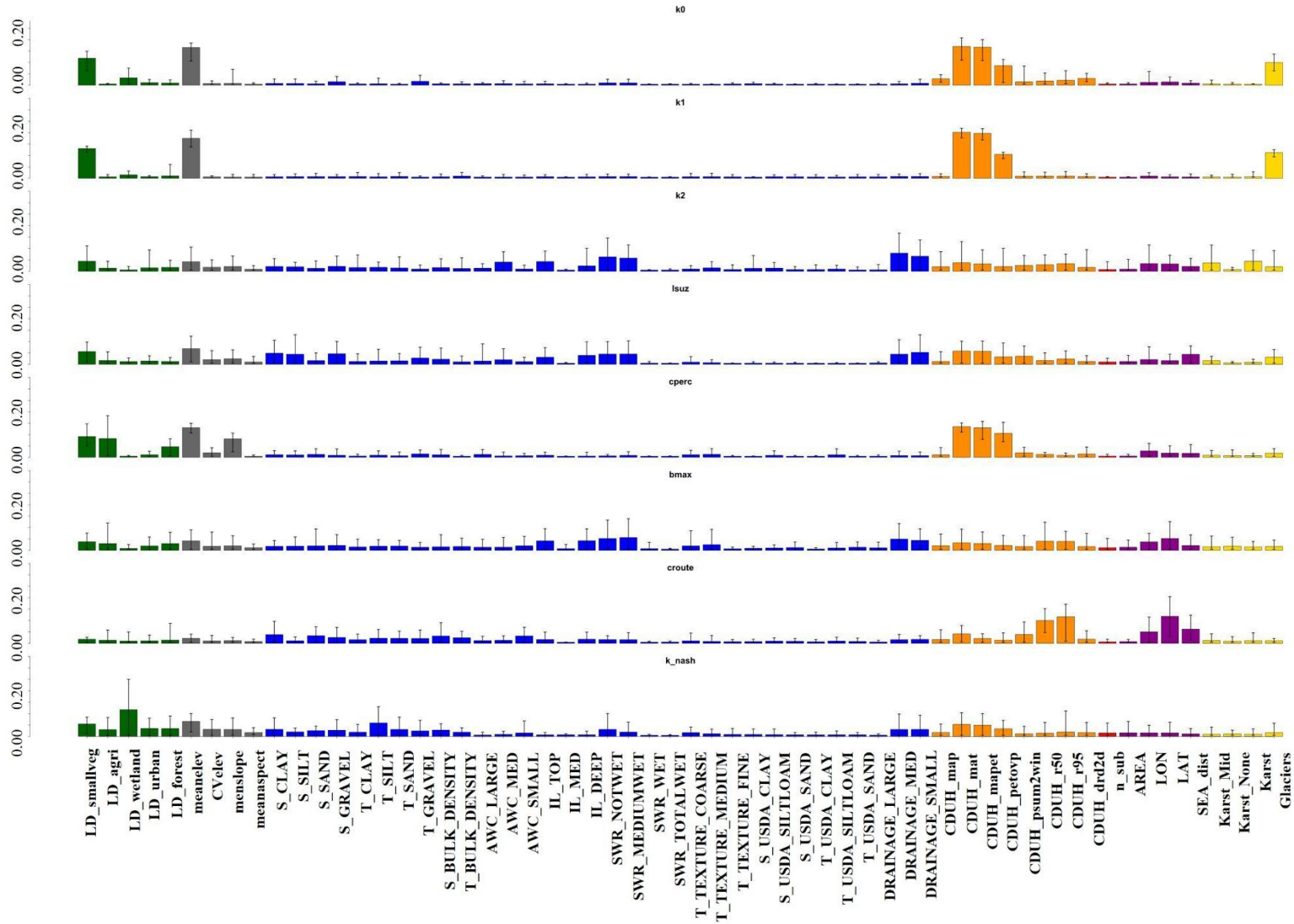


Figure 13. Percentual catchment descriptors relevance in random tree regression (k0, k1, k2, lsuz, cperc, bmax, croute, k_nash) for setup I. Descriptors on the x-axis; colors indicates macro descriptor classes (in green land-use, in gray topography, in blue soil, in orange climate, in red connectivity, in purple position and in yellow karst and presence of glaciers).

2.5 Discussion

2.5.1 Usefulness of the results: are they fit-to-purpose?

In Section 2.4, the results and predictive performance of both the local and regional models have been discussed in detail. Overall, the local simulations during calibration and the predictions during validation can be considered satisfactory, especially considering that the model relies on global input datasets (Section 2.4.2). The development of a tailored Hydrological Units (HUs) dataset (Section 2.4.1) also contributed in improving the overall performance. When calibration data are available, the tuning of model parameters allows the model to adapt reasonably well to local conditions. Nevertheless, even in these local settings, clear limitations emerge. While the model performs well on average, uncertainties become more evident when moving toward extreme conditions. Deviations from observed values become more visible when focusing on the extremes. In particular, the simulations show a consistent tendency to underestimate observed peak flows, revealing a conditional bias. As will be discussed in Chapter 3, such behavior is common in rainfall–runoff modelling and reflects well-known challenges in reproducing high-flow dynamics.

When moving from local to regional simulations, a decline in predictive performance is expected and clearly observed. This deterioration highlights how uncertainties become increasingly relevant when predictions are made in ungauged catchments. At the regional scale, the model provides relatively homogeneous predictive capability across the Greater Alpine Region (GAR), but the results are nevertheless affected by substantial uncertainties. The summary efficiency metrics (wNSE and KGE) indicate that, when considering the entire discharge range, a portion of the catchments is represented reasonably well, although with noticeable degradation in both calibration and validation performance, while a significant number of catchments are poorly reproduced. As illustrated in Figure 6, several basins fall well below commonly accepted efficiency thresholds. These limitations become even more evident when focusing specifically on flood extremes (Figure 9 and Figure 10). The degradation in predictive performance can be attributed to a combination of uncertainty sources already present in local modelling, such as uncertainties in inputs, observations, model structure, and calibrated parameters, together with additional sources introduced by the regionalization process, particularly uncertainty related to parameter transfer. Although a flexible Random Forest approach was used to regionalize model parameters, and equifinality was partially accounted for within the

PASS algorithm, the regionalization process ultimately remains a form of statistical regression aimed at achieving acceptable average performance across the region. Inevitably, such an approach leads to misrepresentation of a number of individual catchments.

Furthermore, the analysis of the relative importance of catchment descriptors in the Random Forest parameter regionalization does not reveal clear and consistent physical relationships. This may partly reflect correlations among descriptors themselves: according to the theory of landscape co-evolution, many catchment characteristics develop in strong interdependence with others (for example, vegetation patterns are closely related to mean annual precipitation). At the same time, it also highlights a more fundamental limitation of similarity-based regionalization approaches, which assume that physical similarity between catchments necessarily implies similarity in hydrological response, an assumption that has been questioned in previous studies (Section 1.1.3).

Despite these limitations, three key points should be emphasized. (I) Even if predictive performance at individual ungauged cross-sections is sometimes unsatisfactory, the regional model remains valuable for analyzing flood events in a spatially distributed manner. Because the model is internally coherent across the entire region, it allows the reconstruction of spatial flood patterns that would be impossible to obtain by relying solely on multiple independent local models. In subsequent studies not included in this thesis, the model was also evaluated in terms of its ability to reproduce the spatial characteristics of major historical floods across the region. The comparison between simulated events and historical flood databases showed encouraging agreement in terms of event duration, intensity, and spatial extent. Such spatially consistent simulations also make it possible, under appropriate assumptions, to explore the potential effects of climate change on regional flood patterns using climate projections. Similar conclusions regarding the usefulness of regional modelling frameworks for distributed flood analysis have recently been reported for the same region (Janzing et al., 2025). (II) The results confirm that calibration and parameter regionalization remain useful tools for regional hydrological modelling. As briefly discussed in Section 1.1.3, alternative approaches based on Newtonian principles attempt to derive model parameters directly from catchment descriptors without calibration. The same study just mentioned (Janzing et al., 2025) applied such an approach to the GAR by using a non-calibrated model (with the exception of the two parameters of a snow-module). Although the results from that study cannot be directly

compared with those presented here, due to differences in catchment selection, input datasets, and calibration/validation periods, their distributed predictive performance was generally lower than that obtained with the calibrated and regionalized model presented in this work. Finally (III), the results presented in this chapter highlight an important message that connects directly to the following chapters of this thesis. When constructing hydrological models, especially when predictions are required in ungauged catchments, uncertainties are not only unavoidable, but can have a substantial impact on the reliability of model outputs. Ignoring these uncertainties may therefore be misleading and, in operational contexts, potentially dangerous. For instance, inaccurate flood predictions could influence emergency decisions and risk management strategies, ultimately affecting real-world outcomes. Despite this, many hydrological modelling studies still focus primarily on deterministic simulations, without explicitly quantifying predictive uncertainty or providing tools for decision-makers to account for it.

Starting from the next chapter, the focus of this thesis therefore shifts toward understanding the aggregated uncertainty of RR models and developing practical approaches to quantify and communicate this uncertainty in a way that can support more informed decision-making.

Chapter 3

(Lombardo et al., 2025)

3. General properties of rainfall-runoff model residuals

3.1 From a model to its uncertainty

3.1.1 About the generality of rainfall-runoff residuals

As anticipated in Chapter 2, the main objective of this thesis is to provide a comprehensive treatment of uncertainties affecting conceptual rainfall–runoff (RR) models and of the methods commonly used to describe them. The large-scale GAR case study presented in Chapter 2 clearly illustrates how RR models, both when applied at gauged locations and when transferred to ungauged catchments, are affected by non-negligible uncertainties. These uncertainties become particularly pronounced under high-flow conditions, which are of primary interest in the context of this work. Before addressing how such issues may be operationally handled, the present chapter is aimed at characterizing the properties of model residuals: how they are ideally expected to behave, how they actually behave in practice, and which challenging features must be accounted for in the development of the subsequent modelling framework.

When dealing with conceptual RR modelling and model residuals, a natural and legitimate question immediately arises: can the statistical properties of residuals be generalized across models, given the existence of hundreds of different RR formulations? Or, conversely, do different models produce fundamentally different residual structures? In addition, catchments to which these models are applied may differ substantially in terms of climatic regime, geomorphological characteristics, seasonality, and degree of intermittency, among many other factors. As a result, even the same model may generate markedly different residuals when applied to different basins. The main difficulty in addressing these preliminary questions lies in the availability of suitable datasets, namely datasets that include RR simulations for a large

number of catchments with heterogeneous climate–morphological characteristics, while at the same time relying on a wide range of different models calibrated over the same locations.

While the use of multiple catchments is common practice in hydrological studies reported in the literature (see McInerney et al., 2017, among many others), the systematic use of multiple RR models is, somewhat surprisingly, far less frequent. Researchers tend to rely on a limited number of models with which they are already familiar, often as a result of long-term experience with a specific modelling framework, and only rarely explore alternative model structures or configurations. When multiple models are considered, they are typically applied to a small number of case studies (e.g., Schoups & Vrugt, 2010; Tajiki et al., 2020), usually involving catchments with relatively limited variability in climate and morphological conditions. To address the questions outlined above, this chapter relies on a recently developed dataset (see Section 3.1.3), which provides discharge simulations for more than 500 catchments distributed across the continental United States.

At this point, a potential criticism may arise: if Chapter 2 identifies the Greater Alpine Region (GAR) as the primary geographical focus of this research, why focusing here on a completely different region? Several reasons motivate this choice. First, snow-dominated catchments characterized by accumulation and melt processes are also present in the United States, particularly in areas close to the Rocky Mountains. Second, to date no comprehensive multi-model dataset is available for the GAR; more generally, even single-model datasets covering a large number of catchments across the entire region are difficult to access. Third, and most importantly, although the initial analyses are carried out on a different geographical domain, all the investigated properties will be later recomputed in Section 4.2.3 for a newly developed test model applied to a subregion of the GAR. It will be shown that the identified residual characteristics largely generalize when moving from one region to another.

Finally, it is worth noting that several of the properties investigated in this chapter have already been discussed in previous studies. However, existing contributions have addressed them in a fragmented manner, and no single work has so far combined all these aspects within a unified framework, particularly given the aforementioned limitations in terms of catchment diversity and model variability. Whenever relevant, references to the existing literature will therefore be provided to place the present analysis in its proper scientific context.

3.1.2 Desired (ideal) properties of model residuals

Before entering into the details of the analysis of conceptual RR model residual properties, it is useful to first discuss the desired properties of models in a more general sense. Although these considerations may appear straightforward, they are often overlooked in practice. In this section, models are intentionally discussed within a broad framework, without reference to any specific hydrological formulation. The properties introduced here are not exclusive to hydrological models but apply to modelling in general, as they arise from fundamental considerations on residuals and their relationship with the variables involved in the modelling process. To illustrate these concepts, let us consider one of the simplest possible examples of a model: the fitting of a linear relationship (a straight line) to a set of observed data points, schematically represented in Figure 14.

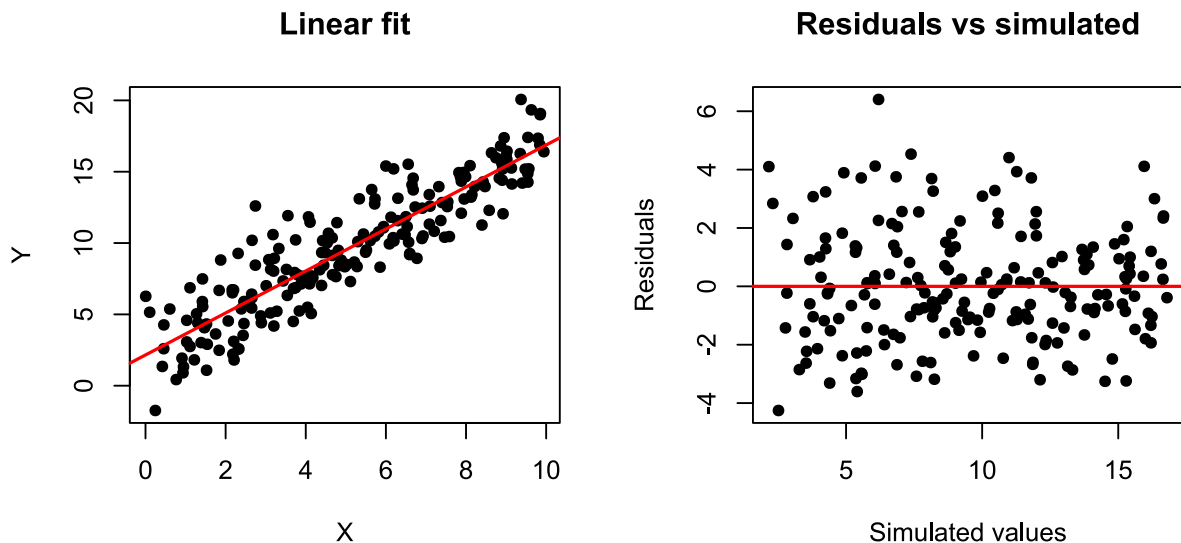


Figure 14. On the left: schematic representation of a simple linear fit between a dependent and an independent variable. In red the best fit with maximum Likelihood. On the right, scatter plot for the model residuals.

Given a dependent variable Y and an independent variable X , the model can be written as

$$Y = \theta_1 + \theta_2 X + \epsilon \quad (\text{Eq. 3.1})$$

where θ_1 and θ_2 are model parameters and ϵ represents the residuals of the linear fit. Several approaches exist for calibrating the model parameters (for example, maximum likelihood estimation), but they all aim, in one way or another, to minimize the discrepancies between model predictions $\hat{Y}$ and observations. Independently of the overall predictive skill of the model, residuals are generally expected to exhibit a number of desirable statistical properties if the model fit is to be considered satisfactory. First, (I) residuals should be unbiased, both in an absolute sense (i.e., mean residuals $\hat{\epsilon} \approx 0$) and in a conditional sense, meaning that the residuals should not display systematic patterns when conditioning on subranges of the independent variable (for example $\hat{\epsilon} \mid X > X_1$ or $\hat{\epsilon} \mid X < X_2$, where X_1 and X_2 define arbitrary subsets of the domain). Second, (II) residuals are often assumed to follow a normal distribution (Figure 15, left panel). This assumption is not a fundamental requirement for a model to be valid, but it is frequently adopted in statistical inference because it simplifies the derivation of analytical results and likelihood functions. Third, (III) residuals should ideally be homoscedastic, meaning that their variance should remain approximately constant across the range of the independent variable (Figure 15, right panel). Finally, (IV) in the case of time series data, residuals should not exhibit serial correlation; that is, the residual at time step t should not systematically influence residuals at subsequent time lags. Serial correlation may arise, for instance, when temporal patterns present in the observed process are not adequately captured by the model structure. It is important to note that these properties describe the statistical behavior expected from a well-specified model. However, satisfying these conditions does not necessarily imply that the model has strong predictive power. For example, it is possible for a model to explain only a limited fraction of the variability of the dependent variable, resulting in large residual variance, while still producing residuals that are unbiased and randomly distributed. In such cases the model may still be statistically well-behaved, even though its explanatory capability remains limited.

In real hydrological systems, such as those discussed in Section 2.3.1, models are evidently far more complex: they are typically nonlinear, often non-differentiable, and characterized by a larger number of parameters θ that must be calibrated. Nevertheless,

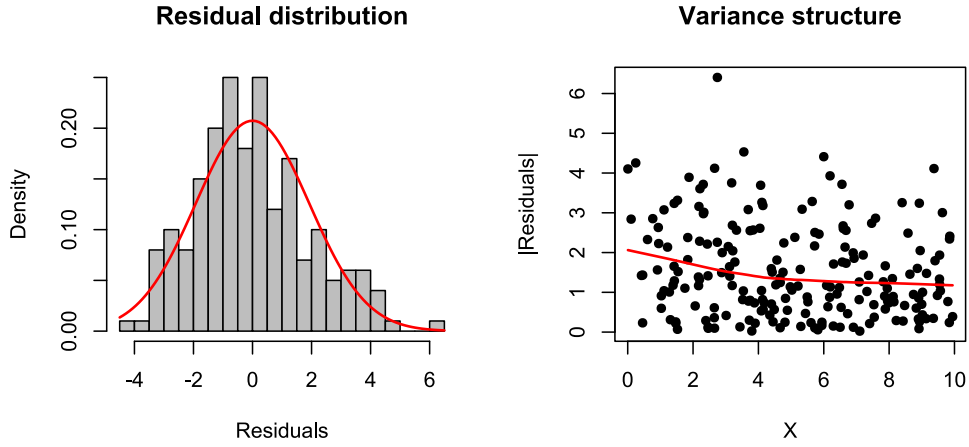


Figure 15. On the left: empirical histogram of model residuals for the linear fit, showing the expected bell-shaped curve of a normal distribution. On the right: representation of model residuals homoskedasticity, with variance nearly constant for any range of the independent variable X .

the same conceptual framework can be generalized to any model, which can be expressed as

$$Y = M(\bar{X}, \bar{\theta}) + \epsilon \quad (\text{Eq. 3.2})$$

where Y represents the target process (e.g., discharge in a RR application), M denotes the model structure (for instance the modified TUWmodel), $\bar{X}$ is a vector of input variables, and $\bar{\theta}$ is the vector of model parameters. The term ϵ again denotes the residual discrepancies between model predictions and observations, which the calibration procedure aims to reduce. Despite the increased complexity of the system, residuals are, in principle, still expected to satisfy the same general properties discussed above. This framework provides a useful basis for analyzing the statistical properties of RR model residuals.

A final remark concerns parameter calibration. One of the most widely used approaches for model fitting is maximum likelihood estimation (MLE). Without entering into excessive detail, the general idea is that, assuming a probability distribution for the residuals (often Gaussian), a likelihood function can be defined and maximized with respect to the model parameters. For instance, under the assumption

of independent and normally distributed residuals with variance σ^2 , the log-likelihood function can be written as

$$\mathcal{LL} = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^N \epsilon_t^2 \quad (\text{Eq. 3.3})$$

where N is the number of observations and ϵ_t denotes the residual at time step t . From this expression, it can be observed that maximizing the likelihood under Gaussian assumptions is equivalent to minimizing the sum of squared residuals. In practice, however, the direct application of likelihood-based methods in conceptual RR modelling can be challenging. This difficulty does not arise from the impossibility of defining likelihood functions for complex models, but rather from the fact that residual distributions in hydrological applications often violate the simplifying assumptions typically adopted in statistical inference, such as independence, homoscedasticity, or normality. Moreover, conceptual RR models may be nonlinear and non-differentiable, making analytical optimization more difficult and often requiring computationally intensive numerical procedures. For these reasons, conceptual RR models are frequently calibrated using alternative approaches based on model efficiency (ME) functions, also referred to as objective functions (OFs), as previously done in Chapter 2. These functions directly quantify the discrepancy between simulated and observed values of the target hydrological variable and are explicitly maximized during calibration. Among them, the Nash–Sutcliffe Efficiency (NSE) and the Kling–Gupta Efficiency (KGE) represent by far the most widely used metrics in hydrological modelling studies, even though several alternative or modified performance measures have been proposed in the literature.

3.1.3 Study Area and multi-model ensemble

This study utilizes a subset of the CAMELS dataset (Addor et al., 2017; Newman et al., 2015), containing data for 559 catchments across the contiguous United States, spanning a wide range of climatic and hydrological characteristics (such subset was obtained from the original 671 catchments dataset after processing for consistency in catchment size and water balance, see Knoben et al., 2020). Simulations (at a daily time scale) are conducted using 78 configurations of the Framework for Understanding Structural Errors (FUSE) software (M. P. Clark et al., 2008), enabling an exploration of differences in hydrological process representation.

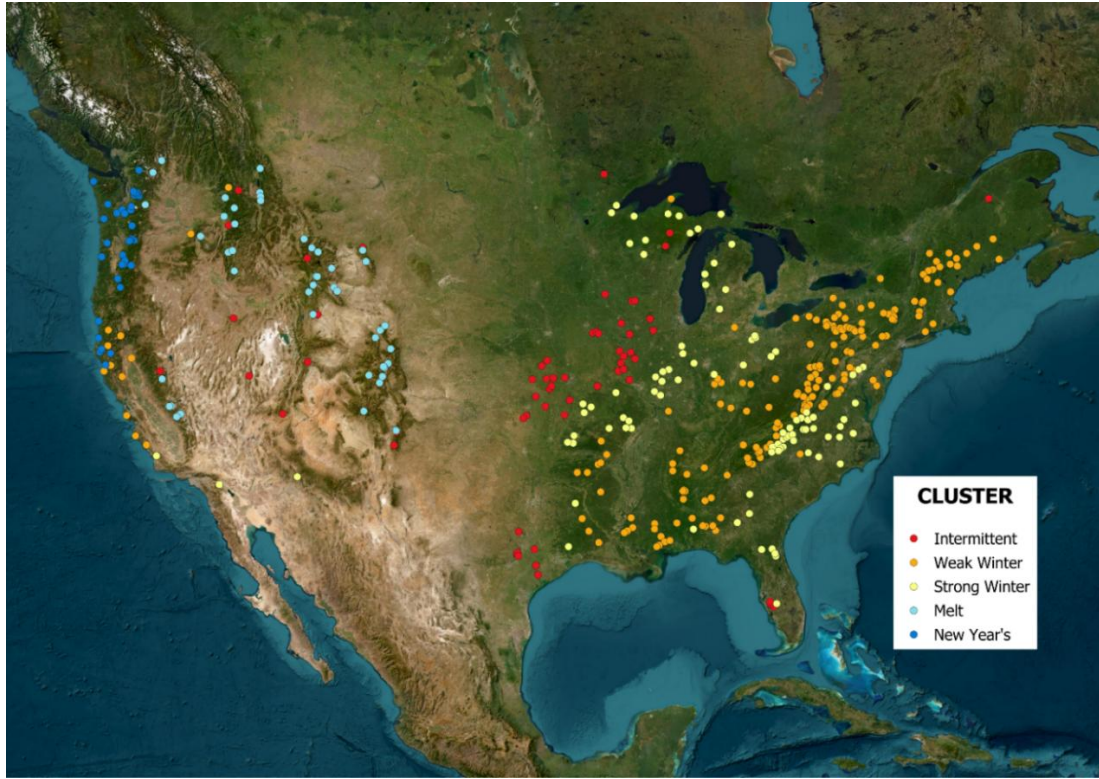


Figure 16. Filtered CAMEL gauging stations over the continental US used in the study. Each color represents one of five discharge regimes following the classification by Brunner et al., 2020.

Each FUSE configuration is a different conceptual hydrological model where the vertical dimension is always discretized into two zones (unsaturated and saturated), but in which the different configurations vary in the definition of the architecture of the zones (the number and arrangement of “buckets” that represent the storage of water) and their parameterization (the equations that represent the model fluxes such as infiltration, vertical percolation, and baseflow, that affect the time evolution of water storage in each of the conceptual buckets).

The dataset leverages recent efforts to calibrate a large ensemble of FUSE models to a large sample of CAMELS catchments (Thébaud, C., 2025). The dataset includes over 43,600 simulated discharge time series (78 FUSE models for each of the 559 catchments) that cover regions from humid eastern climates to arid southwestern zones. Model calibration is performed using the Shuffled Complex Evolution algorithm (SCE-UA; Duan et al., 1992), with 1,000 iterations, and the Kling–Gupta Efficiency as the

objective function. Each of the FUSE models was calibrated over the time period 1989–1998, following a one-year warm-up period (1988), and validated over the period 1999–2009. To maintain robust ensemble analyses, the catchments with intermittent flow are removed: the reason behind this choice is that catchments with observations of zero flow present completely different characteristics from any other catchments (see Section 3.4.4 for a more detailed discussion). This filtering reduces the dataset from 559, to 422 catchments, with residual analysis restricted to the validation period only.

3.1.4 Residuals definition and catchments' clustering

In the context of a model, whether hydrological or not, residuals can typically be defined as:

$$\epsilon_{\mathcal{T}}(t) = \mathcal{T}(Q_s(t)) - \mathcal{T}(Q_o(t)) \quad (\text{Eq. 3.4})$$

Here, ϵ represents the model residuals, Q_s (the letter Q here is used to already refer to river discharge) denotes the simulated values from the model, Q_o refers to the observed values, $\mathcal{T}$ indicates a potential transformation, and t is the time index of the model simulation. Various transformations $\mathcal{T}$ for discharge have been explored in the literature, including logarithmic (Shabestanipour et al., 2023), log-sinh (Wang et al., 2012), and Box-Cox transformations (Box & Cox, 1964; McInerney et al., 2017). Such transformations are commonly employed to stabilize residual properties like heteroskedasticity, reducing skewness, or improving adherence to normality assumptions, thereby enabling more robust statistical analyses. For instance, the Box-Cox transformation has been extensively used to correct variance and distributional issues in residuals (McInerney et al., 2017). If no transformation is applied, $\mathcal{T}(x) = x$, and the residuals remain untransformed. In this study, residual analyses are conducted on untransformed residuals in their "natural" space, alongside two transformations applied to the simulated and observed discharge before computing their difference: (a) the logarithmic transformation $\mathcal{T}(x) = \ln(x)$ is adopted due to its widespread application; (b) a Box-Cox transformation with $\lambda=0.5$ as previous works (McInerney et al., 2017) suggested it as improvement to the simpler log-transformation. To complement the residual analysis, the dataset's catchments are stratified into five distinct hydrological categories based on previously proposed clustering (Brunner et al., 2020). The five clusters are: (1) Intermittent regime (75 catchments): mostly characterized by arid catchments; (2) Weak winter regime (171 catchments): with

minimal seasonality but winter-dominated flows; (3) Strong winter regime (88 catchments): showcasing pronounced seasonality and significant winter peaks; (4) Melt regime (51 catchments): dominated by snowmelt, with spring-driven flows; and (5) New Year's regime (37 catchments): featuring strong seasonal flows peaking around the start of the year (Figure 16). For clarity, the Intermittent Regime (that contained nearly all the intermittent catchment removed from the dataset), is still included due to the presence of arid, but not intermittent, catchments.

3.2 Statistical Characteristics of Residuals

3.2.1 Do residuals display seasonality?

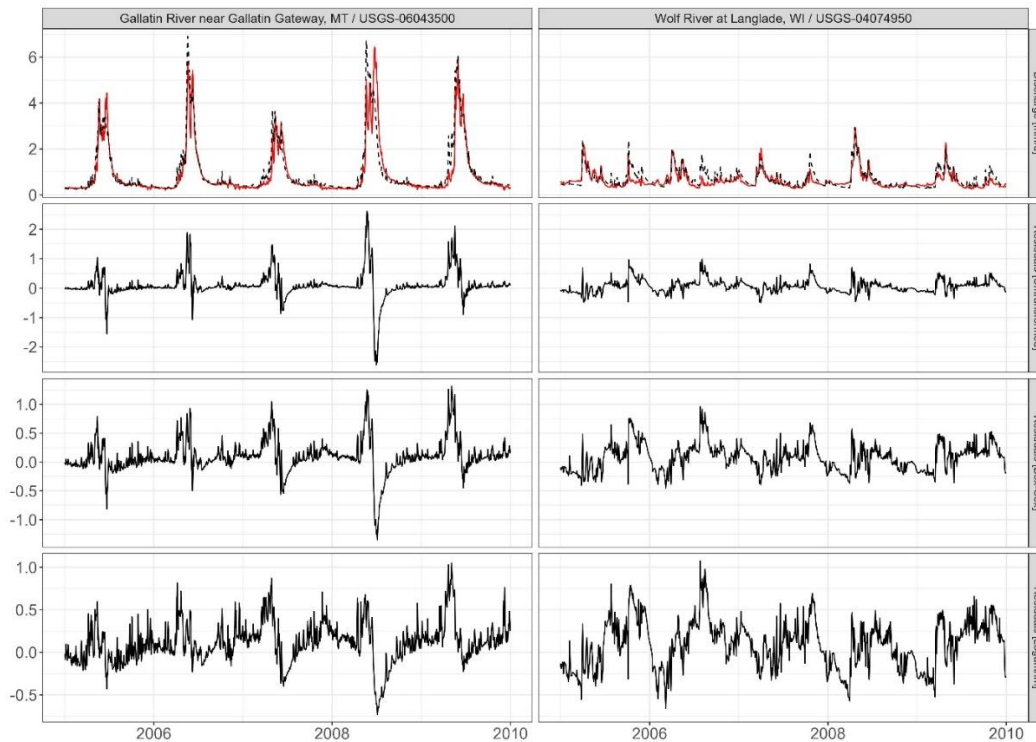


Figure 17. Observed discharge (in black) and simulated discharge (in red), untransformed residuals, residuals after Box-Cox ($\lambda=0.5$) transformation, and residuals after logarithmic transformation for two catchments (left - gauge USGS gauge 06043500, Melt Regime; right - USGS gauge 04074950, Intermittent Regime), for Model n° 1.

River flows are well known to exhibit varying degrees of seasonality, driven by meteorological factors such as precipitation patterns, snow accumulation, and

snowmelt. However, the periodic variations in model residuals have not been thoroughly analyzed in the literature, despite being of interest for residual characterization, and having been implemented in uncertainty estimation method in the past (McInerney et al., 2020). Seasonality persistence in model residuals is directly related to how strong discharge seasonality is for the considered catchment. In Figure 17, two examples are reported for the Gallatin River near Gallatin Gateway catchment (on the left, a catchment belonging to the Melt regime, displaying a strong seasonality), and for the Wolf River at Langlade (on the right, an arid catchment belonging to the Intermittent regime instead, displaying a weak seasonality) for Model n°1 for the sub period of 5 years (2005-2009) of the validation period.

It can be immediately noticed how in Figure 17A, model residuals still show a very clear seasonal component in the signal, less evident instead in Figure 17B. Moreover, such seasonality is not only relevant for untransformed residuals, but it is also persistent after the application of the Box-Cox and logarithmic transformations. Motivated by this evidence, in order to isolate non-seasonal components, a deseasonalization of the residual time series is performed. The seasonal component is estimated by calculating the mean and standard deviation of residuals for each day of the year across the validation period, then standardizing the residuals accordingly (below for the case of residuals in “natural” space):

$$\hat{\epsilon}_{d,j} = \frac{\epsilon_{d,j} - \mu_d}{\sigma_d} \quad (\text{Eq. 3.5})$$

Where d indicates the day of the year, j the specific year, μ_d the mean and σ_d the standard deviation. The same transformation is also applied analogously to the series resulting after transformation of the observed and simulated discharge variables. This operation is basically a normalization of the residuals on a daily basis. While more sophisticated methods, such as Fourier-based deseasonalization or harmonic regression, could be applied (Sang, 2013; Sang et al., 2016) the simplicity and effectiveness of the chosen method make it a practical choice for this study. From this point forward, the deseasonalized residuals will be referred to as the DS time series.

3.2.2 What is the probability distribution of model residuals?

Key features of model residuals are their shape properties, which frequently exhibit heavy tails and asymmetry (Evin et al., 2014; Schoups & Vrugt, 2010; Shabestanipour et al., 2023). These characteristics pose significant challenges for statistical modelling,

particularly when residuals deviate substantially from normality. A simple and intuitive example is shown in Figure 18, where, for each catchment, the percentage of days contributing to 50% of the sum-of-squared error (SSE) is reported (see also Clark et al., 2021). The SSE is defined as:

$$\sum_{t=n-k+1}^n \hat{\epsilon}^2(t) = 0.5 * SSE \quad (\text{Eq. 3.6})$$

$$SSE = \sum_{t=1}^n (Q_s(t) - Q_o(t))^2 = \sum_{t=1}^n \epsilon^2(t) \quad (\text{Eq. 3.7})$$

With n the length of the residual timeseries, $\hat{\epsilon}^2$ the sorted (in ascending order) squared untransformed residuals and k determining a subset of the largest residuals. Once solved Eq.3.7 for k , the corresponding percentage is then easily calculated as k/n . It is immediately clear how just few data points, especially for the westside catchments, can heavily influence the estimate of the MSE, hence the distribution is heavy tailed.

Traditionally, shape properties are quantified using product moment ratios (Hosking, 1990), such as skewness (C_s) and kurtosis (C_k). However, in many practical scenarios, the extreme nature of residual distributions can result in infinite or undefined values for C_s and C_k , thereby limiting their applicability (Vogel et al., 2024). This issue is particularly pronounced in hydrological contexts, where strong variability and outliers are common. To overcome these limitations, L-moments (Hosking, 1990) have been developed as a robust alternative for characterizing heavy-tailed distributions. L-moments are less sensitive to extreme values and provide a more stable framework for analyzing asymmetry and heavy tails. They are defined as a linear combinations of probability-weighted moments (PWMs – Greenwood et al., 1979; Sillitto, 1951) and are denoted by the symbol λ (where λ_1 represents the mean, λ_2 is a measure of dispersion in a probability distribution and so on). Analogous to C_s and C_k , L-skewness and L-kurtosis are defined by the ratios $LC_s = \lambda_3/\lambda_2$ and $LC_k = \lambda_4/\lambda_2$, respectively, and retain the same interpretive meaning as their traditional counterparts. Moreover, L-moment diagrams offer a valuable tool for comparing empirical data samples to theoretical distributions.

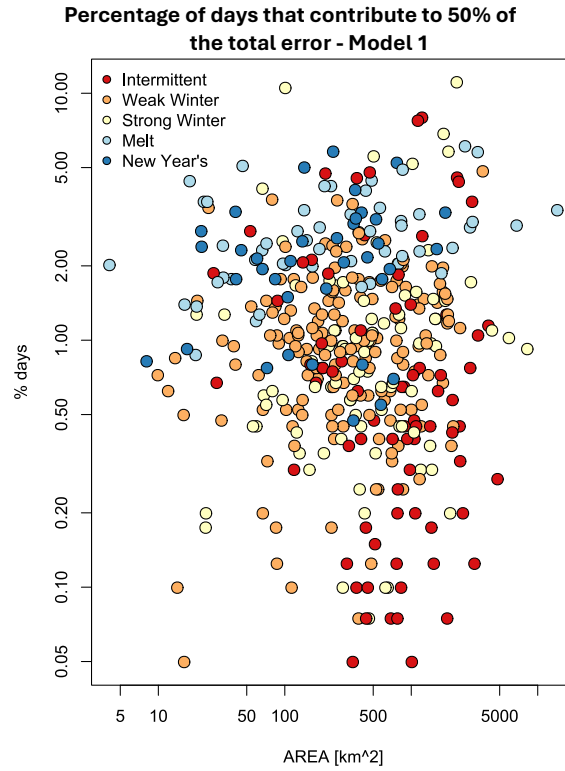


Figure 18. Percentage of days contributing to 50% of the SSE for the 422 discharge gauge stations (model n°1). Regimes: Intermittent in red, Weak Winter in orange, Strong Winter in yellow, Melt in light-blue, and New Year's in blue.

Figure 19 illustrates L-moment diagrams for untransformed residuals and residuals computed after the application of transformations. Each point represents median L-moment values across all models applied to a given catchment, ensuring robustness by avoiding reliance on single-point estimates. Being all the models being forced with the same climate forcings dataset, and assuming most of the residuals are explained by input uncertainty, looking at the median tendency per catchment is interesting because it explores a possible space of how the data can be modelled. For untransformed residuals (Figure 19A), L-kurtosis values frequently exceed 0.4, a threshold suggested in the literature (Vogel et al., 2024) for classifying distributions as extremely heavy-tailed. Such extreme tails can disrupt most traditional statistical analyses. In some catchments, values reach as high as 0.8. These empirical points deviate substantially from the expected curves of theoretical distributions, underscoring the difficulty of modelling such data with conventional methods.

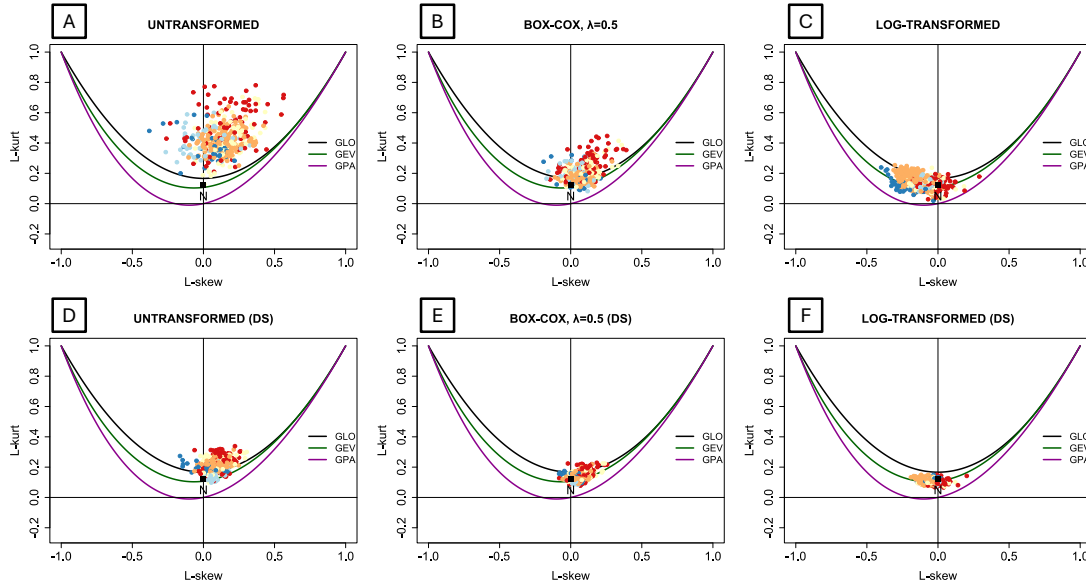


Figure 19. L-moment diagrams (median L-kurtosis vs median L-skewness across models) for different transformations: (A) Untransformed; (B) Box-Cox ($\lambda=0.5$); (C) Log-transformed; (D) Untransformed deseasonalized; (E) Box-Cox ($\lambda=0.5$) deseasonalized; (F) Log-transformed deseasonalized. Expected values for the distributions GLO (generalized logistic), GEV (generalized extreme), GPA (generalized Pareto), and Normal (N). Points colored based on cluster (Intermittent in red, Weak Winter in orange, Strong Winter in yellow, Melt in light-blue, and New Year's in blue). For reference about the expected values for the theoretical distributions see Hosking, 1991.

By contrast, after the application of transformations (Figure 19B and Figure 19C) residuals exhibit substantially lower L-kurtosis, typically remaining below the 0.4 threshold. However, the log transformation, despite further reducing kurtosis with respect to the Box-Cox transformation, also introduces a consistent shift toward negative L-skewness, with values occasionally reaching as low as 0.5.

Following the findings in Section 3.2.1, L-moments are recalculated also for the DS residual time series. Figure 19D shows how much of the variability in L-moment ratios values for untransformed residuals was attributable to seasonality rather than inherent distributional characteristics. For untransformed DS residuals, L-kurtosis values drop significantly, often falling below the 0.4 threshold, although they remain higher on average than those for transformed DS residuals after transformation.

Similarly, L-skewness values demonstrate a reduced range (also after transformation), further stabilizing the distributions.

Still inside the theory of L-moments, Hosking (Hosking, 1999), suggested, for symmetric distribution, for which all odd order moments are by definition equal to zero, to plot instead λ_6/λ_2 vs λ_4/λ_2 to better capture the behavior of the tails.

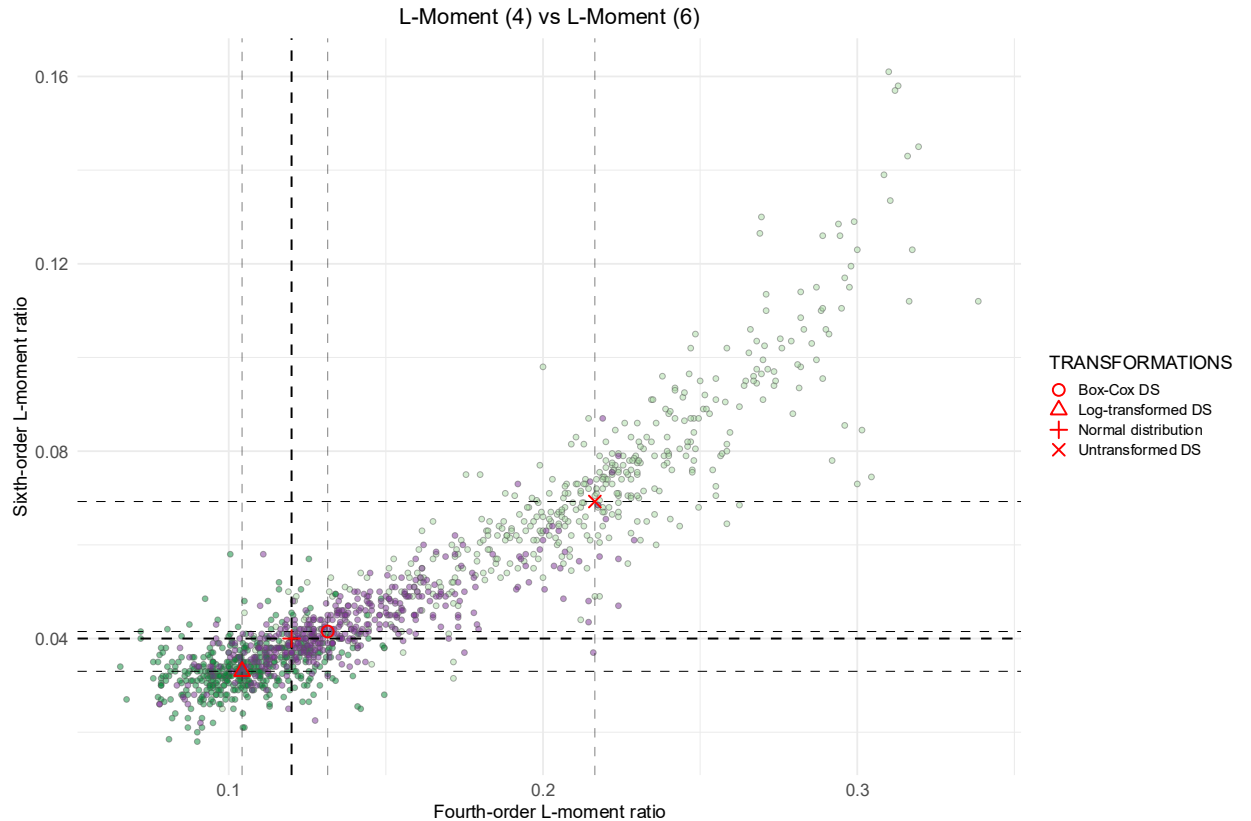


Figure 20. L-moment diagram (λ_6/λ_2 vs λ_4/λ_2) for deseasonalized untransformed, Box-Cox ($\lambda=0.5$) and logarithmic residuals. In red the median L-moments across catchments and models, together with the expected value for a normal distribution.

Figure 20 clearly shows how looking at higher order L-moments ratios allow for further distinction between the behavior of symmetric distribution: in particular for DS residuals after transformation, it can be observed how the two transformations, despite both approximating on average a normal distribution, do it in different ways.

3.2.3 How much does model selection matter in shaping distribution marginal properties?

In Section 3.2.2, the analysis focused on how potential transformations affected the marginal distributional properties of residuals. However, limited attention was given to the variability of these properties across different models, as Figure 19 only presented the median L-moment ratios for each catchment. Building on the diverse set of models included in the dataset, this section provides a more detailed investigation of such variability. Specifically, the analysis explores the spread of L-skewness and L-kurtosis values around their median, aiming to evaluate the extent to which the choice of a hydrological model influences the shape properties of residuals. For each catchment, the interdecile range ($Q_{90}-Q_{10}$) is calculated for both L-skewness and L-kurtosis, for each catchment, across the 78 models. The results, displayed in Figure 21, are again organized by the different transformations already introduced in Section 3.1.4 and Section 3.2.1.

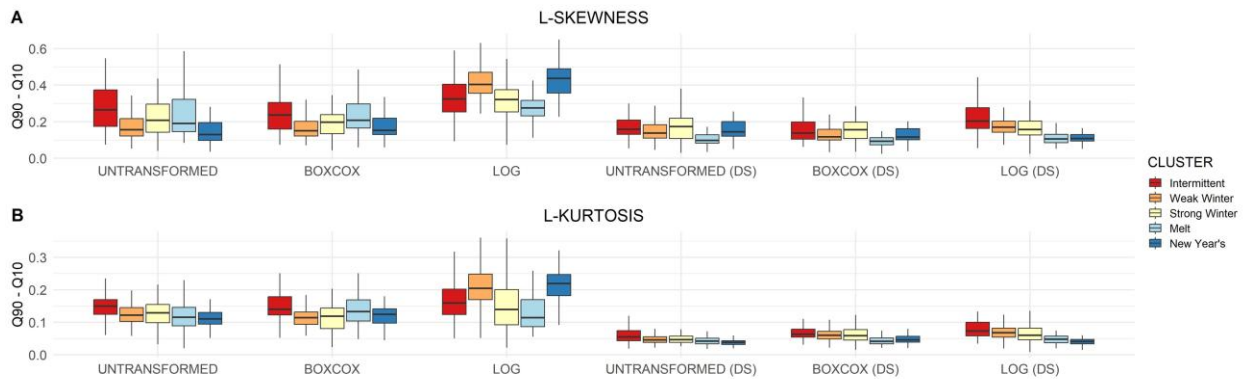


Figure 21. Interdecile range boxplots of L-moment ratios for untransformed residuals and residuals after transformation, with and without seasonal component, for L-skewness (A) and L-kurtosis (B) across models.

Figure 21 highlights how the variability across models depends on the transformation applied. Untransformed residuals, which exhibited on average L-moments ratios with much heavier than normal tails, generally show substantially lower variability compared to the residuals after log-transformation, but comparable to the Box-Cox transformation. Indeed, for both asymmetry and heavy tails, after log-transformation residuals exhibit the greatest variability among all transformations considered. The behavior changes considerably for DS residuals, whether untransformed or not. Here,

the variability in L-skewness and L-kurtosis, especially the latter, is significantly reduced compared to residuals that retain their seasonal component. In certain cases, such as for the Melt and New Year's regimes, the variability is nearly eliminated, indicating that the influence of model choice on the variability of shape properties is minimal.

Residuals that retain their seasonal component therefore introduce additional challenges: if different models produce markedly varying levels of asymmetry and heavy tails, extreme caution is required in practical applications. The variability in residual properties suggests that switching between models could require the use of different error models based on different assumptions; nevertheless, practitioners often rely on familiar models they trust and understand, and rarely consider alternative models, which may limit their ability to account for such variability.

3.2.4 Is heteroskedasticity reduced by a data transformation?

A widely discussed property of hydrological model residuals is their relationship with discharge values, whether observed or simulated. The structure, well-documented in hydrological literature (Evin et al., 2014; McInerney et al., 2017; Schoups & Vrugt, 2010; Sorooshian & Dracup, 1980) is well known to be heteroskedastic. Addressing heteroskedasticity is a primary motivation for introducing transformations, which aim to stabilize variance and approximate a homoscedastic structure, as shown in previous studies (McInerney et al., 2017; Shabestanipour et al., 2023), which highlight the benefits of transformations in improving the interpretability of residuals and reducing variance irregularities. Building on these observations, this section investigates the consistency of heteroskedasticity patterns across catchments, the influence of hydrological clustering, and the effects of removing seasonal components from residual time series. To enable meaningful comparisons, a consistent methodology was applied across all catchment and model combinations:

1. Normalization: each residual time series is normalized by dividing it by its standard deviation. In this way, catchments-model pairs having the same heteroskedastic structure, differing at most for a scaling factor, overlap. In other words, Figure 22 shows the shape of the heteroskedastic structure, not the absolute values of standard deviation for different ranges of simulated discharges;

2. Rearranging: the simulated discharge values are sorted in increasing order; the same sorting is also used to re-order the residual time series;
3. Standard deviation evaluation: using a moving average (window's size of 200 values, ~5% of the data sample), the variance of the newly ordered residual time series is evaluated (in this way a smooth curve is obtained instead of a series of few points for increasing ranges of simulated discharges).

This analysis is conducted for both untransformed residuals and residuals after transformation, with and without seasonal components. The results, summarized in Figure 22, depict shaded areas indicating variance interquartile ranges (25th–75th percentiles, darker shades) and interdecile ranges (10th–90th percentiles, lighter shades) across all catchments and models.

For untransformed residuals, a clear pattern emerges: variance increases linearly at lower quantiles before transitioning to exponential-like growth at higher discharge values. Removing seasonality reduces the overall variance and flattens the initial trend, as reflected in the narrower y-axis range. In contrast, after the log-transformation, residuals display nearly constant variance across discharge quantiles, with only slight deviations at extreme quantiles. As for untransformed residuals, removing seasonality further narrows the y-axis range and marginally reduces variance. The effect of the Box-Cox transformation is instead in the middle of the two, displaying heteroskedasticity mitigation, but resulting less effective than the log-transformation. Variability among hydrological clusters, despite not being reported here, is relatively small, with median trends closely aligned across clusters, except for the Melt regime, which diverges significantly at higher discharge quantiles in all cases. Additionally, differences between hydrological clusters are less pronounced after the application of transformations, reflecting more uniform behavior across the ensemble. It is important to note that while Figure 22 represents ensemble averages across models and catchments, individual cases may deviate significantly from these central tendencies, particularly for untransformed residuals where variability is greater. As such, careful attention is warranted when analyzing specific cases to account for potential deviations from the general trends.

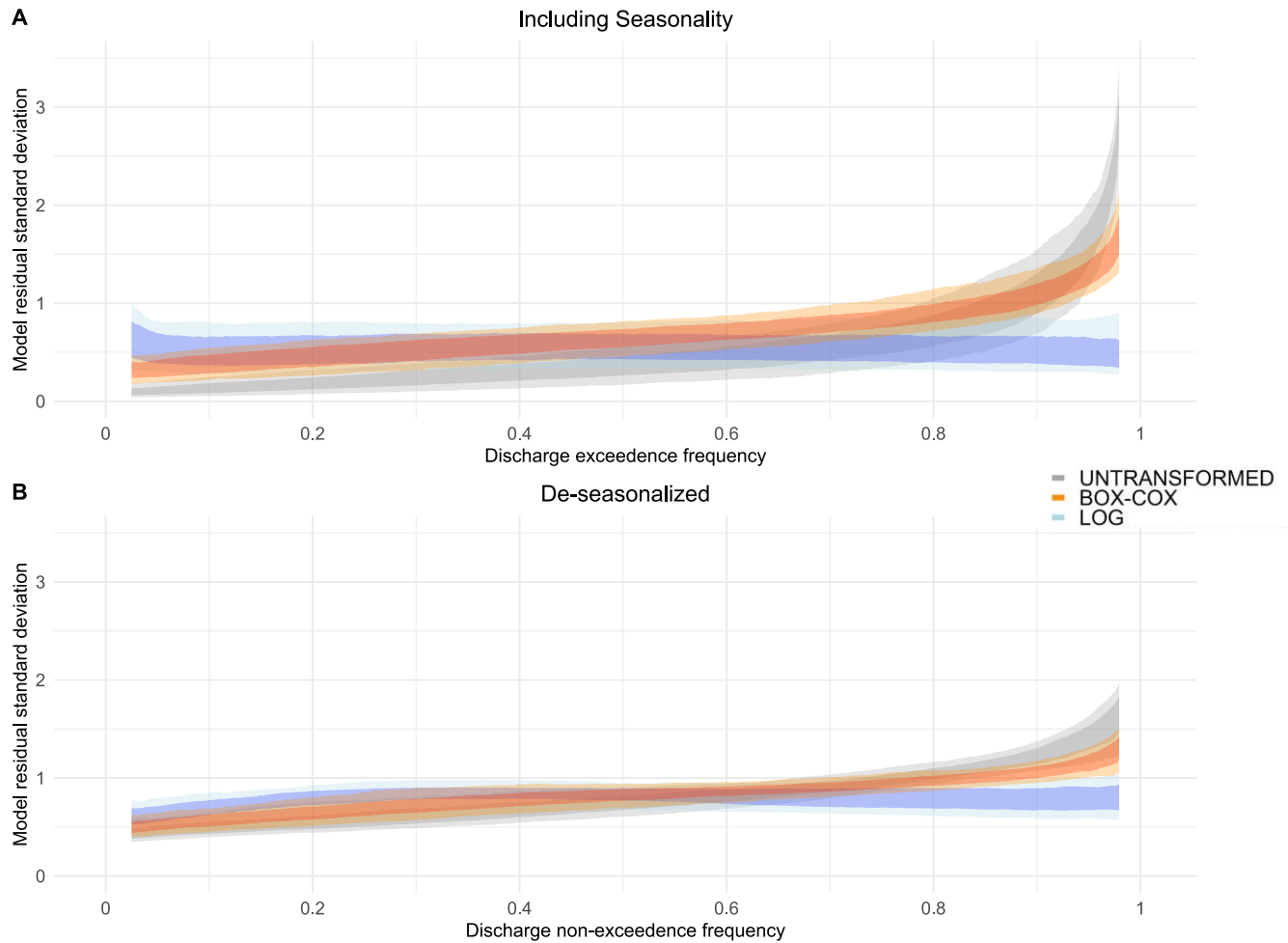


Figure 22. Model residual standard deviation as function of simulated discharge for untransformed residuals, Box-Cox ($\lambda=0.5$) and logarithmic transformation: (A) with seasonality; (B) after deseasonalization.

3.3 Temporal Correlation Characteristics of the Residuals

3.3.1 How persistent are the model errors?

Discharge data are widely recognized for their long memory and strong autocorrelation structures, which persist over time-lags much longer than the typical time steps used in hydrological modelling (e.g., daily or hourly). This phenomenon, extensively documented in the hydrological literature, is attributed to the slow

dynamics of catchment processes and storage mechanisms (Sorooshian & Dracup, 1980). A key question arises: do similar autocorrelation characteristics persist in model residuals? Specifically, when a model produces an error, does that error persist across subsequent time steps? Furthermore, how much of this autocorrelation is influenced by the seasonal component of the signal?

Residual autocorrelation has been observed in previous studies (McInerney et al., 2020; Shabestanipour et al., 2023; Sorooshian & Dracup, 1980), often analyzed using tools such as the autocorrelation function (ACF) and partial autocorrelation function (PACF). These methods measure the correlation of a time series with lagged versions of itself, typically using the Pearson correlation coefficient. However, the application of Pearson correlation in this context has limitations. Pearson's coefficient assumes linearity, normality, and homoscedasticity, assumptions frequently violated by residuals, particularly untransformed or seasonal ones. Barber et al. (Barber et al., 2020) documented the poor performance of Pearson's correlation when used with skewed data and they introduce improved estimators of the Pearson correlation coefficient for highly skewed hydrological data. As outlined in Section 3.2.2 and Section 3.2.3, untransformed residuals often deviate significantly from normality, exhibiting heavy tails and skewness. Even after transformation and deseasonalization, these deviations, while reduced, are not entirely eliminated. Figure 23 highlights challenges associated with Pearson correlation. Scatterplots at lag-1 for residuals before and after transformation are applied show numerous outliers and unequal dispersion around the 1-to-1 line, particularly in seasonal data. This uneven spread reflects non-linear relationships and heteroskedasticity, limiting the validity of Pearson correlation. To address these issues, alternative correlation measures may be necessary for residual analysis. For example, Kendall or Spearman rank correlation, which evaluates rank order rather than magnitude, may offer robustness to outliers and better captures non-linear relationships.

Figure 24C and Figure 24D display a comparison between the median autocorrelation function for the large-ensemble (all catchments and models) calculated with Pearson's correlation (continuous line) and Spearman's correlation (dashed line). The most striking observation is the persistence of autocorrelation across multiple lags for all transformations. This indicates that model errors tend to propagate, with consecutive overestimations or underestimations persisting over several simulation time steps. Additionally, the different transformations enhance overall correlation. Specifically, moving from untransformed residuals to residuals calculated after the

Box-Cox and logarithmic transformations, a noticeable increase in autocorrelation is observed, particularly at lower lags.

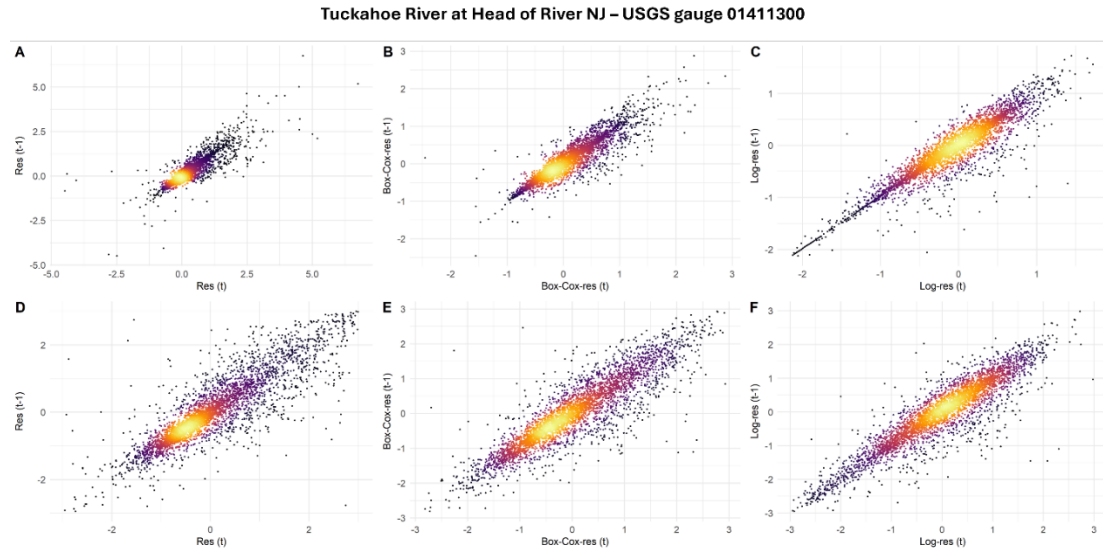


Figure 23. Scatterplots (t vs $t-1$) for untransformed residuals (A), Box-Cox ($\lambda=0.5$) residuals (B), logarithmic residuals (C) (USGS gauge 01411300 - model $n^{\circ}1$). The corresponding deseasonalized residuals are shown in panels D, E, and F. Color scale indicating points density.

A similar pattern is also evident for DS residuals, though the effect is less pronounced. The differences between the two correlation metrics are most noticeable for untransformed residuals where the rank correlation is substantially higher than the sole linear correlation. The increase is observable also for the Box-Cox transformation (to a less degree) while it is nearly negligible for the logarithmic transformation instead. In DS residuals, a general reduction in autocorrelation for the Box-Cox and logarithmic transformations at lower lags is observed, although the correlation remains substantial. For untransformed residuals is instead observed an opposite trend, with a general increase in linear correlation at lower lags. Still in DS residuals, the differences between Pearson and Spearman are negligible.

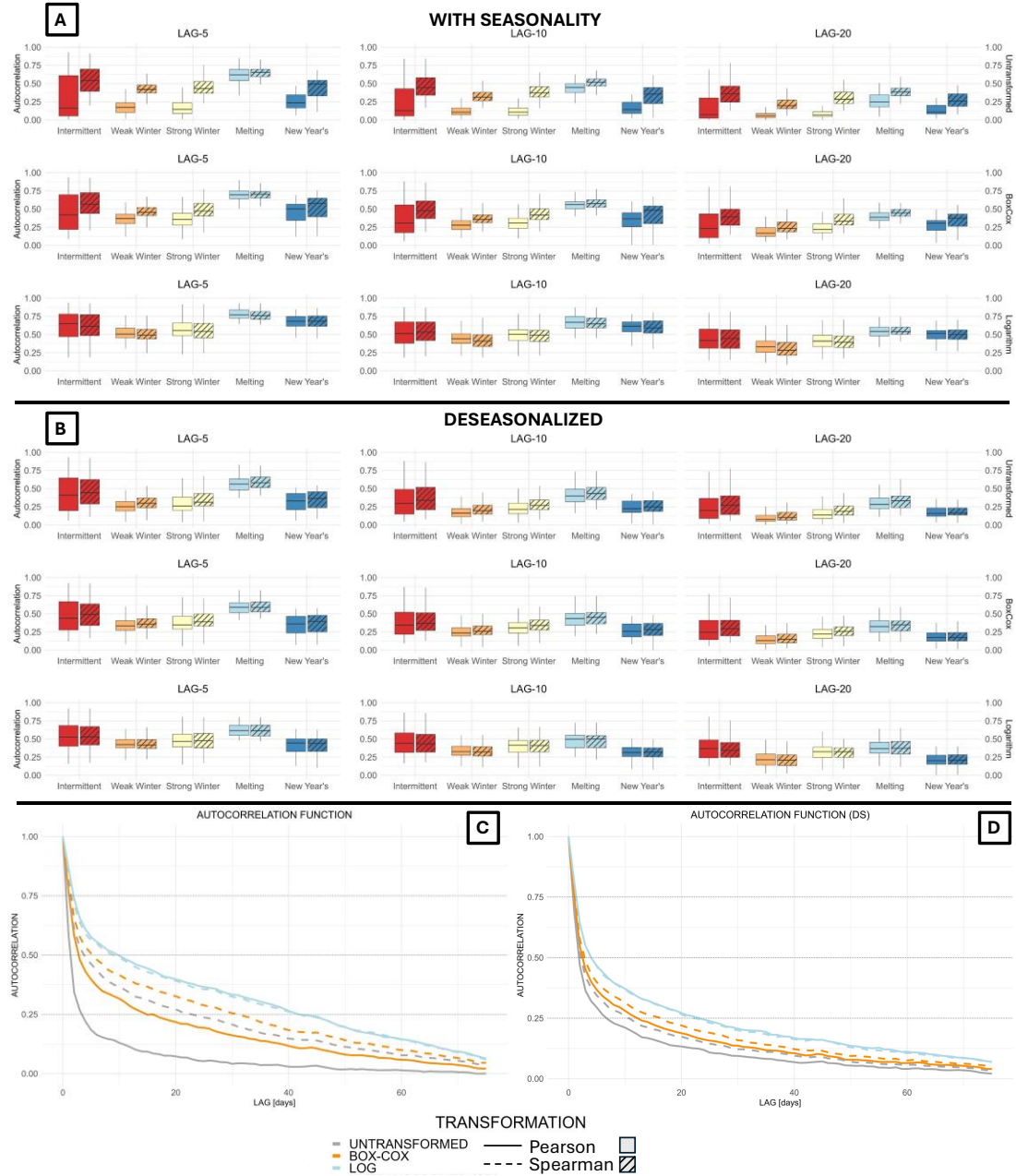


Figure 24. Median autocorrelation function for the large-ensemble for residuals with (C) and without (D) seasonal components for untransformed, Box-Cox and logarithmic transformations up to lag-75. Boxplots (A and B) represent the distribution at lags 5, 10, and 20 across clusters and transformations. The clusters follow the color scheme described in Section 3.1.4. Continuous lines (and boxplots) for Pearson correlation, dashed lines (and boxplots) for Spearman correlation.

Examining differences across clusters (Figure 24A and Figure 24B – uniform boxplots for Pearson, dashed boxplots for Spearman), it is evident that at lags 5, 10, and 20, all groups exhibit similar behaviors, with the exception of the Melt regime (and to a lower extent the New Year’s regime), which consistently shows substantially higher correlation both before and after seasonality removal. In contrast, catchments belonging to the intermittent regime display much greater variability. This variability may be attributed to differences in the amount of “low-flow” days: catchments with a higher frequency of low discharge days exhibit stronger autocorrelation at lower lags due to consecutive close-to-zero values, whereas rivers that experience them with a lower frequency may tend to show lower autocorrelation.

3.3.2 Temporal Correlation’s upper and lower tail dependency

As residual values transition from low to high, their correlation structure may vary substantially. While the correlation coefficient provides useful insights into the overall rank correlation across different lags, tail dependency offers a complementary perspective by focusing on extreme values. Tail dependence coefficients, which quantify the strength of dependence in the upper and lower tails of a distribution, can be estimated non-parametrically as:

$$\lambda_L = \lim_{u \rightarrow 0^+} P(Y \leq F_Y^{-1}(u) | X \leq F_X^{-1}(u)) \quad (\text{Eq. 3.8})$$

$$\lambda_U = \lim_{u \rightarrow 1^-} P(Y > F_Y^{-1}(u) | X > F_X^{-1}(u)) \quad (\text{Eq. 3.9})$$

Where X and Y are two variables with marginal quantiles $F_X^{-1}(u)$ and $F_Y^{-1}(u)$, and $P(Y|X)$ is the conditional probability.

Figure 25 presents the upper and lower tail dependence coefficients as functions of lag for untransformed residuals and residuals after transformations (for Weak Winter and Melt regimes). These two regimes are selected because they display the most different behaviors. Each of the four panels displays the median upper (top half) and lower (bottom half) tail dependence, along with the interquartile ranges (Q25–Q75). The analysis incorporates all catchments within each cluster and all models for each catchment.

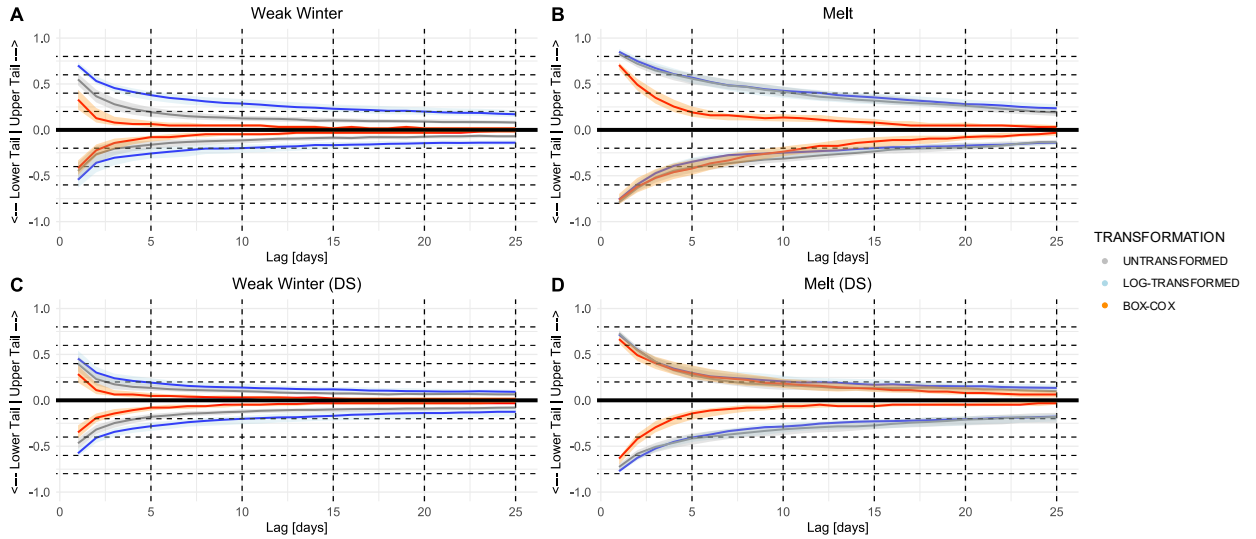


Figure 25. Upper and lower temporal tail correlation up to lag 25 for the different transformations, with and without seasonality, for Weak winter and Melt regimes.

Several key observations can be made: (I) Similar to the overall rank correlation described in Section 3.3.1, both upper and lower tail dependence exhibit persistence across the time series up to lags equal or shorter of 25–35 days for most cases; (II) the tail dependence structure is largely symmetric, with only a slight tendency for the upper tail dependence to be higher than the lower tail dependency; (III) for both seasonal and DS series, the tail dependence coefficients after logarithmic transformation are generally higher, or equal to those of untransformed residuals, while after Box-Cox transformation in several cases tail correlation is instead reduced; (IV) the effect of deseasonalization is more pronounced for smaller lags, but not so much for larger ones.

These findings, along with those reported in Section 3.3.1, underscore how residual correlation structures are not uniform across the entire range of values. Consequently, capturing the full behavior of residual autocorrelation requires approaches that extend beyond single-coefficient summaries (e.g., individual correlation coefficients). One possibility is modelling autocorrelation through bivariate distributions, such as copula models. Figure 26 shows scatterplots analogous to those in Figure 23, focusing on lag-1 autocorrelation, but for pseudo-observations, as used in copula analyses. Pseudo-observations are rank-based transformations of data, typically derived from empirical cumulative distribution functions (ECDFs), that map original values to a (0,1) scale, enabling non-parametric dependence modelling.

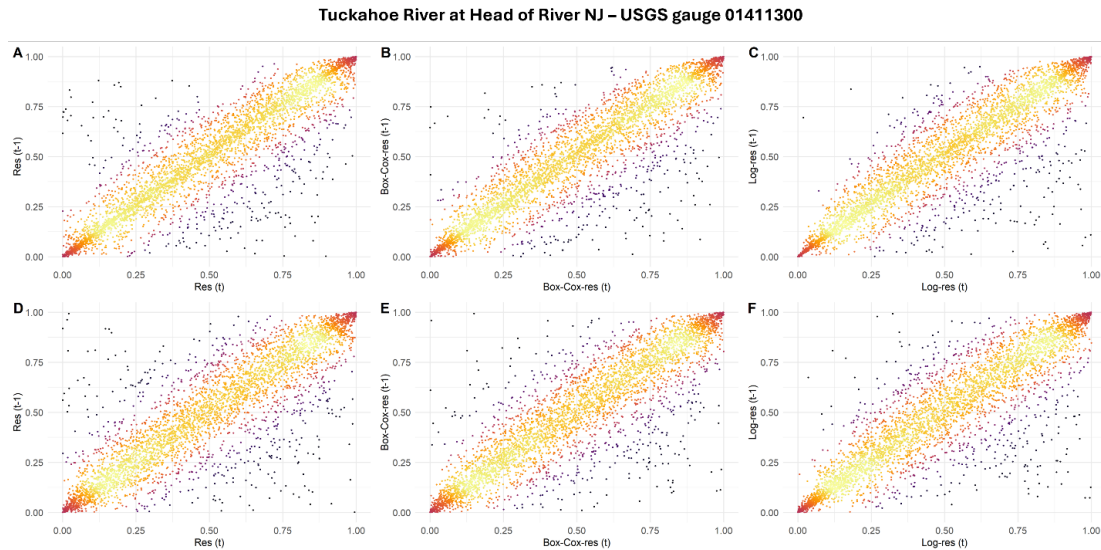


Figure 26. Scatterplots (t vs $t-1$) for pseudo untransformed residuals (A), pseudo Box-Cox ($\lambda=0.5$) residuals (B), pseudo logarithmic residuals (C) (USGS gauge 01411300 - model $n^{\circ}1$). The corresponding deseasonalized residuals are shown in panels D, E, and F. Color scale indicating points density.

A key observation is that the pronounced differences evident in Figure 23 are no longer present in Figure 26. Instead, the residuals exhibit a relatively consistent structure, regardless of whether they were transformed or not. However, the removal of seasonality exerts a clear influence: while values previously clustered along the 1-to-1 line, they become more dispersed after deseasonalization.

Using the VineCopula package in R (Nagler et al., 2024), copula models are fitted to catchment-model combinations for DS cases (not shown here), without imposing predefined structures, allowing the algorithm to select the best-fitting model. The t-copula emerged as the most suitable model in over 98% of cases, reflecting its capacity to capture the observed symmetry in upper and lower tail dependencies. However, goodness-of-fit (GOF) tests revealed that the models are statistically non-significant in most instances. While the t-copula captures the general structure of the data, it falls short of fully representing its complexity. Significant deviations persist, notably instances where low ranks are followed by high ranks (or vice versa) across successive time steps.

3.4 What can we learn: from properties to error models

3.4.1 Difference between regimes and general effect of transformations

The previous sections presented a comprehensive analysis of the statistical properties of conceptual RR models residuals, with the goal of improving the understanding of residuals structures and their implications for uncertainty estimation.

Regarding the marginal distribution properties (Section 3.2.2), the introduction of transformations and normalization clearly shows beneficial effects in reducing tail heaviness. At the same time, some regime-specific differences emerge, particularly for the log transformation (which displayed L-kurtosis values closer to those expected for a normal distribution). In this case, a shift toward negative L-skewness values is observed; however, a closer inspection reveals that this effect is confined to the Weak Winter and New Year's regimes. After deseasonalization, catchments' residuals in the New Year's regime also approximate a symmetric distribution, while those belonging to the Weak Winter regime still tend to display negative asymmetry values. In terms of variability across models (Section 3.2.3), no marked differences among hydrological regimes are evident. Nonetheless, deseasonalization effectively reduces the variability introduced by model selection.

The heteroskedastic structure (noting that in Section 3.2.4 only its shape was considered) is highly consistent across regimes, showing a nonlinear increase with simulated discharge quantiles, although differences in the absolute values of standard deviation remain possible across regimes. This difference may be particularly relevant coming back to the GAR. In agreement with previous findings in the literature, the considered transformations adequately address heteroskedasticity, effectively approximating a homoscedastic structure, with the logarithmic transformation providing the best approximation. A previous study (McInerney et al., 2017) suggested that the Box–Cox transformation could improve upon the logarithmic one; however, in that case the authors proposed as Box-Cox parameter λ values between 0 and 0.2, which are closer to the logarithmic transformation than the $\lambda = 0.5$ adopted in this study. Overall, reducing λ progressively moves the residuals toward a homoscedastic structure, suggesting that in specific cases a sensitivity analysis on the Box–Cox parameter could help optimize the transformation. Deseasonalization in this context reduces overall variability but does not produce noticeable differences across hydrological regimes.

Temporal autocorrelation (Section 3.3) is present in the residuals regardless of the transformation applied. For the non-normalized residuals, particularly at lower lags,

some differences across hydrological regimes are evident (especially for Pearson correlation): catchments in the Intermittent regime exhibit the largest variability among all regimes, with lag-5 correlations ranging from near 0 to values above 0.6 (Figure 24A). Conversely, catchments belonging to the Melt regime, again most clearly at lower lags, display a systematic tendency toward higher autocorrelation values than any other regime. Once again this difference could prove relevant when concentrating on the GAR region. The use of a rank based correlation coefficient (Spearman) highlighted a noticeable increase in temporal correlation for untransformed residuals, while for both Box-Cox and logarithmic transformations, as well as for deseasonalized residuals, the difference between the two metrics is less relevant. Transformations tend to increase overall autocorrelation, while deseasonalization has only a limited effect. The same trends are noticeable also for upper and lower temporal tail correlations, where again the Melt regime (Figure 25B and Figure 25D) display overall higher correlation than the other regimes.

3.4.2 General insights for modelling

Being able to generate realistic ensembles of residuals would fundamentally transform hydrological modelling: instead of a single deterministic trajectory, each model run could be expanded into a full ensemble, providing a direct and quantitative basis for uncertainty assessment and scenario exploration. Yet designing such stochastic models is far from trivial, because residuals do not behave like simple white noise. They carry signatures of seasonality, strong autocorrelation, cross-dependencies (in this study the investigation was limited to the heteroskedastic structure), and non-Gaussian distributional shapes, features that are not constant but may vary across hydrological regimes. Moreover, the deviations in scatterplots in Figure 26, suggest that the residuals may result from at least two overlapping processes: one that governs the primary structure and another introducing anomalies. This study makes a step forward by systematically mapping these properties across a uniquely large multi-model and multi-catchment dataset, allowing us to distinguish common structures from regime-specific differences. These insights are critical: they guide which aspects of residual behavior must be captured explicitly, and which can be simplified without major loss of realism. They also reveal the limits of one-size-fits-all approaches, highlighting the need for general yet flexible stochastic frameworks, such as CoSMoS (Papalexiou, 2018), that can accommodate seasonality, variance stabilization, temporal dependence, and tail behavior in a unified way.

3.4.3 Note about deseasonalization

Accounting for the persistence of seasonality in residual time series clearly improves the stabilization and generalization of residual properties. In the present study, a direct transformation approach was proposed, but alternative strategies could also be considered. Examples include calibrating RR model parameters at monthly or seasonal scales to reduce seasonality in residuals at the source, or developing error models with seasonal-varying parameters. Regardless of whether a direct or indirect approach is adopted, an important issue remains: when attempting to explicitly account for seasonality, the available record length must effectively be reduced. For instance, estimating twelve distinct parameter sets, one for each month, requires partitioning the observed series into twelve sub-samples, each with only one-twelfth of the original data, which are then used for calibration. This raises a key question: how many years of data are required for a robust estimation? In this study, a record length of ten years was considered for the validation period. Whether this is sufficient for reliable parameter estimation is questionable. In general, this situation poses a dilemma: while seasonal approaches offer clear benefits, their applicability is constrained by data availability. There is no straightforward solution, apart from emphasizing the critical importance of long-term data collection and advocating for sustained efforts to maintain and expand discharge monitoring networks, which historically remain less extensive than those for other key hydrological variables such as precipitation or temperature.

3.4.4 Note about intermittent catchments

A general remark concerns intermittent catchments. As noted in Section 3.1.3, although the original dataset included temporary rivers, these were excluded to ensure consistency in the analysis. Arid catchments are already notoriously challenging to model, and the presence of intermittency adds an additional layer of complexity. Consequently, their statistical properties, including those of residuals, are expected to be harder to characterize and more variable, as is already partially evident from the results of the remaining catchments classified within the Intermittent regime (see, for instance, Figure 24A and Figure 24B in Section 3.3.1). A more extensive analysis would therefore be required to provide a comprehensive description. Furthermore, when developing an error model, intermittent catchments impose an additional constraint: the need to correctly reproduce zero-flow statistics. Addressing this requirement likely demands a mixed-type modelling strategy (similar to those proposed

for stochastic rainfall generation), combining one component to represent non-zero flows with another designed to capture the observed degree and temporal structure of intermittency.

3.4.5 Limits in our approach

The dataset used in this study is significantly larger than any previously employed for systematic residual analysis. While it provides extensive coverage of climatic and morphological catchment variability and includes a broad set of models, some caveats are necessary. Clearly, 78 models cannot encompass the full range of possible model architectures and parameterizations; nevertheless, this number was sufficient to highlight key features of model-induced variability in residual properties and to evaluate the effectiveness of the proposed transformations in reducing such variability. Moreover, these models are built from four existing models that are already conceptually quite different (PRMS, Sacramento, VIC, TOPMODEL). The simplifying equations used to represent the processes in these four base models are distinct, and their structural design is as well. The framework can then be considered quite general. Additionally, previous studies, though limited to fewer catchments and narrower sets of properties, have generally employed models outside the FUSE framework. Their results, however, are consistent with the broader findings reported here, at least for the properties shared between the studies.

Another layer of variability not explicitly addressed in this study concerns the role of the model efficiency (ME) function used during calibration. Here, parameters were tuned exclusively with the Kling–Gupta Efficiency (KGE), while other works have also employed different metrics, most notably the Nash–Sutcliffe Efficiency (NSE). However they reported residual properties similar to those observed here. A notable example is Hunter et al., (2021), who specifically investigated the influence of ME selection, although with a focus more on evaluating error-model performance than on a detailed characterization of residual properties.

Finally, an important research gap relates to the scope of the present analysis. This study focused on a purely statistical characterization of aggregated residuals and assessed model selection effects only through an ensemble perspective, considering model outputs without examining in detail how specific structural features might drive residual properties. A more process-oriented approach, directly linking model structure and parameterization to residual behavior, could provide valuable insights, particularly for guiding architectural choices toward models that generate better-behaved residuals.

However, such an endeavor would require substantial additional effort beyond the scope of a single study. Future research could effectively address both the influence of calibration metrics and the process-based connections between model structure and residual characteristics. Moreover, the methods described in the following chapters will be operative, and will just look at model residuals without embedding in the error models consideration about the connection with specific hydrological processes.

The following two chapters will built upon the discussion here presented: Chapter 4 will present an attempt at the development of a simple error model based on residual analysis, together with a comparison with already available methods in the literature, while Chapter 5, will focus on the estimation of confidence intervals following an alternative approach not involving a direct description of model residuals instead. After those, Chapter 6 will shortly explore potential perspective for the regionalization of error models schemes for the estimation of confidence intervals also in ungauged locations.

Chapter 4

4. Post-processing methods

4.1 Deciding a direction

4.1.1 Methods and operativity

From this point onward, the focus shifts to the estimation of uncertainty in conceptual RR models and to the derivation of confidence bounds for model predictions. To clearly distinguish between the original deterministic model outputs and the uncertainty estimates obtained after the application of error modelling schemes, the following nomenclature will be adopted throughout the remainder of this thesis. The term D-model predictions will be used to indicate any deterministic simulation produced by a RR model. These simulations are deterministic in the sense that, once the model structure and parameters are fixed, and given identical input data, the model always produces the same output time series, without explicitly accounting for uncertainty or variability. In contrast, the term S-model predictions will refer to the confidence bounds (or confidence intervals, or envelope) and all related statistics obtained after applying an error model to the deterministic simulations, thus providing a probabilistic characterization of model outputs. The analysis presented in the following chapters is restricted to post-processing (PP) approaches that aggregate all sources of uncertainty into a single error term represented by the model residuals. In these approaches, individual uncertainty contributions, such as parameter uncertainty, input uncertainty, or structural model uncertainty, are not explicitly separated. This choice is motivated by the intention to focus on operationally simple and practically applicable methods rather than on more complex and diagnostically oriented frameworks. From an applied perspective, what ultimately matters is the magnitude and reliability of the total predictive uncertainty, rather than the precise attribution of uncertainty to individual sources. Within this family of post-processing approaches, the available methods can be further subdivided (this is not the only possible

classification) into two conceptually distinct groups. The first group (I) includes methods whose sole objective is the estimation of predictive confidence intervals, thus performing the post-processing transformation from a D-model to an S-model. The second group (II) comprises methods that, in addition to estimating confidence bounds, also aim to generate individual stochastic time series with statistical properties consistent with those of the observed series. A simple schematic representation of these two methodological families is provided in Figure 27.

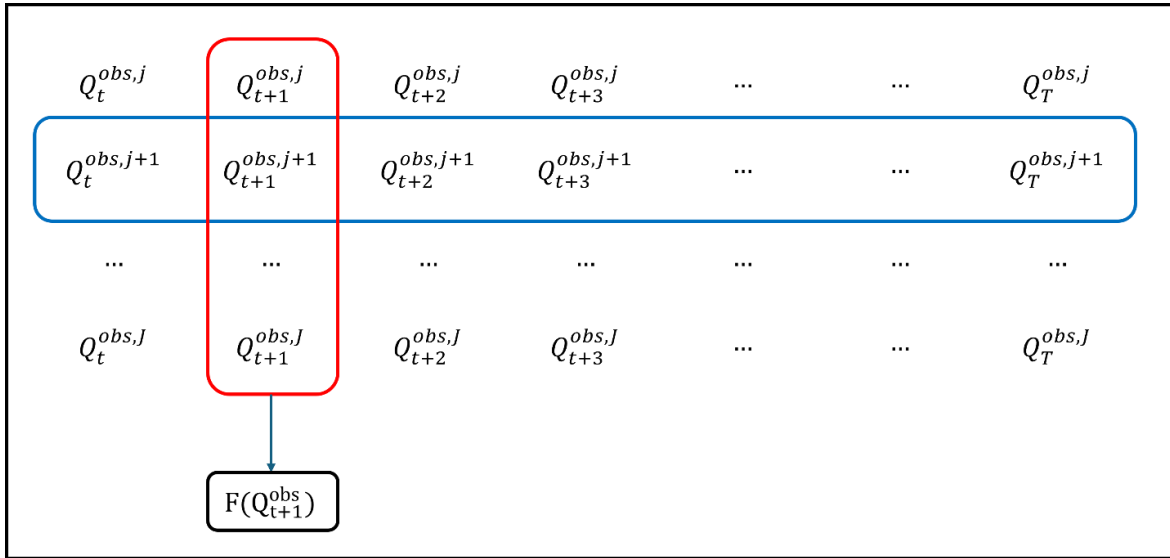


Figure 27. Schematic representation of the j^{th} stochastic discharge series Q (in blue) and of the cumulative density function (CDF) for a D-model prediction at time $t+1$ (in red).

The shared goal of both groups is, for any given time step t , to correctly estimate the predictive distribution $F(Q_t^{obs})$ (represented by the red box in Figure 27). However, only methods belonging to the second group explicitly seek to reproduce the statistical properties of individual stochastic realizations (represented by the blue box in Figure 27). The need for such methods, was highlighted by Vogel (Vogel, 2017), arguing how confidence intervals alone, while interesting and useful, are of little value in approaches like RBDM (Risk-Based Decision Making) because such approaches require the complete set of streamflow ensembles for their implementation, so that the uncertainty intervals alone would not be adequate to apply widely accepted RBDM approaches. From a conceptual standpoint, this second approach is inherently more ambitious and,

consequently, more challenging to implement. By contrast, methods of the first group allow, at least in principle, for simpler modelling schemes, as they do not require an explicit representation of the statistical structure of individual time series. Starting from Section 4.2.1 and for the remainder of Chapter 4, the analysis will initially focus on a modelling approach belonging to the second group.

4.1.2 Parametric vs non-parametric: why are we testing parametric approaches?

Another clarification is necessary. One of the long-term objectives of this thesis is to investigate the transferability of the proposed error modelling framework to ungauged basins through regionalization techniques. In addition, particular emphasis is placed on the high-flow range, which is of primary relevance for many hydrological applications. For these reasons, the error model sought should ideally satisfy a number of additional requirements.

First, it should be parametric, or at least largely parametric. Parametric models offer potential advantages in terms of parsimony, robustness, and suitability for extrapolation beyond the range of observed data, especially when sample sizes are limited. Moreover, their compact representation through a finite set of parameters makes them particularly suitable to regionalization using methodologies that are well established in the hydrological literature, although the feasibility of such transfer must always be assessed on a case-by-case basis. Second, the parametric structures involved should be generalizable across multiple locations, rather than being strongly site-specific. Third, the model should remain relatively simple, as parsimonious formulations, when effective, are generally easier to transfer to ungauged catchments.

At the same time, it is essential to acknowledge that the potential advantages of parametric formulations must be carefully weighed against their ability to correctly represent the underlying data structures. Non-parametric approaches derive their strength precisely from their flexibility and from their ability to reproduce complex dependence patterns observed in the data, including features that may be difficult or impossible to capture using simple parametric forms. Furthermore, non-parametric methods can also be employed within regionalization frameworks, for example by resampling information from hydrologically similar or geographically nearby catchments. As such, the preference for a mostly parametric approach in this thesis should not be interpreted as an a priori assumption of methodological superiority. Rather, the parametric modelling strategy adopted here is motivated by its potential

benefits in terms of transferability and operational applicability, and is therefore treated as a working hypothesis to be rigorously tested. Throughout the following chapters, the proposed parametric methods will be systematically evaluated against existing non-parametric approaches in order to assess whether a tangible improvement in predictive performance and uncertainty characterization is achieved. These considerations will guide several methodological choices made in the subsequent chapters and will be revisited in the final discussion in Chapter 7.

4.2 A new error model

4.2.1 Looking backwards to decide how to move forward

We will start now with the first proposed PP method belonging to the second family previously introduced. The core idea underlying the proposed post-processing error model consists in stochastically reintroducing residual time series into deterministic (D-model) predictions in order to generate predictive uncertainty and confidence bounds. Let Q_t^O denote the observed discharge at time t , and let Q_t^S be the corresponding prediction provided by a calibrated conceptual RR D-model. In this work, model residuals were defined in a transformed space as

$$\epsilon_{\mathcal{T},t} = \mathcal{T}(Q_t^S) - \mathcal{T}(Q_t^O) \quad (\text{Eq. 4.1})$$

where $\mathcal{T}(\cdot)$ denoted a monotonic transformation of discharge. The deterministic model can therefore be equivalently expressed as

$$\mathcal{T}(Q_t^O) = \mathcal{T}(Q_t^S) - \epsilon_{\mathcal{T},t} \quad (\text{Eq. 4.2})$$

The objective of the post-processing (PP) error model is to replace the single deterministic residual series $\{\epsilon_{\mathcal{T},t}\}$ with an ensemble of stochastically generated residual series $\{\epsilon_{\mathcal{T},t}^{(j)}\}$, $j = 1, \dots, J$, whose statistical properties are consistent with those observed in the original residuals. For each ensemble member j , a stochastic discharge trajectory is then obtained as

$$\mathcal{T}(Q_t^{(j)}) = \mathcal{T}(Q_t^S) - \epsilon_{\mathcal{T},t}^{(j)} \quad (\text{Eq. 4.3})$$

which, after inverse transformation, yields

$$Q_t^{(j)} = \mathcal{T}^{-1}(\mathcal{T}(Q_t^S) - \epsilon_{\mathcal{T},t}^{(j)}) \quad (\text{Eq. 4.4})$$

The collection of stochastic realizations $\{Q_t^{(j)}\}_{j=1}^J$ defines, for each time step t , an empirical predictive distribution $F(Q_t)$, from which confidence intervals and other uncertainty metrics can be directly derived.

This general approach is well established in the hydrological literature and has been applied in several previous studies (e.g., Montanari & Brath, 2004; Montanari & Grossi, 2008; Shabestanipour et al., 2023; Sikorska et al., 2015). The main difference between the present work and earlier contributions lies in the way this stochastic reintroduction is implemented. Specifically, the proposed methodology aims to describe the relevant statistical structures using simple and largely parametric formulations, whereas existing approaches often rely either on non-parametric resampling techniques (e.g., bootstrap-based methods) or on parametrizations of relatively high complexity.

The design of the proposed error model is directly informed by the findings presented in Chapter 3. Those results clearly indicate that working with untransformed D-model residuals is not advisable. Raw residuals are characterized by strong autocorrelation, pronounced heteroskedasticity, and heavy-tailed marginal distributions. Attempting to model all these features simultaneously would inevitably lead to complex model structures and would make a fully parametric representation particularly challenging. In contrast, working in a transformed space allows, at least conceptually, for a number of simplifying assumptions. In particular, residuals computed after log-transformation of both simulated and observed discharge exhibit, on average, a variance structure that can be reasonably approximated as homoscedastic. This assumption alone is not sufficient to completely disregard the dependence structure between D-model predictions and residuals, as homoscedastic residuals may still exhibit conditional biases. Moreover the average tendency to homoscedasticity cannot be considered completely general, as individual cases may still display residual conditional variance. However, the rationale behind adopting this initial simplification will be discussed and justified later in Section 4.3.1. Under this working hypothesis, the properties that must be explicitly reproduced by the stochastically generated residuals reduce to two key elements: (I) their temporal dependence structure (i.e., temporal autocorrelation), and (II) their marginal distribution, which is considerably less heavy-tailed than in the untransformed case.

In addition to the logarithmic transformation, deseasonalization (i.e., normalization) was shown to further reduce departures from normality and mitigate residual heteroskedasticity. Some considerations regarding the robustness of the seasonal signal estimation were already discussed in Section 3.4.3. To enhance stability and reduce sensitivity to sampling variability, a monthly, rather than daily, deseasonalization scheme has been adopted here. The same transformation described in Eq. 3.5 is applied, but aggregating at the monthly scale instead that at the daily one.

In the interest of developing a model structure that is as general and transferable as possible, particular attention is given to another result from Chapter 3. Specifically, the dependence structure between residuals at consecutive time steps (ϵ_t vs ϵ_{t-1}) was shown to be remarkably consistent when analyzed in the unit square defined by their empirical non-exceedance probabilities (see Figure 26). Based on this observation, the proposed methodology follows a relatively simple conceptual scheme. Let ϵ_t denote the residuals computed after log-transformation of simulated and observed discharge, and after deseasonalization. These residuals are first mapped to the unit interval through their marginal cumulative distribution function (CDF), yielding the transformed variable

$$u_t = F_\epsilon(\epsilon_t) \quad (\text{Eq. 4.5})$$

where F_ϵ denotes the marginal CDF of the residuals. The dependence structure between u_{t-1} and u_t is then described using a copula model. Once the copula structure has been selected and its parameters inferred for each individual site, stochastic residual time series can be generated iteratively in each catchment. The selection of the model also followed from the general results in Chapter 3.3.2, where a t-copula was found to be the most suited copula to reproduce the structure in most of the cases. Starting from an initial ϵ_{t-1} , the corresponding ϵ_t is sampled conditionally from the fitted copula. The newly generated ϵ_t is then used as conditioning information to generate the residual at time ϵ_{t+1} , and so forth. Repeating this procedure yields a single stochastic residual realization; the process is repeated j times to obtain multiple residual series as illustrated in Figure 27. The starting residual values, under the assumption of independence previously introduced, can be sampled at each iteration randomly from the observed residuals. A schematic representation of the process is reported in Figure 28.

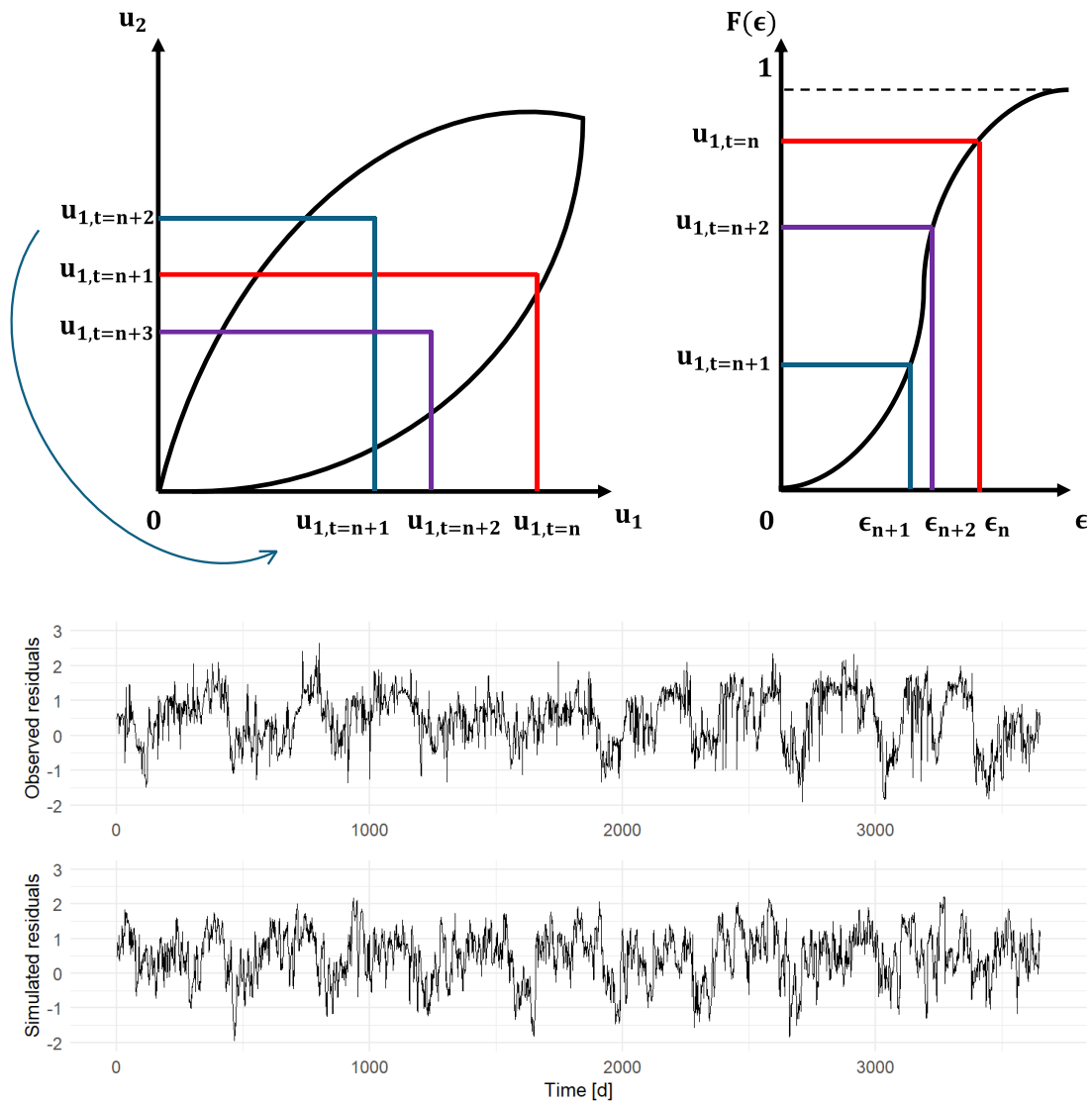


Figure 28. Schematic representation of the C-method. On the top-left, iterative sampling from the fitted copula structure (density schematically represented in black) in the unitary space (u_1 : residual at time $t-1$, u_2 : residual at time t). On the top-right, conversion from unitary to residual space through the residual cumulative distribution function $F(\epsilon)$. On the bottom: observed residual series (after logarithmic transformation, with seasonality) and an example of generated series (after seasonal component reintroduction) for the Cervo a Passobreve catchment for the calibration period.

In order to obtain a fully parametric error modelling framework, the only additional component that should be specified alongside the copula is the marginal distribution of the residuals, represented by their CDF $F_\epsilon(\epsilon)$, which is required to map variables to and back from the unit interval to the residual space. In the initial implementation presented here, this marginal distribution is characterized empirically. The motivation for this choice is pragmatic: before addressing the more challenging task of developing a suitable parametric model for the marginal distribution of log-transformed DS residuals, which, despite exhibiting L-moments closer to those of a normal distribution (

Figure 29C), are not always strictly Gaussian, it is useful to first assess whether the proposed dependence structure is able to adequately reproduce the observed autocorrelation patterns. During the validation period, stationarity is assumed, and the same copula and empirical marginal distribution are used to apply the generation procedure.

It should be noted that by modelling the copula between ϵ_{t-1} and ϵ_t , the proposed approach explicitly constrains only the lag-1 dependence structure. This is a simplification, as previous analyses have shown that residual correlation, particularly in log-transformed space and in catchments dominated by snow accumulation and melt processes, persist over longer time lags. The present formulation should therefore be interpreted as a first step toward a more comprehensive representation, aimed at guiding future methodological developments. Moreover, this initial choice is motivated by a result from the method proposed by Sikorska et. al (Sikorska et al., 2015) later described in detail in Section 4.2.4. Finally, it is important to recall that the proposed scheme does not explicitly account for the dependence structure between D-model predictions and residuals. This limitation is a direct consequence of the simplifying assumptions introduced earlier.

Conceptually, the approach bears some resemblance to an autoregressive model of order p , $AR(p)$, in that the temporal evolution of the residual process is governed by its dependence on past values. However, in a classical autoregressive framework, ϵ_t are expressed as a linear combination of previous residuals, and stochasticity is introduced through an additive innovation term, typically assumed to be independent and identically distributed. In this setting, dependence is encoded exclusively in the conditional mean, while higher-order features such as tail dependence, asymmetry, or state-dependent variability are not explicitly represented. In contrast, the present framework adopts a copula-based formulation in which ϵ_t is directly conditioned on a

single previous residual through a flexible dependence structure defined at the level of the joint distribution. Stochastic variability is not injected via an external innovation term, but instead emerges endogenously from the copula, which governs the full conditional distribution of the residual process. As a result, the model naturally accommodates non-linear dependence, asymmetric behavior, and tail dependence, allowing the propagation of extremes in a manner that cannot be captured by standard autoregressive models. From this perspective, the proposed approach can be viewed as a distributional generalization of an AR(1) process, where dependence is specified in terms of the entire conditional distribution rather than being restricted to the conditional mean.

After the generation of the j residual series (logarithmic, deseasonalized), the constant monthly seasonal signal is added back to each realization, and the original logarithmic transformation is inverted to yield the individual stochastic discharge series. It should be noted how during this retransformation, retransformation biases could be introduced, and a bias correction factor (BCF) should be introduced to account for them (see Section 4.3.5).

4.2.2 General notions about copulas

A copula is a mathematical function that allows the construction of a multivariate joint distribution by coupling univariate marginal distributions with a dependence structure that is independent of the marginals themselves (Nelsen, 2006). They have been used in several field of hydrology in the past with great success (see Genest & Favre, 2007). Formally, according to Sklar's theorem (Sklar, 1959), for any pair of continuous random variables (X, Y) with joint cumulative distribution function $F_{X,Y}(X, Y)$ and marginal distributions $F_X(X)$ and $F_Y(Y)$, there exists a unique copula $C: [0,1]^2 \rightarrow [0,1]$ such that $F_{X,Y}(X, Y) = C(F_X(X), F_Y(Y))$. The copula fully characterizes the dependence between the variables, while the marginals describe their individual behavior. Estimation of a copula from data typically proceeds in two steps. First, the marginal distributions F_X and F_Y are estimated, either parametrically (by fitting a theoretical distribution) or non-parametrically (e.g., using empirical distribution functions). The observed pairs (X_τ, Y_τ) are then transformed into pseudo-observations (u_τ, v_τ) , where $u_\tau = F_X(X_\tau)$ and $v_\tau = F_Y(Y_\tau)$, which lie in the unit square $[0,1]^2$. In a second step, the copula is estimated by fitting a parametric copula family (such as Gaussian, Student-t, Archimedean copulas) or by adopting a non-parametric estimator, using likelihood-based or rank-based methods. Once the copula is estimated,

the joint distribution $F_{X,Y}$ and the corresponding conditional distribution $F_{X|Y}(X|Y)$ can be derived analytically, providing a probabilistic mapping between deterministic model outputs and observations.

4.2.3 Study Area and RR-model

Here we briefly introduce the study area used to test the proposed PP method. Instead of considering the entire GAR, the analysis is restricted to a sub-region of it. This choice was motivated by the need to evaluate the proposed methods under controlled conditions: by focusing on a smaller and more homogeneous area, we reduce the influence of large-scale heterogeneity and avoid unnecessary complexity at this stage of method development. We consider 24 catchments located in the Piedmont region of north-western Italy (see

Figure 29A). The discharge observation data used in calibration (2010-2019) for each catchments are the same already used in Chapter 2. The main difference lies in the input data: in this case, the input and precipitation data are used from the Optima Interpolation (OI), a local high quality database more reliable than the ERA5LAND data used previously. Rather than calibrating a RR model independently for each catchment, a regional modelling strategy was again adopted similarly to Section 2.3.4. Starting from site-specific calibrations, a regionally consistent model parameterization was subsequently derived at each site using again the PASS methodology. The resulting D-models at each site can be regarded as standard RR models; for the purposes of post-processing and confidence-bound estimation, only the simulated discharge time series are required, and the specific model structure leading to the D-model is of secondary importance. The RR model employed is the same modified version of the TUWmodel already previously presented in detail in Section 2.3.1. Further details on model construction are here omitted, as they are not the focus now, and would be redundant with what has already been presented previously. The performance of the D-models, quantified using the Kling–Gupta Efficiency (KGE) are summarized in

Figure 29B for both the calibration and validation (2000-2009) periods, yielding median KGE values of 0.70 and 0.62, respectively. The varying level of performance across catchments, rather than representing a weakness of the test set, provides an opportunity to assess whether the proposed error-model schemes remains robust under suboptimal, yet still reasonable, parameterizations. In

Figure 29C and

Figure 29D are also reported L-moment diagrams and heteroskedasticity plots analogous to the ones presented in Section 3.2.2 for the current catchments.

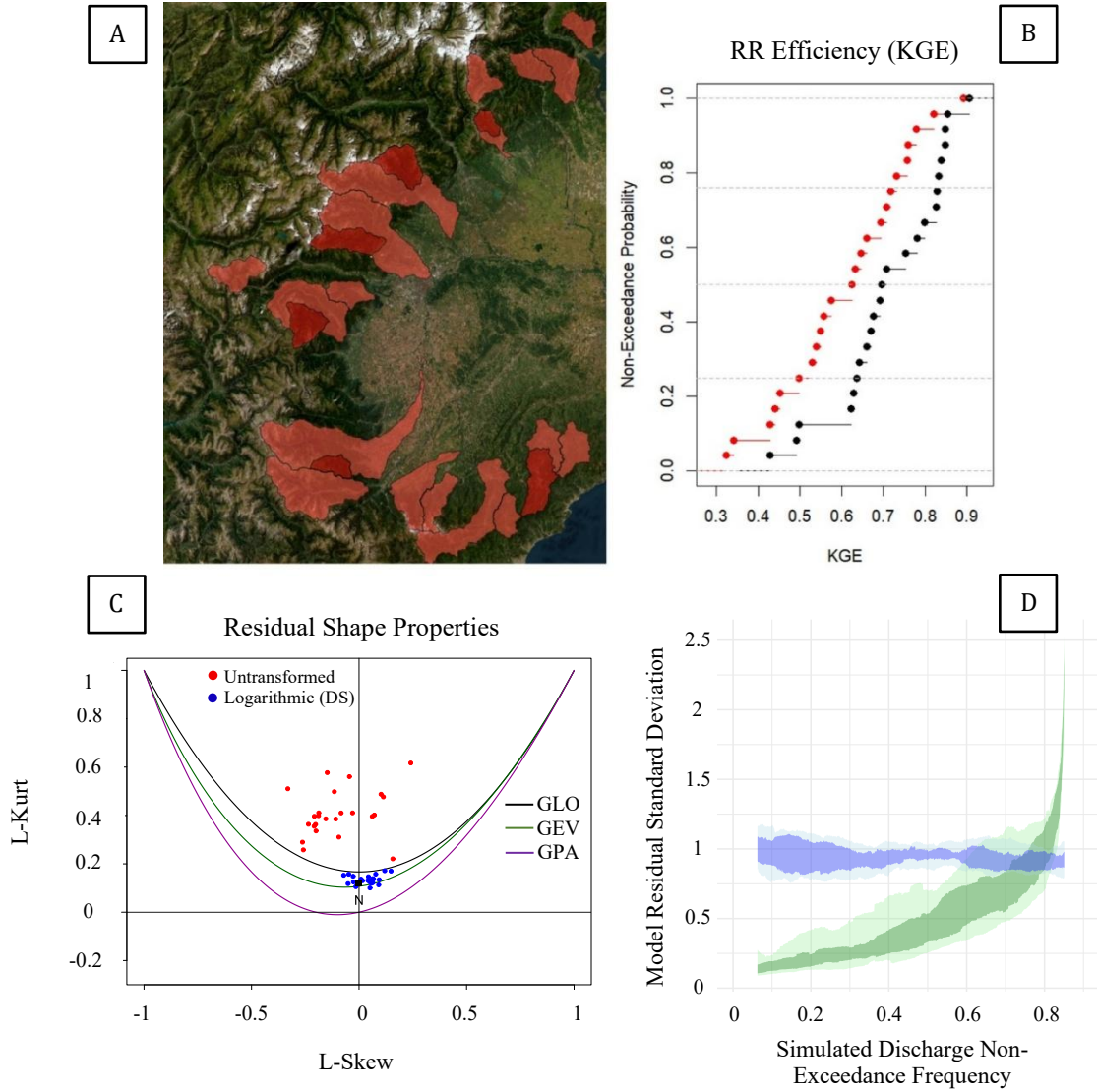


Figure 29. (A) 24 mountainous catchments used in the study. Darker shades indicate nested catchments; (B) KGE empirical distribution for the calibration (2010-2019, in black) and the validation (2000-2009, in red) periods; (C) L-moment diagram for untransformed residual and residuals after logarithmic transformation and monthly deseasonalization; (D) Model residual standard deviation (untransformed and after logarithmic transformation and deseasonalization) as function of simulated discharge.

4.2.4 Benchmark non-parametric methods

Within the class of post-processing (PP) methods that stochastically generate residuals, two approaches from the literature are considered here as benchmarks for comparison with the newly proposed method. Although both methods rely on resampling techniques, they differ substantially in their underlying philosophy and in the way stochastic residual time series are generated. The first benchmark method is the one proposed by Shabestanipour et al. (Shabestanipour et al., 2023). In their approach, residuals from the D-model are first computed as $\epsilon_t = \log(Q_t^S) - \log(Q_t^O)$, where Q_t^S and Q_t^O denote simulated and observed discharge, respectively. An autoregressive model of order p , $AR(p)$, is then fitted to the transformed residual series ϵ_t , yielding innovations ε_t . This first part of their method is indeed parametric.

To account for possible residual heteroskedasticity in the innovations, a K -nearest neighbors (K -NN) bootstrap is applied to the AR residuals, generating at each time step t multiple stochastic realizations $\tilde{\epsilon}_t$. The rationale behind this procedure is to first remove serial correlation from the residuals, since bootstrap techniques are known to break in the presence of temporal dependence. Once resampled, the stochastic innovations $\tilde{\epsilon}_t$ are reintroduced into the AR model, producing a stochastic realization of the transformed residual series $\tilde{\epsilon}_t$. Finally, stochastic discharge realizations $\tilde{Q}_t^O$ are obtained by inverting the initial logarithmic transformation. In the application presented by Shabestanipour et al., an $AR(3)$ model was adopted; the same autoregressive order is used here, following a preliminary verification based on ACF and PACF analyses confirming that the resulting innovations ε_t are temporally uncorrelated at all sites. For the K -NN bootstrap, a value of $k = 100$ was instead selected, following a preliminary sensitivity analysis, because returning overall the best results across all sites.

Two additional aspects of the original framework should be noted. First, Shabestanipour et al. explicitly account for the uncertainty in the AR model parameters; second, they also apply a bias correction factor (BCF) to account for retransformation bias from the logarithmic space, computed under the assumption that the transformed residuals ϵ_t follow a normal distribution. Both the application and the omission of the BCF are investigated and subsequently compared in Section 4.3.5. A schematic representation of the complete original framework is provided in Figure 30. This method was originally designed for the sole application in a calibration period (to then be used for long-term water management planning), and not for an application to an

independent validation period. Therefore, it will not be applied during the validation phase in the comparative analysis, but for the calibration period only.

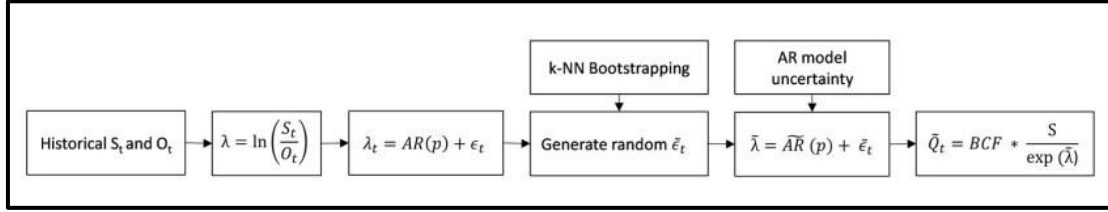


Figure 30. Schematic representation of the framework proposed by Shabestanipour et al. Figure taken from the original paper (Shabestanipour et al., 2023).

The second benchmark PP method considered is the resampling-based approach proposed by Sikorska et al. (Sikorska et al., 2015). In this case, residuals are defined without any transformation as $\epsilon_t = Q_t^S - Q_t^O$. Unlike the AR-based approach, no explicit model is used to describe the temporal dependence of the residual series. Instead, a K-NN bootstrap (conditioned on D-model predictions) is directly applied at each time step t to resample multiple alternative residual values $\tilde{\epsilon}_t$. These resampled residuals are then added to the corresponding deterministic simulations Q_t^S , yielding multiple stochastic discharge realizations $\tilde{Q}_t^O$ at each time step. The parameter k has to be decided: in the current application a value of $k=100$ has been selected after some sensitivity analysis tests. In general, in agreement with the results reported by Sikorska, the selection of k introduces small variations in the results, with noticeable differences (degradation) noticeable only when k is selected too small or too large compared to the total amount of available data points (similarly to the selection of k for the AR-method). In validation, assuming stationary, the procedure is applied identically.

A key distinction between this method and the AR-based approach lies in the way stochastic trajectories are generated. In the AR method, each stochastic realization corresponds to a full residual time series evolving from time t to T , explicitly preserving temporal dependence through the autoregressive structure. In contrast, this second resampling-based method generates, at each individual time step, an ensemble of independent residual values, without explicitly modelling their temporal evolution. As a result, the statistical properties of full stochastic trajectories from time t to T are obtained indirectly, through the aggregation of pointwise resampling outcomes, rather than being modelled explicitly.

At first glance, it may appear counterintuitive to use this second method as a benchmark for approaches aimed at reproducing the temporal autocorrelation structure of residuals, since no explicit mechanism is introduced to preserve autocorrelation. However, the authors reported that their approach was nevertheless able, somewhat surprisingly, to reproduce the observed residual autocorrelation structure. They attributed this result to the fact that since residual statistics depend on Q_t^S , temporal dependence is implicitly transferred to the residuals, thereby emulating the statistical behaviour of the observed errors. This result, also justified the initial assumption to develop the copula based approach considering only the dependence between ϵ_t and ϵ_{t-1} : in a first attempt to keep the model as simple as possible, we wanted to test if also in our case, by constraining only the lag-1 autocorrelation, we could be able to indirectly reproduce in the series also longer lags correlation. It should be noted that the case study presented by Sikorska et al. was based on hourly synthetic discharge data, whereas the present application considers daily time series. Although the calibration period spans a comparable duration (approximately ten years), the hourly resolution results in a substantially larger number of data points. The transferability of their findings to the daily case is therefore explicitly evaluated in the following comparative analysis. For the remainder of this chapter, the three frameworks will be referred to as the C-method (the proposed copula-based approach), the AR-method (autoregressive-based), and the R-method (resampling-based).

4.2.5 Note on the calibration and validation periods selection

A cautionary note is required regarding the selection and use of calibration and validation periods. From an operational perspective, the most appropriate strategy would be to divide the available data into three distinct sub-periods: (I) a first period used to calibrate the D-model; (II) a second period in which the D-model is applied in validation and the resulting residuals are used to calibrate the error model; and (III) a third period devoted to full validation, where both the D-model and the error model are applied outside their respective calibration ranges. The motivation for separating the calibration of the error model from that of the D-model lies in the well-known fact that D-model performance typically degrades when moving from calibration to validation. Consequently, the statistical properties of the residuals obtained during validation may differ from those observed during calibration. Calibrating the error model on a period in which the D-model is already operating in extrapolation allows this effect to be explicitly accounted for.

In the present comparative study, however, the available data are divided into only two periods: a calibration period, used for both the D-model regional calibration and the error models calibration, and a validation period. As a result, the aforementioned effect is not explicitly addressed. This choice is dictated by data availability. Dividing the approximately 20-year-long time series into three sub-periods would lead to an insufficient number of data points in one or more of them, undermining the reliability of both calibration and validation procedures. As a direct consequence, the error models are calibrated under conditions in which residuals are expected to be smaller than those encountered during the subsequent validation period. This implicitly assumes that the statistical properties of the residuals are stationary and independent of changes in D-model performance across sub-periods. Although this assumption is not ideal, and is imposed by data limitations, the same procedure is applied consistently across all catchments and for all post-processing methods. While this leads to sub-optimal calibration and validation in an absolute sense, it ensures that the comparative analysis remains internally consistent and that differences in performance can be attributed to the methods themselves rather than to inconsistencies in the “experimental” setup.

4.2.6 Validation metrics

The evaluation of the performance of the different methods will be assessed by looking at some statistical properties of model residuals as well as at some validation metrics for the estimated S-models. First of all, L-moment diagrams will be used to show how well the marginal properties of model residual are reproduced in the calibration period within the individual stochastic series, as well as ACF plots to show how well the serial correlation in the residual time series is reproduced instead.

To evaluate instead the reliability of the estimated confidence bounds, following standard practice in hydrology (e.g., Montanari & Koutsoyiannis, 2012, and others) we will employ the Predictive Probability Plot (PPP, following the nomenclature of Koutsoyiannis & Montanari, 2022a), as introduced by Laio & Tamea, 2007. This metric evaluates the overall quality of the estimated confidence bands by comparing the theoretical distribution of observed discharge quantiles within the confidence bounds, expected to be uniform under perfect uncertainty estimation, with the corresponding empirical distribution. PPP is in fact a plot of the empirical distribution function $F_z(z)$ of a stochastic variable $\bar{z}$ where the latter also represents probability, that is, a conditional non-exceedance probability, namely:

$$\bar{z}_Q := F_{q|Q}(\bar{q}) \quad (\text{Eq. 4.6})$$

In other words, $\bar{z}$ is the distribution function of the prediction evaluated for the observed value of the predictand. A schematic representation for the interpretation of a PP-plot is reported in Figure 31.

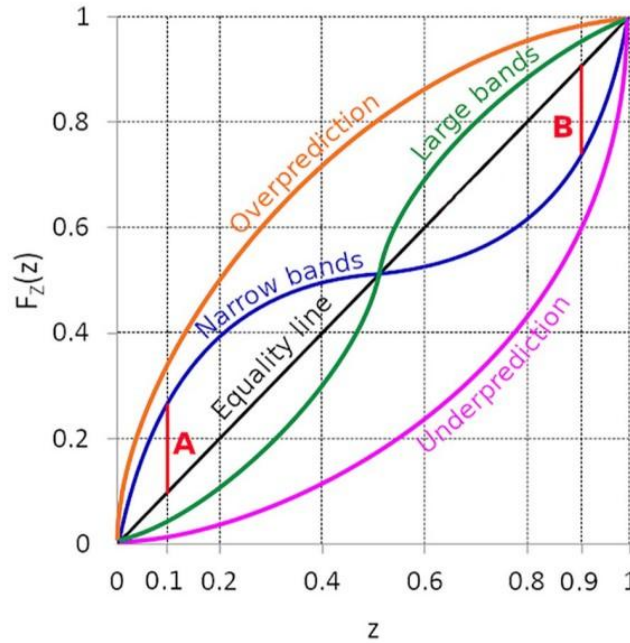


Figure 31. Interpretation scheme for a Predictive Probability Plot (PPP – taken from Koutsoyiannis and Montanari, 2022).

Deviations from the 1:1 line in the PPP indicate departures from the theoretical behavior, such as systematic under- or overestimation, or confidence bands that are excessively wide or narrow. In addition to the global PPP, a separate analysis is conducted by restricting the evaluation to high-flow simulated quantiles ($Q_s > Q_{70}$), in order to verify that the confidence bands adequately represent uncertainty within the flow ranges most relevant for flood estimation. As anticipated before, these two metrics will be shown in calibration for all three methods, while only for the C-method and the R-method for the validation period. It is implicitly implied that during the validation period stationarity is assumed, as discussed previously in Section 4.2.5.

4.3 Comparative results

4.3.1 Is the independence assumption between D-model predictions and transformed residuals reasonable?

A key assumption underlying the proposed simplified parametric C-method framework is that the dependence structure between D-model predictions and the transformed residuals, after logarithmic transformation and deseasonalization, can be reasonably approximated by independence (i.e., by an independent copula). Although the approximately homoscedastic pattern, strongly suggested by

Figure 29D, supports this hypothesis, homoscedasticity alone is not sufficient to guarantee independence. Figure 32 presents illustrative examples of the empirical joint structure during the calibration period between D-model simulations and transformed residuals for selected catchments.

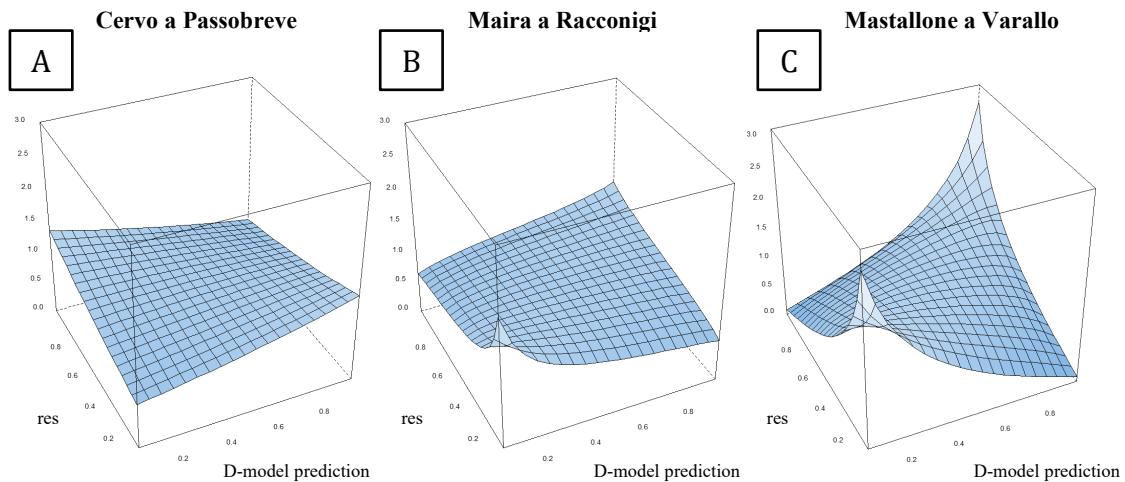


Figure 32. Examples of empirical structures (in the unitary space) in the calibration period between D-model predictions and transformed residuals (logarithmic transformation and monthly deseasonalization) for the (A) Cervo a Passobreve, (B) Maira a Racconigi, and (C) Mastallone a Varallo catchments.

In each case, a copula was fitted automatically via maximum likelihood estimation. For some sites (e.g., Cervo a Passobreve and Maira a Racconigi), the independence assumption appears broadly acceptable. For others (e.g., Mastallone a Varallo), discernible dependence patterns remain evident, presenting residual tail correlation.

Overall, among the 24 analyzed catchments, approximately six exhibit clear residual dependence structures, while in the remaining eighteen the observable deviations from strict independence, mostly concentrated in the tails, are comparatively modest and the simplifying assumption may be considered reasonable as a first-order approximation. These limitations should be kept in mind when interpreting the results presented in the following sections.

4.3.2 Marginal and temporal correlation properties of the stochastically generated residuals

Starting from the marginal distributional properties of the stochastically generated model residuals, Figure 33 reports L-moment diagrams for the calibration period and for the different methods at three representative sites: Cervo a Passobreve (C1), Germanasca a Perrero (C7), and Tanaro a Piantorre (C22). In each diagram, the red point represents the target marginal properties derived from the observed residuals, gray points correspond to the individual stochastic series, and the crossing point of the dashed blue lines indicate the average statistics of the generated series. When interpreting the results, it should be recalled that the R-method uses residuals in their original (untransformed) form, which explains why its empirical shape properties differ markedly from those of the AR-method and the C-method. For the C-method, the statistics are computed on the stochastic series after reintroducing the estimated seasonal component; they are therefore compared with observed residuals after logarithmic transformation, consistently with the AR-method.

Overall, both the R-method and the newly proposed C-method reproduce the target marginal distribution reasonably well, showing strong agreement with the observed L-skewness and L-kurtosis. The AR-method, in contrast, tends to slightly miss both statistics, a pattern that is also observed for the remaining test catchments. These differences can be explained by considering the mechanisms underlying the three approaches. The R-method relies on a direct K-NN resampling of model residuals; consequently, the residuals introduced in the stochastic simulations are directly drawn from the empirical sample, which naturally facilitates the reproduction of the marginal distribution. The C-method also preserves information about the original marginal distribution through the empirical residual CDF used to map between the discharge space and the pseudo-observation space. As long as the generated pseudo-values in a realization are approximately uniformly distributed, the marginal distribution of residuals is correctly reproduced when transforming back to the original space.

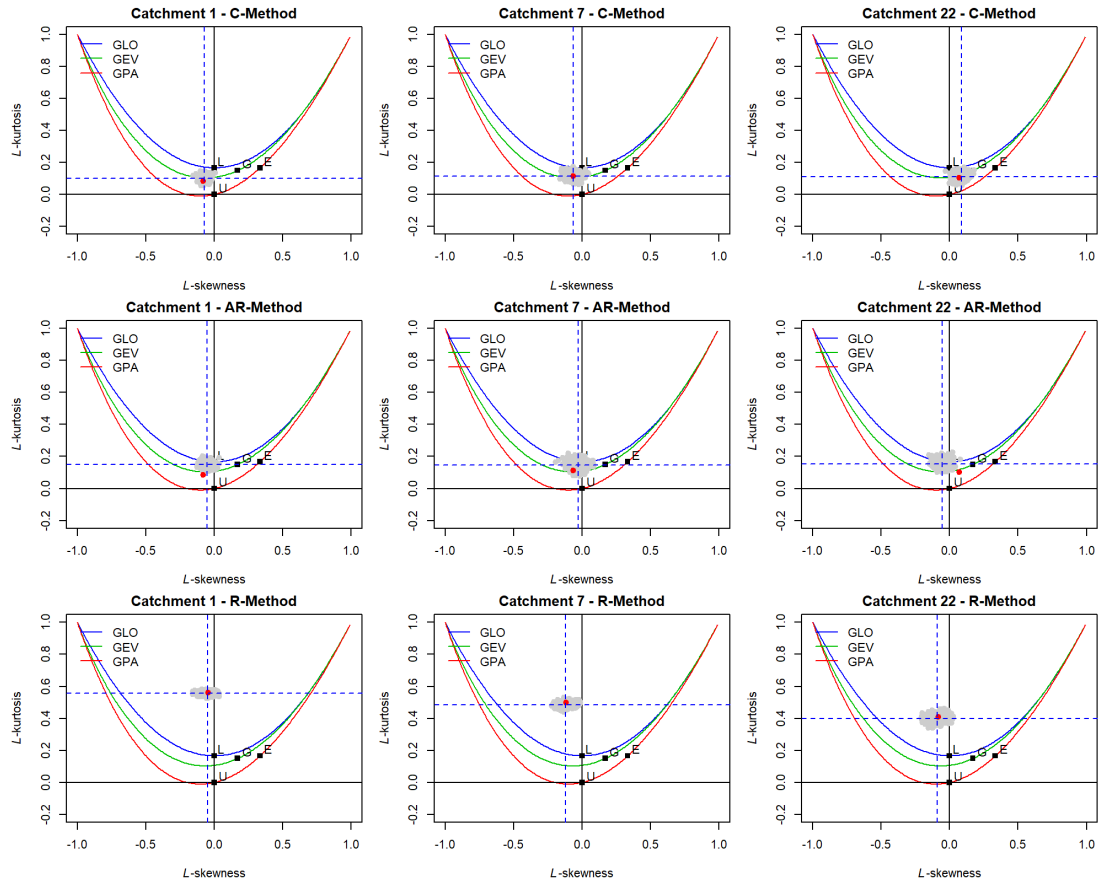


Figure 33. L-diagram plots (residuals - calibration period) for three sites: Cervo a Passobreve (C1), Germanasca a Perrero (C7), and Tanaro a Piantorre (C22). In red target statistics, in gray point for individual stochastic generated series, in dashed blue lines, the average of the stochastic series. Target reference statistics of the observed residual series: untransformed residuals for the R-method; after logarithmic transformation for the C-method and AR-method.

This assumption is consistent with the independence hypothesis underlying the C-method: by assuming independence between D-model predictions and residuals, any value from the original residual distribution can be sampled at time t , constrained only by the residual at the previous time step. The AR-method is the only approach that does not explicitly incorporate information about the empirical marginal distribution of the residuals. Instead, it relies on the autoregressive structure and on the reintroduction of stochastic innovations to reproduce the marginal properties. In practice, however, this

mechanism proves less effective than the other two approaches in reproducing the observed distributional shape.

Turning to the more challenging task of reproducing residual autocorrelation, which represents the main objective of the proposed approaches, the results for the same sites are shown in Figure 34.

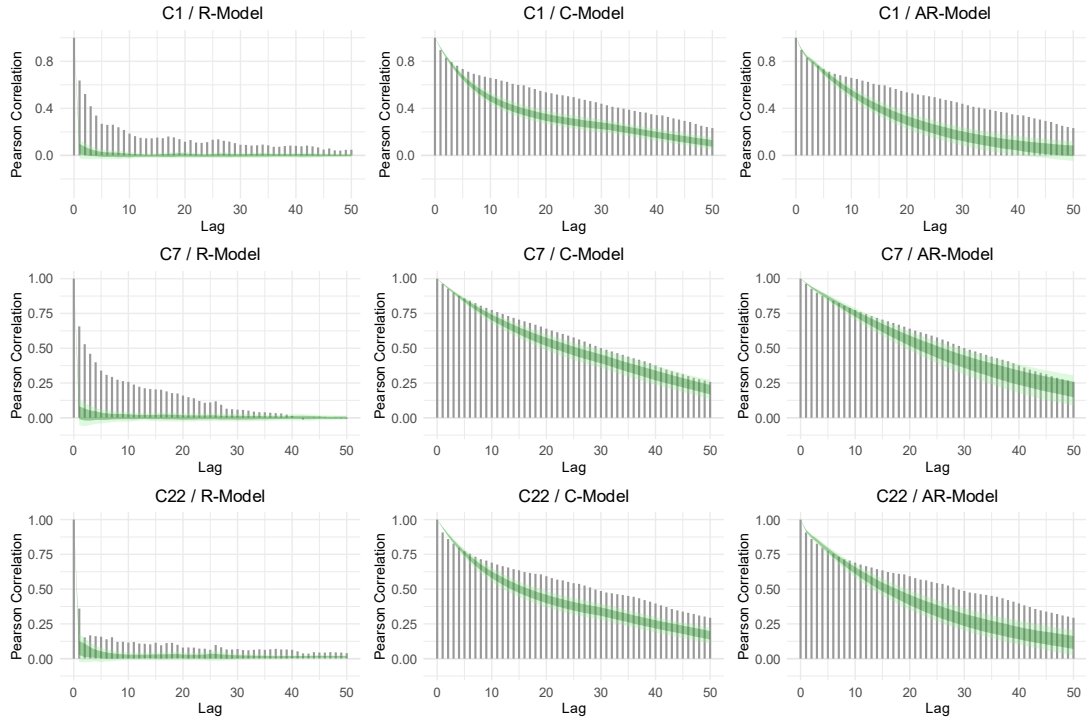


Figure 34. Pearson correlation up to lag 50 for the calibration period for the residuals at Cervo a Passobreve (C1), Germanasca a Perrero (C7), and Tanaro a Piantorre (C22). Darker-shades 50% interval; lighter-shades 80% interval. Target reference statistics of the observed residual series indicated by the gray bars: untransformed residuals for the R-method; after logarithmic transformation for the C-method and AR-method.

Considering the AR-method and the C-method, which explicitly aim to reproduce temporal dependence, the case of Cervo a Passobreve (C1) shows that the autocorrelation function (ACF) is reproduced reasonably well up to about lag-5. However, from lag-10 onward, and increasingly at larger lags, the stochastic series clearly underestimate the observed autocorrelation. This behavior suggests that, for this catchment, neither method can be considered fully successful in reproducing the temporal dependence structure. A similar pattern is observed for the Tanaro a Piantorre

catchment (C22). In contrast, for the Germanasca a Perrero catchment (C7), both methods reproduce the observed autocorrelation much more effectively, capturing the ACF up to lag-50 and beyond, although with a slight overestimation at the smallest lags. More generally, the effectiveness of both approaches in reproducing autocorrelation appears to be strongly site-dependent: the ACF is satisfactorily reproduced in roughly half of the catchments, whereas in the remaining cases the autocorrelation tends to be underestimated at larger lags. It is also noteworthy that, despite the AR-method explicitly constraining residual dependence up to lag-3 (in this implementation) and the C-method only up to lag-1, the two approaches exhibit overall similar performances across the different sites.

The results obtained with the R-method differ markedly. Although the original study by Sikorska et al. reported that autocorrelation was indirectly reproduced, in the present analysis the stochastic residual series generated by this method exhibit almost no temporal dependence. For all catchments, not only the examples shown in Figure 34, the autocorrelation is already close to zero at lag-1 or lag-2. On one hand, this behavior is consistent with the structure of the method, which does not explicitly attempt to reproduce temporal dependence and therefore generates essentially uncorrelated residual series. On the other hand, this finding makes it somewhat surprising that the original study, based on 10 years of synthetic hourly data, reported a satisfactory indirect reproduction of autocorrelation. Investigating the reasons for this discrepancy lies beyond the scope of the present work. Nevertheless, it raises interesting questions regarding the influence of model temporal resolution and of the post-processing procedures on the ability of different methods to reproduce temporal dependence, which may deserve further investigation in future studies.

4.3.3 Marginal and temporal correlation properties of the stochastically generated discharge series

Similarly to Section 4.3.2, this section examines the ability of the different approaches to reproduce the statistical properties of the observed discharge series during the calibration period after reintroducing the stochastically generated residuals. The results for the AR-method and the C-method are reported only for the configuration in which the bias correction factor (BCF) is not applied. Since the BCF acts as a multiplicative term, it does not affect either the marginal shape properties or the autocorrelation structure of the series; therefore, the statistical properties of the series remain unchanged whether the BCF is included or not. We begin by analyzing

the marginal properties of the reconstructed discharge series. The corresponding L-moment diagrams are shown in Figure 35.

Both the AR-method and the C-method exhibit a systematic tendency to reproduce L-skewness and L-kurtosis values larger than those observed in the discharge data.

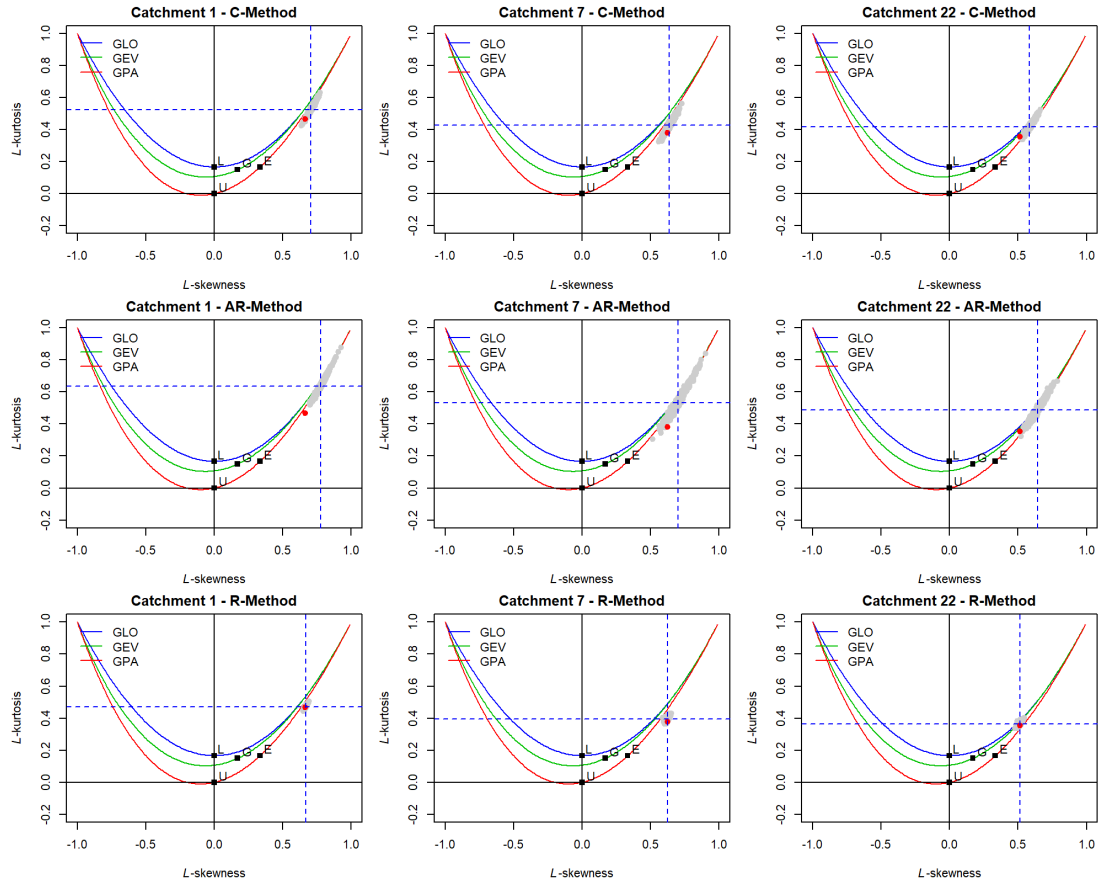


Figure 35. L-diagram plots (discharge - calibration period) for three sites: Cervo a Passobreve (C1), Germanasca a Perrero (C7), and Tanaro a Piantorre (C22). In red target statistics, in gray point for individual stochastic generated series, in dashed blue lines, the average of the stochastic series.

In contrast, the R-method consistently produces values that are much closer to the target marginal shape properties, thus outperforming the other two approaches in this respect. Since all three methods were previously shown to reproduce the marginal properties of their respective residual series satisfactorily, these discrepancies must originate from the way residuals are defined and transformed prior to stochastic

generation. The R-method operates on residuals computed directly in the discharge space, whereas the AR-method and the C-method rely on residuals obtained after applying a logarithmic transformation to both the D-model predictions and the observed discharge.

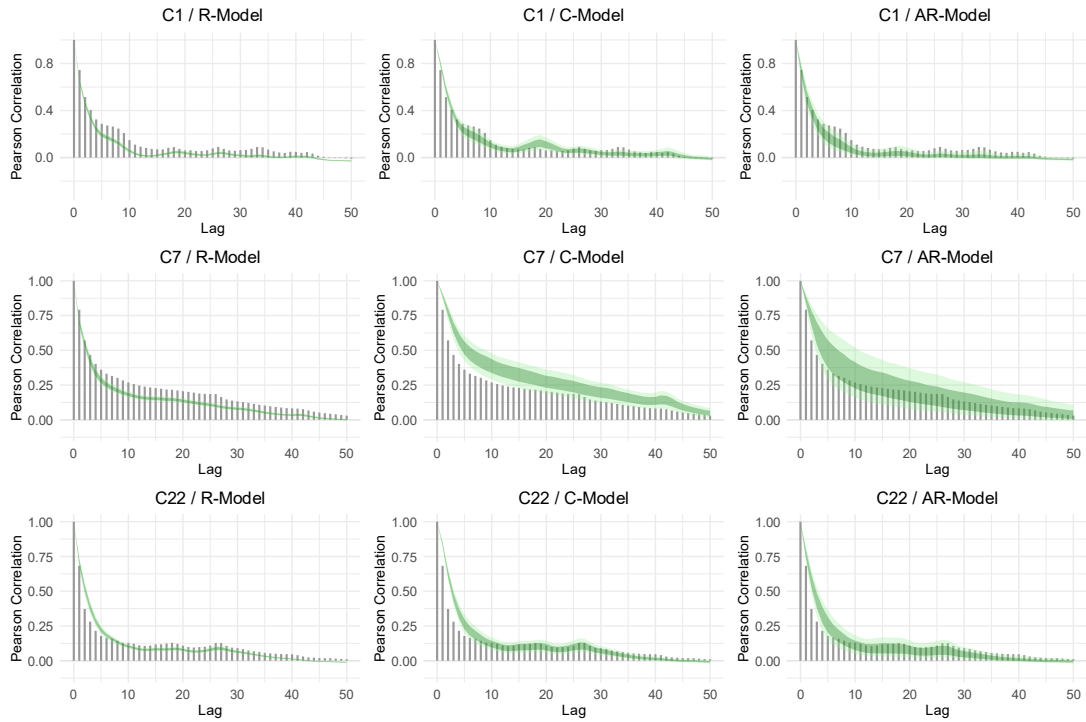


Figure 36. Pearson correlation up to lag 50 for the calibration period for the discharge at Cervo a Passobreve (C1), Germanasca a Perrero (C7), and Tanaro a Piantorre (C22). Darker-shades 50% interval; lighter-shades 80% interval. Target reference statistics of the observed discharge series indicated by the gray bars.

Turning to the temporal dependence structure (Figure 36) the autocorrelation of the reconstructed discharge series is generally reproduced reasonably well for the three catchments considered, although with some dispersion among the individual stochastic realizations. Two aspects are worth highlighting. First, for the Cervo a Passobreve catchment (C1), the residual autocorrelation was poorly reproduced at larger lags, yet the autocorrelation of the reconstructed discharge series is still well captured. This apparent inconsistency arises because the conversion of residuals back to discharge involves the use of the deterministic D-model predictions, which are themselves autocorrelated.

Second, the opposite situation can also occur. In some catchments, the residual autocorrelation is reproduced relatively well, yet the autocorrelation of the reconstructed discharge series remains poorly captured. This highlights that the relationship between residual and discharge autocorrelation is not straightforward and depends on the interaction between the stochastic residual component and the deterministic predictions from the rainfall–runoff model. This behavior can be understood by examining once again the structure of the reconstruction equation (below starting from the case of untransformed residuals, but later generalized for the logarithmic transformation):

$$\widetilde{Q}_t^0 = Q_t^S - \tilde{\epsilon}_t, \quad (\text{Eq. 4.7})$$

The autocovariance function of the reconstructed discharge is given by

$$\gamma_{\widetilde{Q}^0}(k) = \gamma_{Q^S}(k) + \gamma_{\tilde{\epsilon}}(k) - \text{Cov}(Q_t^S, \tilde{\epsilon}_{t-k}) - \text{Cov}(\tilde{\epsilon}_t, Q_{t-k}^S) \quad (\text{Eq. 4.8})$$

Therefore, the autocorrelation of the reconstructed discharge does not depend solely on the autocorrelation of the residual process. It also depends on: (I) the autocorrelation structure of the deterministic predictions Q_t^S , (II) the variance ratio between Q^S and $\tilde{\epsilon}$, and (III), crucially, the cross-covariance structure between Q^S and $\tilde{\epsilon}$. In other words, residuals are not independent of Q^S ; rather, they inherit a specific cross-correlation structure with the model predictions. When residuals are generated stochastically, even if their marginal distribution and temporal autocorrelation are accurately reproduced, their cross-correlation with Q^S is generally not preserved. As a consequence, the interaction terms in the above expression may differ substantially from those observed in reality. In the present framework, however, residuals are defined in logarithmic space (after the seasonal component is reintroduced):

$$\epsilon_t = \log(Q_t^S) - \log(Q_t^0) \quad (\text{Eq. 4.9})$$

and the reconstructed discharge is

$$Q_t^0 = Q_t^S e^{-\tilde{\epsilon}_t} \quad (\text{Eq. 4.10})$$

Where ϵ_t is the observed log-residual, $\tilde{\epsilon}_t$ is the simulated residual, and $e^{-\tilde{\epsilon}_t}$ is the multiplicative stochastic factor. Define:

$$X_t = e^{-\tilde{\epsilon}_t} \quad (\text{Eq. 4.11})$$

Then:

$$\widetilde{Q}_t^0 = Q_t^S X_t \quad (\text{Eq. 4.12})$$

The autocovariance of the reconstructed discharge becomes

$$\gamma_{\widetilde{Q}^0}(k) = \text{Cov}(Q_t^S X_t, Q_{t-k}^S X_{t-k}) \quad (\text{Eq. 4.13})$$

This expression is fundamentally nonlinear and cannot be reduced to a simple sum of components, as in the additive case. Even if Q_t^S and ϵ_t were independent, we would obtain

$$\gamma_{\widetilde{Q}^0}(k) = \gamma_{Q^S}(k) E[X_t]E[X_{t-k}] + E[Q_t^S]E[Q_{t-k}^S] \gamma_X(k) \quad (\text{Eq. 4.14})$$

where $\gamma_X(k) = \text{Cov}(X_t, X_{t-k})$. Thus, the discharge autocovariance depends on both the deterministic structure $\gamma_{Q^S}(k)$, and the transformed residual autocovariance $\gamma_X(k)$. A crucial mathematical result explains why good reproduction of residual autocorrelation does not guarantee good reproduction of discharge autocorrelation. Assume for illustration that $\tilde{\epsilon}_t$ is Gaussian with variance σ_ϵ^2 and autocorrelation $\rho_\epsilon(k)$. Then the autocorrelation of $X_t = e^{-\tilde{\epsilon}_t}$ is:

$$\rho_X(k) = \frac{\exp(\sigma_\epsilon^2 \rho_\epsilon(k)) - 1}{\exp(\sigma_\epsilon^2) - 1} \quad (\text{Eq. 4.15})$$

This shows that autocorrelation is not preserved under the exponential transformation. The relationship is nonlinear and depends on the variance σ_ϵ^2 . When σ_ϵ^2 is non-negligible, even small errors in $\rho_R(k)$ can be amplified or dampened in $\rho_X(k)$. Therefore, good reproduction of $\rho_R(k)$ does not imply good reproduction of $\rho_X(k)$. Even perfect reproduction of $\rho_X(k)$ does not ensure correct reproduction of $\rho_{\widetilde{Q}^0}(k)$, because Q_t^S modulates the process multiplicatively. Two effects may then occur. If the autocorrelation of Q_t^S dominates the total variance, errors in reproducing residual autocorrelation may have limited impact on the discharge autocorrelation, which becomes largely driven by Q_t^S . Conversely, even a well-reproduced residual autocorrelation may lead to discrepancies in the discharge autocorrelation if the

exponential transformation distorts the dependence structure, or the cross-correlation structure between Q_t^S and the stochastic component is not preserved. In other words, accurately reproducing the marginal temporal structure of residuals is not sufficient to guarantee accurate reproduction of the temporal structure of the reconstructed discharge. The nonlinear back-transformation and multiplicative coupling with Q_t^S fundamentally alter the dependence structure. This explains why improvements (or deteriorations) in the residual autocorrelation do not translate monotonically into improvements (or deteriorations) in the discharge autocorrelation: the reconstructed series is the result of a coupled deterministic-stochastic system in which nonlinear transformation and interaction effects jointly determine the final dependence structure.

4.3.4 S-model reliability

The resulting stochastic discharge series can also be used to generate prediction envelopes, which allow the confidence intervals of the S-models to be assessed. The reliability of the estimated confidence bounds over the full range of D-model predictions is shown in Figure 37. Results are reported for the calibration period for all methods, and for the validation period only for the R-method and the C-method. For

the C-method and the AR-method, PPP are first presented for the case in which the bias correction factor (BCF) is not applied.

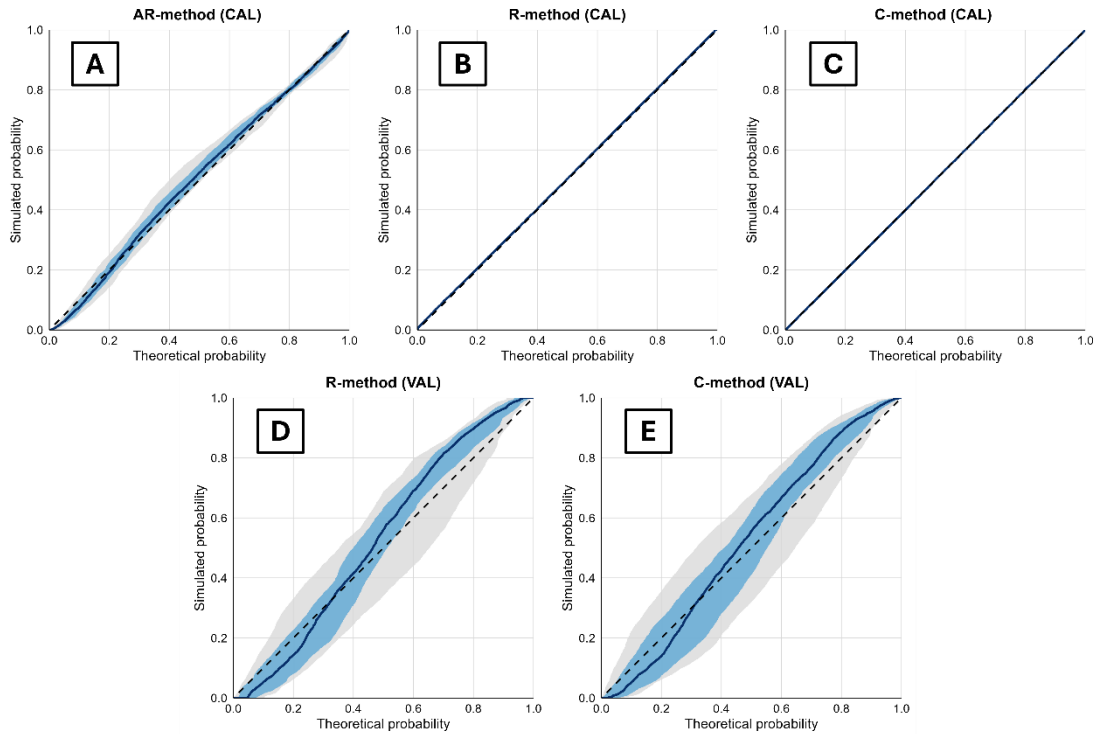


Figure 37. Predictive Probability Plots (PPPs) for the entire D-model predictions discharge range. Calibration: (A) AR-method; (B) R-method; (C) C-method. Validation: (D) R-method; (E) C-method. Blue shades: 50% catchments coverage. Gray shades: 80% catchments coverage.

During calibration, all methods produce highly reliable confidence intervals. The AR-method shows slight deviations from the theoretical 1:1 line, although these remain minor and practically negligible. It should also be noted that the PPP display coverage intervals (50% in blue shades and 80% in gray shades) aggregated across the 24 catchments. The objective here was primarily to illustrate the overall performance across all sites rather than the behavior at individual locations. In the validation period, the different methods again show comparable performance.

Although some deviations from the theoretical alignment are present, the estimated confidence bounds can still be considered overall satisfactory. Figure 38 instead presents PPP computed for a restricted range of D-model predictions, considering only values greater than Q_{70}^S .

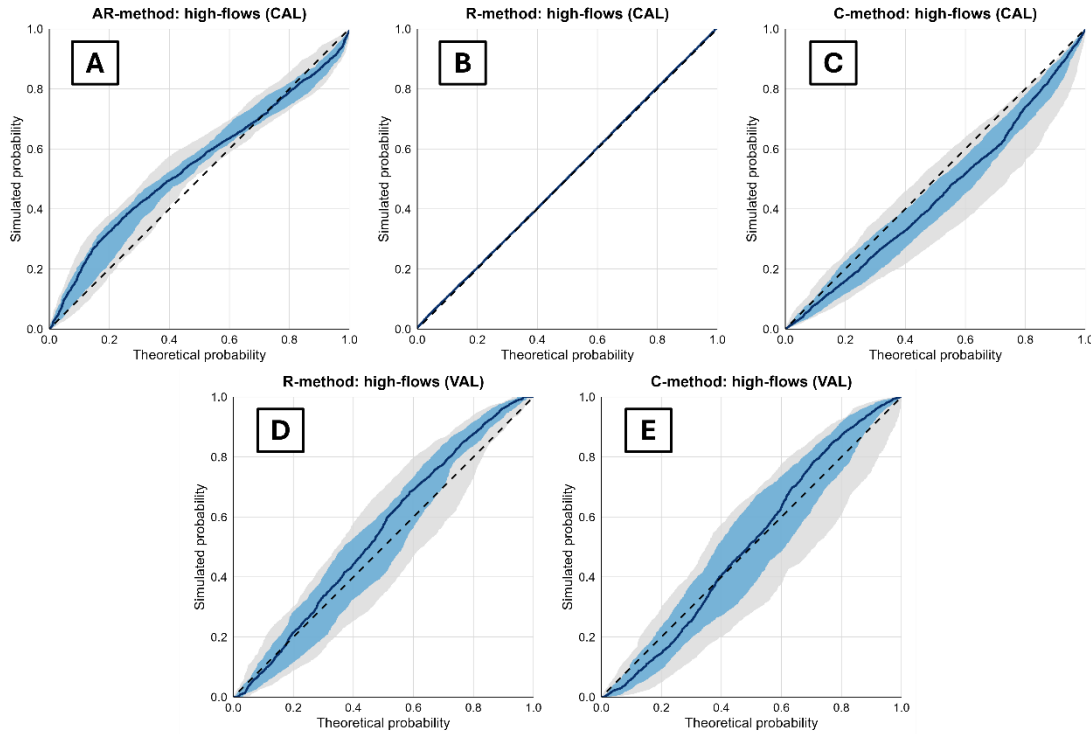


Figure 38. Predictive Probability Plots (PPPs) for the D-model predictions greater than Q_{70}^S . Calibration: (A) AR-method; (B) R-method; (C) C-method. Validation: (D) R-method; (E) C-method. Blue shades: 50% catchments coverage. Gray shades: 80% catchments coverage.

In this case the analysis focuses exclusively on predicted high-flow discharge quantiles. Under these conditions, clearer differences between methods emerge. During calibration, the R-method clearly outperforms the alternatives. Both the AR-method and the C-method show noticeable deviations: the former tends to produce systematic overprediction, whereas the latter exhibits a larger spread and variability. In the validation period the R-method also shows some deviations, but these remain, on average, smaller than those observed for the C-method.

4.3.5 Back Correction Factor (BCF)

In Section 4.2.4, it was noted that in the original AR-method a Bias Correction Factor (BCF) was also introduced. The $\widetilde{Q}^0$ is in fact reconstructed from deterministic model simulations Q^S through a multiplicative error structure defined in logarithmic space, yielding the same inverse transformation as in Eq. 4.11. In the stochastic setting, N synthetic residual series $\tilde{\epsilon}_t^{(j)}$, with $j = 1, \dots, N$, are generated, producing the corresponding stochastic discharge realizations

$$\widetilde{Q}_t^{0(j)} = Q_t^S e^{-\tilde{\epsilon}_t^{(j)}} \quad (\text{Eq. 4.16})$$

Because the exponential function is strictly convex, Jensen's inequality implies that, in general,

$$E[e^{-\epsilon_t}] \neq e^{-E[\epsilon_t]} \quad (\text{Eq. 4.17})$$

Specifically, whenever $\text{Var}(\epsilon_t) > 0$,

$$E[e^{-\epsilon_t}] > e^{-E[\epsilon_t]} \quad (\text{Eq. 4.18})$$

which implies that back-transformation from log-space introduces a systematic multiplicative bias in the mean of the reconstructed discharge. This phenomenon is commonly referred to as Jensen bias. To remove this distortion, a multiplicative correction term, is introduced and defined as

$$\text{BCF} = E[e^{-\epsilon_t}] \quad (\text{Eq. 4.19})$$

The bias-corrected discharge is therefore expressed as

$$\widetilde{Q}_{t,\text{corr}}^{0(j)} = \frac{Q_t^S e^{-\tilde{\epsilon}_t^{(j)}}}{\text{BCF}} \quad (\text{Eq. 4.20})$$

where $E[\cdot]$ denotes expectation with respect to the distribution of the residual process. Under the assumption that the residual process is Gaussian (assumption done by Shabestanipour et al.),

$$\epsilon_t \sim \mathcal{N}(\mu_\epsilon, \sigma_\epsilon^2) \quad (\text{Eq. 4.21})$$

where $\mu_\epsilon = E[\epsilon_t]$ is the residual mean and $\sigma_\epsilon^2 = \text{Var}(\epsilon_t)$ its variance, the moment-generating function of the normal distribution provides a closed-form expression:

$$E[e^{-\epsilon_t}] = \exp\left(-\mu_\epsilon + \frac{1}{2}\sigma_\epsilon^2\right) \quad (\text{Eq. 4.22})$$

Hence, under normality, the Back-Correction Factor becomes

$$\text{BCF}_{\text{norm}} = \exp\left(-\mu_\epsilon + \frac{1}{2}\sigma_\epsilon^2\right). \quad (\text{Eq. 4.23})$$

If, in addition, the residual process is independent of the deterministic simulation Q_t^S , then

$$E[\widetilde{Q}_{t,\text{corr}}^{(j)}] = E[Q_t^S] \quad (\text{Eq. 4.24})$$

so that the first moment of the reconstructed discharge coincides with that of the deterministic model. It is important to note that this correction addresses only the bias introduced by the nonlinear back-transformation and does not correct any structural bias inherent in Q_t^S . From a probabilistic perspective, however, the BCF acts as a global multiplicative rescaling of the entire stochastic ensemble:

$$\widetilde{Q}_{t,\text{corr}}^{(j)} = \widetilde{Q}_t^{(j)} / \text{BCF} \quad (\text{Eq. 4.25})$$

Since this factor is constant across time and realizations, it modifies both the central tendency and the dispersion of the predictive distribution at each time step. Let $\hat{F}_t(x) = P(\widetilde{Q}_t \leq x)$ denote the predictive cumulative distribution function at time t . Predictive reliability is commonly assessed through the probability integral transform

$$u_t = \hat{F}_t(Q_t^0) \quad (\text{Eq. 4.26})$$

which should follow a Uniform (0,1) distribution under perfect reliability. A global multiplicative adjustment of the ensemble shifts the relative position of the observation Q_t^O within the predictive distribution. Consequently, even if the BCF enforces individual series unconditional mean consistency, it may distort the time-wise predictive distribution and degrade reliability diagnostics such as predictive probability plots. In particular, if the deterministic component Q_t^S already compensates for part of the nonlinear bias through its interaction with the residual process, the additional multiplicative correction may over-adjust the ensemble, leading to systematic deviations from uniformity in the probability integral transform. Therefore, while the BCF is theoretically required to correct the Jensen bias under lognormal assumptions, its practical effect on predictive reliability depends on the joint structure of Q_t^S and ϵ_t , and on whether independence and distributional assumptions are satisfied in the stochastic generation procedure.

The effect of applying the BCF under the normality assumption was evaluated for both the AR-method and the C-method, i.e., the two approaches based on logarithmic transformation (results are shown in Figure 39 for the calibration period, where the impact is already clearly visible).

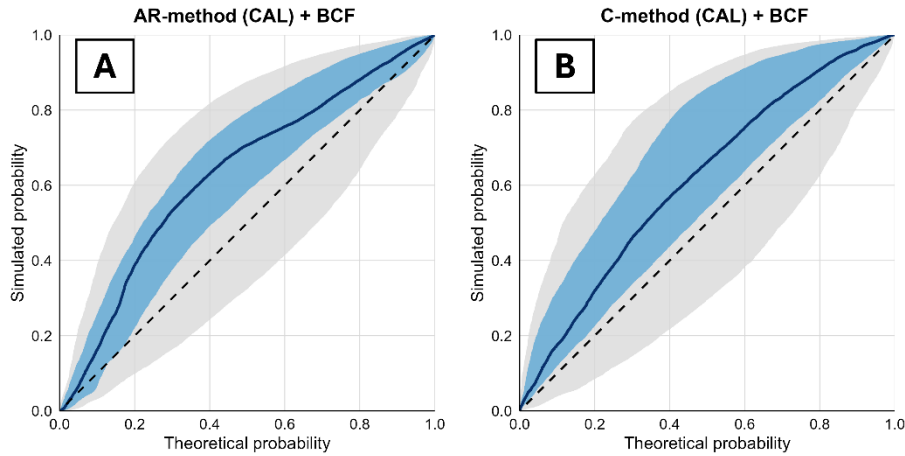


Figure 39. Predictive Probability Plots (PPPs) for the entire range of D-model predictions in the calibration period applying the BCF derived under normality assumption. (A) AR-method; (B) C-method. Validation: (D) R-method; (E) C-method. Darker shades: 50% catchments coverage. Lighter shades: 80% catchments coverage.

The comparison with the corresponding results obtained without applying the BCF (Figure 37A and Figure 37C) shows a marked deterioration in the reliability of the S-models when the correction is introduced. This evidence suggests that, when the primary objective is the reliability of the predictive envelope rather than the exact reproduction of the unconditional mean of the process of the individual stochastic series generated, omitting the BCF may represent the preferable choice.

4.4 Results assessment: how to move forward?

4.4.1 Is the assumption of independence really reasonable for the C-model?

It is now necessary to critically assess the performance of the newly proposed C-method. In calibration, when considering the entire discharge range, its average performance was broadly comparable to that of the R-method. However, when focusing specifically on the high-flow regime, a clear degradation emerged. This discrepancy calls into question one of the core working assumptions underlying the method, namely the assumption of independence structure between D-model predictions and transformed residuals. While it is true that the logarithmic transformation together with the deseasonalization substantially reduced the strong heteroskedastic structure originally present in the untransformed residuals, the results indicate how for some catchments, as already noted in Figure 32, structures remain non-negligible. The larger deviations observed for high flows are therefore plausibly attributable to a misrepresentation of residual data structures in this range. This does not imply that the C-method could not, in principle, be extended to account explicitly for residual dependence. However, incorporating such refinements would conflict with the additional requirements outlined in Section 4.1.2 for a method intended to support regionalization. Increasing model complexity to address conditional variance would reduce structural parsimony, limit transferability, and complicate parameter estimation across multiple catchments.

One possible way to remain within a copula-based framework while addressing these issues would be to expand the dependence structure to include additional variables. A natural extension would be to model jointly the dependence between D-model predictions and residuals, together with the temporal dependence of the residuals themselves, within a higher-dimensional copula structure (e.g., including ϵ_t , ϵ_{t-1} , and Q^S). However, this approach presents substantial theoretical and practical challenges. First, while multivariate copulas of dimension greater than two can certainly be

defined, flexible specification becomes difficult when pairwise dependence structures differ in form and strength. Fully parametric multivariate copulas are typically tractable only under restrictive assumptions (e.g., elliptical structures such as the Gaussian or Student-t copula). In situations where dependence structures vary across pairs and may belong to different copula families, a more general construction is required (Joe, 2014; Nelsen, 2006). A possible solution is provided by vine copulas. Vine copulas (or pair-copula constructions - Aas et al., 2009; Bedford & Cooke, 2001) decompose a multivariate copula into a cascade of bivariate copulas arranged according to a graphical structure (C- or D-Vine). This allows each pair of variables to be modelled using potentially different copula families, thereby offering considerable flexibility in representing complex dependence patterns. While powerful, vine copulas are inherently high-dimensional, parameter-rich structures. Their specification requires selecting a vine structure, choosing appropriate families for each pair, and estimating multiple dependence parameters. Such flexibility, although attractive in local applications, results in models that are substantially more complex and computationally demanding. Second, even assuming that a vine-copula framework could be successfully implemented and potentially regionalized, a fundamental issue remains: consistency of pairwise dependence structures across catchments. While the dependence between ϵ_t and ϵ_{t-1} has repeatedly shown a stable structure (often adequately represented by a t-copula), the same cannot be said for the dependence between D-model predictions and residuals. In some catchments, residual heteroskedasticity appears negligible, implying weak or near-independent dependence. In others, a clear residual conditional structure is present, and the appropriate copula family may vary from case to case. Such variability may be accommodated in site-specific modelling, where the dependence structure can be tailored individually, but it is fundamentally at odds with the objective of developing a parsimonious and transferable regional approach.

These considerations must also be viewed in light of another key result: residual autocorrelation, the very property that motivated these first group of modelling strategy, is not consistently well reproduced by the simple C-method proposed here (nor, in fact, by the other methods tested). If the key property that the framework was designed to capture cannot be robustly reproduced across catchments, it becomes legitimate to question whether further increasing structural complexity is justified. In our view, the answer is negative within the operational scope of this work. Our primary objective remains the reliable estimation of confidence intervals (and only secondarily

the reproduction of the statistical properties of the individual stochastic series). For this purpose, simpler approaches, as presented in Chapter 5, may prove sufficient and more robust. Pursuing increasingly complex copula constructions would not provide proportional benefits, while substantially increasing modelling effort, parameter uncertainty, and challenges for regionalization. Importantly, this decision should not be interpreted as a general criticism of methods aiming at stochastically reproducing individual residual series. With additional methodological development, explicit parametric modelling of cross-correlation structures, and site-specific calibration, such methods could indeed prove effective in other contexts. Starting from the result presented, they are simply not the most appropriate tools for the objectives and constraints of the present thesis: it is in fact also clear how the simpler non-parametric R-method, when focusing on reliability of the produced S-model, provide equal performance; hence the application of the C-method in this current form does not provide enough additional benefits to justify its application. For the same reason, we do not further pursue a full parametric characterization of the marginal distribution of model residuals (the missing element for a fully parametric formulation of the C-method, as described in Section 4.2.1). Although a parametric representation could be theoretically possible, it would likely require complex and potentially non-standard distributions, introducing additional estimation challenges without contributing materially to the goals of this work. Since the alternative approach adopted in the next chapter does not require an explicit parametric specification of the residual marginals, further effort in that direction would not be aligned with the intended scope (still remaining an interesting research gap for future research).

Chapter 5 therefore shifts perspective and explores a different methodological strategy, based on the more focused objective of directly estimating confidence intervals rather than reproducing the full stochastic structure of the residual process.

Chapter 5

5. Beyond residual modelling: an alternative approach to stochastic rainfall-runoff modelling

5.1 A different philosophy

5.1.1 Bias correction and direct confidence bound estimation: BLUECAT and its parametric evolution

In the previous chapter, we attempted to construct an S-model and estimate confidence bounds starting from the constraining premises that the individual series composing the S-model envelope should have maintained the same marginal distribution, correlation, bias and heteroskedasticity of the observed discharge. As discussed in the concluding remarks of Chapter 4, this strategy proved theoretically appealing but practically problematic. The development of a simple parametric model, potentially suitable for regionalization, was hindered by inconsistencies in cross-correlation structures between simulated discharge and transformed residuals and by unsatisfactory performance in the high-flow range, which is of particular relevance for this thesis. Most importantly, even if a more sophisticated stochastic formulation could in principle have been developed to address these shortcomings, such a model would likely have required a level of structural complexity incompatible with the primary objective of this research, namely the development of an approach suitable for practical and operational regionalization. In other words, increasing model realism through additional parametric refinements would not necessarily translate into a method that is transferable, robust, and parsimonious at the regional scale. For these reasons, rather than further pursuing increasingly complex formulations aimed at reproducing the full

stochastic structure of residual time series, we redirect the focus toward a different strategy. The shift in perspective is conceptual before being technical: instead of attempting to jointly reproduce all statistical properties of the observed discharge (including marginal shape, heteroskedasticity, and autocorrelation), we now concentrate exclusively on the estimation of reliable confidence intervals. In this revised framework, the success of the method is evaluated solely in terms of the adequacy of the estimated confidence bounds; the reproduction of other properties, most notably the temporal autocorrelation of individual stochastic realizations, is no longer imposed as a modelling constraint.

In 2022, Montanari and Koutsoyiannis (Koutsoyiannis & Montanari, 2022a) proposed a new method to approach the problem of the definition of confidence bounds they called BLUECAT. Their method was (again) non-parametric, while here the main idea is to develop an evolved version that is parametric instead. The details about the BLUECAT methodology are reported in the original paper by Koutsoyiannis and Montanari, so the reader is invited to find there the full discussion. Here we will limit the explanation of their method to the main concepts to highlight where the differences between the non-parametric and the newly proposed parametric approach arise. To facilitate the connection between this chapter and their original paper, we will adopt the same terminology whenever possible. BLUECAT aims at transforming point prediction obtained by any deterministic RR model, into stochastic predictions (D-model upgraded to S-model). In general, we can consider any hydrologic D-model as a function G transforming input x_τ at discrete time τ into the deterministic output discharge Q_τ :

$$Q_\tau = G(x_\tau) \quad (\text{Eq. 5.1})$$

The function G can be widely different depending on the specific model applied, and usually involves multiple state variables internal to the model architecture. The input vector x_τ can also be widely different depending on the requirements of the model. The methodology does not distinguish between the different processes across different models, because to perform the upgrade from D-model to S-model, only the output Q_τ is required. To perform this upgrade, the D-model output Q_τ is compared to the observed data q_τ , effectively estimating the probability distribution of the observed data conditioned on the D-model predictions:

$$F_{q|Q}(q|Q) = P\{\underline{q} \leq q \mid \underline{Q} = Q\} \quad (\text{Eq. 5.2})$$

obtaining the corresponding S-model output. Here q and Q are assumed concurrent and referring to discrete time τ . The underlined variables indicate they are stochastic. The use of the observed discharge, implies a correction of biases if present in the D-model predictions. From the S-model output, confidence bands as well as useful statistics (e.g., mean, median, quantiles) can be extracted. The main “ingredient” to be determined is the conditional distribution itself. In the original work, the authors decided for the fully data driven (non-parametric) approach, while other possibilities were already suggested but not explored in the study. One of those suggestion was to estimate the conditional distribution in eq. 2 by assuming a copula $C(F_q(q), F_Q(Q))$ and to determine $F_{q|Q}(q|Q)$ as:

$$F_{q|Q}(q|Q) = \frac{F_{q|Q}(q, Q)}{f_Q(Q)} \quad (\text{Eq. 5.3})$$

With

$$F_{q|Q}(q, Q) = C(F_q(q), F_Q(Q)) \quad (\text{Eq. 5.4})$$

Here we follow exactly this approach, by fitting and defining a copula between the D-model and the predictand (observations), defining the conditional distribution to upgrade the D-model to the S-model. All other assumptions behind the original methodology still apply here, namely the assumption of stationarity between the calibration period and the validation (or application period), and the assumption of calibration period length sufficient to ensure that enough information is available to correctly perform the upgrade.

5.1.2 Main operational differences

Aside from the theoretical formulation of the methodology, both the non-parametric and the parametric approaches follow in practice a similar operational logic. During the calibration period, in the non-parametric BLUECAT framework, at each simulation time τ , a deterministic point estimate Q_τ is obtained from the calibrated D-model. A subset of simulated values Q of size $m_1 + m_2 + 1$ is then selected, consisting of the value closest to Q_τ , together with m_1 values smaller than Q_τ and m_2 values larger

than it. The corresponding observed discharges q_i , $i = 1, \dots, m_1 + m_2 + 1$, are subsequently identified, and this sample is used to estimate confidence intervals and summary statistics. While an advantage of the non-parametric approach is that resampling is performed exclusively from observed discharge, thereby avoiding the introduction of unrealistic values, the selection of m_1 and m_2 is not straightforward, and extrapolation beyond the observed data is inherently impossible. To illustrate this issue, consider the case $m_1 = m_2 = 100$, resulting in a sample of 201 values. If the D-model prediction Q_τ is one of the simulated extremes in the calibration period, for example let us assume the highest value, it is not possible to identify any simulated value larger than Q_τ . Consequently, $m_2 = 0$, and the effective sample size must be reduced. More generally, the choice of m_1 and m_2 involves a trade-off between the representativeness of the sample and its numerosity. On the one hand, a sufficiently large sample is required to compute meaningful statistics; on the other, including values that are too distant from Q_τ reduces the local representativeness of the sample and leads to wider confidence bands. Increasing the sample size improves numerical robustness but incorporates increasingly distant values, whereas reducing it preserves locality at the expense of statistical reliability. This trade-off is also heavily influenced by the length of the calibration period: for the same values of m_1 and m_2 , a calibration period of 10 years will necessarily include more distant points from the considered D-model prediction than a period of 15 years. These effects become increasingly pronounced near the extremes of the distribution, and when the number of available years is limited, it may be impossible to identify a satisfactory compromise between these competing requirements. A schematic representation of these effects is provided in Figure 40. In contrast, during the calibration period the parametric approach based on copulas relies on the quality of the fitted dependence structure. If the copula adequately represents the dependence between the D-model predictions and the observed discharges, the limitations encountered in the non-parametric approach are inherently avoided. After defining a suitable copula between the pseudo-D-model predictions and the pseudo-observed discharges, each deterministic D-model prediction is transformed into the $[0,1]$ domain using its empirical cumulative distribution function (e.g., yielding a value of 0.8). The conditional probability distribution function is then derived analytically from the copula. From this conditional distribution, an arbitrarily large number of points can be generated, resolving issues related to sample size, while remaining within the $[0,1]$ interval.

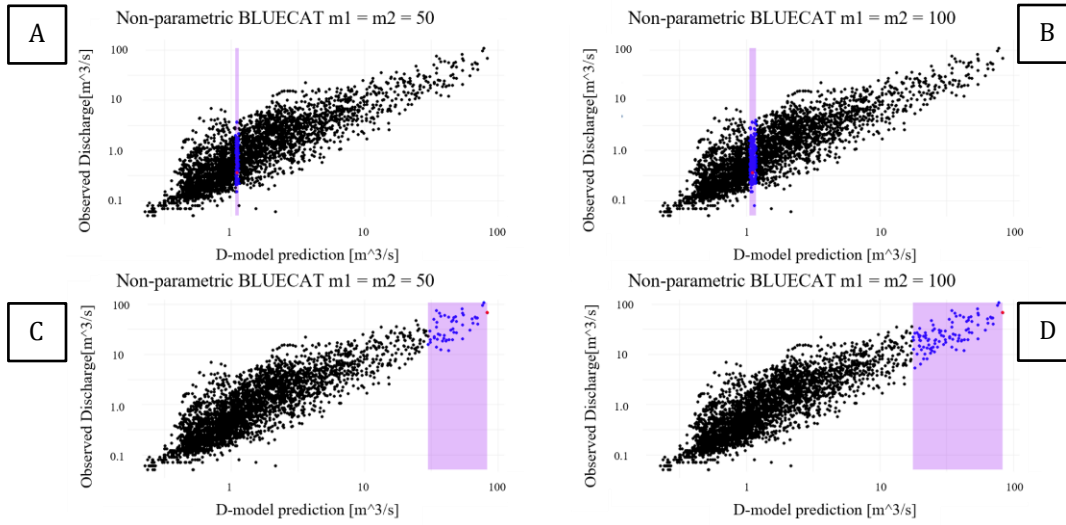


Figure 40. Examples of trade-off between sample representativeness and sample size in the non-parametric BLUECAT approach. Median D-model prediction (A; $m_1 = m_2 = 50$) and (B; $m_1 = m_2 = 100$); Maximum D-model prediction (C; $m_1 = m_2 = 50$) and (D; $m_1 = m_2 = 100$). In red the current point; corresponding observed discharge subset indicated by the blue points inside the purple shaded area.

These samples are subsequently mapped back to the discharge domain using the empirical CDF of the observed data (from now on indicated as O-CDF to distinguish it from the S-CDF of the D-model predictions). Near the extremes, for example, in the case of the largest D-model prediction in the calibration period, the same procedure can be applied without any modification, provided that the fitted copula still accurately represents the underlying data structure. In this case, no issues arise with respect to either sample numerosity or representativeness. Finally, the parametric approach can also address a further limitation of the original non-parametric method, namely the inherent upper and lower bounds of the estimated confidence bands near extremes. Consider again the case of the largest D-model prediction: regardless of the choice of m_1 and m_2 , the corresponding sample used to estimate confidence bands can include at most the largest observed discharge value and cannot exceed it. As a result, the estimated confidence bands for any quantile are necessarily bounded by the historical observations and cannot extrapolate beyond them. This represents a substantial limitation, as it precludes the possibility that the true discharge may exceed the observed maximum due to model uncertainty. While in this example lower-than-predicted values can be represented within the confidence bands, higher-than-observed values cannot. In its original non-parametric formulation, this latter scenario is

therefore not reproducible. In a subsequent work (Koutsoyiannis & Montanari, 2022b), despite mainly focusing on applications on climate projections, the authors actually tried to address this original limitation, but still in a non-parametric way by extrapolating beyond the range of observed discharges using linear interpolations.

In the parametric copula-based approach, this limitation can be addressed instead by modifying the empirical cumulative distribution functions of the D-model predictions and of the observed discharges. Specifically, the empirical CDFs are replaced by mixed CDFs, in which the central part of the distribution is represented empirically, while the lower and/or upper tails are modelled parametrically to allow for extrapolation beyond the range of available data. A full parameterization of the complete CDF could also be a possible choice. In the present study, we focus explicitly on high flows; therefore, the CDFs parametric modelling is applied exclusively to the upper tails. This procedure is well grounded in Extreme Value Theory, which states that exceedances over a sufficiently high threshold from a sample drawn from any underlying distribution asymptotically follow a Generalized Pareto distribution (Balkema & Haan, 1974; Katz et al., 2002; Pickands, 1975). In practice, the mixed CDFs are used respectively to convert D-model predictions into pseudo-values within the $[0,1]$ square and to map the pseudo-observations sampled from the conditional distribution function derived from the fitted copula, back into the physical discharge space. For quantiles below a threshold Q_{TR} , the mixed CDFs coincides with the empirical CDFs, whereas for higher quantiles they both follow the parametric tails models. A graphical illustration of the entire copula-based process is shown in Figure 41.

During the validation period, both the non-parametric and the parametric approaches rely on the assumption of stationarity and ergodicity between the calibration and validation periods. In the non-parametric case, for each D-model prediction in the validation period, the $m_1 + m_2 + 1$ closest D-model predictions from the calibration period are identified, and the corresponding observed discharges are used to construct confidence bands. In the parametric approach, the D-model simulations in the validation period are first transformed into pseudo-values using the mixed S-CDF defined in the calibration period. These pseudo-values are then employed to extract the conditional distribution functions from the copula fitted during calibration; then pseudo-observed discharges are sampled from them. Finally, these samples are mapped back to the real discharge space using the same mixed O-CDF for the observations defined, again, for the calibration period.

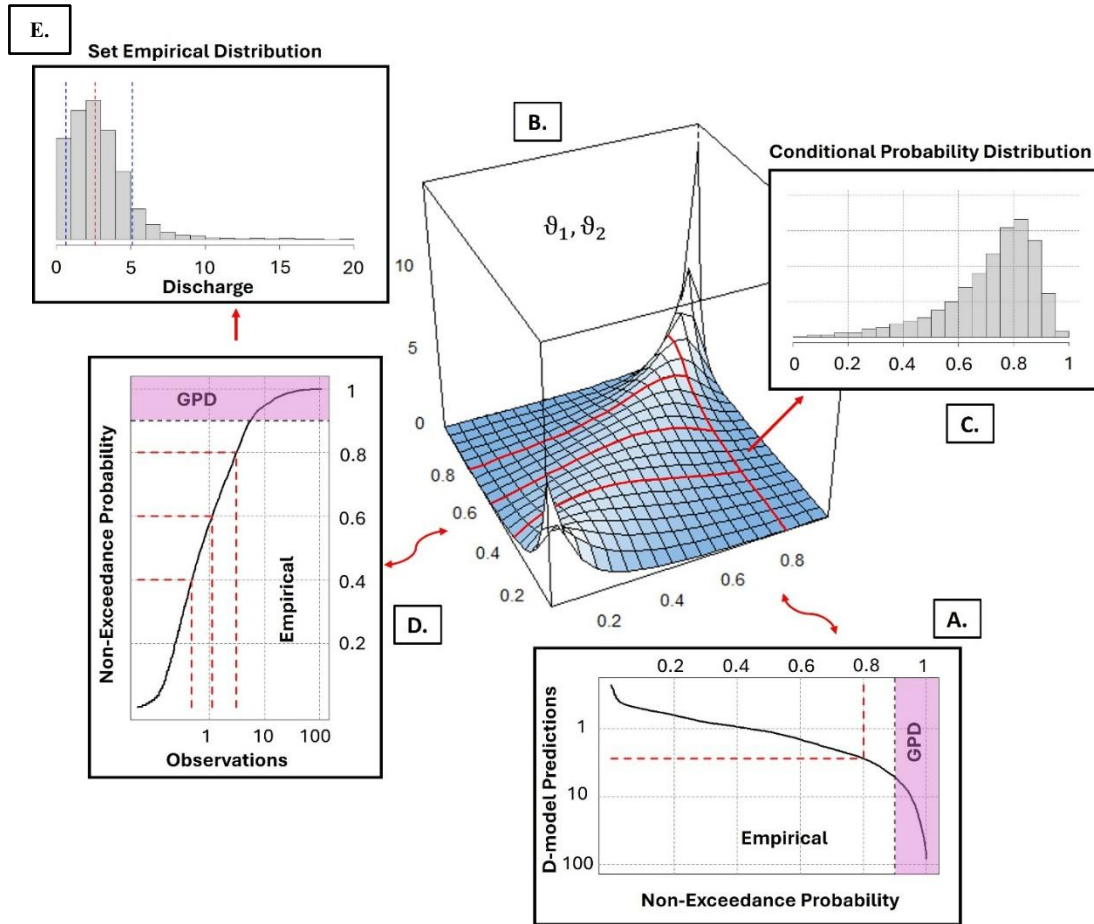


Figure 41. Schematic representation of the application of the “parametric-BLUECAT” with a two parameters (θ_1 and θ_2) BB7 copula for a point prediction of the D-model. (A) Transformation of predicted discharge into pseudo-value using the D-model mixed CDF; (B) fitted copula model between D-model predictions and observations; (C) empirical conditional probability density function; (D) re-transformation from uniform to real space through the observed discharge mixed CDF; (E) Empirical probability distribution function for the “observations” set from which confidence interval are estimated with order statistics.

A couple of final remarks: in their original work, together with the estimation of CI with order statistics, the authors proposed a second approach based on K-moments (Koutsoyiannis, 2019). For clarity, in the following comparisons, the order statistics have been used, and not this alternative approach. A second note is about the extended-BLUECAT (including the non-parametric extrapolation on the tails) later proposed by

the same authors: in the following comparison the original method will be used, without allowing for the non-parametric extrapolation on the tails of the distribution.

5.1.3 Parametric copula model tailored for high flows

In addition to the proposed parametric version of the BLUECAT model described in Section 5.1.2, a second modelling approach is here considered and investigated. This approach is conceptually analogous to the first, but is thought to be specifically tailored to the high-flow range. The core idea is to fit a parametric copula to describe the dependence between D-model predictions and observed discharge during the calibration period, using only data pairs corresponding to D-model predictions exceeding a user-defined quantile threshold (e.g., Q_{70}^S). By limiting the fitting procedure to this subset of the data, the aim is to potentially achieve a more accurate representation of the dependence structure in the high-flow range than would be possible when modeling the entire range of predictions with a single copula. The selection of the threshold quantile is inherently somewhat arbitrary, as the method can in principle be applied to any threshold. However, it must account for the amount of available data: for instance, selecting a threshold of Q_{70}^S implies that only 30% of the calibration data are available for copula fitting, which may be insufficient for robust parameter estimation depending on the length of the calibration period. As in the previous approach, the use of a mixed-type empirical distribution function remains necessary to enable extrapolation in the upper tail. In this framework, the resulting confidence bands are defined only for predictions falling within the selected high-flow range and are not available for predictions outside this interval.

The application of the method during the validation period follows principles similar to those described in Section 5.1.2. Only D-model predictions exceeding the same quantile threshold, now defined based on the validation-period data, are considered. These predictions are transformed into pseudo-observations using the S-CDF from the calibration period, after which the remaining steps of the procedure are applied analogously to the calibration phase.

As discussed in greater detail in Section 5.2.2, restricting the range of applicability of the model has some important implications for the choice of suitable copula families. In particular, when focusing exclusively on high-flows, it is no longer necessary to capture both lower and upper tail dependence. Instead, only upper tail dependence needs to be represented, which allows for the use of simpler and more tractable copula structures.

5.1.4 Validation metrics for the S-model

No matter which of the two approaches is used, the S-models obtained must be validated to check that they are useful for practical applications. The validation of the model outputs is carried out using multiple performance metrics. Some metrics are designed to assess the overall reliability of the estimated confidence bands, while others focus on evaluating the quality of selected reference statistics (e.g., the ensemble median) in comparison with the original D-model predictions. All metrics are computed and reported for the same set of test catchments already previously presented in Section 4.2.3, considering the original non-parametric BLUECAT (selecting different values of $m_1 = m_2$) and both parametric versions introduced in Section 5.1.2 and Section 5.1.3, and are evaluated separately for the calibration and validation periods.

The first metric, to assess overall ensemble reliability, is again the PPP as already introduced in Section 4.2.6. Plot will be produced looking at the entire D-model prediction discharge range as well as the case where only D-model predictions are larger than Q_{70}^S to have clearer insights about the CI reliability in the high-flow range. As a second metric, the mean Interval Score (IS) is adopted, a strictly proper scoring rule widely used for evaluating predictive distributions (Gneiting & Raftery, 2007; Wilks, 2006). The IS measures the quality of central prediction intervals by combining information on their width (sharpness) and their ability to cover the observed values (reliability). For a central $(1-\alpha)$ prediction interval $[l, u]$ and an observation y , the IS is defined as:

$$IS_{\alpha}(y, l, u) = (u - l) + \frac{2}{\alpha}(l - y)1_{y < l} + \frac{2}{\alpha}(y - u)1_{y > u} \quad (\text{Eq. 5.5})$$

where the first term penalizes wide intervals, and the last two terms penalize observations falling outside the interval, with a magnitude inversely proportional to the nominal coverage $(1-\alpha)$. In this study, the 80% interval ($Q_{90} - Q_{10}$) is considered, and the IS provides a single, quantitative measure summarizing both interval width and reliability. This measure was originally introduced for applications to climate models, but has been used in the past in hydrologic application (e.g., Papacharalampous et al., 2019). Because the Interval Score is not scale-invariant, larger river discharge model residuals contribute more strongly to the score when calculated in linear scale. It could be possible to account for this effect and to provide a measure of relative predictive accuracy across the entire range of flows, by computing the IS on the logarithm of

positive river discharge model residuals, following standard practice in hydrology and atmospheric sciences (Scheuerer & Hamill, 2015; Wilks, 2006). In that case, for a positive residual $y > 0$ and interval bounds $l, u > 0$, the log-transformed IS is defined as:

$$\begin{aligned} \text{IS}_\alpha^{\log}(y, l, u) = & (\log(u) - \log(l)) + \frac{2}{\alpha} (\log(l) - \log(y)) 1_{y < l} \\ & + \frac{2}{\alpha} (\log(y) - \log(u)) 1_{y > u} \end{aligned} \quad (\text{Eq. 5.6})$$

This transformation would ensure that under or overestimation would be penalized proportionally to the relative error, providing a more balanced evaluation between low and high flows, while preserving the proper scoring rule properties. Being in this case mostly interested in high-flows, the original IS score is preferred. The final IS is then defined as the average of all $\text{IS}_\alpha(y, l, u)$ values.

The third metric consists in re-evaluating the Kling–Gupta Efficiency, the objective function used to calibrate the reference D-model, for the median of the estimated S-models produced by each method. This validation metric is used to assess if additionally to the uncertainty estimation, the central measure of the S-model could be considered and used as a valid prediction time series not differently from the D-model prediction one. This re-evaluation is not performed for the parametric BLUECAT approach tailored to high flows, as the ensemble median in this case is only defined over a restricted range of D-model discharge predictions.

Lastly, a focused control of extreme values will be performed; some of the previous measures were already more focused on the high-flow discharge range, but not specifically on flood extremes (were the effect of the parametric tails for the copula-based approaches triggers). On average, 3 peak events (peak-over-threshold) each year are extracted from the observed discharge series, from the D-model prediction series and from the ensemble median series respectively; after an initial preprocessing to exclude interdependent events (a minimum time lag of 2 days is set between peaks; if two peaks are closer in time, only the highest is considered) they are sorted and compared (effectively showing a comparison of the empirical distribution of peak-over-threshold values). It should be noted how for this validation metric the values from each series do not necessarily come from the same times τ , but it is rather a comparison closer to traditional flood frequency analysis and flood frequency curves.

5.2 Results analysis

5.2.1 Parametric tail and mixed cumulative distribution functions

With regard to the mixed cumulative distribution functions, the results of fitting a Generalized Pareto distribution (GPD) to peak-over-threshold values of D-model predictions and observed discharge are here presented. The probability density function of the Generalized Pareto distribution is given by:

$$P(x) = \frac{1}{\sigma} \left(1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right)^{-\left(\frac{1}{\xi+1}\right)} \quad (\text{Eq. 5.7})$$

where $\mu \in (-\infty, \infty)$ is the location parameter, $\sigma \in (0, \infty)$ is the scale parameter, and $\xi \in (-\infty, \infty)$ is the shape parameter.

For both parametric copula approaches, the model applied to the entire discharge range and the model tailored to high-flows, and for both D-model predictions and observed discharges, the parametric tail was fitted using subsets of data of equal numerosity (~ 110 data points). Specifically, when considering the full discharge range, the fit was performed on the upper 3% of the data, corresponding to observations greater than or equal to the observed discharge Q_{97}^O and D-model predictions Q_{97}^S . When restricting the analysis to D-model predictions exceeding Q_{70}^S (value selected for the high-flow tailored application), the parametric tail was instead fitted to the upper 10% of the subset of the original observed and predicted discharge. Note how for the observed discharge, despite most of the data points in the samples for a given catchment for the two cases are the same, the samples are not identical being the peak over threshold sub setting being performed based on the D-model predictions flow magnitude and not on the observed discharge flow magnitude. It is also important to note that, during the calibration period over which the parametric tail was estimated, the discharge records are complete and homogeneous across all catchments.

Table 2. Observed discharge parametric tail (Generalized Pareto Distribution) adaptation tests outcomes for the 24 sites considered in the study (full D-model predictions discharge range). Results reported for the Kolmogorov-Smirnov test and the Anderson-Darling test ($\alpha = 5\%$). “obs” = sample statistic; p-value calculated from Monte-Carlo bootstrap (10000 iterations).

CATCHMENT	KS_obs	p-value	KS_PASS	AD_obs	p-value	AD_pass
Cervo a Passobreve	0.06	0.88	✓	0.51	0.29	✓
Cervo a Vignano Biellese	0.05	0.96	✓	0.26	0.79	✓
Dora Di Bardonecchia a Beaulard	0.10	0.60	✓	1.44	0.01	X
Grana a Monterosso	0.07	0.91	✓	0.43	0.38	✓
San Bernardino a San Bernardino Santino	0.07	0.88	✓	0.60	0.20	✓
Sangone a Trana	0.07	0.79	✓	0.51	0.24	✓
Germanasca a Perrero	0.08	0.76	✓	0.61	0.12	✓
Erro a Cartosio	0.05	0.95	✓	0.35	0.52	✓
Bormida di Spigno a Piana Crixia	0.03	0.99	✓	0.15	0.98	✓
Bormida di Millesimo a Murialdo	0.09	0.65	✓	0.89	0.04	X
Ellero a Mondovì	0.07	0.93	✓	0.64	0.13	✓
Sesia a Campertogno	0.04	0.96	✓	0.28	0.76	✓
Mastallone a Varallo	0.09	0.75	✓	0.91	0.06	✓
Soana a Pont Canavese	0.09	0.67	✓	0.56	0.20	✓
Bormida di Spigno a Mombaldone	0.07	0.86	✓	0.46	0.31	✓
Pesio a Carrù	0.05	0.94	✓	0.21	0.89	✓
Strona di Omegna a Gravellona Toce	0.10	0.63	✓	0.78	0.11	✓
Stura di Viù a Germagnano	0.09	0.71	✓	0.48	0.29	✓
Stura di Lanzo a Torino	0.06	0.90	✓	0.53	0.22	✓
Stura di Demonte a Gaiola	0.06	0.96	✓	0.39	0.48	✓
Chisone a San Martino	0.05	0.93	✓	0.28	0.70	✓
Tanaro a Piantorre	0.03	1.00	✓	0.11	1.00	✓
Orco a San Benigno Canavese	0.04	0.98	✓	0.18	0.95	✓
Maira a Racconigi	0.05	0.94	✓	0.24	0.83	✓

Table 2, Table 3, and Table 4 summarize the goodness-of-fit results for the parametric GPD tail model applied to the 24 study catchments, respectively for the observed discharge (considering the case of full D-model predictions discharge range), the observed discharge (considering the case of D-model predictions larger than Q_{70}^S), and for the D-model predictions. Model adequacy was assessed using both the Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) tests at a significance level

$\alpha = 5\%$. Since the distribution parameters were estimated from the data, critical values for both tests were obtained through a Monte Carlo parametric bootstrap procedure (10,000 iterations).

Table 3. Observed discharge parametric tail (Generalized Pareto Distribution) adaptation tests outcomes for the 24 sites considered in the study (D-model predictions greater than Q_{70}^S discharge range). Results reported for the Kolmogorov-Smirnov test and the Anderson-Darling test ($\alpha = 5\%$). “obs” = sample statistic; p-value calculated from Monte-Carlo bootstrap (10000 iterations).

CATCHMENT	KS_obs	p-value	KS_PASS	AD_obs	p-value	AD_pass
Cervo a Passobreve	0.06	0.88	✓	0.51	0.29	✓
Cervo a Vignano Biellese	0.05	0.96	✓	0.26	0.78	✓
Dora Di Bardonecchia a Beaulard	0.10	0.62	✓	1.49	0.02	X
Grana a Monterosso	0.07	0.91	✓	0.44	0.37	✓
San Bernardino a San Bernardino Santino	0.07	0.88	✓	0.60	0.21	✓
Sangone a Trana	0.07	0.78	✓	0.52	0.23	✓
Germanasca a Perrero	0.08	0.73	✓	0.89	0.03	X
Erro a Cartosio	0.05	0.95	✓	0.35	0.53	✓
Bormida di Spigno a Piana Crixia	0.03	0.99	✓	0.15	0.98	✓
Bormida di Millesimo a Murialdo	0.09	0.65	✓	0.90	0.04	X
Ellero a Mondovì	0.06	0.95	✓	0.58	0.17	✓
Sesia a Campertogno	0.05	0.84	✓	0.36	0.60	✓
Mastallone a Varallo	0.09	0.74	✓	0.91	0.06	✓
Soana a Pont Canavese	0.09	0.65	✓	0.56	0.21	✓
Bormida di Spigno a Mombaldone	0.07	0.86	✓	0.45	0.32	✓
Pesio a Carrù	0.05	0.98	✓	0.23	0.85	✓
Strona di Omegna a Gravellona Toce	0.10	0.62	✓	0.78	0.11	✓
Stura di Viù a Germagnano	0.09	0.73	✓	0.48	0.29	✓
Stura di Lanzo a Torino	0.06	0.90	✓	0.53	0.22	✓
Stura di Demonte a Gaiola	0.06	0.96	✓	0.39	0.46	✓
Chisone a San Martino	0.05	0.93	✓	0.28	0.71	✓
Tanaro a Piantorre	0.03	1.00	✓	0.11	1.00	✓
Orco a San Benigno Canavese	0.04	0.97	✓	0.18	0.94	✓
Maira a Racconigi	0.05	0.94	✓	0.24	0.83	✓

As it can be seen, the Kolmogorov-Smirnov test never rejected the null hypothesis of adaptation to the GPD for the selected confidence level. The Anderson-darling test

rejected instead the null hypothesis in a limited number of cases across the different sites; at the same time, it can be noted how in such cases the p-value is always close to 0.05, indicating how in all cases the null hypothesis would instead never be rejected by selecting a confidence level of 1% instead. Considering these result, the parametric description of the tails has been considered satisfactory and will be applied in the following simulations.

Table 4. D-model prediction discharge parametric tail (Generalized Pareto Distribution) adaptation tests outcomes for the 24 sites considered in the study. Results reported for the Kolmogorov-Smirnov test and the Anderson-Darling test ($\alpha = 5\%$). “obs” = sample statistic; p-value calculated from Monte-Carlo bootstrap (10000 iterations).

CATCHMENT	KS_obs	p-value	KS_PASS	AD_obs	p-value	AD_pass
Cervo a Passobreve	0.07	0.71	✓	0.37	0.57	✓
Cervo a Vignano Biellese	0.08	0.68	✓	0.80	0.11	✓
Dora Di Bardonecchia a Beaulard	0.07	0.81	✓	0.56	0.20	✓
Grana a Monterosso	0.06	0.91	✓	0.53	0.23	✓
San Bernardino a San Bernardino Santino	0.10	0.64	✓	0.98	0.05	X
Sangone a Trana	0.05	0.82	✓	0.66	0.13	✓
Germanasca a Perrero	0.07	0.80	✓	0.64	0.13	✓
Erro a Cartosio	0.06	0.79	✓	0.34	0.56	✓
Bormida di Spigno a Piana Crixia	0.06	0.88	✓	0.39	0.47	✓
Bormida di Millesimo a Murialdo	0.03	0.99	✓	0.10	1.00	✓
Ellero a Mondovì	0.06	0.88	✓	0.59	0.17	✓
Sesia a Campertogno	0.09	0.83	✓	1.26	0.02	X
Mastallone a Varallo	0.09	0.60	✓	1.21	0.03	X
Soana a Pont Canavese	0.07	0.71	✓	0.49	0.32	✓
Bormida di Spigno a Mombaldone	0.09	0.69	✓	0.57	0.20	✓
Pesio a Carrù	0.06	0.80	✓	0.60	0.17	✓
Strona di Omegna a Gravellona Toce	0.06	0.86	✓	0.50	0.34	✓
Stura di Viù a Germagnano	0.05	0.91	✓	0.43	0.36	✓
Stura di Lanzo a Torino	0.08	0.82	✓	1.01	0.03	X
Stura di Demonte a Gaiola	0.07	0.80	✓	1.23	0.01	X
Chisone a San Martino	0.04	0.98	✓	0.23	0.84	✓
Tanaro a Piantorre	0.05	0.92	✓	0.47	0.33	✓
Orco a San Benigno Canavese	0.07	0.70	✓	0.54	0.29	✓
Maira a Racconigi	0.06	0.70	✓	0.32	0.60	✓

5.2.2 Which parametric copula can model the data structure?

Before presenting the results obtained with the validation metrics introduced in Section 5.1.4, it is still necessary to address the choice of an appropriate parametric copula to model the dependence structure between D-model predictions and observed discharges. A first indication of suitable copula families is obtained from the empirical tail dependence coefficients, namely the lower- and upper-tail dependence measures (see Section 3.3.2, but here reported again for clarity) λ_L and λ_U , defined as:

$$\lambda_L = \lim_{u \rightarrow 0^+} P\left(Y \leq F_Y^{-1}(u) \mid X \leq F_X^{-1}(u)\right) \quad (\text{Eq. 5.8})$$

$$\lambda_U = \lim_{u \rightarrow 1^-} P\left(Y > F_Y^{-1}(u) \mid X > F_X^{-1}(u)\right) \quad (\text{Eq. 5.9})$$

where X and Y denote two generic random variables (here the D-model predictions and the observed discharges) with marginal quantile functions $F_X^{-1}(u)$ and $F_Y^{-1}(u)$, respectively, and $P(X|Y)$ denotes conditional probability. When considering the full range of D-model predictions, encompassing both low- and high-flow conditions, both lower- and upper-tail dependence will be taken into account. In contrast, for the parametric approach specifically tailored to high-flows, only the upper-tail dependence coefficient λ_U will result relevant.

Figure 42A clearly shows that, when the entire range of D-model discharge predictions is considered, upper- and lower-tail dependence are markedly different and strongly asymmetric. Upper-tail dependence is substantially stronger, with most catchments exhibiting values around 0.7 and only a few cases showing weaker dependence. In contrast, lower-tail dependence is generally weaker but displays a much wider variability across catchments, ranging from values close to zero up to values exceeding 0.4, with a small number of cases exhibiting stronger lower-tail dependence approaching 0.6. As a consequence, for the parametric copula model applied to the full range of D-model predictions, a flexible copula capable of modelling upper- and lower-tail dependence independently is required. Such copulas are typically more complex than traditional single-parameter copulas and are characterized by two parameters (θ_1 and θ_2). Within this class of copulas, commonly referred to as the BB family (Joe H., 1997), the copula selected for the present study is the BB7 copula (also known as the Joe–Clayton copula, see Joe, 1996).

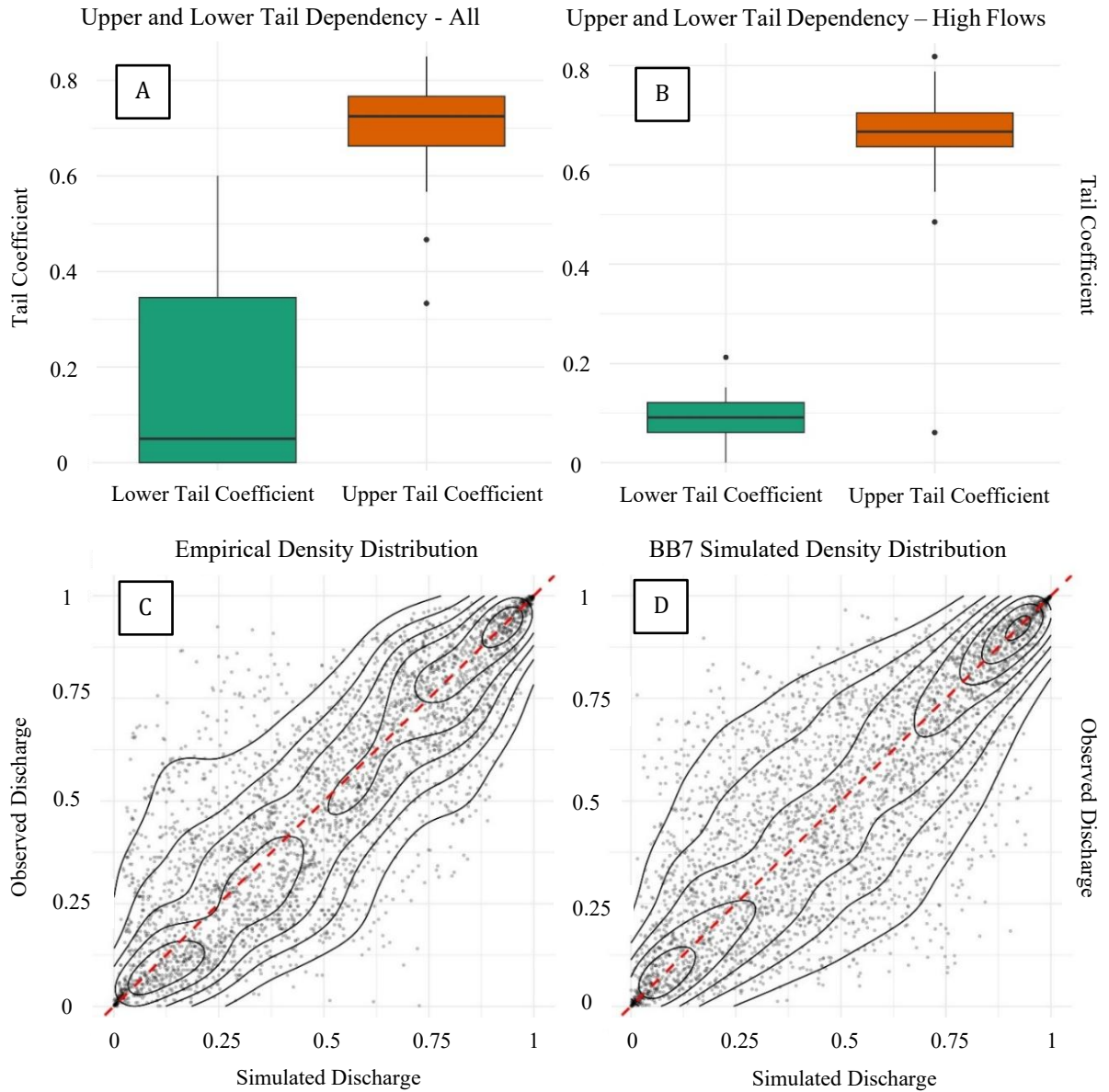


Figure 42. Boxplots for the lower (in green) and upper (in brown) tail dependency coefficients for the 24 catchments of the study: (A) Entire D-model prediction discharge range; (B) D-model prediction discharge range greater than Q_{70} . 2D density plots for the “Cervo a Passobreve” catchment: (C) empirical 2D density for the entire discharge range; (D) simulated 2D density from a simulated sample of equal size from the BB7 copula fitted to the entire discharge range.

This Archimedean copula allows independent control of upper- and lower-tail dependence through its two parameters θ_1 and θ_2 . Its joint cumulative distribution function and tail dependence coefficients are given by:

$$C_{BB7}(u, v) = 1 - \left[1 - ((1 - u)^{\theta_1} + (1 - v)^{\theta_1} - (1 - u)^{\theta_1}(1 - v)^{\theta_1})^{\theta_2} \right]^{\frac{1}{\theta_2}} \quad (\text{Eq. 5.10})$$

$$\lambda_{U, BB7} = 2 - 2^{\frac{1}{\theta_1}} \quad (\text{Eq. 5.11})$$

$$\lambda_{L, BB7} = 2^{-\frac{1}{\theta_2}} \quad (\text{Eq. 5.12})$$

The ability to explicitly control both upper and lower tail dependence makes the BB7 copula particularly attractive in applications where extreme co-movements may differ between high and low values of the variables. Within the BB family, the BB1 copula (Clayton–Gumbel) represents another two-parameter Archimedean copula capable of modelling dependence in both tails. However, the range of tail dependence structures that can be represented by BB1 is more restricted, whereas BB7 provides greater flexibility in capturing strongly asymmetric tail behavior (Durante & Sempi, 2010). The remaining members of the BB family (BB2–BB8) exhibit different tail dependence properties: some allow dependence only in one tail, while others can represent both tails but with more constrained dependence structures. For these reasons, BB7 provides a convenient and flexible specification for modelling asymmetric tail dependence in the present study. An example for the Cervo a Passobreve catchment respectively for the empirical 2D density distribution between D-model predictions and observed discharges and the resulting 2D density distribution from a sample simulated from the fitted BB7 copula is shown in Figure 42C and Figure 42D.

When restricting the analysis to D-model predictions exceeding the threshold Q_{70} (Figure 42B), lower-tail dependence is found to be close to zero for all catchments, while upper-tail dependence remains significant. Based on this observation, a different copula choice was adopted for the parametric approach tailored to high flows. Specifically, a simpler single-parameter copula (θ) was selected instead of a two-parameter BB-family copula. In this case, the Joe copula, another Archimedean copula

capable of modelling only upper-tail dependence, was chosen, as it exhibits stronger upper-tail dependence than other single-parameter copulas such as the Gumbel copula. Its joint distribution and upper-tail dependence coefficient are given by:

$$C_{\text{Joe}}(u, v) = 1 - \left[(1 - u)^\theta + (1 - v)^\theta - (1 - u)^\theta (1 - v)^\theta \right]^{\frac{1}{\theta}} \quad (\text{Eq. 5.13})$$

$$\lambda_{\text{U,Joe}} = 2 - 2^{\frac{1}{\theta}} \quad (\text{Eq. 5.14})$$

All copulas were fitted in R using the “VineCopula” package (Nagler et al., 2024) by maximum likelihood estimation. Model performance was subsequently evaluated by examining directly the resulting S-models using the validation metrics described in Section 5.1.4, rather than by conducting formal goodness-of-fit tests for individual copulas at each catchment.

5.2.3 Predictive Probability Plots and Interval Score

Figure 43 and Figure 44 present the predictive probability plots (PPPs) for the calibration and validation periods respectively. In both Figures, results are already reported (as in Section 4.3) for all catchments simultaneously showing coverages (50% in blue shades and 80% in gray shades) to provide a global idea of the methods performances at multiple sites. Focusing first on Figure 43, which refers to the calibration period, the original non-parametric BLUECAT generally exhibits better performance than the parametric counterpart based on the BB7 copula. However, it should be noted that this behavior is largely a consequence of the construction of the non-parametric method. Because the non-parametric approach resamples exclusively from observed discharge values, it will necessarily exhibit a “perfect” diagonal line in calibration (see the bottom of Section 4.2 of the original paper by Koutsoyiannis and Montanari). In contrast, the copula-based parametric model may already display deviations from the 1:1 line during calibration, depending on how well the fitted copula represents the dependence structure in the data. For this reason, the relative performance of the two approaches should primarily be assessed during the validation period. Nevertheless, the deviations observed for the parametric approach during calibration remain limited, and the reliability of the estimated confidence bands can be considered more than satisfactory for practical applications. At the same time, the parametric approach provides the additional advantages discussed in Section 5.1.2.

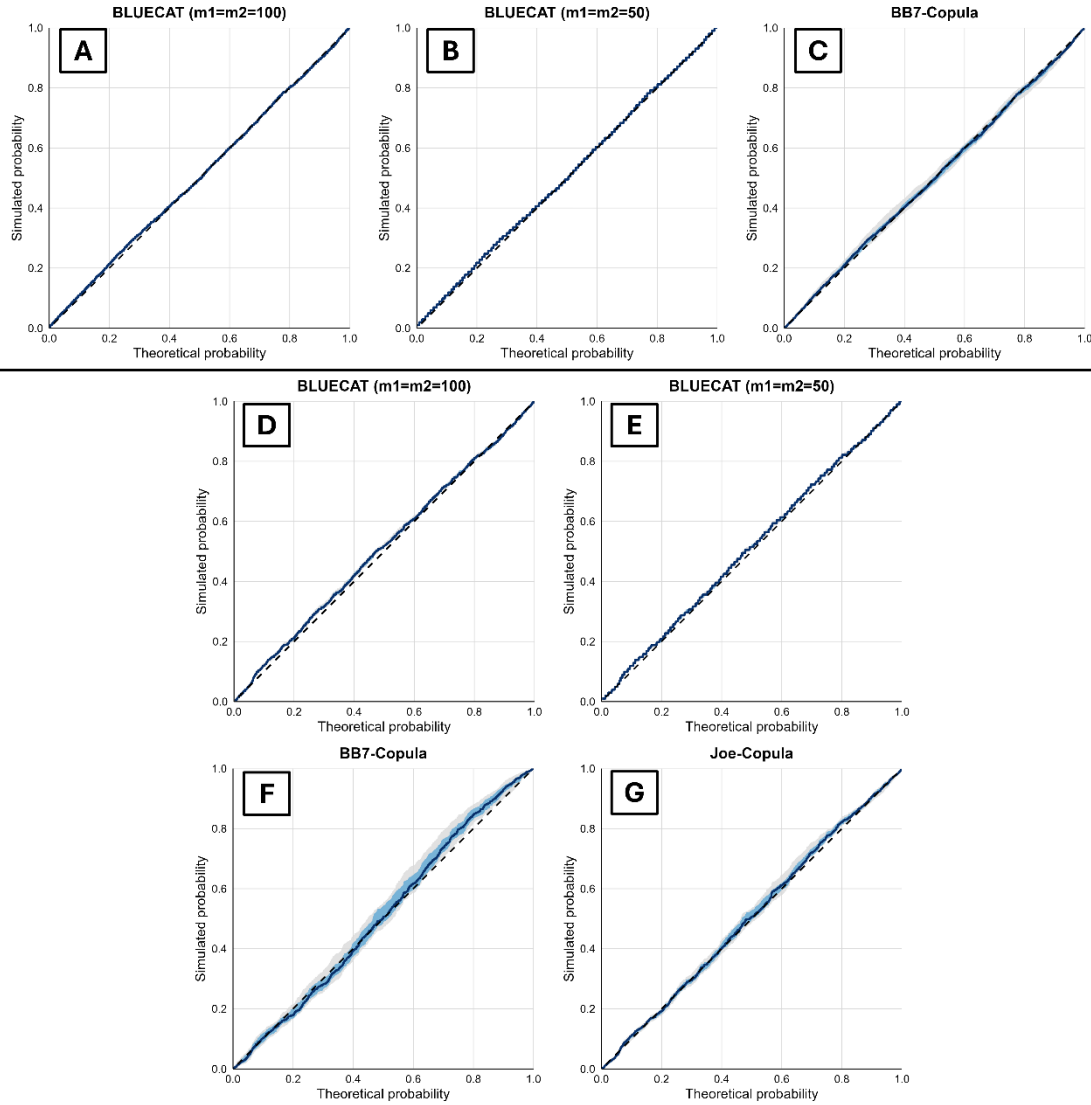


Figure 43. Predictive probability plots (PPPs) for all catchments for the calibration period (2010-2019). (A) non-parametric BLUECAT with $m_1 = m_2 = 100$; (B) non-parametric BLUECAT with $m_1 = m_2 = 50$; (C) parametric copula (BB7, Joe-Clayton); (D) non-parametric BLUECAT with $m_1 = m_2 = 100$, only for D-model $\geq Q_{70}$; (E) non-parametric BLUECAT with $m_1 = m_2 = 50$, only for D-model $\geq Q_{70}$; (F) parametric copula (BB7, Joe-Clayton), only for D-model $\geq Q_{70}$; (G) parametric copula (Joe) tailored for high-flows. Blue shades: 50% catchments coverage. Gray shades: 80% catchments coverage.

Panels (D), (E), (F), and (G) of Figure 43 allow a more detailed analysis of confidence-band reliability in the high-flow range. Here again, the non-parametric method generally performs better, for the same reasons discussed above. However, the parametric method specifically tailored to high flows, based on a single-parameter copula, exhibits performance comparable to that of the non-parametric approach. The differences observed between the two parametric methods can be attributed to the quality of the copula representativeness in the high-flow range. Overall, these results indicate that when the application of interest is restricted to a limited range of D-model predictions, a tailored copula-based approach may provide some benefits. Figure 44 summarizes the performance of the different methods during the validation period. In this case, also the non-parametric approach exhibits deviations from the theoretical 1:1 line (panels A and B). Depending on the catchment, these deviations correspond to either underprediction or overprediction of the confidence bounds. The two-parameter copula-based approach (panel C) shows similar deviations, indicating equal robustness to the non-parametric method. Panels (D), (E), (F), and (G) of Figure 44 show the PPPs restricted to high-flow conditions. In this case, both the non-parametric method and the two-parameter copula-based approach exhibit substantial deviations from the 1:1 line, with the nature of the deviations (under- or overestimation) similar between the two. The single-parameter copula approach tailored to high flows, also presents deviation of similar entity, despite a central tendency slightly more aligned to the 1:1 line, not displaying the desired improvements with respect to the method encompassing the entire discharge range.

Similar results can also be observed in Figure 45, where are reported the interval scores (for the 80% confidence interval) for the entire discharge range (and so the Joe-copula based method is omitted), in this case only for the validation period.

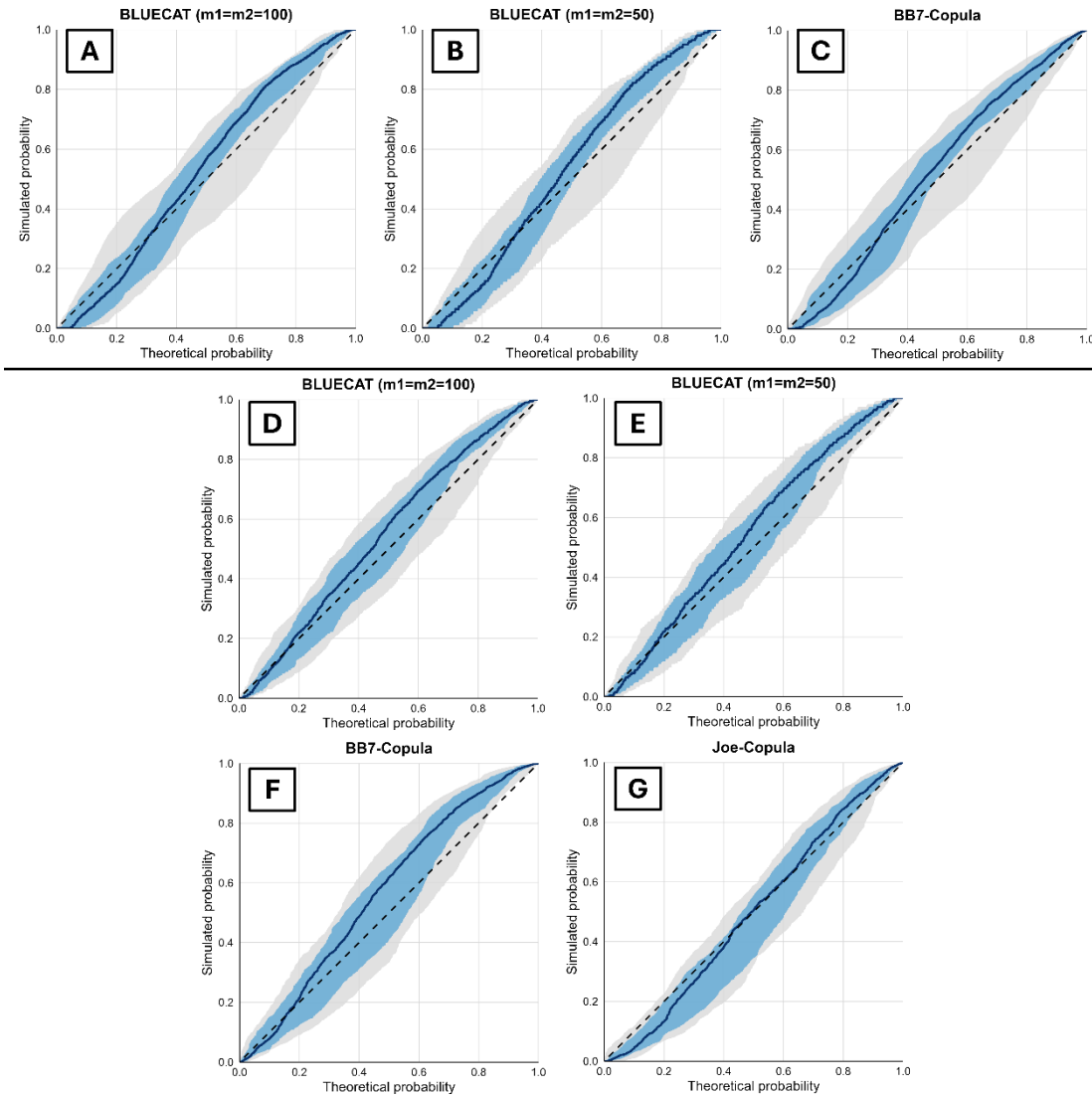


Figure 44. Predictive probability plots (PPPs) for all catchments for the validation period (2000-2009). (A) non-parametric BLUECAT with $m_1 = m_2 = 100$; (B) non-parametric BLUECAT with $m_1 = m_2 = 50$; (C) parametric copula (BB7, Joe-Clayton); (D) non-parametric BLUECAT with $m_1 = m_2 = 100$, only for D-model $> Q_{70}$; (E) non-parametric BLUECAT with $m_1 = m_2 = 50$, only for D-model $> Q_{70}$; (F) parametric copula (BB7, Joe-Clayton), only for D-model $> Q_{70}$; (G) parametric copula (Joe) tailored for high-flows. Blue shades: 50% catchments coverage. Gray shades: 80% catchments coverage.

Remembering how the Interval Score internally gives already more weight to the reliability and sharpness of the CI for high-flows, it should not result surprising that the non-parametric BLUECAT with $m_1 = m_2 = 100$ clearly emerges as the worst performing model in most cases. Despite having a similar reliability to the others, close to extremes, the representativeness of the CIs, and so their sharpness, due to the bounding issues described in Section 5.1.2, is quite low. A similar issue is present also for the case where $m_1 = m_2 = 50$, but to a lesser degree. The copula models instead emerge on top with a slight edge in performance. It should also be noted how the differences between the top model in each catchments usually are very thin, so the results should be interpreted critically, considering how in most cases the classifying between best and second best model is the result of differences in score of just 0.1.

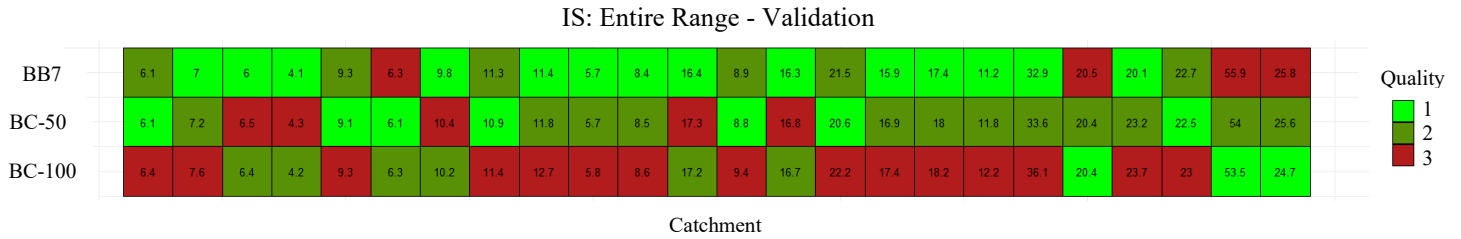
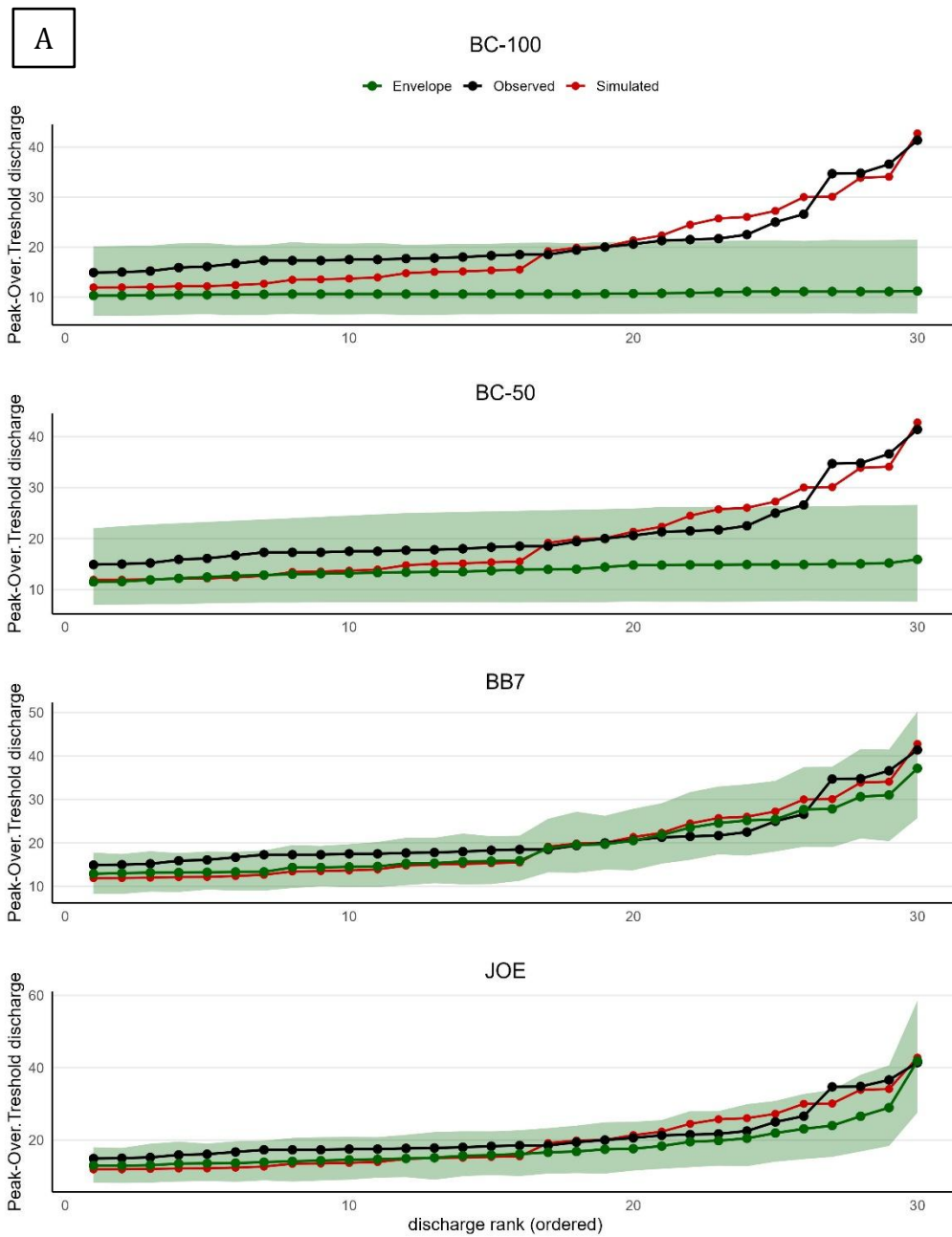


Figure 45. Interval scores (IS) for the 80% confidence interval (CI) for the validation period for the entire predicted discharge range. Colors (from 1, green = “best”, to 3, red = “worst”) indicating Interval Score ranking.

5.2.4 Focus on the peak-over-threshold

Focusing now explicitly on the flood extremes, we can analyze more in detail the effect due to the introduction of the parametric GPD tails for the mixed cumulative distribution functions. It is in fact only for value very close to the highest quantiles that the parametric tails triggers (the metrics for the D-model prediction higher than Q_{70}^S still includes mostly values for which the empirical parts of the mixed CDF are used).



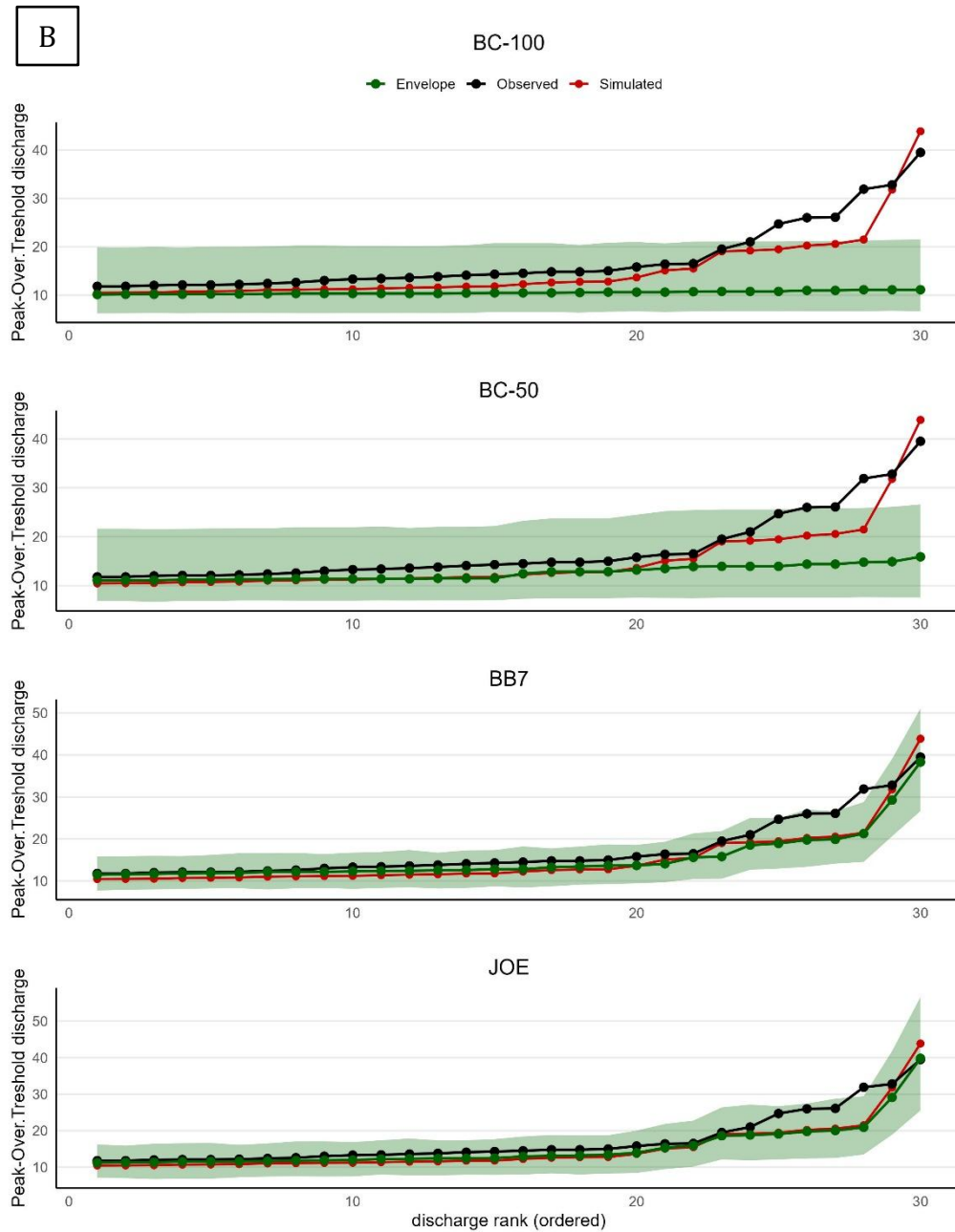


Figure 46. Peak-Over-Threshold (POT) comparison (average of three events for year) between observed discharge (in black), D-model predictions (in red) and S-model predictions (in green, shades indicating confidence interval of 80%) for the BC-100, BC-50, BB7, and Joe methods. (A) Calibration (2010-2019) period; (B) validation (2000-2009) period, for the Grana a Monterosso catchment.

In Figure 46A and Figure 46B is reported an example for the Grana a Monterosso catchment (in calibration and validation). Already in the calibration period, more explicitly than for the sole reliability reported in Figure 44, where it seemed that all methods resulted in similar performances, it can be clearly seen how the original BLUECAT approach resents from the upper boundary issue previously described in Section 5.1.2. The case with $m_1 = m_2 = 100$ resents it more greatly than the case with $m_1 = m_2 = 50$ due to inclusion of further away points from the D-model prediction considered, losing local representatives. The copula based approaches, thanks to the introduced parametric tails, does not suffer instead from this issue, keeping an high reliability also for extreme discharge quantiles, showing informative CI. A similar effect is also visible in the validation period. It must be noted how in validation, despite the bounding effect will always be a factor only for the original non-parametric BLUECAT, there is an additional element that must be considered. We are assuming stationarity of all processes, hence also stationarity of the O-CDF. The direct consequence is that if this assumption (that especially on extremes may be more easily violated, also considering we are only using 10 years in calibration) is not satisfied, we may still have also for the copula-based BLUECAT a lower representatives on extremes due to a poor correction provided by the cumulative distribution function. In the proposed example is not the case (the assumption of stationarity seems to hold nicely), but in other cases it may still be a factor playing a defining role. In calibration instead this fact will never be relevant, because the correction provided is always by definition the real, correct one.

5.2.5 Properties of the ensemble median

We can now examine in greater detail the properties of the ensemble median, defined as the Q_{50} of the estimated confidence bands. Figure 47A and Figure 47B show the Kling–Gupta Efficiency (KGE) recalculated using the ensemble median for the calibration and validation periods, respectively, for the different methods considered. The single-parameter copula approach is excluded from this analysis because it does not generate an envelope over the full discharge range, and therefore the median is not continuously defined.

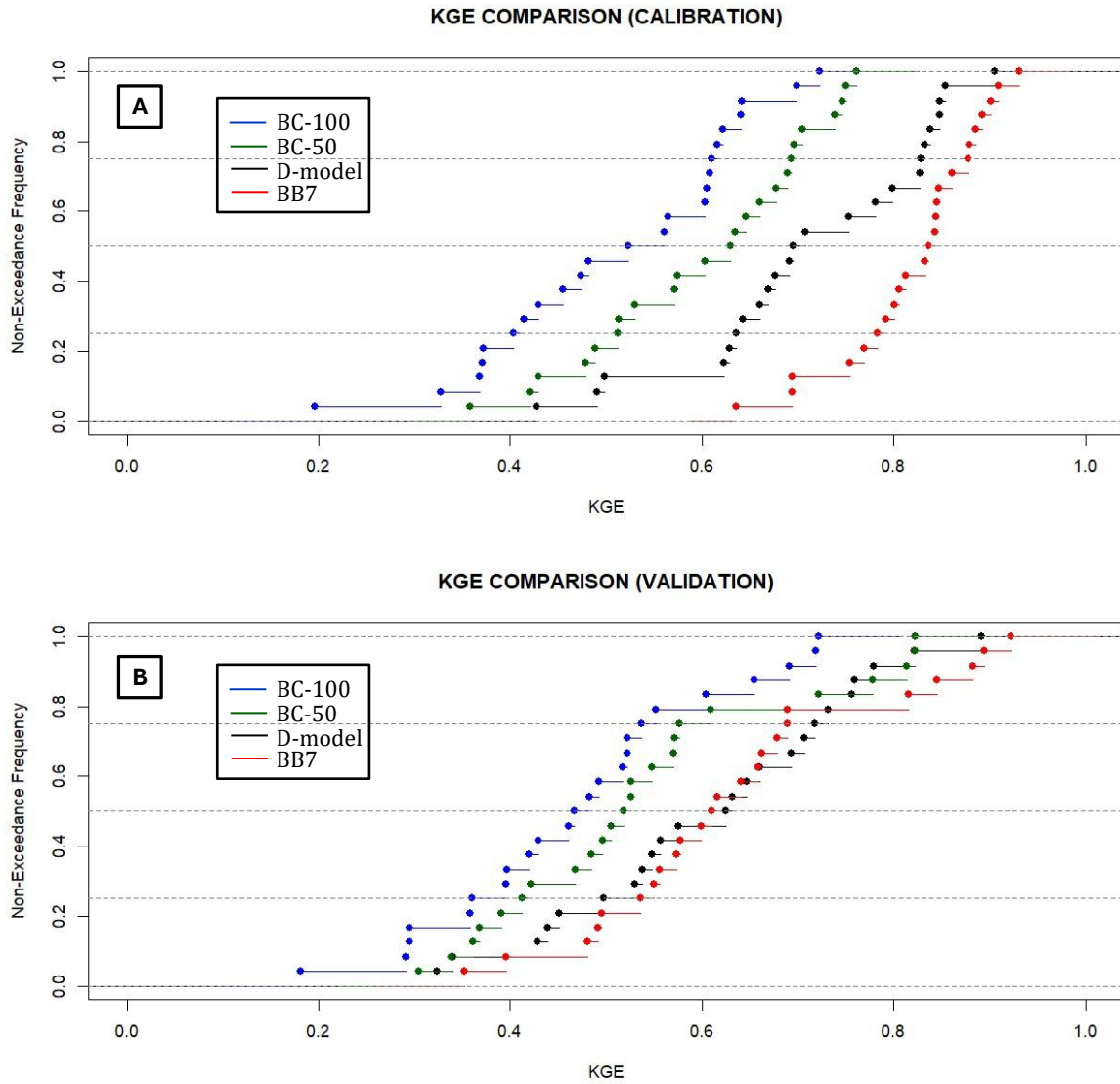


Figure 47. KGE cumulative distribution function for the 24 catchments using the D-model predictions (in black), the S-model median from non-parametric BLUECAT for different $m_1 = m_2$ (blue, 100; and green, 50), and the S-model median from the two parameter BB7 model parametric BLUECAT (in red). (A) calibration period (2010-2019); (B) validation period (2000-2009).

The results for the original non-parametric BLUECAT clearly indicate a degradation in performance, as measured by KGE, when compared to the original D-model predictions prior to the estimation of confidence bands. This reduction in performance

can again be attributed to the bounding issue discussed in Section 5.1.2. By construction, the non-parametric approach cannot include in the ensemble values exceeding the maximum observed discharge in the calibration record. As a direct consequence, the ensemble median tends to strongly underestimate observed extreme flows. Because the KGE (and similarly other efficiency metrics like the NSE) is particularly sensitive to the reproduction of high flows, penalizing errors in extreme high-flow conditions more strongly than errors in low-flow conditions, this degradation in performance is an expected outcome, at least given the available length of the observed record. In contrast, the parametric BLUECAT approach clearly outperforms the non-parametric one in this respect. A stronger improvement is observed in the calibration period (because the correction provided by the O-CDF is the real one), while in the validation period KGE performances between D-model predictions and median S-model predictions are instead similar. This result can be interpreted as an additional indication of the advantage of employing the S-model generated from the parametric approach to the sole D-model: in the worst performing period (validation), the median of the S-model still has, when considered as a standalone series analogous to the D-model, equal performance, but the S-model provides the additional information about the estimated CI. This was exactly the objective: there is no reason to employ the D-model instead of the S-model, because in all cases (for all signatures), the latter has equal or better performances, allowing overall for better informed applications.

5.3 Did the new paradigm pay off?

5.3.1 Not better, not worst (but for floods)

Overall, the results in Section 5.2 highlight both the strengths and limitations of the different BLUECAT formulations, and clarify the conditions under which each approach is expected to perform best. The analysis of tail dependence coefficients revealed a pronounced asymmetry between upper and lower tail dependence when considering the full range of D-model predictions, indicating that flexible copula families capable of independently modelling both tails are required. This finding justifies the adoption of a two-parameter BB7 copula for the general parametric formulation. Conversely, when restricting the analysis to high-flow conditions, lower tail dependence becomes negligible across all catchments, while upper-tail dependence

remains significant, supporting the use of a simpler single-parameter copula specifically tailored to the upper tail.

Regarding instead the parametric fit of the upper tail of the cumulative distribution function of the observations, the Kolmogorov-Smirnov test did not reject the null hypothesis at any of the 24 sites, in any case considered. In contrast, the Anderson-Darling test rejected the null hypothesis in few cases across sites. For those instances, reassessing the Anderson-Darling statistic using a more stringent significance level ($\alpha = 1\%$) resulted in non-rejection of the null hypothesis instead. These results overall indicated how the parametric fit can be considered satisfactory and representative for the extrapolation beyond observed discharge maxima.

The evaluation based on predictive probability plots confirms that the non-parametric BLUECAT exhibits near-ideal behavior during calibration, but this result is largely structural, as the method resamples directly from the observed data and therefore enforces a perfect probabilistic consistency by construction. As a consequence, calibration results alone are not informative for comparing parametric and non-parametric approaches. Nevertheless, confidence bounds reliability in validation show similar performance for the parametric approaches, with a slight improvement for high D-model predictions quantiles for the model tailored for high-flows. Similar conclusions are supported by the analysis of interval scores.

In general, the non-parametric approach systematically suffers from the bounded-resampling issue, which leads to an underestimation of extreme discharges and, consequently, to a degradation of performance metrics that are sensitive to high flows, such as the KGE or the peak over threshold validation. In contrast, the parametric BLUECAT framework overcomes this limitation by allowing extrapolation beyond the observed range, resulting in a substantial improvement in the representation of extreme events and in the performance of the S-model median. Taken together, these results demonstrate that while the non-parametric approach remains attractive for its simplicity, the parametric copula-based formulations provide a more robust and physically consistent characterization of predictive uncertainty, especially when the application focuses on high-flow conditions or on out-of-sample performance.

5.3.2 About the generality of the copula structures

The presented analysis was restricted to a set of alpine catchments that were relatively homogeneous from both a climatological and a geomorphological perspective. As a consequence, the dependence structures inferred through the copula-

based BLUECAT formulation should be interpreted within this specific context, and their generality beyond mountain environments cannot be assumed a priori. In particular, all the investigated basins exhibit perennial flow regimes, while ephemeral catchments were deliberately excluded from the analysis. In such systems, the presence of a large proportion of zero or near-zero discharges would likely induce strong features in the lower-left quadrant of the copula, potentially requiring dedicated mixed or discrete–continuous formulations to adequately represent the joint behavior of simulated and observed flows. For basins characterized by markedly different hydro-climatic conditions (e.g. snow-free regimes, catchments with weaker seasonality or arid catchments), changes in the dominant sources of uncertainty and in the error structure of the D-model may result in different dependence patterns between predictions and observations. Although these aspects are not explicitly explored here, they are conceptually consistent with the previous findings on the statistical properties of hydrological model residuals across contrasting catchment types explored in Chapter 3. While that analysis focused on residual distributions rather than on the joint structure between simulated and observed flows, both perspectives are intrinsically linked, as the copula implicitly encapsulates information on model bias, heteroscedasticity, and tail dependence. In this sense, the copula families identified as suitable in the present study, mainly flexible two-parameter Archimedean models, should be regarded as effective within the explored domain rather than universally optimal. Extending the methodology to more diverse catchment classes would likely require reassessing the appropriate copula family and, potentially, adopting different dependence structures for distinct hydrological regimes. Nevertheless, the general copula-based BLUECAT framework remains fully applicable, as it does not impose any restriction on the copula choice itself, thus preserving methodological flexibility while allowing the dependence structure to adapt to the statistical characteristics of the target catchment. Future research could explicitly address this aspect by applying the proposed framework across contrasting hydro-climatic regimes and flow persistence conditions, with the aim of systematically comparing the resulting dependence structures and assessing their degree of generality or regime-specific behavior.

5.3.3 Computational trade-offs

From a purely practical point of view, one of the strengths of the original non-parametric BLUECAT methodology lies in its extremely light computational requirements. In that formulation, for any point D-model prediction, once the m1 and

m_2 values are selected, the subset of observations used to estimate confidence intervals or reference statistics can be identified directly, without the need for iterative procedures or repeated sampling. As a result, uncertainty quantification is obtained at a negligible additional computational cost with respect to the deterministic simulation.

The copula-based version, on the other hand, is inherently more demanding from a computational perspective. In this case, uncertainty estimation relies on the explicit modelling of the joint distribution between D-model predictions and observed discharges, and on the derivation of the corresponding conditional predictive distribution for each individual D-model prediction. For a given D-model prediction, this typically requires generating a sample from the conditional copula distribution, from which confidence intervals and summary statistics are then inferred. This Monte-Carlo sampling step is unavoidable for most copula families, as the conditional probability density function or its quantiles are generally not available in closed analytical form. The computational burden further depends on the specific copula family adopted. Single-parameter copulas, such as Joe or other Archimedean families, allow relatively fast conditional simulation due to their simpler functional structure. In contrast, two-parameter copulas from the BB family, which are particularly effective in representing asymmetric tail dependence, require more complex numerical operations to generate conditional samples, resulting in longer execution times. This increase in cost is therefore the direct consequence of the greater flexibility of these copulas in reproducing extreme-value dependence, rather than of an inefficient implementation.

Despite these higher requirements, the overall computational effort of the copula-based approaches remains moderate in practical applications. Execution times are typically well below those that would limit operational uses, and are easily compatible with offline analyses or near-real-time forecasting frameworks. In this respect, the additional cost can be considered acceptable when weighed against the improved reliability of the estimated confidence bands observed especially for the high-flow tailored method. The copula-based formulations thus represent a reasonable trade-off between computational efficiency and the need for a more reliable characterization of predictive uncertainty.

5.3.4 Toward regionalization

Following the results and the discussion above, this final section wants to be the bridge between the local applications and the regional application explored in the

following Chapter 6. The alternative parametric approach explored with the copula-based BLUECAT, in contrast to the C-method proposed in Chapter 4, presents more desirable features thinking about a possible regionalization. First, its reliability in defining CI, especially in validation, is comparable with the results obtained with the non-parametric approach (with improvements for flood-extremes). If on one hand it would have been desirable to improve on the non-parametric BLEUCAT approach (something that we need to consider disproved at least as far as it concerns the cases here examined), on the other hand not having worst performances opens up to the idea of regionalizing this parametric model.

Remembering the additional requirements listed in Section 4.1.2, we can now see how this method satisfies all of them: (I) the model is largely parametric, or mixed-parametric for the involved CDFs; (I) the model structure is rather simple, just composed by two cumulative distribution functions (O-CDF and S-CDF) and the two (or one) parameter copula; (III) the structure is generalizable locally across multiple sites: in fact all catchments have been reasonably described by the same BB7 (or Joe for the high-flow tailored application) copula. The fact all these requirements are satisfied, is well different from saying regionalizing the model is simple and/or feasible, but the challenges involved in this process will now be explored in the following Chapter 6.

Chapter 6

6.Regional estimation of conceptual RR model uncertainty

6.1 PP-methods regionalization perspectives

6.1.1 A short history of uncertainty estimation in ungauged catchments

Up to this point, Chapter 4 and Chapter 5 have focused on the development, implementation, and validation of non-parametric and parametric approaches for estimating confidence intervals associated with deterministic conceptual RR model predictions in gauged catchments. The present chapter shifts the perspective to the regional estimation of CIs at ungauged sites. Predictions in ungauged basins have been at the core of hydrological research for several decades, most prominently within the framework of the PUB (Blöschl et al., 2013) initiative of the International Association of Hydrological Sciences. However, while substantial progress has been achieved in the regionalization of full conceptual RR models and of specific hydrological signatures, the explicit regional assessment of predictive uncertainty at ungauged locations remains comparatively underexplored. This is somewhat surprising, as providing uncertainty information together with regionalized deterministic (D-model) predictions would represent a natural and necessary step toward a complete characterization of predictive performance at sites where no discharge measurements are available.

As with any regionalization problem, the underlying assumption is that catchments within a region share common behaviors and governing principles that can be transferred in space. This assumption is intuitively acceptable for physically based processes, such as infiltration, percolation, or snowmelt, and for their relationships with physiographic catchment attributes. It may appear less straightforward when applied to

residual properties, which are not physical processes per se. Nevertheless, as shown in the previous chapters, residual behavior is closely linked to hydrological dynamics and model structural deficiencies. From this perspective, the idea of regionalizing residual properties and their statistical structure is not conceptually inconsistent. In this case, the underlying (strong) assumption is that ungauged catchments will also produce similar error structures to donor catchments based on proximity or similarity criteria.

Only few studies have explicitly addressed the estimation of confidence bounds for regionalized D-model predictions at ungauged sites. A GLUE-type approach consisting of transferring ensembles of parameter sets obtained on donor (gauged) catchments to target (ungauged) catchments was developed by McIntyre (McIntyre et al., 2005a). An alternative approach is instead based on a framework based on constrained parameter sets (Yadav et al., 2007; Zhang et al., 2008). Bayesian approaches can also be used (Bulygina et al., 2011, 2012). All the mentioned methods, as noted by some authors (McIntyre et al., 2005b; Winsemius et al., 2009), suffer from a difficult interpretation of the uncertainty bounds, as they cannot be easily interpreted as confidence intervals, and therefore their practical use remain problematic. Moreover the reported examples rely uniquely on ensemble of model parameters, not accounting for the several additional sources of uncertainties discussed in Chapter 1.

Within the methods that try to use PP approaches to estimate uncertainty in ungauged catchments, to the best of our knowledge, only one approach is available in the literature, namely Bourgin et al., 2015, later employed in other works (Farmer & Levin, 2018; Levin & Farmer, 2020). In their work, the authors proposed a multiplicative error scheme in which empirical error series from neighboring donor catchments were resampled and transferred to the ungauged site. As later discussed in Section 6.3.1, their approach was fully non-parametric and did not involve the transfer of parameters from a parametric error model through regional regression or similar techniques. When presenting their results, the authors acknowledged the intrinsic difficulty of evaluating the quality of the estimated confidence bands, given the absence of established benchmarks for reliability and sharpness of CIs in ungauged settings.

In line with the developments of the previous chapters, the objective here is instead to explore the regional estimation of CIs at ungauged sites starting from a (partially) parametric PP framework, in which model parameters are transferred through conventional regionalization tools (e.g., regional regression relationships). Within the methodologies investigated in this thesis, the best suitable candidate for such an approach is the parametric BLUECAT (from now on referred to as P-BLUECAT)

framework. As previously demonstrated in Chapter 4, the other alternative, represented by the C-method, while proved equally competitive than non-parametric approaches when constrained to relatively simple and generalizable structures, still presented several limitations in reproducing the desired residuals statistical properties. Although improvements would in principle be possible, they would require substantially more complex model formulations, which would be difficult to generalize across sites and, consequently, problematic to regionalize. Despite their conceptual simplicity relative to fully stochastic residual error models, the P-BLUECAT approaches still entail non-trivial challenges when considered from a regionalization perspective. This chapter does not aim to provide a definitive or fully operational solution to the problem. Rather, it seeks to outline the main conceptual and practical issues involved, discuss the potential for regional transferability, and present preliminary results that may serve as a foundation for future developments.

The remainder of the chapter is organized as follows: Section 6.2 introduces the general framework and key challenges associated with the regionalization of the P-BLUECAT methodology. Section 6.3 will briefly present the approach of Bourgin et al., which is subsequently implemented and used as a benchmark for comparative analysis, following the same rationale adopted in Chapter 4. Section 6.5 reports the results of the regionalization experiments, and Section 6.6 concludes with a broader discussion of the findings and their implications.

6.2 Regional P-BLUECAT model

6.2.1 Model components and challenges

The regionalization of the P-BLUECAT framework, without, at this stage, distinguishing between the one-parameter and two-parameter copula formulations, given their shared operational structure, can be decomposed into three main components. These correspond to the sequential elements required for model application, as previously schematized in Figure 41. First, when moving to ungauged locations, the deterministic discharge predictions must be obtained through a regional model, since local calibration is not feasible in the absence of observed discharge data. From a methodological standpoint, this represents the least problematic step. Throughout this thesis, it has been shown repeatedly that regional RR models can be derived and applied in a consistent manner. In the present chapter, preliminary results will be shown for a regional model developed again using the PASS algorithm, with the key difference that it is implemented within a Leave-One-Out (LOO) framework.

The rationale for adopting a different model in this context will be detailed in Section 6.2.3.

Second, the copula structure linking deterministic predictions and observations must also be regionalized. In ungauged settings, although D-model predictions can be obtained from a regional model, observed discharge is unavailable; consequently, the dependence structure between predictions and observations cannot be locally estimated. Within a parametric framework, the regionalization of the copula can be approached in a conceptually straightforward manner. Following standard practice in regional studies, one may attempt to relate copula parameters to catchment descriptors (e.g., climatic and physiographic attributes) and derive regression relationships to transfer parameter estimates from gauged to ungauged sites. While this strategy is methodologically clear, its practical success is not guaranteed. Meaningful regressions may not exist, or may lack sufficient robustness for reliable transfer. Implicitly, this step assumes that the error structure captured by the copula parameters is, at least partially, controlled by measurable catchment characteristics and therefore exhibits spatial organization within the study region. Section 6.2.4 further examines this issue.

The third component concerns the regionalization of the observed discharge cumulative distribution function (O-CDF), which is required to transform samples drawn from the conditional copula distribution, defined in the unit interval, back to discharge values in physical space. This step introduces both operational and conceptual challenges. As will be discussed later from Section 6.2.5 to Section 6.2.7, it likely represents the most critical and impactful element of the entire regionalization procedure, since it directly affects the shape and reliability of the resulting confidence intervals.

6.2.2 Intertwining effects on S-model location and width

Among the three components introduced in Section 6.2.1, the regionalization of the copula structure and of the O-CDF directly controls both the location and the width of the resulting confidence intervals. With respect to the S-model location (e.g., its mean and thus its bias), the O-CDF plays a central role. As discussed previously in Chapter 5, in the local (gauged) configuration the empirical O-CDF derived from observed discharges provides the reference distribution required to correct potential biases in the D-model predictions. Assuming reliable discharge measurements, this correction can be regarded as statistically consistent. In a regional context, however, the O-CDF itself must be estimated and transferred to the ungauged site. Any global or

conditional bias affecting this regionalized O-CDF will inevitably propagate to the S-model outputs, thereby shifting the estimated confidence bounds.

Concerning the width of the confidence intervals, both the regionalized copula parameters and the regionalized O-CDF exert a controlling influence. The role of the copula parameters is relatively straightforward. Different parameter values define different dependence structures and thus different joint probability density functions. This, in turn, leads to different conditional distributions and consequently to different spreads in the samples drawn from them. The influence of the O-CDF on interval width is more subtle and relates to the local slope of the cumulative distribution function. Consider, for example, a case in which samples drawn from the conditional distribution in the unit interval lie between 0.6 and 0.8. Suppose two alternative O-CDFs share the same mean (i.e., no overall bias, or the same absolute bias), but differ in their local slope within that probability range. An O-CDF with a steeper slope in that interval will map the sampled probability range into a wider range of discharge values, resulting in broader confidence bands. Conversely, an O-CDF with a flatter local slope will compress the same probability range into a narrower discharge interval, producing tighter bands. Disentangling, in a stepwise manner, the individual effects of the different regionalized components, particularly the copula parameters and the O-CDF, is not straightforward. The width of the confidence intervals is jointly controlled by both elements, and their effects are intrinsically intertwined. Nevertheless, decomposing the overall regionalization problem into its constituent parts is conceptually useful, as it allows one to isolate and better understand the specific contribution of each component.

To this end, Section 6.5 will present the results of a “thought experiment” designed to partially separate these effects. In a first step, only the impact of copula regionalization is considered. This is achieved by assuming the existence of a hypothetical, perfectly regionalized O-CDF. Under this assumption, within the regional framework only the D-model predictions and the copula parameters are inferred, while the empirical CDF derived from the actual observations is retained for the back-transformation from the unit interval to discharge space. Clearly, this assumption is not realistic. The regionalization of a cumulative distribution function (or equivalently of a flow duration curve) is itself a well-known and non-trivial problem, extensively discussed in the literature (Blöschl et al., 2013; Castellarin et al., 2007; Castellarin, Galeati, et al., 2004; Castellarin, Vogel, et al., 2004), with documented limitations and sources of uncertainty. The assumption of a perfectly known O-CDF is therefore

adopted solely as an intermediate analytical step, before subsequently evaluating the additional impact introduced by its explicit regionalization.

6.2.3 Leave-One-Out (LOO) regional model

The regionalization experiments are carried out on the same set of 24 mountainous catchments introduced in Section 4.2.3 and shown in

Figure 29A. In the present context, however, each catchment must be considered alternately as gauged or ungauged depending on the specific model component being regionalized. Although this distinction may initially appear counterintuitive, it is essential for ensuring conceptual consistency in the transfer procedure. To clarify, consider catchment C1 as the target ungauged site and the remaining 23 catchments as potential donors. When regionalizing the D-model, the donor catchments are treated as gauged: their model parameters are estimated through local calibration and subsequently transferred to the target site using the selected regionalization approach. The same logic applies to the O-CDF component. Since the objective is to provide the most reliable possible correction to the S-model, the empirical O-CDF derived from local observations represents the most informative basis for parameter estimation at donor sites.

The situation differs when regionalizing the copula parameters. One might be tempted to estimate copula parameters at donor catchments under locally calibrated (gauged) conditions and then transfer them to the target site. However, this would introduce an inconsistency. It is well established that D-model performance in ungauged conditions is generally inferior to that obtained through local calibration (see Chapter 2). Consequently, the dependence structure between D-model predictions and observations, on which the copula parameters are based, will differ between gauged and ungauged configurations. In ungauged conditions, the scatter around the 1:1 line will be typically larger, reflecting increased predictive uncertainty and, therefore, wider confidence intervals. Transferring copula parameters estimated under locally calibrated conditions would implicitly assume a stronger dependence structure than the one that can realistically be expected at an ungauged site. This would result in overly narrow confidence intervals and a systematic underestimation of predictive uncertainty. To avoid this inconsistency, copula parameters must be estimated at donor catchments under ungauged-like conditions. Operationally, this means that each donor catchment is treated as if it were ungauged, by using out of sample regional D-model predictions (the configuration used in Chapter 4 and Chapter 5 was regional, but not out of sample)

rather than locally calibrated ones, while still relying on observed discharge for parameter estimation. In other words, the copula parameters must reflect the dependence structure expected in ungauged applications, not in locally calibrated settings. A similar rationale was adopted by Bourgin et al., and the reader is referred to Section 3.2 of their paper for further details. Conceptually, this reasoning parallels the temporal separation between calibration and validation discussed in Section 4.2.5, although here the separation is applied in space rather than in time. These considerations imply the need for multiple regional RR models. Specifically, 24 distinct regional models are constructed, each excluding one catchment in turn. In each case, the omitted catchment is treated as ungauged, and the resulting model provides ungauged D-model predictions for that site. Collectively, these Leave-One-Out (LOO) models yield a complete set of ungauged D-model predictions for all 24 catchments.

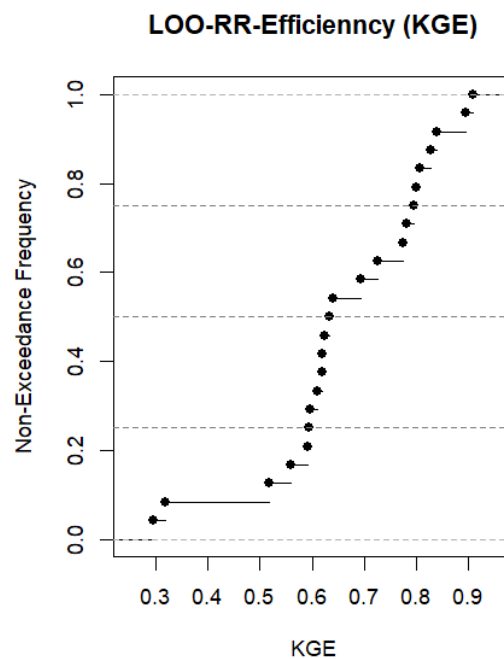


Figure 48. Leave-One-Out (LOO) rainfall-runoff model efficiency (KGE) for the calibration period (2010-2019). Each point corresponding to the performance of the out-of-sample D-model predictions at each ungauged site obtained using the remaining 23 catchments as donors.

At the same time, each regional model also produces D-model simulations for the other 23 catchments included in its calibration; however, these simulations correspond to

regional (not out of sample) predictions. Depending on the model component under consideration (D-model, copula parameters, or O-CDF), the appropriate set of predictions (gauged or ungauged) will be used. In the remainder of this chapter, all results will refer to this LOO regional modeling framework. The performance of the 24 LOO regional D-models is summarized in Figure 48. For all the sections related to the regionalization procedure, we will only focus on the calibration period (2010-2019). Extending the results for the validation period would be possible following similar operational practices as for the previous sections, but considering here the focus is on the perspectives and feasibility of the regionalization, we limited the analysis on the less challenging scenario of just spatial, and not simultaneously temporal, extrapolation.

6.2.4 Regionalization of the copula parameters

The first component to be addressed for a fully operational regional implementation of the copula-based uncertainty framework is the regionalization of the dependence structure linking D-model predictions and observed discharge. To assess the feasibility of meaningfully regionalizing the parameters of the copula, an approach analogous to that adopted in Section 2.3.4 is followed. Specifically, random trees, already employed within the PASS algorithm for the regionalization of conceptual RR model parameters, are used to infer the copula parameters based on catchment similarity. Although this strategy does not yield explicit analytical regressions between copula parameters and CDs, it offers the advantage of capturing potentially complex and non-linear relationships that would remain unrepresented under simple linear regression schemes. In this application, random trees are not implemented within an iterative optimization framework aimed at maximizing a predefined efficiency metric. Unlike the regionalization of RR model parameters, where multiple candidate parameter sets can be explored and shuffled, here only one copula parameter set is estimated per catchment. Consequently, no iterative reshuffling procedure is feasible. The operational adequacy of the regionalization is therefore evaluated indirectly, by examining the reliability of the confidence intervals obtained at ungauged sites. It is worth recalling that, within the thought experiment introduced earlier, the influence of copula parameter regionalization can be isolated from that of the O-CDF regionalization. This allows the effect of transferring copula parameters alone to be assessed, without the confounding contribution of errors arising from the regionalization of the discharge distribution.

6.2.5 Conceptual dilemma of regional O-CDF

If the first two elements of the uncertainty estimation copula-based model to be addressed for a full operational regional application are “trivial” (regional D-model and copula-structure), the second (regionalization of the O-CDF) is instead quite complex. First, it necessary to discuss a conceptual dilemma: does it make any sense to regionalize the O-CDF in the first place? One could argument that correcting any bias (absolute or conditional) in the D-model predictions through a regional O-CDF is somewhat counterintuitive, because any existing bias in the regional model is the direct result of a fault in the regionalization procedure, and should be fixed during RR model regionalization instead of being accounted for in the post-processing phase. In other words it may be considered odd to introduce an additional external model to fix the bias of the first one. In the local application this operation was less counterintuitive because the correction was performed using the real observation O-CDF, and the bias correction where limited to conditional biases, assuming the employed RR models where reasonable and well calibrated and so did not presented any absolute bias. If this position is somewhat reasonable from a theoretical point of view (we need to improve our RR model regionalization knowledge and performances at the source), it struggles against current practical applications. If it was possible to fix biases (at least absolute) during the regionalization procedure, we would do so, but the best methods today available often fail in such effort. The RR model regionalization is based on the regional optimization of the same metrics, like the KGE, used in calibration, still carrying on all their limitations. If you attempt to find regional regressions to relate biases in LOO predictions at the different sites to catchment descriptors, you would not find any meaningful one. In other words, at the current stage of practice, there is no clear hydrologically based reason why in regionalization some sites may suffer from absolute positive biases, why some other from absolute negative biases, and why some others from no absolute biases at all. In this context, despite not being the theoretical most elegant solution, the use of an external model to better infer the O-CDF to provide a more robust bias correction to the LOO D-model regional predictions seems reasonable. The idea is that the regional LOO D-model will provide the information about the temporal evolution of the series, the regional copula structure part of the information about the uncertainty, and the regional O-CDF from the external model a more robust correction than not applying none (that is equivalent to use the D-model prediction CDF as observation O-CDF for the correction). The D-model provides something that a regional O-CDF alone would not provide (temporal patterns), while

the regional O-CDF could provide what the RR regional model fail to reproduce accurately (biases correction). We use the conditional because this hypothesis must be tested (it is indeed possible that the regional O-CDF does not perform much better than the regional RR model predictions CDF in ungauged basins), but has the advantage to be able to rely to a large literature regarding such problem. The problem of regionalizing a flow duration curve ($FDC = 1 - CDF$, from now on the two will be used interchangeably) is in fact being long studied and several studies are already available, differently from regional uncertainty estimation for which very few are available instead. We do not need to develop from zero such external model for the CDF, we can borrow from previous knowledge. Following this approach, Section 6.5 will investigate to what extent the regionalization of the O-CDF with a dedicated model could better perform compared to the use, as O-CDF, of the regional LOO D-model predictions CDF.

6.2.6 Regional correction: which CDF?

The next question to be asked is which CDF must be considered. Several CDFs are usually considered in hydrology, from mean annual CDFs to period-of-record. In theory, the CDF we would need to regionalize is the period-of-record because we want to keep track of all values (especially extremes). At the same time, regionalization method for a CDF are well known to be more robust for mean annual CDFs. Unfortunately, a mean annual CDF will behave somewhat similarly to a period-of-record CDF only for mid-flow discharge ranges, but will severely overestimate and underestimate respectively the low-flow and high-flow ranges, especially the closer you are to the extremes (inducing strong conditional biases in the resulting S-model). Again, here two possibilities are open, depending on the goal you want to achieve with the regional uncertainty model: if the focus is about average conditions, a regional mean annual CDF may be reasonable and suggested (be aware of the inevitable misrepresentation close to extremes), but in our case, where the interest is also on floods, we cannot rely on the simpler CDF, but we need to attempt the regionalization of the period-of-record, that will be the focus from now on.

6.2.7 BB7 copula model: regionalization of the period-of-record CDF

The cases for the BB7 and Joe copula models will be now discussed separately. For the BB7 copula model we need to regionalize the full period-of-record (from now on POR) O-CDF. There is an extensive body of literature addressing the regionalization

of flow duration curves, proposing alternative methodological philosophies and a variety of techniques within each framework. Comparative assessments of regionalization strategies have been conducted, among others, by Castellarin (Castellarin et al., 2004) and within the PUB synthesis volume edited by Günter Blöschl (Blöschl et al., 2013). Depending on the adopted classification scheme, regionalization methods can be broadly grouped into parametric, statistical, and graphical approaches. A key outcome of the PUB initiative was that the overall performance of these approaches tends to be comparable, with no consistent evidence of clear superiority of more complex or fully parametric methods over simpler alternatives, with some improvements for more complex methods limited to the low-flow discharge range. Similar conclusions were drawn by Castellarin et al. in their comparative analysis. In light of these findings, the results presented here rely on the regionalization approach proposed by Smakhtin (Smakhtin et al., 1997). Although this method is not fully parametric, it provides a transparent and straightforward implementation while delivering performance levels that, according to the comparative literature, are expected to be similar to those achievable with more statistically sophisticated techniques. The choice of this approach is therefore motivated by parsimony and operational simplicity, rather than by expectations of improved predictive skill from more complex models.

The method represents the POR-CDF at an ungauged site as the product of two components: (I) a regional dimensionless POR-FDC, and (II) a dimensional scaling factor, here taken as the mean annual flow (MAF) of the target catchment. Both components are inferred from a set of donor catchments.

The regional dimensionless POR-CDF is obtained through the following steps:

- For each donor site, the empirical POR-CDF and the corresponding mean annual flow are first computed. A dimensionless POR-CDF is then derived by normalizing the POR-CDF by the local MAF;
- The regional dimensionless curve for the ungauged site is subsequently estimated as either the arithmetic mean or a weighted mean of the donors dimensionless curves. When weighting is applied, weights are assigned according to catchment similarity: the Euclidean distance between the target and each donor is computed in the space of selected catchment descriptors, and weights are defined as inversely proportional to this distance;

- The MAF at the ungauged site is estimated separately using a simple linear regression model. The final regionalized POR-CDF is then reconstructed by multiplying the regional dimensionless curve by the estimated MAF;

Two alternative regionalization configurations are tested. In the first, the 24 catchments within the study area are used exclusively as potential donors. Each site is treated in turn as ungauged, while the remaining 23 serve as donors; the dimensionless POR-CDF and the MAF are regionalized accordingly, and the procedure is repeated for all sites. In the second configuration, the donor pool is expanded to include all catchments available within the broader GAR dataset used in Chapter 2. The rationale behind this choice is that increasing the amount of available information can enhance regionalization performance. In hydrology, the transfer of information from hydrologically similar catchments, even if geographically distant, has often proven effective in improving predictive skill. To limit the influence of too hydrologically dissimilar basins when expanding the donor pool, a preliminary clustering was performed.

1. Catchments within the GAR region were grouped using a standard k-means algorithm based on 14 catchment descriptors (CDs), including four morphologic variables, eight climatological indicators, glacier coverage, and catchment area. The procedure resulted in $k = 3$ clusters;
2. When a given site was treated as ungauged, only catchments belonging to the same cluster were considered as potential donors;
3. Only the 30% most similar cluster members are considered. These selected basins were then used both for the regional estimation of the MAF and for constructing the regional dimensionless POR-CDF. Weights for the aggregation of donor curves were defined as:

$$w_i = \frac{d_i^{-E}}{\sum_{i=1}^N d_i^{-E}} \quad (\text{Eq. 6.1})$$

where w_i denotes the weight assigned to the i -th donor, d_i is its distance from the target site in the descriptor space, N is the total number of selected donors, and E is a tuning exponent (set here equal to 2).

An example of the donor selected following the two methods for the Cervo a Passobreve catchment is presented in Figure 49:

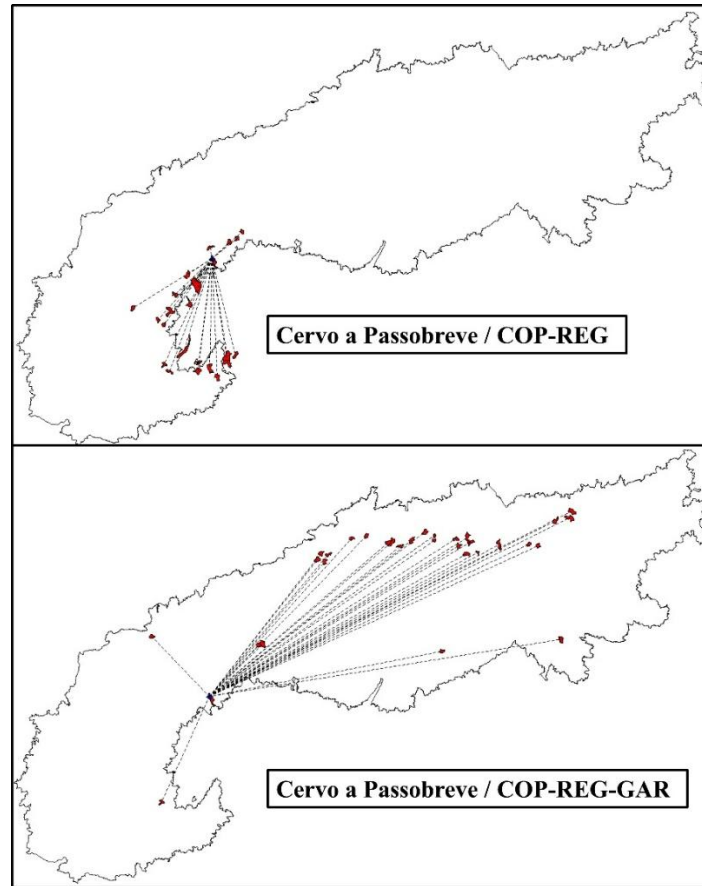


Figure 49. Difference in donor catchments selection, for the Cervo a Passobreve catchment, between the COP-REG and the COP-REG-GAR pulling methods. In red are reported only the hydrological units (see Chapter 2) containing the gauge station (not full catchments). In blue the target catchment considered ungauged.

The mean annual flow at the ungauged sites was estimated through linear regression models. Leave-one-out cross-validation was performed systematically, treating each site in turn as ungauged. Satisfactory predictive performance was obtained using only

two covariates: the logarithm of mean annual precipitation (MAP) and the logarithm of catchment area (A). An illustrative example for the Grana a Monterosso catchment is shown in Figure 50.

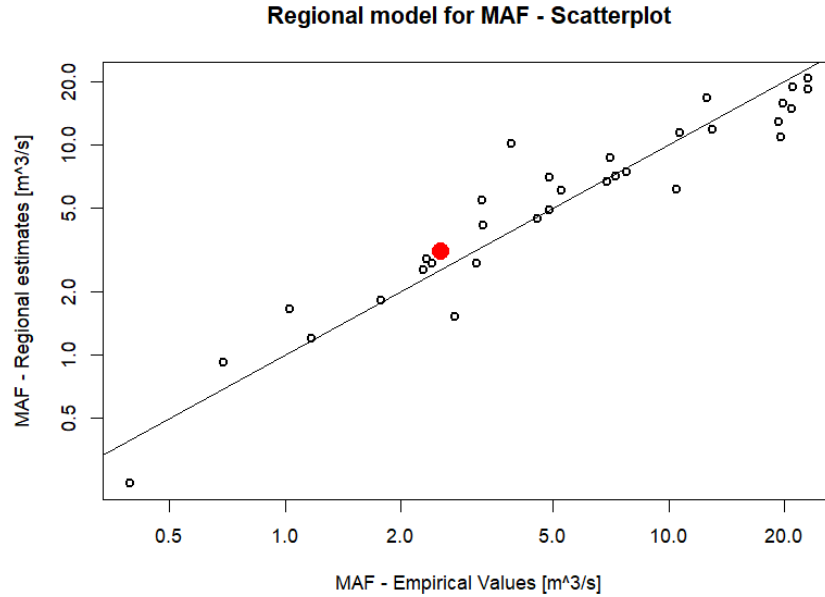


Figure 50. Empirical mean annual flow (MAF) vs regional mean annual flow estimates for the Grana a Monterosso catchment. In red the leave-one-out MAF prediction for the target ungauged catchment.

6.2.8 Joe copula model: regionalization of the “period-of-record” CDF

For the second copula model, specifically tailored to high flows, the regionalization of the period-of-record “O-CDF” requires additional clarifications. The quotation marks are intentional: even in the local implementation, this empirical “O-CDF” was not constructed from all observed discharges exceeding a fixed observational threshold. Instead, it was derived by selecting the observed values corresponding to days when the D-model predictions exceeded the simulated threshold Q_{70}^S . As a consequence, this empirical distribution may include not only high observed discharges but also moderate or even low observed values. This occurs because the D-model can occasionally produce high-flow predictions on days when the actual discharge is substantially lower, reflecting overestimation errors of varying magnitude. Therefore, the resulting “O-CDF” does not coincide with the upper tail of the observed discharge distribution (i.e., values exceeding Q_{70}^0), but rather represents the distribution of

observations conditionally sampled based on high simulated flows. This distinction becomes critical when attempting regionalization at an ungauged site. As in the two-parameter copula case, two conceptual options could be considered: (I) constructing the “O-CDF” directly from regional D-model predictions, or (II) relying on a regional estimate of the distribution. However, the first option raises an immediate issue: how should the subset corresponding to the upper 30% of regional D-model predictions be defined? Selecting all simulated values exceeding Q_{70}^S would ignore the fact that the model is known to occasionally overestimate discharge, thereby failing to replicate the conditional structure embedded in the locally defined “O-CDF”, and a conditional sampling is not possible because the observation are unavailable. Alternative selection strategies were explored, but no robust or theoretically sound solution was identified. As a result, for the Joe copula model, a fully regional construction of this particular “O-CDF” would be the only conceptually consistent option. To the best of our knowledge, no precedent exists in the literature for the regionalization of such a conditionally defined empirical distribution.

Given these methodological ambiguities, also considering how the in Section 5.2.4 was showed how the BB7-copula model performed similarly to the tailored Joe-copula model for extremes, and considering that the two-parameter copula model already poses substantial challenges, we decided not to pursue the regionalization of the single-parameter copula model further. The subsequent analysis therefore focuses exclusively on the two-parameter specification.

6.3 Benchmark and validation metrics

6.3.1 Benchmark non-parametric PP regionalization method

As already done in Chapter 4, an existing regionalization method from the literature is applied and used as a benchmark to compare the results obtained with the proposed approach to a fully non-parametric regionalization procedure. Bourgin et al. (2015) introduced a non-parametric regionalization technique based on the resampling of model residuals, rather than directly on discharge, following a block-bootstrap strategy, similar to some extent to the method previously proposed by Sikorska Sikorska et al. (2015). Once a target ungauged catchment is selected, the four closest catchments (in geographical space) are selected as donors. At each donor catchment, residuals (defined in this case as relative errors $\epsilon_t = Q_t^O / Q_t^S$) are computed and subsequently clustered into ten sub-groups based on the deciles of the empirical

cumulative distribution function (CDF) of the regional D-model predicted discharges. As in our proposed approach, donor catchments are treated as ungauged during this phase. For each decile, a super-group of residuals is then constructed by merging the corresponding sub-groups from all donor catchments. At the ungauged target site, each regional D-model prediction is assigned to a decile based on the empirical CDF of the D-model discharge predictions at that site. For each prediction, j residuals $\tilde{\epsilon}_j$ are sampled from the corresponding donor super-group. The associated discharge realizations used to define confidence intervals are then obtained as $\widetilde{Q}_{t,j}^O = \tilde{\epsilon}_j * Q_t^S$. Repeating this procedure for all regional D-model predictions yields confidence intervals for the entire discharge time series. The use of a block-bootstrap scheme allows the method to account for potentially complex dependence structures between D-model predictions and residuals, which are expected to be strong in the case of relative errors, similarly to what is observed for untransformed residuals in Section 3.2.4 (at least for residual heteroscedasticity). In their original study, Bourgin et al. adopted a different regionalization strategy for the RR model than the one used in this thesis; however, this difference does not affect the applicability of the residual-based uncertainty estimation method itself. The definition of residual super-groups for a target ungauged catchment will just be defined using the LOO D-model prediction at the selected donor catchments.

The only methodological difference in the present application from their original method concerns the selection of the donor catchments. While four donor catchments are selected for each target ungauged basin, two alternative selection strategies are considered: one based on spatial proximity, using the four geographically closest catchments (as done in the original article), and a second based on distance in the descriptor space, and therefore on hydrological similarity. Because the total number of available catchments (24) is relatively small, the two strategies often identify donor sets with overlapping catchments, although some differences are always present. The purpose of this comparison is to evaluate whether the commonly preferred similarity-based criterion actually improves the regionalization results compared with the simpler proximity-based approach, which has been supported by some recent findings in the literature (Levin & Farmer, 2020). These two strategies define two distinct regionalization scenarios from now on denoted as Bourgin-proximity (BOU-PROX), and Bourgin-descriptors (BOU-CD).

6.4 Validation Metrics for regional S-model

As previously done for the applications at gauged locations, a set of validation metrics is used here to assess the performance of the regional uncertainty estimation. Predictive probability plots are again employed, both over the entire discharge range and for LOO D-model predictions exceeding Q_{70}^S , to provide a general assessment of S-model reliability. These are complemented by an additional metric, not previously used, designed to evaluate the joint performance in terms of reliability and sharpness. This metric is an alternative form of the Interval Score (see Section 5.1.4), introduced by Bourgin et al. (2015), in which the IS is evaluated relative to a “climatological” benchmark. The purpose is to quantify whether the regional uncertainty estimation improves upon a simple reference uncertainty that could be adopted in the absence of any regional post-processing model. The benchmark is defined using the range between Q_{90}^S and Q_{10}^S (i.e., the 80% interval) of the LOO D-model predictions at the ungauged site. In practice, this corresponds to defining constant confidence bounds based on the empirical distribution of regional model predictions. Based on this reference, the Interval Skill Score (ISS) is defined as:

$$ISS = 1 - \frac{IS}{IS_{\text{clim}}} \quad (\text{Eq. 6.2})$$

where IS is the Interval Score computed for a catchment as in Eq. 5.5, and IS_{clim} is the reference Interval Score obtained using the constant “climatological” confidence bounds. The ISS is positive when the regional uncertainty model improves upon the benchmark and negative otherwise. As noted by Bourgin et al., this is not a particularly challenging benchmark; however, it represents the level of uncertainty estimation that could be obtained without applying any of the more sophisticated regional uncertainty models considered here. Unlike the IS, the ISS is a relative metric and can therefore be directly compared across catchments.

6.5 Results and Comparative assessment

6.5.1 Copula parameters regression

As anticipated in Section 6.2.4, random trees have been applied to attempt the regionalization of the two BB7 copula parameters to the ungauged sites. Leave-one-out tests have been conducted considering each catchment alternatively ungauged. No

regression are reported for the simple reasons than for both the θ_1 and θ_2 parameters of the BB7 copula, no meaningful regression was found. In all cases when considering the 23 donor catchments, regressions with very high correlations ($R^2 > 0.97$) were found for both the parameters controlling the upper and lower tail correlation; when used instead in out-of-sample for the ungauged site, the regression demonstrated, in most cases a very weak predictive capability. Despite the very weak explanatory power, the regionalization by mean of random trees have been carried on and used anyway, and in the following results will be from now on assumed its application.

6.5.2 Predictive Probability Plots

As previously done for the local configuration, predictive probability plots are presented here to assess the performance of the different regional methodological alternatives introduced in Section 6.2. For clarity, the following nomenclature is from now on adopted: (I) COP-OBS (regionalized copula + real observation CDF as O-CDF); (II) COP-SIM (regionalized copula + regional D-model predictions CDF as O-CDF); (III) COP-REG (regionalized copula + regionalized CDF as O-CDF); (IV) COP-REG-GAR (regionalized copula + regionalized CDF, pulling donors from the entire GAR, as O-CDF); (V) BOUR-PROX (method proposed by Bourgin considering the closest donors in space) and (VI) BOUR-CD (method proposed by Bourgin considering the closest donors in the descriptors space).

Inspection of Figure 51 already allows several considerations. Focusing first on the regional copula-based approaches, a marked contrast emerges between the COP-OBS configuration and the COP-SIM, COP-REG, and COP-REG-GAR configurations. When the only source of uncertainty in the regional confidence interval estimation stems from the regionalization of the copula parameters (COP-OBS), the overall reliability across the 24 catchments can be considered satisfactory and potentially suitable for practical purposes. However, once the additional uncertainty associated with the O-CDF regionalization is introduced, a clear deterioration in reliability becomes evident across all configurations.

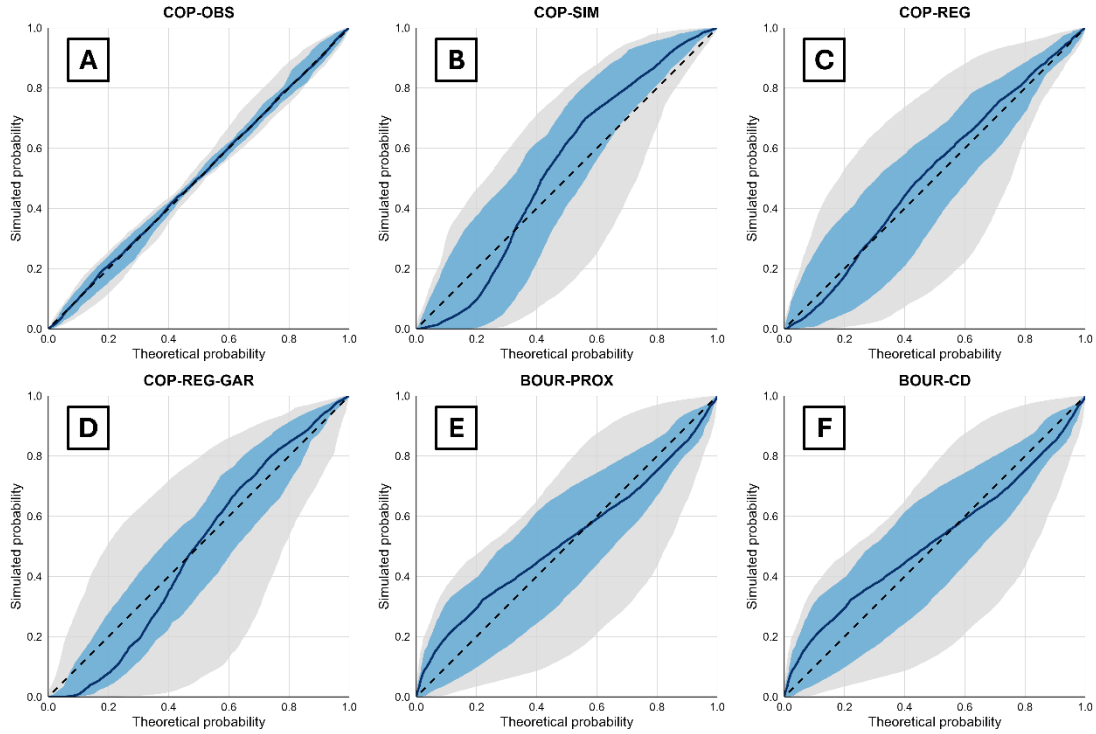


Figure 51. Predictive probability plots (PPPs) for the 24 catchments in ungauged conditions considering the entire regional D-model predictions discharge range. (A) COP-OBS; (B) COP-SIM; (C) COP-REG; (D) COP-REG-GAR; (E) BOUR-PROX; (F) BOUR-CD. In blue shades coverage for 50% of the sites; in gray shades coverage for 80% of the sites.

In particular, the central tendency of the predictive probability plots for the COP-SIM case deviates more noticeably from the 1:1 line than for the COP-REG and COP-REG-GAR cases. Between the latter two, the COP-REG-GAR configuration presents narrower 50% coverage interval, but simultaneously a wider 80% coverage. Moreover, central tendency deviations are stronger than in the COP-GAR case.

When comparing the regional copula-based approaches with the non-parametric resampling strategy proposed by Bourgin et al., the overall performance across the 24 catchments appears surprisingly similar, notwithstanding the fundamentally different methodological rationale. In terms of average reliability, the COP-REG configurations exhibit central tendencies comparable to those obtained with BOUR-PROX and BOUR-CD. Furthermore, no meaningful differences can be discerned between BOUR-PROX and BOUR-CD. Figure 52 also reports analogous diagnostics restricted to

regional D-model predictions exceeding Q_{70}^S , thereby focusing on higher-flow conditions.

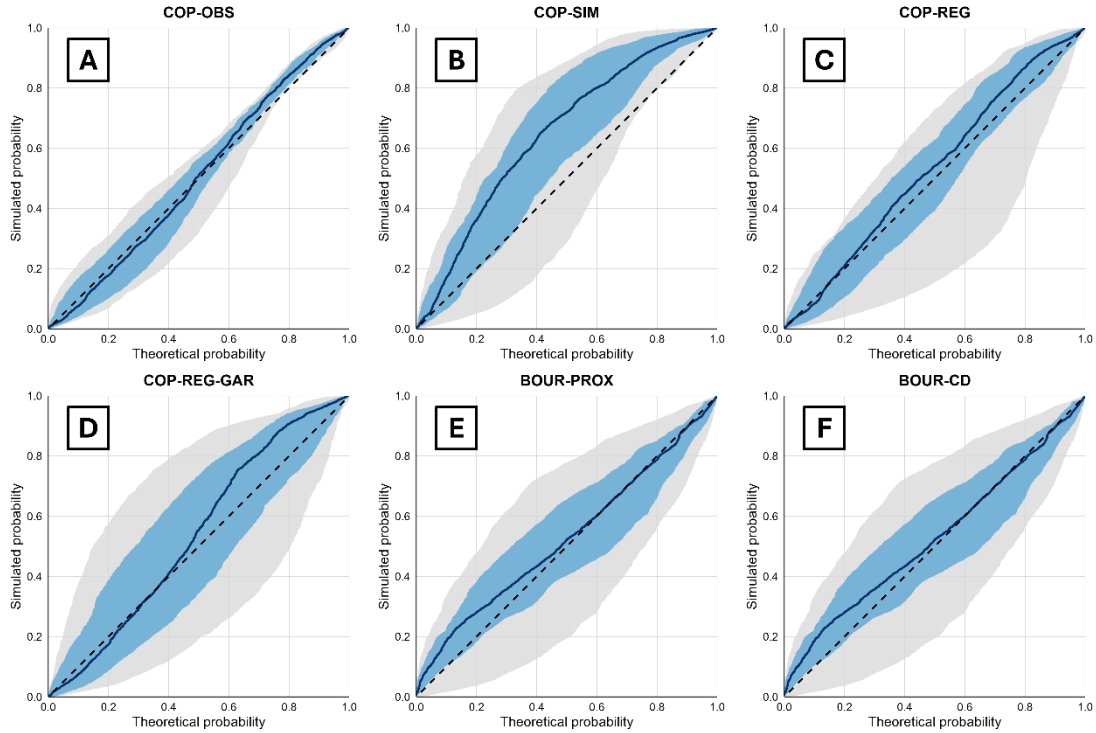


Figure 52. Predictive probability plots (PPPs) for the 24 catchments in ungauged conditions, for regional D-model predictions larger than Q_{70}^S . (A) COP-OBS; (B) COP-SIM; (C) COP-REG; (D) COP-REG-GAR; (E) BOUR-PROX; (F) BOUR-CD. In blue shades coverage for 50% of the sites; in gray shades coverage for 80% of the sites.

Within the copula-based approaches, the COP-OBS configuration again clearly outperforms the other setups, confirming that uncertainty arising solely from copula parameter regionalization is less detrimental than that associated with the O-CDF transfer. In contrast, the COP-SIM method exhibits a pronounced deviation of the central tendency from the 1:1 line, indicating systematic overestimation and representing the weakest performance in the high-flow range. More evidently than in the full discharge range analysis, a substantial difference is visible between the COP-REG and COP-REG-GAR configurations, with the first clearly improving, especially for the 50% coverage rate, over the second, suggesting that extending the potential donor catchments degrades performance compared to the use of a spatial contiguous subset. Comparing copula-based methods, specifically the COP-GAR configuration, with the non-parametric resampling approaches under high-flow conditions reveals

similar performance in terms of 80% coverage rates. For the 50% coverage level, the COP-REG configuration produce instead slightly improved intervals than BOUR-PROX and BOUR-CD.

6.5.3 Interval Skill Score (ISS)

Figure 53 presents the Interval Skill Score (ISS) statistics for the different methods. Overall, all approaches show a substantial improvement over the “climatological” benchmark, with ISS values often exceeding 60%. Compared to the percentage improvements reported by Bourgin et al. in their study over France (average ~60%), which included more than 900 sites, the improvements observed for the test catchments considered here are, on average, higher.

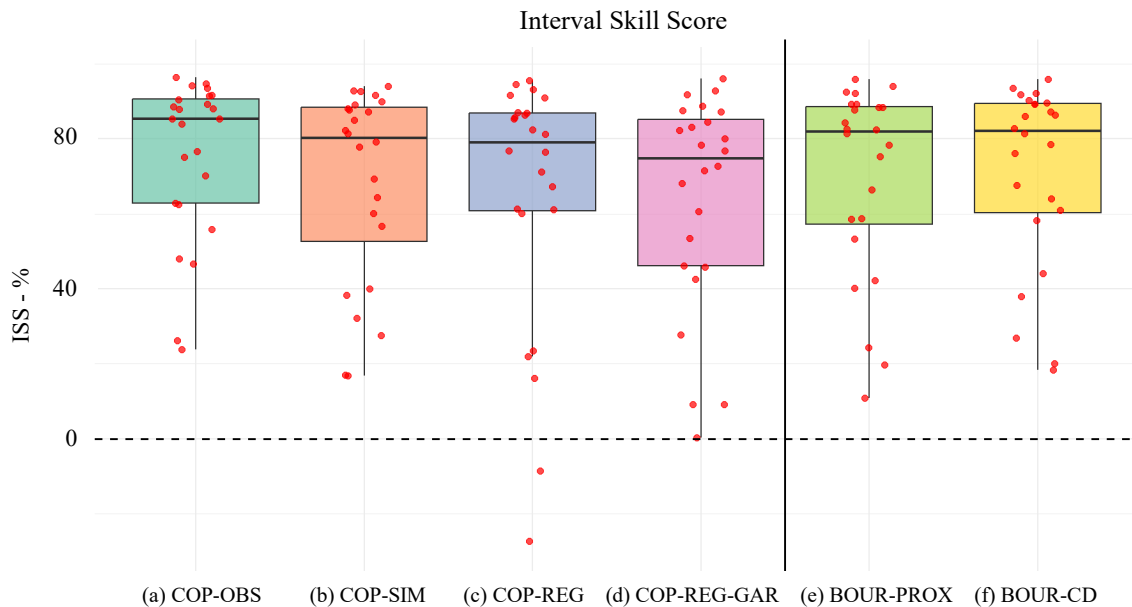


Figure 53. Interval Skill Score (ISS) for the COP-OBS, COP-SIM, COP-REG, COP-REG-GAR, BOUR-PROX and BOUR-CD methods. Each boxplot includes all catchments (considered ungauged) of the study area. In red, points for individual sites.

Among the tested configurations, only the COP-REG method produces negative ISS values for two catchments, indicating a degradation relative to the constant confidence bounds. These cases are associated with situations in which the regionalization of the O-CDF was particularly unrepresentative of the actual flow duration curve at the ungauged target site. Overall, all methods can be considered

useful, as they consistently outperform the simple benchmark and show no major differences in performance among the various alternatives.

6.6 About the usefulness of regional uncertainty estimation

6.6.1 Where does most of the uncertainty lay?

A first relevant point of discussion emerges from a closer examination of the COP-OBS configuration. Previous analyses have shown that the regional transfer of copula parameters cannot be supported by robust and meaningfully predictive regressions, nor by clearly interpretable random tree structures. In particular, no stable relationship with catchment descriptors appears to exist when the objective is to predict copula parameters out of sample. Despite this apparent lack of predictive structure, when the copula parameters are nevertheless regionalized, disregarding the limited explanatory power of the transfer relationships, the additional uncertainty introduced in the resulting S-models remains, on average, acceptable. The derived confidence intervals still exhibit a reasonably good level of reliability across the 24 catchments.

To further explore this aspect, an alternative configuration of the COP-OBS setup, hereafter referred to as COP-OBS-C, was tested. In this variant, instead of transferring θ_1 and θ_2 through a regionalization scheme, both parameters were assumed constant across all ungauged target sites and set equal to their average values ($\overline{\theta}_1$ and $\overline{\theta}_2$), computed from the locally fitted copulas obtained using regional D-model predictions (i.e., using regional LOO-D-model prediction and real observations for the 23 available donor catchments). Figure 54 reports, alongside the original COP-OBS results, the predictive probability plots corresponding to this simplified configuration. No substantial degradation in the average S-model performance across the 24 catchments (considered ungauged) can be observed. From an operational standpoint, this finding is noteworthy. On the one hand, it confirms the difficulty of deriving reliable regional transfer rules for copula parameters governing the dependence structure between regional D-model simulations and observed discharge. On the other hand, it suggests that, at least for the catchments analyzed here, the confidence interval reliability is not overly sensitive to moderate inaccuracies in the copula parameter specification. As long as a structurally reasonable dependence model is adopted, characterized, for instance, by strong upper-tail dependence, weaker lower-tail dependence, an appropriate copula family, and parameter values within plausible range, the resulting confidence intervals remain broadly acceptable.

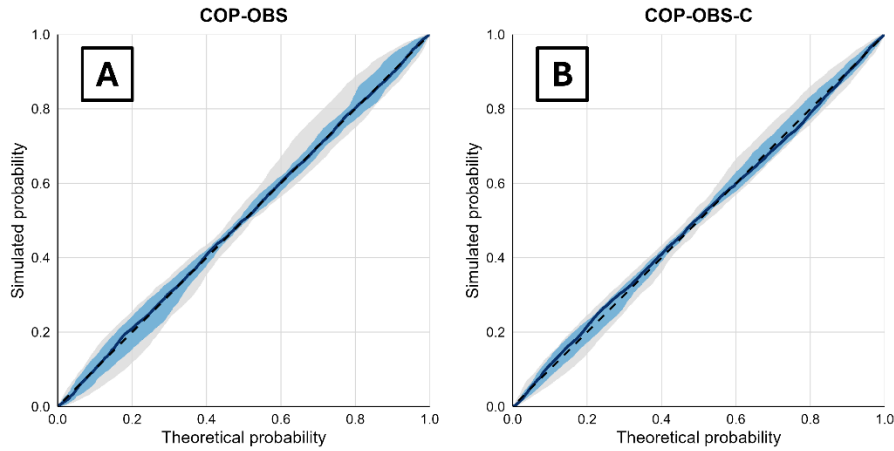


Figure 54. Predictive probability plots (PPPs) for the 24 catchments in ungauged conditions, for the entire range of regional D-model predictions. (A) COP-OBS; (B) COP-OBS-C. In blue shades coverage for 50% of the sites; in gray shades coverage for 80% of the sites.

Although a more accurate parameter estimation does improve overall reliability (as evident from the comparison between Figure 43 and Figure 54), deviations from the optimal parameter values do not appear to compromise operational applicability in a critical manner.

In contrast, a substantial portion of the observed unreliability originates from the regional estimation of the correction introduced via the O-CDF. The results clearly indicate that improvements in the regionalization of the POR-FDC at ungauged sites (whose theoretical limit is represented by the COP-OBS case) would translate directly into more reliable S-model confidence bounds. It should be emphasized that only a single regionalization strategy for the O-CDF has been explored in the present analysis. Numerous alternative approaches remain to be investigated. For this reason, the results presented in this chapter were anticipated as preliminary. There is considerable scope for enhancing the overall modeling framework by focusing specifically on this component, either by testing alternative transfer techniques or by developing regionally tailored methodologies, potentially better suited to the hydrological characteristics of the study area than more general approaches. For instance, a regional model already existed for the Piedmont region (Ganora et al., 2016), but was developed for the regionalization of mean annual FDC, and not for POR-FDC.

This direction represents the most promising avenue for future research, and it will be essential for determining whether a level of performance compatible with operational implementation at ungauged sites can (or cannot) ultimately be achieved.

6.6.2 Regional uncertainty for high-flow conditions: potential avenues

Once it is recognized that most of the uncertainty in regional estimation of confidence bounds arises from the O-CDF regionalization, it is notable that this uncertainty is not uniform across the discharge range. A closer look at Figure 51 and Figure 52 shows that, on average, the regional S-models perform better in the high-flow range than across the full range of discharges. Although a few catchments remain poorly represented, causing pronounced deviations in the COP-GAR configuration at the 80% coverage, the 50% coverage highlights improved overall reliability, despite inevitable departures from the 1:1 line. This pattern may be explained by two factors. First, the regionalization of the θ_1 copula parameter, which governs upper-tail dependence, may be less sensitive than the regionalization of θ_2 , controlling lower-tail dependence. Second, the CDF regionalization method itself might inherently perform better for high-flow conditions than for low-flow conditions. Comparing the S-model reliability in the high-flow range between the COP and BOUR methods shows higher performance for the COP approach. This observation suggests a potential avenue for further development: when the focus is on high-flow or flood conditions, regionalization could be applied selectively to the upper quantiles of the CDF. Such an approach might improve reliability relative to attempting to capture the entire discharge distribution at ungauged sites. Given that the high-flow range is already relatively better represented in the current results, despite no specific calibration targeting high discharges, focusing on upper-tail regionalization could provide a more operationally effective solution, particularly for the most critical discharge ranges in flood management.

Chapter 7

7. Discussion and conclusion

7.1 Fitting the pieces together

In the previous chapters, several applications of rainfall–runoff models under both gauged and ungauged conditions, together with the associated uncertainty estimation procedures, have been presented. Taken together, these chapters can be viewed as a complete modelling framework, illustrating the full process required to construct an informative hydrological model, starting from deterministic predictions and subsequently introducing probabilistic uncertainty estimation.

Chapter 2 presented an example of a large-scale regional rainfall–runoff model developed for large flood reconstruction across the Greater Alpine Region. This application clearly demonstrated how deterministic predictions alone, particularly in ungauged catchments and when focusing on hydrological extremes, are inherently affected by significant uncertainties. These results showed that uncertainty quantification should not be considered optional, but rather a necessary component of hydrological modelling in order to avoid misrepresentation of extreme events. The regional model performance in ungauged catchments was consistent with findings commonly reported in the hydrological literature: some catchments showed acceptable deterministic performance, while others were characterized by much lower model efficiency. A further difficulty emerged from the regionalization procedure itself. Although model parameters were regionalized using regression tree methods based on similarity criteria, no clear relationship between model performance and catchment characteristics could be identified. Model performance appeared spatially homogeneous and without clear sub-regional patterns, achieving one of the main goal of the regionalization, while also meaning that regionalization approach alone cannot guarantee reliable predictions even for catchments that appear similar to others in the region.

These findings highlighted the need for explicit uncertainty quantification. To move in this direction, Chapter 3 investigated the general properties of rainfall–runoff model residuals. The study, conducted on a large and morphologically and climatologically diverse set of catchments in the United States, and using multiple model structures, showed that many residual properties are broadly consistent across sites, largely independent of catchment type and model structure. At the same time, some catchment-specific behaviors were identified, such as stronger temporal autocorrelation in snow-dominated catchments and higher residual variability in arid basins. This analysis made it possible to identify residual properties that are sufficiently general to support the development of widely applicable post-processing error models. It also showed how certain residual transformations can simplify residual behavior, for example by reducing tail heaviness or approximating homoscedasticity, thereby enabling the use of simpler modelling approaches.

Chapters 4 and 5 then focused on post-processing methods for estimating aggregated rainfall–runoff model uncertainty and generating confidence intervals, thus moving from a purely deterministic framework to a probabilistic one. In Chapter 4, building on the findings regarding residual properties and transformations, a new (mostly) parametric method was proposed (the C-method). The method was applied to residuals after logarithmic transformation and included a representation of monthly seasonality. The main objective was to develop a method as simple as possible in view of future regionalization. The method was designed to generate individual stochastic residual series, allowing the production of both ensemble simulations and confidence intervals. A comparative study conducted on a sub-region of the Greater Alpine Region, against two alternative approaches (the R-method and the AR-method), showed that while simple methods can generate realistic residual and discharge series at some sites, they perform poorly at others, lacking robustness in performance. Moreover, obtaining simultaneously realistic individual series and reliable confidence intervals resulted challenging. This indicated that more complex error model structures may be required for robust performance across different catchments, although increased model complexity would make regionalization more difficult.

Based on these results, Chapter 5 introduced an alternative post-processing approach aimed specifically at estimating confidence intervals rather than generating individual stochastic series, allowing for a simpler error model structure. The method, a copula-based parametric development of the BLUECAT methodology, proved to be as effective as the original non-parametric approach in producing reliable confidence

bounds, while substantially improving uncertainty estimation for extreme flows thanks to the parametric representation of the tails of the discharge distribution. Although performance improvements were limited to discharge peaks, together with the comparable performance across the full range of predicted discharge with respect to the original non-parametric method, made regionalization of this model structure feasible.

Finally, Chapter 6 presented the last step of the framework: the regionalization of a post-processing error model for rainfall–runoff uncertainty estimation in ungauged catchments. Preliminary results were presented through a comparative analysis with the non-parametric resample-based methodology proposed by Bourgin et al. (2015). The regionalization of the different model components (rainfall–runoff model, copula structure, and discharge distribution) showed that most of the uncertainty arises from the regional estimation of the discharge cumulative distribution function, whereas uncertainties associated with the regional estimation of the dependence structure between deterministic model predictions and observed discharges are comparatively smaller and less problematic. Overall, similar performance between the non-parametric and the parametric approach have been again achieved. Moreover, the reliability of the estimated confidence intervals in ungauged catchments remain largely unfit, indicating how this research area still requires developments.

7.2 Extended approaches comparison

7.2.1 Is there a “best” model based on the Interval Score?

Across Chapter 4 and 5, several methods were presented and applied locally to the same set of test catchments. In this final chapter, a broader discussion is required to compare all these approaches, including those based on different methodological philosophies, and to address a practical question: which method should be preferred? Any conclusions must, however, be interpreted with caution, as they are based on a limited sample of 24 mountainous catchments and cannot be considered fully general.

The different methods are here quickly recalled to ease the interpretation of the following tables: among methods aiming at stochastically reproducing individual series, the method proposed in Chapter 4 have been the copula based C-method, the resampling R-method by Sikorska et al. (2015), and the autoregressive-based AR-method by Shabestanipour et al. (2023). Among the methods aiming at the sole estimation of confidence intervals, in Chapter 5 the BLUECAT (Koutsoyiannis & Montanari, 2022a) method was considered, for two selections of the k parameter equal

100 and 50, (BC-100 and BC-50), together with its parametric evolution (P-BLUECAT; BB7 copula-based and Joe copula-based tailored for high-flows).

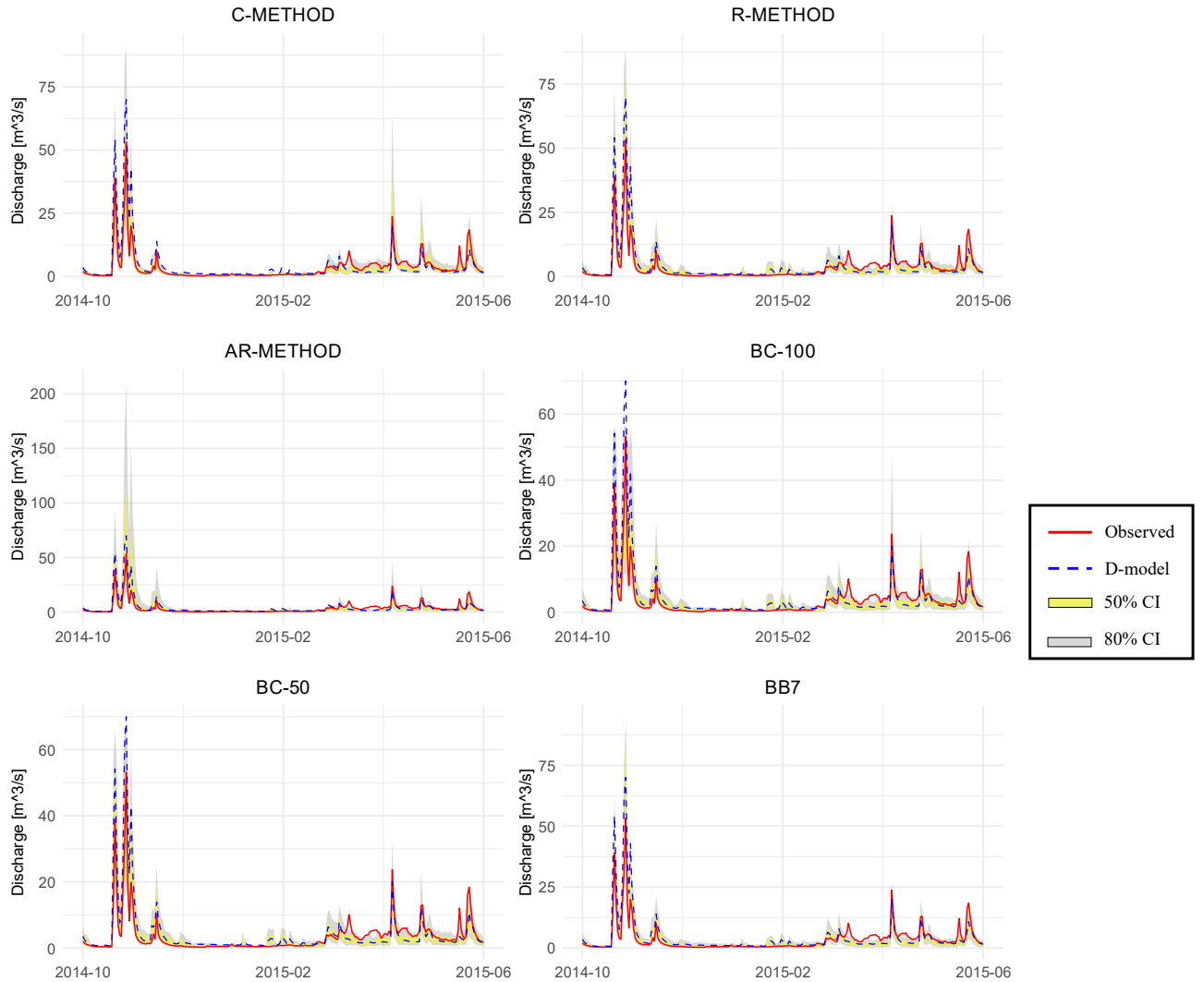


Figure 55. S-model for the different methods (C-method, R-method, AR-method, BC-100, BC-50, P-BLUECAT-BB7) for a subset of the calibration period (from 2014-10-16 to 2015-06-23) for the Cervo a Passobreve catchment. In yellow 50% confidence interval; in gray 80% confidence interval. Dashed blue line: D-model prediction; red continuous line: observed discharge.

To summarize the results compactly, Table 5 and Table 6 report the Interval Score (IS; see Section 5.1.4) for all methods across the full range of deterministic model (D-model) predictions, for both calibration and validation periods. The IS is adopted as the main comparison metric because it combines reliability and sharpness into a single numeric measure, thus complementing the previous analyses, which focused primarily (with the exception of Chapter 5) on reliability alone through predictive probability plots. The Joe P-BLUECAT method is excluded for the same reason discussed in Section 5.1.4 (it produces confidence intervals only for a limited range of D-model predictions), while the AR-method is omitted from the validation period results following the considerations in Section 4.2.4 instead, being the method originally designed for the sole application in a calibration period (to then be used for long-term water management planning). Additionally, in Figure 55, for the Cervo a Passobrevé catchment, the full S-models together with the observed discharge and D-model predictions are reported, for a window of the calibration period, for the different methods.

In the tables, rows correspond to catchments and columns to methods, with each cell reporting the IS value. Colors are scaled within each row: the best-performing model (lowest IS) is shown in dark green, the worst (highest IS) in red, while intermediate models color is scaled depending on the proximity between the two extremes, having greenish color if closer to the best, and reddish colors if closer to the worst. The best model for each catchment is also highlighted in bold white.

For the calibration period (Table 5), the main observation is that the AR-method performs consistently worse than all other approaches. The remaining methods generally yield comparable results, with the C-method and R-method achieving the highest number of best performances. The BC-50 and BB7 P-BLUECAT methods show similar behavior, while BC-100 tends to produce slightly higher IS values, though still markedly better than the AR-method.

The validation period (Table 6), which provides the most meaningful assessment, reveals more distinct differences. First (I), the non-parametric BLUECAT with $m_1 = m_2 = 100$ emerges as the weakest method, frequently producing the highest or second-highest IS values. This confirms (as already discussed in Section 5.1.2) that using large m values with limited data leads to poor representation of extremes, which strongly influence the IS.

Table 5. Summary Interval Score table (calibration period) for all methods applied at all the catchments. Colors (from dark green = best, to red = worst) are scaled independently for each row based on relative performance (the closer to 1 = the closer to the worst, the closer to 0 = the closer to the best). Best model for each site also highlighted in white.

Catchment	BC-100	BC-50	BB7	C-METHOD	AR-METHOD	R-METHOD
C24	24	23.3	24.4	23.1	38.3	23.6
C23	36.1	34.4	33.9	30.5	49.2	32.1
C22	21.2	19.6	18.7	17	23.4	16.8
C21	22.8	21	21.7	18.1	25.9	19.6
C20	20	19.4	22.9	19.4	23.5	19.2
C19	40.3	35.6	34.3	29.2	39.4	32.9
C18	10.2	9.38	8.87	7.6	10.8	8.43
C17	20.3	19.7	19.8	27.1	32.9	19.6
C16	15.8	15.3	15.4	18.8	19.1	15
C15	23.7	20.6	20.5	25.5	36.3	19.4
C14	8.28	8.11	9.07	10	9.6	8.15
C13	11	9.71	9.36	12.6	13.2	8.87
C12	14.2	14.1	15	12	16.3	14.2
C11	8.11	7.64	7.39	6.04	8.59	7.13
C10	5.87	5.57	5.39	5.47	7.84	4.94
C9	11.1	9.13	8.18	9.37	38.2	8.15
C8	11.1	10.2	9.52	9.39	17.8	8.87
C7	7.43	6.79	7.51	5.78	8.79	6.26
C6	6.53	6.16	5.85	7.46	8.5	5.72
C5	12.6	11.4	10.5	12.8	17.3	10.6
C4	4.08	3.97	4.12	3.18	4.35	3.94
C3	7.85	7.73	8.61	7.37	9.55	7.8
C2	9.51	8.48	8.34	8.37	12.3	7.9
C1	6.19	5.79	5.93	5.78	7	5.65

The BC-50 configuration performs better but still ranks among the less effective approaches, reinforcing the difficulty of balancing set size and representativeness when data availability is limited. Second (II), the C-method appears to be the best-performing approach overall. This result is somewhat unexpected, as earlier discussion suggested reliability limitations related to its simplified representation of residual autocorrelation (that was the main aim of the model). However, when evaluated solely in terms of prediction intervals (with no bias correction factor), it provides the most favorable

trade-off between reliability and sharpness. Third (III), the R-method and BB7 P-BLUECAT show intermediate performance, with similar average efficiency across catchments, although not necessarily for the same sites.

Table 6. Summary Interval Score table (validation period) for all methods applied at all the catchments. Colors (from dark green = best, to red = worst) are scaled independently for each row based on relative performance (the closer to 1 = the closer to the worst, the closer to 0 = the closer to the best). Best model for each site also highlighted in white.

Catchment	BC-100	BC-50	BB7	C-METHOD	AR-METHOD	R-METHOD
C24	24.7	25.6	25.9	25.6	/	25.3
C23	53.5	54	56	53.6	/	53.3
C22	23	22.5	23	18.7	/	21.6
C21	23.7	23.2	19.8	16.8	/	22.3
C20	20.4	20.4	20.6	19.2	/	19.7
C19	36.1	33.6	32.5	32.6	/	33.1
C18	12.2	11.8	11.3	9.44	/	11.7
C17	18.2	18	17.6	26.6	/	17.2
C16	17.4	16.9	15.9	17.6	/	16.7
C15	22.2	20.6	21.2	21.6	/	20.6
C14	16.7	16.8	16.2	14.8	/	16.9
C13	9.36	8.77	8.97	13.4	/	8.76
C12	17.2	17.3	16.4	13.6	/	16.8
C11	8.59	8.47	8.37	7.85	/	8.38
C10	5.78	5.73	5.7	4.82	/	5.69
C9	12.7	11.8	11.5	13	/	11.9
C8	11.4	10.9	11.2	11	/	11.1
C7	10.2	10.4	9.8	7.71	/	9.68
C6	6.28	6.12	6.28	6.83	/	6.23
C5	9.3	9.11	9.36	10.7	/	9.03
C4	4.21	4.27	4.11	3.84	/	4.3
C3	6.43	6.48	6.03	6.08	/	6.48
C2	7.64	7.21	7.07	7.05	/	6.92
C1	6.43	6.08	6	5.68	/	5.83

Relative performance

Method

Finally (IV), it is important to emphasize the presence of exceptions. Even the BC-100 method, despite its generally poor performance, occasionally outperforms all alternatives, while the C-method can be among the worst in some cases. These

deviations reflect the underlying assumptions of each method. For example, the C-method relies on the independence between D-model predictions and residuals, which may not hold everywhere, while the P-BLUECAT depends on accurately reproducing the copula structure between predictions and observations.

Overall, while some general tendencies can be identified, no single method consistently outperforms the others in all cases. When possible, multiple approaches should therefore be tested on a case-by-case basis. This is feasible in practice, as all considered methods are relatively simple to implement and computationally efficient, which was one of the main motivations for focusing on post-processing uncertainty estimation methods in this work. The possibility to jointly consider multiple approaches also remains open: an alternative could be to apply multiple individual PP uncertainty estimation methods, producing multiple confidence intervals, and later to weight average them based on some performance metric (for example, the same IS). Consequently, some degree of arbitrariness in the way the weight would need to be defined would be included.

7.2.2 Is there an optimal error model for the flood extremes?

Throughout this thesis, particular focus has been posed multiple times on floods; it come then natural to ask, among the post-processing models considered, if any can be considered optimal in estimating uncertainty for extreme predicted discharge.

The analysis of the Interval Score (Section 7.2.1) for the validation period identifies three best-performing post-processing methods: the BB7 P-BLUECAT, the C-method, and the R-method. For BB7 P-BLUECAT, Section 5.2.4 already showed how the introduction of the mixed D-model predictions cumulative distribution function (S-CDF) and mixed observed discharge cumulative distribution function (O-CDF) enables extrapolation beyond observed extremes, allowing for informative confidence intervals even for the most extreme D-model predictions in Peak-over-Threshold (POT) analyses. In contrast, the behavior of the other two methods under extreme conditions has not yet been examined in detail. To provide further insight into their comparative performance, the POT analysis is extended here to include all methods introduced in Chapter 4. Figure 56, presents a representative example for the Grana at Monterosso catchment during the calibration period, following the same procedure, for the same site, described in Section 5.1.4. All three methods are able to produce reasonable confidence intervals for extreme discharges, although with distinct characteristics.

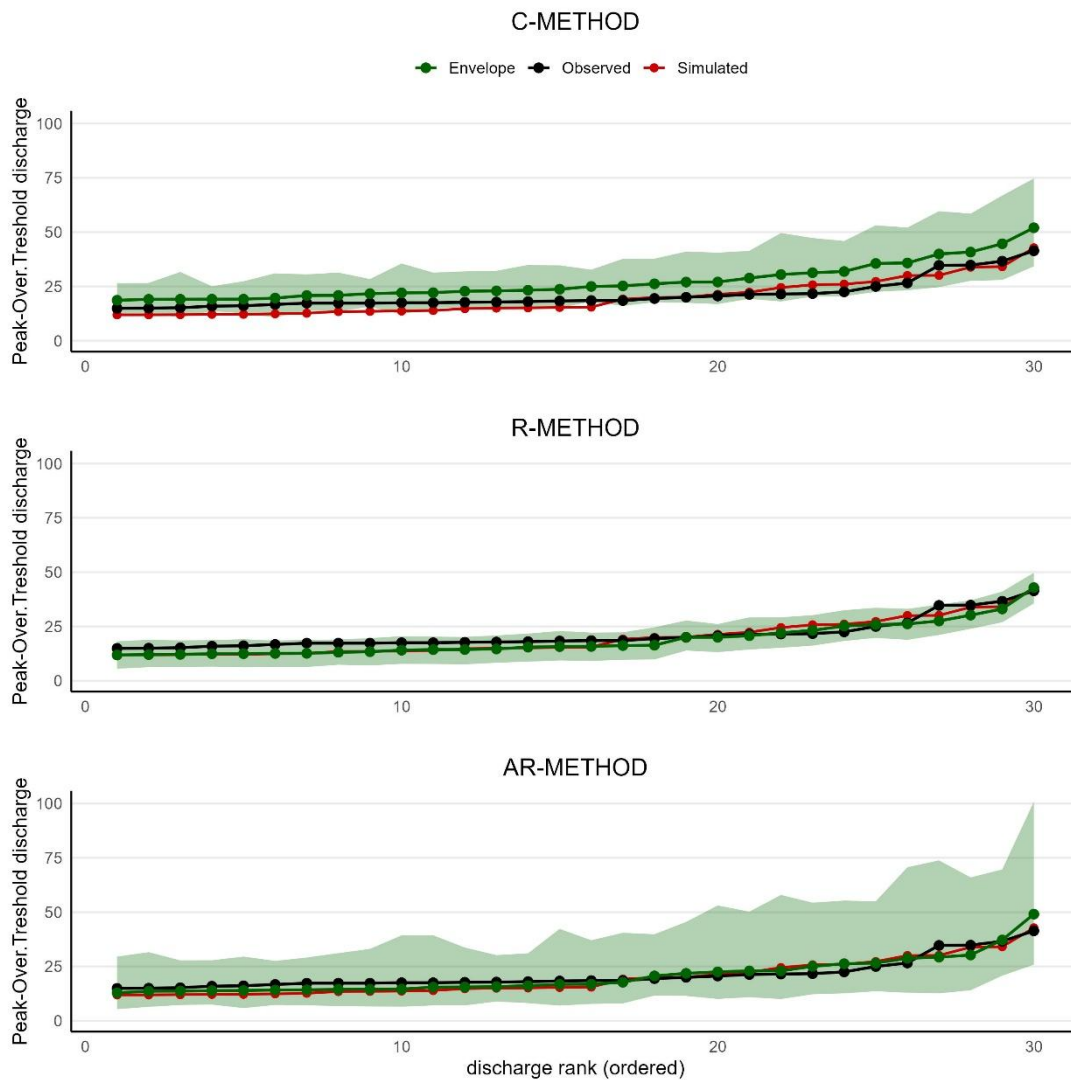


Figure 56. Peak-Over-Threshold (POT) comparison (average of three events for year) between observed discharge (in black), D-model predictions (in red) and S-model predictions (in green, shades indicating confidence interval of 80%) for the C-method, the R-method, and the AR-method, for the calibration period, for the Grana a Monterosso catchment.

The AR-method, despite performing worst in terms of IS, generates intervals that are generally well centered around the observed extremes, but with large widths, indicating low sharpness. The C-method produces narrower intervals, but observed values tends to concentrate to the lower envelope probabilities (this is consistent with the deviations

in the predictive probability plots for predictions larger than Q_{70}^S in Figure 38). The R-method appears to offer the best compromise, providing intervals that are both sharp and well centered. Although illustrated for a single catchment, this behavior is consistent across the full set of 24 basins. When considered together with the IS results, this suggests that the R-method, already among the top-performing approaches, is additionally also effective in representing uncertainty in the upper tail of the discharge distribution. More generally, all methods presented in Chapter 4 are capable of producing informative confidence intervals for extreme flows. The only clear limitation was observed for the original non-parametric BLUECAT, which suffered from bounding issues due to its construction.

A final remark concerns the AR-method: its poor performance in terms of IS, despite not being affected by bounding issues, indicates a substantial degradation in the balance between reliability and sharpness, particularly at higher flow quantiles. It is important, however, not to draw overly general conclusions from this result. The AR-method was originally developed and presented (Shabestanipour et al., 2023) on a single case study (the Squannacook River, USA), characterized by climatic and morphological conditions very different from those of the Alpine catchments here considered. Moreover, in the original application, 37 years of daily discharge data were available, considerably more than the 10 years used in the current application. Its performance may therefore depend strongly on catchment characteristics and available observations record length. This highlights both the limitations of the present analysis and the need for further investigation across a wider range of hydrological settings, as later discussed in Section 7.4.1, as well as the need for further testing, considering varying model calibration period lengths (see Section 7.2.3).

7.2.3 Parametric vs non-parametric: are there real advantages?

Another important point of discussion concerns the general performance comparison between non-parametric and parametric approaches. In Section 4.1.2, the focus on parametric approaches was motivated by methodological preferences and potential advantages (i.e., benefits in extrapolation during the validation period and the more interpretable model regionalization), but it was also explicitly stated that this choice did not imply any assumed a priori superiority. Instead, the potential benefits of parametric methods had to be tested rigorously. This was done in Chapters 4 and 5 by comparing the proposed (mostly) parametric approaches with existing resampling-based non-parametric methods.

Overall, for the tested conditions, the hypothesis that parametric approaches would outperform non-parametric ones must be considered unsupported. This applies both to methods designed to generate individual stochastic discharge series and to methods aimed solely at estimating confidence intervals. In most cases, the parametric approaches did not outperform the non-parametric alternatives, but rather produced comparable results. This outcome was somewhat unexpected, since parametric methods are theoretically better suited for extrapolation. It was therefore reasonable to expect better performance in temporal validation, where uncertainty estimation is performed out of sample, also considering that only 10 years of daily observed discharge data were employed in calibration. One possible explanation is related to the length of the available data record. Although a 10-year daily record could initially appear relatively short, it may instead still be sufficiently long to allow effective application of non-parametric resampling methods, thereby limiting the potential advantages of parametric approaches. In this context, an interesting open question concerns how the relative performance of the different approaches would change if shorter or longer calibration periods were used. Since uncertainty estimation performance generally degrades with shorter calibration records, it is possible that this degradation would occur more rapidly for non-parametric methods than for parametric ones. This hypothesis could be tested by repeating the analysis using progressively shorter calibration periods (for example 7, 5, and 3 years), while keeping the validation period fixed (e.g., the same 10 years used in this thesis). Such an analysis could help identify a threshold record length below which parametric methods clearly outperform non-parametric ones, which would justify their use in data-scarce environments. This analysis was not conducted here, but it represents a straightforward and potentially valuable direction for future research, especially considering that all post-processing methods considered are computationally efficient and easy to apply across multiple catchments.

Parametric approaches remain instead particularly attractive for ungauged applications. Although the comparison between parametric and non-parametric approaches in Chapter 6 again showed broadly comparable performance, the analysis of the idealized COP-OBS case, where the regional flow duration curve at the ungauged site was assumed to be perfectly known, demonstrated that improvements in the regionalization of this specific component could lead to substantial improvements in uncertainty estimation reliability. In other words, there are still clear opportunities for improving the parametric framework, particularly by improving the regional

estimation of the flow duration curve. Such improvement pathways are easier to conceptualize within a parametric framework than within a resampling-based non-parametric one. Moreover, parametric approaches offer an additional advantage in terms of interpretability, particularly in regionalization contexts. With a parsimonious parameterization, it is possible to interpret model behavior and attempt to link uncertainty structures to hydrological processes. This type of process-based interpretation is instead much more difficult to achieve with non-parametric resampling approaches.

7.3 Other operative setups: forecast

An important aspect that still deserves clarification concerns the distinction between the uncertainty analysis developed in this work and the problem of forecasting in a strict sense. This thesis has focused on post-processing approaches to uncertainty, namely on the statistical and probabilistic characterization of model errors and variability, without explicitly addressing the generation of operational forecasts. This choice should not be interpreted as a limitation, but rather as a deliberate methodological focus: understanding the structure of uncertainty represents a fundamental prerequisite for any reliable predictive system.

In hydrological forecasting, the objective is to estimate near future quantities (e.g., streamflow) based on information available at the present time, inherently involving additional sources of uncertainty related to initial conditions, meteorological inputs, and model structure. In this context, the uncertainty characterized in this thesis constitutes a necessary, though not exhaustive, component of the overall predictive problem. While forecasting explicitly targets future states, post-processing focuses on extracting statistically consistent representations of uncertainty from model outputs; the two perspectives are therefore complementary rather than alternative.

The methodologies developed in this work can be naturally transferred to a forecasting framework by embedding them within the prediction chain as calibration layers acting on model outputs. In particular, the statistical structures identified for residuals and errors can be used to construct predictive distributions conditional on model simulations. A key aspect in this transfer is the explicit dependence on the forecasting horizon. Error properties are known to vary significantly with lead time: short-term forecasts (e.g., 1–3 hours ahead) are typically dominated by errors in initial conditions and fast dynamics, whereas longer horizons increasingly reflect structural model deficiencies and input uncertainty. Consequently, post-processing models

should not be assumed to be invariant across lead times, but rather calibrated separately for different forecasting horizons. This implies, for example, that distinct probabilistic models may be required for 1-hour, 3-hour, and daily forecasts, each characterized by its own error distribution, variance structure, and potential bias correction. The framework developed in this thesis can support such an approach by providing flexible tools to model heteroscedasticity, non-Gaussian behavior, and other relevant statistical features of the residuals. Moreover, the identification of scaling properties and dependence structures can inform how uncertainty propagates across time, allowing for more coherent multi-horizon predictive systems.

In operational terms, this translates into a modular architecture in which deterministic or ensemble forecasts are first generated and then post-processed using horizon-specific statistical models, yielding calibrated predictive distributions. Such an approach enhances not only the accuracy but also the reliability and interpretability of forecasts, as it explicitly accounts for the evolution of uncertainty with lead time. Therefore, although an operational forecasting system has not been developed, this thesis contributes directly to one of its essential components: the rigorous and structured representation of uncertainty. The results obtained here provide a foundation upon which forecasting applications can be built, particularly in contexts where the accurate quantification and simple communication of predictive uncertainty are of primary importance.

7.4 About the urgency of synthesis

7.4.1 Generality and synthesis projects efforts

Throughout this thesis, the application and comparison of uncertainty estimation methods were limited primarily to mountainous catchments, with relatively little attention given to other morpho-climatic conditions. However, Chapter 3 already demonstrated the importance of extending the analysis beyond a single catchment group. In that chapter, residual properties were analyzed across hydrological clusters, highlighting both similarities and differences among catchment types and clearly showing the benefits of stratified analyses. To extend the findings presented in the previous chapters and move toward more general conclusions, a larger study would be necessary, including a wider range of catchments and covering the broadest possible variability in morpho-climatic conditions.

More generally, a review of the existing literature on rainfall–runoff uncertainty estimation reveals a significant gap in the synthesis of existing knowledge. Many studies propose and test new methods (e.g., Koutsoyiannis & Montanari, 2022;

Montanari & Brath, 2004; Montanari & Grossi, 2008; Morawietz et al., 2011; Pianosi & Raso, 2012; Shabestanipour et al., 2023; Sikorska et al., 2015; Smith et al., 2012, and others), but often lack in clear comparisons with already existing approaches, and/or are often applied to a limited number of catchments, which prevents the results from being considered general. As a consequence, most studies include caveats regarding the specific catchments and conditions analyzed, and broader conclusions are frequently deferred to future research that is rarely carried out. At the current stage of research on rainfall–runoff model uncertainty, numerous new methods will certainly continue to be proposed; however, the research field may now have reached a point where a comprehensive synthesis of existing knowledge would not only be useful, but could be necessary. In this sense, the current situation resembles that of research on prediction in ungauged basins approximately twenty years ago: a large amount of knowledge existed, but it was fragmented and difficult to integrate into a coherent framework, making it less accessible and less usable, particularly for practitioners and researchers outside the dedicated research community. The Prediction in Ungauged Basins (PUB - Hrachowitz et al., 2013) initiative played a crucial role in addressing this issue by defining clear research directions and providing a collaborative framework that brought together multiple research groups. Considering the success of the PUB initiative, a similar collaborative effort focused specifically on uncertainty estimation would likely be highly beneficial. Ideally, such an initiative should not be limited to post-processing approaches like those investigated in this thesis, but should also include Bayesian frameworks and GLUE-type methods. The objective would be to consolidate existing knowledge and create a common framework for method comparison. One possible approach would be to define a set of agreed-upon test catchments, similar to what was done in the MOPEX project, to serve as standard benchmarks for testing both existing and newly developed methods. These catchments should cover a wide range of hydrological, climatological, and morphological conditions (for example small and large basins, dry and wet climates, snow-dominated and rainfall-dominated regimes).

Such an effort would provide several important benefits. First (I), having a common set of test catchments would make it much easier to compare different methods and could help reveal patterns that are currently hidden by the limited scope of individual studies. For example, it might become clear that certain methods perform better in arid catchments while others perform better in humid or snow-dominated environments. Second (II), a unified testing framework would make it possible to agree

on a common set of validation metrics and to establish standard benchmarks for uncertainty estimation performance. At present, different studies use different evaluation metrics (e.g., predictive probability plots, Interval Score, and other measures) making comparisons difficult. Establishing a set of commonly accepted metrics would allow methods to be evaluated on the same basis and would provide a consistent benchmark for future developments. It would also be possible to define different metrics for different objectives, such as metrics for the full discharge range and metrics specifically tailored to high-flow or low-flow conditions.

A collaborative framework of this type would be particularly “easy” to implement for post-processing methods, since these typically require only observed discharge and deterministic model predictions, and are therefore largely independent of the specific rainfall–runoff model used to generate the predictions. Developing a similar framework for methods based on different modelling philosophies would be more challenging. Approaches such as GLUE and Bayesian methods are directly linked to the rainfall–runoff model structure and require the hydrological model to be run as part of the uncertainty estimation procedure. Nevertheless, including these approaches in a unified testing and benchmarking framework would be essential in order to provide a comprehensive comparison of uncertainty estimation methodologies across different modelling paradigms.

7.5 Rainfall-runoff model and uncertainty: but for who?

I would like to close this thesis with a final reflection, reconnecting to the ideas introduced in the preface. The guiding principle behind this work has always been simple: to do something because it could be useful, not just for the sake of publishing in a scientific journal.

Too often, the details of rainfall–runoff modelling, especially regarding uncertainty estimation, remain accessible only to a small group of experts. Important results are frequently presented in ways that make them difficult to understand and use for the people who actually need them. At the same time, conceptual RR models are being used more and more widely across hydrology, and even beyond the field. However, they are increasingly used not as an end goal, but as a tool within broader studies, for example, to provide input to an irrigation model or to support ecological research. These are just a few cases, but similar applications can be found across many different fields.

Despite this widespread use, there is still a clear gap in how knowledge is transferred. As a result, RR models are often treated as simple deterministic tools, and not always used in the most appropriate way. This leads to a broader issue: as experts in this field, we have a responsibility. We cannot just stand aside and notice that these models are often used superficially. Nor can we limit our role to pointing out flaws when models are applied incorrectly or when their limitations, like those discussed in this thesis, are ignored. Instead, it is our role to recognize this problem and actively contribute to a solution. If most users of RR models are not experts, then it is our responsibility to provide them with tools that allow for correct use and interpretation. We need to make these tools accessible, simplified, but not oversimplified or trivialized. This is not just about being helpful; it is a necessity. Otherwise, these models will continue to be misused, affecting not only individual studies but also the quality of a large body of research built on top of them. Since these models often sit at the very beginning of the research chain, improving their use is essential.

I like imagining a near future in which these systems are used more consciously and responsibly. A future where individual ambitions are, at least in part, set aside in favor of collaborative efforts, aimed at developing tools that are genuinely useful for everyone, especially for non-experts, so that their application can improve significantly. Only the future will give us an answer, but for now we must be the first to commit ourselves in this direction.

Reference

- Aas, K., Czado, C., Frigessi, A., & Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2), 182–198. <https://doi.org/10.1016/j.insmatheco.2007.02.001>
- Abbott, M. B., Bathurst, J. C., Cunge, J. A., O’Connell, P. E., & Rasmussen, J. (1986a). An introduction to the European Hydrological System — Systeme Hydrologique Europeen, “SHE”, 1: History and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology*, 87(1), 45–59. [https://doi.org/10.1016/0022-1694\(86\)90114-9](https://doi.org/10.1016/0022-1694(86)90114-9)
- Abbott, M. B., Bathurst, J. C., Cunge, J. A., O’Connell, P. E., & Rasmussen, J. (1986b). 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](https://doi.org/10.1016/0022-1694(86)90115-0)
- 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>
- Ao, T., Ishidaira, H., Takeuchi, K., Kiem, A., Yoshitani, J., Fukami, K., & Magome, J. (2006). Relating BTOPMC model parameters to physical features of MOPEX basins. *Journal of Hydrology*, 320, 84–102. <https://doi.org/10.1016/j.jhydrol.2005.07.006>
- Balkema, A. A., & Haan, L. de. (1974). Residual Life Time at Great Age. *The Annals of Probability*, 2(5), 792–804. <https://doi.org/10.1214/aop/1176996548>
- Barber, C., Lamontagne, J. R., & Vogel, R. M. (2020). Improved estimators of correlation and R2 for skewed hydrologic data. *Hydrological Sciences Journal*, 65(1), 87–101. <https://doi.org/10.1080/02626667.2019.1686639>
- Bedford, T., & Cooke, R. M. (2001). Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines. *Annals of Mathematics and Artificial Intelligence*, 32(1), 245–268. <https://doi.org/10.1023/A:1016725902970>
- Beven, K. (2012). *Rainfall-Runoff Modelling: The Primer*. *Rainfall-Runoff Modelling: The Primer: Second Edition* (Vol. 15). <https://doi.org/10.1002/9781119951001>

- 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., Smith, P., & Freer, J. (2007). Comment on “Hydrological Forecasting Uncertainty Assessment: Incoherence of the GLUE Methodology” by Pietro Mantovan and Ezio Todini. *Journal of Hydrology*, 338, 315–318. <https://doi.org/10.1016/j.jhydrol.2007.02.023>
- Beven, K., Smith, P., & Freer, J. (2008). So Just Why Would a Modeller Choose to Be Incoherent? *Journal of Hydrology*, 354, 15–32. <https://doi.org/10.1016/j.jhydrol.2008.02.007>
- Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology / Un modèle à base physique de zone d’appel variable de l’hydrologie du bassin versant. *Hydrological Sciences Bulletin*, 24(1), 43–69. <https://doi.org/10.1080/02626667909491834>
- Blöschl, G. (2006). Rainfall-Runoff Modeling of Ungauged Catchments. In *Encyclopedia of Hydrological Sciences*. <https://doi.org/10.1002/0470848944.hsa140>
- Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., & Savenije, H. (Eds.). (2013). *Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139235761>
- Boughton, W., & Chiew, F. (2007). Estimating runoff in ungauged catchments from rainfall, PET and the AWBM model. *Environmental Modelling & Software*, 22(4), 476–487. <https://doi.org/10.1016/j.envsoft.2006.01.009>
- Bourgin, F., Andréassian, V., Perrin, C., & Oudin, L. (2015). Transferring global uncertainty estimates from gauged to ungauged catchments. *Hydrology and Earth System Sciences*, 19(5), 2535–2546. <https://doi.org/10.5194/hess-19-2535-2015>
- Box, G. E. P., & Cox, D. R. (1964). *An Analysis of Transformations*.
- Brunner, M., Melsen, L. A., Newman, A., Wood, A., & Clark, M. P. (2020). Future streamflow regime changes in the United States: assessment using functional classification. *Hydrology and Earth System Sciences*, 24, 3951–3966. <https://doi.org/10.5194/hess-24-3951-2020>
- Bulygina, N., McIntyre, N., & Wheeler, H. (2011). Bayesian conditioning of a rainfall-runoff model for predicting flows in ungauged catchments and under land use

- changes. *Water Resources Research*, 47(2).
<https://doi.org/10.1029/2010WR009240>
- Bulygina, N., Ballard, C., McIntyre, N., O'Donnell, G., & Wheeler, H. (2012). Integrating different types of information into hydrological model parameter estimation: Application to ungauged catchments and land use scenario analysis. *Water Resources Research*, 48(6). <https://doi.org/10.1029/2011WR011207>
- Castellarin, A., Galeati, G., Brandimarte, L., Montanari, A., & Brath, A. (2004). Regional flow-duration curves: reliability for ungauged basins. *Advances in Water Resources*, 27(10), 953–965.
<https://doi.org/10.1016/j.advwatres.2004.08.005>
- Castellarin, A., Vogel, R. M., & Brath, A. (2004). A stochastic index flow model of flow duration curves. *Water Resources Research*, 40(3).
<https://doi.org/10.1029/2003WR002524>
- Castellarin, A., Camorani, G., & Brath, A. (2007). Predicting annual and long-term flow-duration curves in ungauged basins. *Advances in Water Resources*, 30(4), 937–953. <https://doi.org/10.1016/j.advwatres.2006.08.006>
- Clark, M., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research - WATER RESOUR RES*, 47. <https://doi.org/10.1029/2010WR009827>
- Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, 44(12). <https://doi.org/10.1029/2007WR006735>
- 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., Kavetski, D., & Fenicia, F. (2012). Reply to comment by K. Beven et al. on “Pursuing the method of multiple working hypotheses for hydrological modeling.” *Water Resources Research*, 48(11).
<https://doi.org/10.1029/2012WR012547>
- Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., et al. (2021). The Abuse of Popular Performance Metrics in Hydrologic Modeling. *Water Resources Research*, 57(9), e2020WR029001.
<https://doi.org/10.1029/2020WR029001>

- Dooge, J. C. I. (1987). Looking for hydrologic laws. *Journal of Hydrology*, 96(1–4), 3–4. [https://doi.org/10.1016/0022-1694\(87\)90138-7](https://doi.org/10.1016/0022-1694(87)90138-7)
- Dooge J. C. I. (1957). The rational method for estimating flood peaks. *Engineering*, 184:311–313, 374–377
- Draper, D. (1997). Assessment and Propagation of Model Uncertainty. *J R Stat Soc Series B*, 57. <https://doi.org/10.2307/2346087>
- Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., et al. (2006). Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *Journal of Hydrology*, 320(1), 3–17. <https://doi.org/10.1016/j.jhydrol.2005.07.031>
- Duan, Qingyun, Sorooshian, S., & Gupta, V. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, 28(4), 1015–1031. <https://doi.org/10.1029/91WR02985>
- Durante, F., & Sempi, C. (2010). Copula Theory: An Introduction. In P. Jaworski, F. Durante, W. K. Härdle, & T. Rychlik (Eds.), *Copula Theory and Its Applications* (pp. 3–31). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-12465-5_1
- Evin, G., Kavetski, D., Thyer, M., & Kuczera, G. (2013). Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. *Water Resources Research*, 49, 4518–4524. <https://doi.org/10.1002/wrcr.20284>
- Evin, G., Thyer, M., Kavetski, D., McInerney, D., & Kuczera, G. (2014). Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. *Water Resources Research*. <https://doi.org/10.1002/2013WR014185>
- Farmer, W. H., & Levin, S. (2018). Characterizing Uncertainty in Daily Streamflow Estimates at Ungauged Locations for the Massachusetts Sustainable Yield Estimator. *JAWRA Journal of the American Water Resources Association*, 54(1), 198–210. <https://doi.org/10.1111/1752-1688.12603>
- Fernandez, W., Vogel, R. M., & Sankarasubramanian, A. (2000). Regional calibration of a watershed model. *Hydrological Sciences Journal*, 45(5), 689–707. <https://doi.org/10.1080/02626660009492371>
- Feyen, L., Vrugt, J., Nualláin, B., Knijff, J., & Roo, A. (2007). Parameter optimisation and uncertainty assessment for large-scale streamflow simulation with the

- LISFLOOD model. *Journal of Hydrology*, 332, 276–289. <https://doi.org/10.1016/j.jhydrol.2006.07.004>
- Ganora, D., Laio, F., Masoero, A., & Claps, P. (2016). Spatially-smooth regionalization of flow duration curves in non-pristine basins. In *Proceedings of IAHS* (Vol. 373, pp. 73–80). Copernicus GmbH. <https://doi.org/10.5194/piahs-373-73-2016>
- Genest, C., & Favre, A.-C. (2007). Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. *Journal of Hydrologic Engineering - J HYDROL ENG*, 12. [https://doi.org/10.1061/\(ASCE\)1084-0699\(2007\)12:4\(347\)](https://doi.org/10.1061/(ASCE)1084-0699(2007)12:4(347))
- Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477), 359–378. <https://doi.org/10.1198/016214506000001437>
- Godet, J., Gaume, E., Javelle, P., Nicolle, P., & Payrastre, O. (2024). Technical note: Comparing three different methods for allocating river points to coarse-resolution hydrological modelling grid cells. *Hydrology and Earth System Sciences*, 28(6), 1403–1413. <https://doi.org/10.5194/hess-28-1403-2024>
- Greenwood, J. A., Landwehr, J. M., Matalas, N. C., & Wallis, J. R. (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressible in inverse form. *Water Resources Research*, 15(5), 1049–1054. <https://doi.org/10.1029/WR015i005p01049>
- 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), 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>
- He, Y., Bárdossy, A., & Zehe, E. (2011). A review of regionalisation for continuous streamflow simulation. *Hydrology and Earth System Sciences*, 15(11), 3539–3553. <https://doi.org/10.5194/hess-15-3539-2011>
- Hosking, J. R. M. (1990). L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1), 105–124. <https://doi.org/10.1111/j.2517-6161.1990.tb01775.x>
- Hosking, J. R. M. (1991). Approximations for Use in Constructing L-moment Ratio Diagrams. IBM Research Division, T.J. Watson Research Center. Retrieved from <https://books.google.it/books?id=ddqXPgAACAAJ>

- Hosking, J. R. M. (1999). L-moments and their applications in the analysis of financial data.
- Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal*, 58(6), 1198–1255. <https://doi.org/10.1080/02626667.2013.803183>
- Huang, M., & Liang, X. (2006). On the assessment of the impact of reducing parameters and identification of parameter uncertainties for a hydrologic model with applications to ungauged basins. *Journal of Hydrology*, 320(1), 37–61. <https://doi.org/10.1016/j.jhydrol.2005.07.010>
- Hundecha, Y., & Bárdossy, A. (2004). Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model. *Journal of Hydrology*, 292(1), 281–295. <https://doi.org/10.1016/j.jhydrol.2004.01.002>
- Hundecha, Y., Ouarda, T. B. M. J., & Bárdossy, A. (2008). Regional estimation of parameters of a rainfall-runoff model at ungauged watersheds using the “spatial” structures of the parameters within a canonical physiographic-climatic space. *Water Resources Research*, 44(1). <https://doi.org/10.1029/2006WR005439>
- Hunter, J., Thyer, M., McInerney, D., & Kavetski, D. (2021). Achieving high-quality probabilistic predictions from hydrological models calibrated with a wide range of objective functions. *Journal of Hydrology*, 603, 126578. <https://doi.org/10.1016/j.jhydrol.2021.126578>
- Jakeman, A. J., & Hornberger, G. M. (1993). How much complexity is warranted in a rainfall-runoff model? *Water Resources Research*, 29(8), 2637–2649. <https://doi.org/10.1029/93WR00877>
- Janzing, J., Wanders, N., van Tiel, M., van Jaarsveld, B., Karger, D. N., & Brunner, M. I. (2025). Hyper-resolution large-scale hydrological modelling benefits from improved process representation in mountain regions. *Hydrology and Earth System Sciences*, 29(23), 7041–7071. <https://doi.org/10.5194/hess-29-7041-2025>
- Joe, H. (1996). Families of m -variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In *Distributions with fixed marginals and related topics* (Vol. 28, pp. 120–142). Institute of Mathematical Statistics. <https://doi.org/10.1214/lnms/1215452614>

- Joe, H. (2014). *Dependence Modeling with Copulas*. New York: Chapman and Hall/CRC. <https://doi.org/10.1201/b17116>
- Katz, R. W., Parlange, M. B., & Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, 25(8), 1287–1304. [https://doi.org/10.1016/S0309-1708\(02\)00056-8](https://doi.org/10.1016/S0309-1708(02)00056-8)
- Kelly, K. S., & Krzysztofowicz, R. (1997). A bivariate meta-Gaussian density for use in hydrology. *Stochastic Hydrology and Hydraulics*, 11(1), 17–31. <https://doi.org/10.1007/BF02428423>
- 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, Wouter 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>
- Koutsoyiannis, D. (2019). Knowable moments for high-order stochastic characterization and modelling of hydrological processes. *Hydrological Sciences Journal*, 64(1), 19–33. <https://doi.org/10.1080/02626667.2018.1556794>
- Koutsoyiannis, D., & Montanari, A. (2022a). Bluecat: A Local Uncertainty Estimator for Deterministic Simulations and Predictions. *Water Resources Research*, 58. <https://doi.org/10.1029/2021WR031215>
- Koutsoyiannis, D., & Montanari, A. (2022b). Climate Extrapolations in Hydrology: The Expanded Bluecat Methodology. *Hydrology*, 9(5), 86. <https://doi.org/10.3390/hydrology9050086>
- Kuczera, G., Kavetski, D., Franks, S., & Thyer, M. (2006). Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*, 331(1), 161–177. <https://doi.org/10.1016/j.jhydrol.2006.05.010>
- Laio, F., & Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11(4), 1267–1277. <https://doi.org/10.5194/hess-11-1267-2007>
- Levin, S. B., & Farmer, W. H. (2020). Evaluation of Uncertainty Intervals for Daily, Statistically Derived Streamflow Estimates at Ungaged Basins across the Continental U.S. *Water*, 12(5), 1390. <https://doi.org/10.3390/w12051390>

- Loague K. (2010), editor. Benchmark Papers in Hydrology: Rainfall–Runoff Modelling. IAHS Press, Wallingford, UK
- Lombardo, L., Papalexiou, S. M., Thébault, C., Clark, M., Vogel, R., & Viglione, A. (2025). Residual Dynamics in Hydrological Models: Insights from a large sample of catchments and models. *Advances in Water Resources*, 206, 105165. <https://doi.org/10.1016/j.advwatres.2025.105165>
- McInerney, D., Thyer, M., Kavetski, D., Lerat, J., & Kuczera, G. (2017). Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research*, 53. <https://doi.org/10.1002/2016WR019168>
- McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Tuteja, N., & Kuczera, G. (2020). Multi-temporal Hydrological Residual Error Modeling for Seamless Subseasonal Streamflow Forecasting. *Water Resources Research*, 56. <https://doi.org/10.1029/2019WR026979>
- McIntyre, N., Lee, H., Wheeler, H., Young, A., & Wagener, T. (2005a). Ensemble predictions of runoff in ungauged catchments. *Water Resources Research*, 41(12). <https://doi.org/10.1029/2005WR004289>
- McIntyre, N., Lee, H., Wheeler, H., Young, A., & Wagener, T. (2005b). Ensemble predictions of runoff in ungauged catchments. *Water Resources Research*, 41(12). <https://doi.org/10.1029/2005WR004289>
- Merz, B., Hall, J., Disse, M., & Schumann, A. (2010). Fluvial flood risk management in a changing world. *Natural Hazards and Earth System Sciences*, 10. <https://doi.org/10.5194/nhess-10-509-2010>
- Merz, R., & Blöschl, G. (2004). Regionalisation of catchment model parameters. *Journal of Hydrology*, 287(1), 95–123. <https://doi.org/10.1016/j.jhydrol.2003.09.028>
- Merz, R., Tarasova, L., & Basso, S. (2020). Parameter’s Controls of Distributed Catchment Models—How Much Information is in Conventional Catchment Descriptors? *Water Resources Research*, 56(2), e2019WR026008. <https://doi.org/10.1029/2019WR026008>
- Montanari, A. (2005). Large sample behaviors of GLUE in assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research - WATER RESOURCES*, 410. <https://doi.org/10.1029/2004WR003826>

- Montanari, A., & Brath, A. (2004). A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research*, 40(1). <https://doi.org/10.1029/2003WR002540>
- 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., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, 48(9). <https://doi.org/10.1029/2011WR011412>
- Morawietz, M., Xu, C.Y., Gottschalk, L., Tallaksen, L.M., (2011). Systematic evaluation of autoregressive error models as post-processors for a probabilistic streamflow forecast system. *Journal of Hydrology* 407, 58–72. doi:10.1016/j.jhydrol.2011.07.007.
- Mullen, K. M., Ardia, D., Gil, D. L., Windover, D., & Cline, J. (2011). DEoptim: An R Package for Global Optimization by Differential Evolution. *Journal of Statistical Software*, 40, 1–26. <https://doi.org/10.18637/jss.v040.i06>
- Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Erhardt, T., et al. (2024, December 17). VineCopula: Statistical Inference of Vine Copulas (Version 2.6.0). Retrieved from <https://cran.r-project.org/web/packages/VineCopula/index.html>
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3), 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
- Nelsen, R. B. (2006). *An introduction to copulas* (2. ed). New York: Springer.
- Newman, A., Clark, M., Sampson, K., Wood, A., Hay, L., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, 19, 209–223. <https://doi.org/10.5194/hess-19-209-2015>
- Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., & Loumagne, C. (2005). Which potential evapotranspiration input for a lumped rainfall-runoff model? *Journal of Hydrology*, 303(1–4), 290–306. <https://doi.org/10.1016/j.jhydrol.2004.08.026>
- Oudin, L., Andréassian, V., Perrin, C., Michel, C., & Le Moine, N. (2008). Spatial proximity, physical similarity, regression and ungaged catchments: A

- comparison of regionalization approaches based on 913 French catchments. *Water Resources Research*, 44(3). <https://doi.org/10.1029/2007WR006240>
- Oudin, L., Kay, A., Andréassian, V., & Perrin, C. (2010). Are seemingly physically similar catchments truly hydrologically similar? *Water Resources Research*, 46(W11558), 15. <https://doi.org/10.1029/2009WR008887>
- Papacharalampous, G., Tyralis, H., Langousis, A., Jayawardena, A. W., Sivakumar, B., Mamassis, N., et al. (2019). Probabilistic Hydrological Post-Processing at Scale: Why and How to Apply Machine-Learning Quantile Regression Algorithms. *Water*, 11(10), 2126. <https://doi.org/10.3390/w11102126>
- Papalexiou, S. M. (2018). Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal distributions, correlation structures, and intermittency. *Advances in Water Resources*, 115, 234–252. <https://doi.org/10.1016/j.advwatres.2018.02.013>
- Parajka, J., Merz, R., & Blöschl, G. (2007). Uncertainty and multiple objective calibration in regional water balance modelling: Case study in 320 Austrian catchments. *Hydrological Processes*, 21, 435–446. <https://doi.org/10.1002/hyp.6253>
- Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, 279(1–4), 275–289. [https://doi.org/10.1016/S0022-1694\(03\)00225-7](https://doi.org/10.1016/S0022-1694(03)00225-7)
- Pianosi, F., Raso, L., (2012). Dynamic modeling of predictive uncertainty by regression on absolute errors. *Water Resources Research* 48, W03516. doi:10.1029/2011wr010603
- Pickands, J. (1975). Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 3(1), 119–131. <https://doi.org/10.1214/aos/1176343003>
- Plate, E. J., Ihringer, J., & Lutz, W. (1988). Operational models for flood calculations. *Journal of Hydrology*, 100, 489–506. [https://doi.org/10.1016/0022-1694\(88\)90198-9](https://doi.org/10.1016/0022-1694(88)90198-9)
- Quilty, J., & Adamowski, J. (2020). A stochastic wavelet-based data-driven framework for forecasting uncertain multiscale hydrological and water resources processes. *Environmental Modelling & Software*, 130, 104718. <https://doi.org/10.1016/j.envsoft.2020.104718>
- 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>

- Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., & Franks, S. W. (2011). Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. *Water Resources Research*, 47(11). <https://doi.org/10.1029/2011WR010643>
- Sang, Y. (2013). A review on the applications of wavelet transform in hydrology time series analysis. *Atmospheric Research*, 122, 8–15. <https://doi.org/10.1016/j.atmosres.2012.11.003>
- Sang, Y., Singh, V., Sun, F., Yaning, C., Liu, Y., & Yang, M. (2016). Wavelet-Based Hydrological Time Series Forecasting. *Journal of Hydrologic Engineering*, 21, 06016001. [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001347](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001347)
- Schaeffli, B., Talamba, D. B., & Musy, A. (2007). Quantifying hydrological modeling errors through a mixture of normal distributions. *Journal of Hydrology*, 332(3–4), 303–315. <https://doi.org/10.1016/j.jhydrol.2006.07.005>
- Scheuerer, M., & Hamill, T. M. (2015). Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities. *Monthly Weather Review*, 143(4), 1321–1334. <https://doi.org/10.1175/MWR-D-14-00269.1>
- Schoups, G., & Vrugt, J. (2010). A Formal Likelihood Function for Parameter and Predictive Inference of Hydrologic Models With Correlated, Heteroscedastic, and Non-Gaussian Errors. *Water Resources Research*, 46. <https://doi.org/10.1029/2009WR008933>
- Seibert, J., & Bergström, S. (2022). A retrospective on hydrological catchment modelling based on half a century with the HBV model. *Hydrology and Earth System Sciences*, 26(5), 1371–1388. <https://doi.org/10.5194/hess-26-1371-2022>
- Shabestanipour, G., Brodeur, Z., Farmer, W., Steinschneider, S., & Lamontagne, J. (2023). Stochastic Watershed Model Ensembles for Long-Range Planning: Verification and Validation. *Water Resources Research*, 59. <https://doi.org/10.1029/2022WR032201>
- Sikorska, A. E., Montanari, A., & Koutsoyiannis, D. (2015). Estimating the Uncertainty of Hydrological Predictions through Data-Driven Resampling Techniques. *Journal of Hydrologic Engineering*, 20(1), A4014009. [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0000926](https://doi.org/10.1061/(ASCE)HE.1943-5584.0000926)
- Sillitto, G. P. (1951). interrelations between certain linear systematic statistics of samples from any continuous population. *Biometrika*, 38(3–4), 377–382. <https://doi.org/10.1093/biomet/38.3-4.377>

- Simonovic, S. (2012). Managing Water Resources: Methods and Tools for a Systems Approach. *Managing Water Resources: Methods and Tools for a Systems Approach*, 1–640. <https://doi.org/10.4324/9781849771917>
- Singh V. P. (1995). *Computer Models of Watershed Hydrology*. Water Resource Publications, Highlands Ranch, CO
- Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., et al. (2003). IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, 48(6), 857–880. <https://doi.org/10.1623/hysj.48.6.857.51421>
- Sklar, M. (1959). Fonctions de répartition à N dimensions et leurs marges.
- Smakhtin, V., Hughes, D., & Creuse-Naudin, E. (1997). Regionalization of Daily Flow Characteristics in Part of the Eastern Cape, South Africa. *Hydrological Sciences Journal*, 42, 919–936. <https://doi.org/10.1080/02626669709492088>
- Smith, P.J., Beven, K.J., Weerts, A.H., Leedal, D., (2012). Adaptive correction of deterministic models to produce probabilistic forecasts. *Hydrology and Earth System Sciences* 16, 2783–2799. doi:10.5194/hess-16-2783-2012.
- Sorooshian, S., & Dracup, J. A. (1980). Stochastic Parameter Estimation Procedures for Hydrologic Rainfall-Runoff Models: Correlated and Heteroscedastic Error Cases. *Water Resources Research*, 16, 430–442. <https://doi.org/10.1029/WR016i002p00430>
- Stedinger, J. R., Vogel, R. M., Lee, S. U., & Batchelder, R. (2008). Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resources Research*, 44(12). <https://doi.org/10.1029/2008WR006822>
- Storn, R., & Price, K. (1997). Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *Journal of Global Optimization*, 11, 341–359. <https://doi.org/10.1023/A:1008202821328>
- Tajiki, M., Schoups, G., Hendricks Franssen, H. J., Najafinejad, A., & Bahremand, A. (2020). Recursive Bayesian Estimation of Conceptual Rainfall-Runoff Model Errors in Real-Time Prediction of Streamflow. *Water Resources Research*, 56(2), e2019WR025237. <https://doi.org/10.1029/2019WR025237>
- Valdez, E. S., Anctil, F., & Ramos, M.-H. (2022). Choosing between post-processing precipitation forecasts or chaining several uncertainty quantification tools in hydrological forecasting systems. *Hydrology and Earth System Sciences*, 26(1), 197–220. <https://doi.org/10.5194/hess-26-197-2022>
- Vogel, R. M. (2005). Regional Calibration of Watershed Models. In *Watershed Models* (pp. 47–71). <https://doi.org/10.1201/9781420037432.ch3>

- Vogel, R. M. (2017). Stochastic watershed models for hydrologic risk management. *Water Security*, 1, 28–35. <https://doi.org/10.1016/j.wasec.2017.06.001>
- Vogel, R. M., Papalexiou, S. M., Lamontagne, J. R., & Dolan, F. C. (2024). When Heavy Tails Disrupt Statistical Inference. *The American Statistician*, 1–15. <https://doi.org/10.1080/00031305.2024.2402898>
- Wagener, T., Wheeler, H., & Gupta, H. (2004). Rainfall-Runoff Modelling In Gauged And Ungauged Catchments. <https://doi.org/10.1142/9781860945397>
- Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment Classification and Hydrologic Similarity. *Geography Compass*, 1(4), 901–931. <https://doi.org/10.1111/j.1749-8198.2007.00039.x>
- Wang, Q. J., Shrestha, D. L., Robertson, D. E., & Pokhrel, P. (2012). A log-sinh transformation for data normalization and variance stabilization. *Water Resources Research*, 48(5). <https://doi.org/10.1029/2011WR010973>
- Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences* (2nd ed). Amsterdam Paris: Elsevier.
- Winsemius, H. C., Schaefli, B., Montanari, A., & Savenije, H. H. G. (2009). On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. *Water Resources Research*, 45(12). <https://doi.org/10.1029/2009WR007706>
- Xie, Z., Yuan, J., Duan, Q., Zheng, X., Liang, M., & Chen, F. (2007). Regional Parameter Estimation of the VIC Land Surface Model: Methodology and Application to River Basins in China. *Journal of Hydrometeorology*, 8, 447–468. <https://doi.org/10.1175/JHM568.1>
- Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, 30(8), 1756–1774. <https://doi.org/10.1016/j.advwatres.2007.01.005>
- Zhang, Z., Wagener, T., Reed, P., & Bhushan, R. (2008). Reducing uncertainty in predictions in ungauged basins by combining hydrologic indices regionalization and multiobjective optimization. *Water Resources Research*, 44(12). <https://doi.org/10.1029/2008WR006833>

DATASETS

European Environmental Agency (2012) ECRINS – European Catchments and Rivers

- Network System. EEA Technical Report No. 9/2012.
- FAO, IIASA, ISRIC, ISSCAS, and JRC (2012). Harmonized World Soil Database (version 1.2). FAO, Rome, Italy and IIASA, Laxenburg, Austria.
- Copernicus Land Monitoring Service (2018). CORINE Land Cover (CLC) 2018, Version 20. European Environment Agency (EEA).
- Stevanović, Z., Vigna, B., Galve, J. P., Andreo, B., et al. (2015). Mediterranean Karst Aquifer Map (MEDKAM). International Association of Hydrogeologists (IAH), UNESCO.
- RGI Consortium (2017). Randolph Glacier Inventory – A Dataset of Global Glacier Outlines: Version 6.0. Global Land Ice Measurements from Space (GLIMS).
- Pfeffer, W. T., Arendt, A. A., Bliss, A., et al. (2014). The Randolph Glacier Inventory: a globally complete inventory of glaciers. *Journal of Glaciology*, 60(221), 537–552. <https://doi.org/10.3189/2014JoG13J176>
- European Space Agency (ESA) (2021). Copernicus DEM – Global Digital Elevation Model (GLO-90), Version 3.0. Copernicus Programme.
- Thébault, C. (2025). FUSE simulations, HydroShare, <http://www.hydroshare.org/resource/d450bfdfdb94458ac0577f04dd60529>

APPENDIX

Parameter	Units	Database	Description
LD_smallveg	%	CLC	Land Cover Fraction (Small Vegetation)
LD_agri	%	CLC	Land Cover Fraction (agricultural fields)
LD_wetland	%	CLC	Land Cover Fraction (Wetland)
LD_urban	%	CLC	Land Cover Fraction (Urban areas)
LD_forest	%	CLC	Land Cover Fraction (Forest)
meanelev	m	ESA GLO-90	Mean Elevation
CVelev	-	ESA GLO-90	Elevation Coefficient of Variation
meanslope	%	ESA GLO-90	Mean slope
meanaspect	°	ESA GLO-90	Mean Aspect
S_CLAY	%	HWSD	Subsoil Clay Fraction
S_SILT	%	HWSD	Subsoil Silt Fraction
S_SAND	%	HWSD	Subsoil Sand Fraction
S_GRAVEL	%	HWSD	Subsoil Gravel Content
T_CLAY	%	HWSD	Topsoil Clay Fraction
T_SILT	%	HWSD	Topsoil Silt Fraction
T_SAND	%	HWSD	Topsoil Sand Fraction
T_GRAVEL	%	HWSD	Topsoil Gravel Content
S_BULK_DENSITY	kg/dm ³	HWSD	Density Subsoil Bulk Density
T_BULK_DENSITY	kg/dm ³	HWSD	Density Topsoil Bulk Density
AWC_LARGE	%	HWSD	Available Water Storage Capacity (Large) Fraction
AWC_MED	%	HWSD	Available Water Storage Capacity (Medium) Fraction
AWC_SMALL	%	HWSD	Available Water Storage Capacity (Small) Fraction
IL_TOP	%	HWSD	Shallow Impermeable Layer Fraction
IL_MED	%	HWSD	Medium-Deep Impermeable Layer Fraction
IL_DEEP	%	HWSD	Deep Impermeable Layer Fraction
SWR_NOTWET	%	HWSD	Not Wet Soil Water Regime Fraction (Code 1)
SWR_MEDIUMWET	%	HWSD	Not Wet Soil Water Regime Fraction (Code 2)
SWR_WET	%	HWSD	Not Wet Soil Water Regime Fraction (Code 3)
SWR_TOTALWET	%	HWSD	Not Wet Soil Water Regime Fraction (Code 4)
T_TEXTURE_COARSE	%	HWSD	Coarse Topsoil Texture (Code 1) Fraction
T_TEXTURE_MEDIUM	%	HWSD	Medium Topsoil Texture (Code 2) Fraction
T_TEXTURE_FINE	%	HWSD	Fine Topsoil Texture (Code 3) Fraction
S_USDA_CLAY	%	HWSD	Subsoil USDA Clay Class Fraction

Parameter	Units	Database	Description
S_USDA_SILTLOAM	%	HWSD	Subsoil USDA Silt Class Fraction
S_USDA_SAND	%	HWSD	Subsoil USDA Sand Class Fraction
T_USDA_CLAY	%	HWSD	Topsoil USDA Clay Class Fraction
T_USDA_SILTLOAM	%	HWSD	Topsoil USDA Silt Class Fraction
T_USDA_SAND	%	HWSD	Topsoil USDA Sand Class Fraction
DRAINAGE_LARGE	%	HWSD	Large Soil Drainage Fraction
DRAINAGE_MED	%	HWSD	Medium Soil Drainage Fraction
DRAINAGE_SMALL	%	HWSD	Small Soil Drainage Fraction
CDUH_map	mm	ERA5LAND	Long-Term Mean Annual Precipitation
CDUH_mat	°C	ERA5LAND	Long-Term Mean Annual Temperature
CDUH_mapet	mm	ERA5LAND	Long-Term Mean Annual Potential Evapotranspiration
CDUH_petovp	-	ERA5LAND	Aridity Index
CDUH_psum2win	-	ERA5LAND	Ratio of Long-Term Summer Precipitation (May-Oct) and Winter Precipitation (Nov-Apr)
CDUH_r50	mm/day	ERA5LAND	Long-Term Median Maximum Daily Precipitation
CDUH_r95	mm/day	ERA5LAND	Long-Term 95th Percentile of Maximum Daily Precipitation
CDUH_drd2d	mm	ERA5LAND	Long-Term Mean Absolute Difference of Rainfall Amount Between Two Consecutive Days
n_sub	-	UH_dataset	Number of HU constituting the (sub)catchment
AREA	km ²	UH_dataset	(Sub)catchment area
LON	m	UH_dataset	Longitude
LAT	m	UH_dataset	Latitude
SEA_dist	m	UH_dataset	Distance from the sea
Karst_Mid	%	MEDKAM	Medium Karstic Fraction
Karst_None	%	MEDKAM	No Karstic Fraction
Karst	%	MEDKAM	Strong Karstic Fraction
Glaciers	%	RGI	Catchment Area Fraction Covered by Permanent Glaciers

R SCRIPTS

A. TUWmodel (GAR) calibration

```
#####
# 1. USER SETTINGS
#####

# Calibration mode
mode <- "independent"    # options: "independent" or "network"

# Objective Function
of <- "KGE"              # options: "KGE" or "wNSE"

# Saving path
save_path <- "C:/YOUR/SAVE/PATH/"
save_name <- "calibration_results.RData"

#####
# 2. LIBRARIES
#####

library(sf)
library(raster)
library(zoo)
library(TUWmodel)
library(DEoptim)
library(hydroGOF)

#####
# 3. FUNCTIONS
#####

GAR_level <- function(GAR) {
  level <- vector()
  for(i in 1:dim(GAR)[1]){
    temp <- GAR$ZHYD[i]
    flag <- FALSE
    count <- 1
    while(flag == FALSE){
      temp <- GAR$ZHYD[which(GAR$NextDown %in% temp)]
      if(length(temp) == 0){
        flag <- TRUE
      }
      if(length(temp) != 0){

```

```

        count <- count + 1
      }
    }
    level[i] <- count
  }
  L_cat <- as.data.frame(cbind(GAR$ZHYD, level))
  L_cat[,2] <- as.numeric(L_cat[,2])
  colnames(L_cat) <- c("ZHYD", "Level")
  return(L_cat)
}

#####
# 4. LOAD BASIN DATA
#####

setwd("PATH_TO_CLUSTER_FILES")
A <- readLines("ZHYDcluster1.txt")
B <- readLines("ZHYDcluster2.txt")
AB <- c(A,B)      #ID of the HU in the model attributed to a gauge
setwd("PATH_TO_GAR_SHAPEFILE")
GAR <- st_read("GAR_constructed_new.shp")
L_cat <- GAR_level(GAR)

#####
# 5. LOAD STATION-BASIN LINKS
#####

setwd("PATH_TO_POINTTOCAT")
PointToCat <- read.table("PointToCat_SL3_10.txt", header=TRUE)
level <- vector()
for(i in 1:dim(PointToCat)[1]){
  dummy <- which(L_cat$ZHYD == PointToCat$ZHYD[i])
  level[i] <- L_cat$Level[dummy]
}
PointToCat <- cbind(PointToCat, level)
dummy <- order(PointToCat$level, decreasing=FALSE)
PointToCat <- PointToCat[dummy,]

#####
# 6. LOAD DISCHARGE DATA
#####

setwd("PATH_TO_DISCHARGE")
load("Q_SL3_10.RData")
data_list <- obs
stations <- names(data_list)

if(length(which((names(data_list) %in% PointToCat$ORIGINA) ==
FALSE)) != 0){

```

```

    data_list <- data_list[~which((names(data_list) %in%
PointToCat$ORIGINA) == FALSE)]
}
if(length(which(PointToCat$ORIGINA %in% names(data_list) == FALSE))
!= 0){
    PointToCat <- PointToCat[~which(PointToCat$ORIGINA %in%
names(data_list) == FALSE),]
}

dummy <- vector()
for(i in 1:length(data_list)){
    dummy[i] <- which(names(data_list) == PointToCat$ORIGINA[i])
}
data_list <- data_list[dummy]
obs <- data_list

for(i in 1:length(obs)){
    dummy <- obs[[i]]
    dummy[dummy == -999] <- NA
    obs[[i]] <- dummy
}

#####
# 7. FILTER BASINS
#####

PointToCat <- PointToCat[PointToCat$ORIGINA %in% stations,]
PointToCat <- PointToCat$ZHYD
obs <- obs[!(PointToCat %in% AB)]
PointToCat <- PointToCat[!(PointToCat %in% AB)]

#####
# 8. AREA FILTER
#####

load("area_cats_GAR.RData")
AREA <- vector()
for(i in 1:length(PointToCat)){
    AREA[i] <- area_cats_GAR[[which(names(area_cats_GAR) ==
PointToCat[i])]]
}
obs <- obs[which(AREA/1000000 < 1000)]
PointToCat <- PointToCat[which(AREA/1000000 < 1000)]

#####
# 9. CLIMATE INPUTS
#####

```

```

setwd("PATH_TO_CLIMATE_DATA")

load("P_short_red.RData")
load("T_short_red.RData")
load("PET_short_red.RData")
CI <- list()
colnames(PET) <- colnames(P)
CI$P <- P
CI$T <- T
CI$PET <- PET

#####
# 10. PARAMETER INITIALIZATION
#####

param_mat <- matrix(NA, ncol=length(GAR$geometry), nrow=16)
for(i in 1:length(GAR$geometry)){
  param_mat[,i] <- c(1.2, 2.5, 2, -
1, 0, 0.5, 300, 10, 1, 15, 125, 50, 4, 15, 25, 0.8)
}

# Indicates if a basin has already been calibrated (used in network
mode)
CalYet <- rep(FALSE, length(GAR$geometry))

# Indicates if observed discharge must be converted from mm/d to
m³/s
mm_to_m3 <- rep(FALSE, length(PointToCat))

#####
# 11. GRID
#####

setwd("PATH_TO_GRID")
grid <- st_read("grid_new.shp")
colnames(grid)[1] <- "Px1"
grid$Px1 <- c(1:length(grid$Px1))
grid <- st_transform(grid, crs(GAR))
GAR <- st_make_valid(GAR)
grid <- st_make_valid(grid)
area_px1 <- as.numeric(st_area(grid$geometry))

#####
# 12. CALIBRATION SETUP
#####

calibr <- vector('list', length(PointToCat))
names(calibr) <- as.numeric(GAR$ZHYD[PointToCat])
obs_cal <- obs

```

```

for(mm in 1:length(PointToCat)){
  cat("Basin ",mm," out of ",length(PointToCat),"\n",sep="")
  obs <- obs_cal[[mm]]
  dummy <- CI[[1]][,1]
  obs <- window(obs, start=as.Date("2010-01-01"), end=as.Date("2019-
07-31"))
  obs <- apply(merge.zoo(obs,dummy),1,sum) - dummy

#####
# Identify upstream basins
#####

Z_temp <- GAR$ZHYD[PointToCat[mm]]
flag <- FALSE
temp <- Z_temp
while(flag == FALSE){
  temp <- GAR$ZHYD[which(GAR$NextDown %in% temp)]
  if(length(temp) != 0){
    Z_temp <- c(Z_temp, temp)
  }
  if(length(temp) == 0){
    flag <- TRUE
  }
}

#####
# Build topology
#####

flag <- vector()
for(i in 1:length(Z_temp)){
  flag[i] <- which(GAR$ZHYD == Z_temp[i])
}
down <- vector()
for(i in 1:length(flag)){
  down[i] <- GAR$NextDown[flag[i]]
}
Z_temp <- as.data.frame(cbind(Z_temp, (1:length(down))))
Z_temp[,2] <- as.numeric(Z_temp[,2])

for(i in 1:dim(Z_temp)[1]){
  if((down[i] %in% Z_temp[,1]) == FALSE){
    down[i] <- NA
  }
  if((down[i] %in% Z_temp[,1]) == TRUE){
    down[i] <- Z_temp[which(Z_temp[,1] == down[i]),2]
  }
}

```

```

down <- as.numeric(down)
#####
# SUB-BASIN AREAS
#####

area_cats <- vector()
for(i in 1:length(flag)){
  area_cats[i] <- as.numeric(st_area(GAR$geometry[flag[i]]))
}

#####
# OBS UNIT CONVERSION
#####

if(mm_to_m3[mm] == TRUE){
  obs <- obs*((10^-3)*sum(area_cats)/86400)
}

#####
# LEVELS
#####

Level <- rep(0, length(down))
L <- unique(c(1:length(down))['c(1:length(down)) %in% down])
Level[L] <- 1
flag_l <- 2
k <- FALSE
while(k==FALSE){
  L <- unique(down[L])
  Level[L] <- flag_l
  flag_l <- flag_l + 1
  if(sum(!is.na(L))==0){
    k <- TRUE
  }
}
Z_temp <- cbind(Z_temp, Level, down)

#####
# PIXEL INTERSECTIONS
#####

topo <- vector('list', dim(Z_temp)[1])
cat_area <-
as.numeric(st_area(st_union(GAR$geometry[as.numeric(Z_temp$Z_temp)]))
))
for(i in 1:dim(Z_temp)[1]){
  dummy <- st_intersection(grid, GAR$geometry[which(GAR$ZHYD ==
Z_temp[i,1])])
  topo[[i]]$pxl <- as.numeric(rownames(dummy))
}

```

```

      topo[[i]]$weights <-
as.numeric(st_area(dummy$geometry))/cat_area
      topo[[i]]$zones <- as.numeric(st_area(dummy$geometry))
      topo[[i]]$weights <- topo[[i]]$weights/sum(topo[[i]]$weights)
    }
    names(topo) <- Z_temp$Z_temp

ME1 <- function(param, prec, tmean, pet, Z_temp, topo, disc, of) {

  ##### PRE-CALCULATE UH FOR NASH #####
  inter <- matrix(NA, ncol = dim(Z_temp)[1], nrow = length(disc))
  UH <- vector()
  for(t in 1:5000){
    UH[t] <- (1/((param[16])*factorial(0)))*(t^(0))*exp(-
(t/param[16]))
  }
  UH <- UH[1:1000]
  UH <- UH/sum(UH)
  for(pp in 1:max(Z_temp$Level)){
    qualicat <- which(Z_temp$Level == pp)
    for(jj in 1:length(qualicat)){
      TUW <- TUWmodel(prec = as.matrix(prec[,which(colnames(prec)
%in% topo[[qualicat[jj]]$pxl]]),
      airt =
as.matrix(tmean[,which(colnames(tmean) %in%
topo[[qualicat[jj]]$pxl]]),
      ep = as.matrix(pet[,which(colnames(pet) %in%
topo[[qualicat[jj]]$pxl]]),
      area = topo[[qualicat[jj]]$weights], param =
param[1:15])

      FC <- vector()
      for(conv in 1:length(topo[[qualicat[jj]]$pxl])){
        FC[conv] <- topo[[qualicat[jj]]$zones[conv]*(10^-3)/86400
      }
      flag_dummy <- FALSE
      if(dim(TUW$qzones)[1] == 1){
        TUW <- t(TUW$qzones)*FC
        flag_dummy <- TRUE
      }
      if(flag_dummy == FALSE){
        for(conv in 1:dim(TUW$qzones)[2]){
          TUW$qzones[,conv] <- TUW$qzones[,conv]*FC[conv]
        }
        TUW <- apply(TUW$qzones,1,sum)
      }
      if(Z_temp$Level[qualicat[jj]] != 1){
        dummy <- which(Z_temp$down == qualicat[jj])
        if(is.null(dim(inter[,dummy])) == TRUE){

```

```

      add <- inter[,dummy]
      add[is.na(add)] <- 0
      TUW <- TUW + add
    }
    if(is.null(dim(inter[,dummy])) == FALSE){
      add <- apply(inter[,dummy], 1, sum, na.rm=TRUE)
      TUW <- TUW + add
    }
  }
  TUW[is.na(TUW)] <- 0
  Q_out <- convolve(TUW, rev(UH), type='open')
  Q_out <- Q_out[1:length(TUW)]
  inter[,qualicat[jj]] <- Q_out
  TUW[is.na(TUW)] <- 0
  Q_out <- convolve(TUW, rev(UH), type='open')
  Q_out <- Q_out[1:length(TUW)]
  inter[,qualicat[jj]] <- Q_out
}
}

simu <- inter[,which(Z_temp$Level == max(Z_temp$Level))]
simu[is.nan(simu)] <- -999
simu <- zoo(simu, time(disc))
simu <- simu[-c(1:303)]
obse <- disc[-c(1:303)]
if(of == "KGE"){
  eff <- KGE(simu, obse, na.rm=TRUE)
  eff <- -eff
}
if(of == "wNSE"){
  eff <- wNSE(simu, obse, na.rm=TRUE)
  eff <- -eff
}
return(eff)
}

#### MODEL EFFICIENCY 2 - SOME HUS ALREADY CALIBRATED ####

ME2 <- function(param, Done_Yet , CalYet_2, prec, tmean, pet,
Z_temp, topo, disc, of) {
  inter <- matrix(NA, ncol = dim(Z_temp)[1], nrow = length(disc))

  #### PRE-CALCULATE UH FOR NASH ####
  UH <- vector()
  for(t in 1:5000){
    UH[t] <- (1/((param[16])*factorial(0)))*(t^(0))*exp(-(t/param[16]))
  }
  UH <- UH[1:1000]

```

```

UH <- UH/sum(UH)
for(pp in 1:max(Z_temp$Level)){
  qualicat <- which(Z_temp$Level == pp)
  for(jj in 1:length(qualicat)){
    if(CalYet_2[qualicat[jj]] == FALSE){
      TUV <- TUVmodel(prec =
as.matrix(prec[,which(colnames(prec) %in%
topo[[qualicat[jj]]$pxl]]),
          airt =
as.matrix(tmean[,which(colnames(tmean) %in%
topo[[qualicat[jj]]$pxl]]),
          ep = as.matrix(pet[,which(colnames(pet)
%in% topo[[qualicat[jj]]$pxl]]),
          area = topo[[qualicat[jj]]$weights, param
= param[1:15]))

    FC <- vector()
    for(conv in 1:length(topo[[qualicat[jj]]$pxl])){
      FC[conv] <- topo[[qualicat[jj]]$zones[conv]*(10^-
3)/86400
    }
    flag_dummy <- FALSE
    if(dim(TUV$qzones)[1] == 1){
      TUV <- t(TUV$qzones)*FC
      flag_dummy <- TRUE
    }
    if(flag_dummy == FALSE){
      for(conv in 1:dim(TUV$qzones)[2]){
        TUV$qzones[,conv] <- TUV$qzones[,conv]*FC[conv]
      }
      TUV <- apply(TUV$qzones,1,sum)
    }
    if(Z_temp$Level[qualicat[jj]] != 1){
      dummy <- which(Z_temp$down == qualicat[jj])
      if(is.null(dim(inter[,dummy])) == TRUE){
        add <- inter[,dummy]
        add[is.na(add)] <- 0
        TUV <- TUV + add
      }
      if(is.null(dim(inter[,dummy])) == FALSE){
        add <- apply(inter[,dummy], 1, sum, na.rm=TRUE)
        TUV <- TUV + add
      }
    }
    TUV[is.na(TUV)] <- 0
    Q_out <- convolve(TUV, rev(UH), type='open')
    Q_out <- Q_out[1:length(TUV)]
    inter[,qualicat[jj]] <- Q_out
  }
}

```

```

        if(CalYet_2[qualicat[jj]] == TRUE){
            inter[,qualicat[jj]] <- Done_Yet[,qualicat[jj]]
        }
    }
}
simu <- inter[,which(Z_temp$Level == max(Z_temp$Level))]
simu[is.nan(simu)] <- -999
simu <- simu[-c(1:303)]
obse <- disc[-c(1:303)]
if(of == "KGE"){
    eff <- KGE(simu, obse, na.rm=TRUE)
}if(of == "KGE"){
    eff <- KGE(simu, obse, na.rm=TRUE)
    eff <- -eff
}
if(of == "wNSE"){
    eff <- wNSE(simu, obse, na.rm=TRUE)
    eff <- -eff
}
return(eff)
}
cat("Basin ",mm," out of ",length(PointToCat)," / ",length(topo),"
sub-basins \n",sep="")

quali_pixel_current <- vector()
for(cycle in 1:length(topo)){
    quali_pixel_current <- c(quali_pixel_current, topo[[cycle]]$pxl)
}
quali_pixel_current <- unique(quali_pixel_current)

cal_flag <- FALSE

if(sum(CalYet[which(colnames(param_mat)%in%Z_temp$Z_temp)])==0){
    timestamp()
    calibr_temp <- try(DEoptim(fn=ME1, lower=c(SCF=0.9, DDF=0.0,
Tr=1.0, Ts=-3.0, Tm=-2.0, LPrat=0.0, FC=0.0, BETA=0.0, k0=0.0,
k1=2.0, k2=30.0, cperc=0.0, bmax=0.0, croute=0.0, k_nash = 0.1),
upper=c(SCF=1.5, DDF=5.0, Tr=3.0, Ts=1.0, Tm=2.0, LPrat=1.0,
FC=600.0, BETA=20.0, k0=2.0, k1=30.0, k2=250.0, lsuz=100.0,
cperc=8.0, bmax=30.0, croute=50.0, k_nash = 1),
control=DEoptim.control(NP=NA, itermax=200, reltol=1e-3, steptol=10,
trace=5, parallelType=1, packages=c('TUWmodel','hydroGOF')),
prec = CI[[1]][,which(colnames(CI[[1]]) %in% quali_pixel_current)],
tmean=CI[[2]][,which(colnames(CI[[1]]) %in% quali_pixel_current)],
pet=CI[[3]][,which(colnames(CI[[1]]) %in% quali_pixel_current)],
Z_temp = Z_temp, topo = topo,
disc=obs, of = of))
    calibr[[mm]] <- calibr_temp

```

```

timestamp()

param_mat[1:16, which(GAR$ZHYD %in% Z_temp$Z_temp)] <-
as.numeric(calibr[[mm]]$optim$bestmem)
if(mode == "independent"){
  CalYet[which(GAR$ZHYD %in% Z_temp$Z_temp)] <- FALSE
}
if(mode == "network"){
  CalYet[which(GAR$ZHYD %in% Z_temp$Z_temp)] <- TRUE
}
cal_flag <- TRUE
}

#####
if(cal_flag == FALSE){

  cat("Start pre-stuff\n")

  Done_Yet <- matrix(NA, ncol = dim(Z_temp)[1], nrow = length(obs))
  dummy <- vector()
  for(rpm in 1:length(Z_temp$Z_temp)){
    dummy[rpm] <- which(colnames(param_mat) == Z_temp$Z_temp[rpm])
  }
  ParYet <- param_mat[, dummy]

  for(pp in 1:max(Z_temp$Level)){
    qualicat <- which(Z_temp$Level == pp)
    for(jj in 1:length(qualicat)){
      TUV <- TUVmodel(prec =
as.matrix(CI[[1]][, topo[[qualicat[jj]]]$pxl]),
            airt =
as.matrix(CI[[2]][, topo[[qualicat[jj]]]$pxl]),
            ep =
as.matrix(CI[[3]][, topo[[qualicat[jj]]]$pxl]),
            area = topo[[qualicat[jj]]]$weights, param =
ParYet[1:15, qualicat[jj]])

      FC <- vector()
      for(conv in 1:length(topo[[qualicat[jj]]]$pxl)){
        FC[conv] <- topo[[qualicat[jj]]]$zones[conv]*(10^-3)/86400
      }

      flag_dummy <- FALSE
      if(dim(TUV$qzones)[1] == 1){
        TUV <- t(TUV$qzones)*FC
        flag_dummy <- TRUE
      }
      if(flag_dummy == FALSE){

```

```

    for(conv in 1:dim(TUW$qzones)[2]){
      TUW$qzones[,conv] <- TUW$qzones[,conv]*FC[conv]
    }
    TUW <- apply(TUW$qzones,1,sum)
  }

  if(Z_temp$Level[qualicat[jj]] != 1){
    dummy <- which(Z_temp$down == qualicat[jj])
    if(is.null(dim(Done_Yet[,dummy])) == TRUE){
      add <- Done_Yet[,dummy]
      add[is.na(add)] <- 0
      TUW <- TUW + add
    }
    if(is.null(dim(Done_Yet[,dummy])) == FALSE){
      add <- apply(Done_Yet[,dummy], 1, sum, na.rm=TRUE)
      TUW <- TUW + add
    }
  }

  UH <- vector()
  for(t in 1:5000){
    UH[t] <-
    (1/((ParYet[16,qualicat[jj]])*factorial(0)))*(t^(0))*exp(-
    (t/ParYet[16,qualicat[jj]]))
  }
  UH <- UH[1:1000]
  UH <- UH/sum(UH)

  TUW[is.na(TUW)] <- 0
  Q_out <- convolve(TUW, rev(UH), type='open')
  Q_out <- Q_out[1:length(TUW)]

  Done_Yet[,qualicat[jj]] <- Q_out
}

CalYet_2 <- vector()
for(rpm in 1:length(Z_temp$Z_temp)){
  CalYet_2[rpm] <- CalYet[which(GAR$ZHYD == Z_temp$Z_temp[rpm])]
}

cat("End pre-stuff\n")
cat("Start calibration\n")

## SITEMA ORDINE PARYET ##

```

```

    cat(mm, " out of ", length(PointToCat), "\n", sep="")
    timestamp()
    calibr_temp <- try(DEoptim(fn=ME2, lower=c(SCF=0.9, DDF=0.0,
Tr=1.0, Ts=-3.0, Tm=-2.0,
                                LPrat=0.0, FC=0.0,
BETA=0.0, k0=0.0, k1=2.0, k2=30.0,
                                lsuz=1.0,
cperc=0.0, bmax=0.0, croute=0.0, k_nash = 0.1),
                        upper=c(SCF=1.5, DDF=5.0, Tr=3.0,
Ts=1.0, Tm=2.0,
                                LPrat=1.0, FC=600.0,
BETA=20.0, k0=2.0, k1=30.0, k2=250.0,
                                lsuz=100.0, cperc=8.0,
bmax=30.0, croute=50.0, k_nash = 1),
                        control=DEoptim.control(NP=NA,
itermax=200, reltol=1e-3, steptol=10, trace=5, parallelType=1,
packages=c('TUWmodel', 'hydroGOF')),
    prec = CI[[1]][,quali_pixel_current],
    tmean=CI[[2]][,quali_pixel_current],
    pet=CI[[3]][,quali_pixel_current], Z_temp = Z_temp, topo = topo,
    disc=obs, Done_Yet = Done_Yet, CalYet_2 = CalYet_2, of= of))
    calibr[[mm]] <- calibr_temp
    timestamp()

    param_mat[1:16,which(GAR$ZHYD %in%
Z_temp$Z_temp)[as.numeric(!CalYet[which(GAR$ZHYD %in%
Z_temp$Z_temp)])*which(GAR$ZHYD %in% Z_temp$Z_temp)!=0]] <-
as.numeric(calibr[[mm]]$optim$bestmem)
    if(mode == "independent"){
      CalYet[which(GAR$ZHYD %in% Z_temp$Z_temp)] <- FALSE
    }
    if(mode == "network"){
      CalYet[which(GAR$ZHYD %in% Z_temp$Z_temp)] <- TRUE
    }
  }

#####
# SAVE INTERMEDIATE RESULTS
#####

save(calibr, file=file.path(save_path, save_name))

}

```

B. Random Tree

```
.DT.app <- function(par.in, catch_CDs, grds_CDs) {
  iparam <- ncol(par.in)

  catch_CDs <- as.data.frame(catch_CDs)
  grds_CDs <- as.data.frame(grds_CDs)
  DT.save <- list()
  par.pred <- array(NA, dim=c(nrow(catch_CDs), iparam))
  grd.par.pred <- array(NA, dim=c(nrow(grds_CDs), iparam))

  ### account for NAs in predictors
  grds_CDs.temp <- grds_CDs
  goodgrds <- complete.cases(grds_CDs.temp)
  grds_CDs.temp[is.na(grds_CDs.temp)] <- -9999.99 # does not matter
  par.in.temp <- as.data.frame(par.in)
  for (ipar in 1:iparam) {
    trainX <- cbind(par.in.temp[,ipar], catch_CDs)
    colnames(trainX) <- c("Par", colnames(catch_CDs))
    DT.pred <- rpart(Par ~ ., data=trainX,
                     control=rpart.control(minsplit=min(20,
nrow(trainX)/2)))
    DT.save[[ipar]] <- DT.pred
    ### cats
    par.pred[,ipar] <- predict(DT.pred, catch_CDs)
    ### grds
    grd.par.pred[,ipar] <- predict(DT.pred, grds_CDs.temp)
  }
  grd.par.pred[!goodgrds,] <- NA
  DT.app.back <- list()
  DT.app.back$cat.par.pred <- par.pred
  DT.app.back$grd.par.pred <- grd.par.pred
  DT.app.back$tree <- DT.save
  return(DT.app.back)
} ### DT.app
```

C. PASS options

```
PASS.options <- function(maxLoops=100, nGroups=10, REGloops=5,
                          generalized.mean.power=-1,
                          proportion.max.eff.update=0.95,
                          sampling='random', optim.subset.cat=0.7) {
  list(maxLoops=maxLoops, nGroups=nGroups, REGloops=REGloops,
       generalized.mean.power=generalized.mean.power,
       proportion.max.eff.update=proportion.max.eff.update,
       sampling=sampling, optim.subset.cat=optim.subset.cat)
}
```

D. PASS Print

```
print.PASS <- function (x, ...) {
  cat('Output of the PArAmeter Set Shuffling algorithm:\n\n')
  cat(' number of catchments:',
nrow(x$groups[[1]]$regionalized.parameters$cat.par.pred), '\n')
  cat(' number of model units (e.g. pixels):',
nrow(x$groups[[1]]$regionalized.parameters$grd.par.pred), '\n')
  cat('sampling strategy:', x$PASS.options$sampling, '\n')
  cat(' number of loops:', x$PASS.options$maxLoops, '\n')
  cat(' number of groups:', x$PASS.options$nGroups, '\n')
  cat(' number of loops for regional consistency:',
x$PASS.options$REGloops, '\n')
  if (x$PASS.options$sampling == 'optim') {
    cat(' proportion of randomized parameters (when optim):',
x$PASS.options$optim.subset.cat, '\n')
  }
  cat('lumped regionalization efficiencies:', '\n')
  print(round(apply(sapply(x$groups, function(x) x$cat.eff.lump), 2,
summary), 3))
  cat('distributed regionalization efficiencies:', '\n')
  print(round(apply(sapply(x$groups, function(x) x$cat.eff.dist), 2,
summary), 3))
  cat(' updated number of train parameter sets:',
sum(sapply(x$train.parameters.updated, nrow)), '\n')
}
```

E. PASS (Regionalization)

```
PASS <- function (Y, # list (N.cat) of data.frames (XXX x
N.par) with locally lumped calibrated model parameters OR previously
obtained PASS output
                  X.cat, # matrix or data.frame (N.cat x
N.dsc) of catchment descriptors
                  X.grd, # matrix or data.frame (N.grd x
N.dsc) of model unit/pixel descriptors
                  wcat,
                  L_cat, # dataframe (cat and Level)
                  model.eff.fn, # function to calculate the regional
efficiency
                  model.eff.fn2, # function to calculate the lumped
efficiency
                  lower, # two vectors specifying scalar real
lower and upper bounds on each parameter to be optimized, so that
the i-th element
```

```

        upper,          # of 'lower' and 'upper' applies to
the i-th parameter.
        options=PASS.options(), # options for PASS
        GAR,           # (sf) - $ZHYD = ID HUs; $NextDown =
HUs immediatly downstream;
        CI,            # climate input (list of NxM
dataframes, N=timesteps, M=number of HUs), [[1]] daily
precipitation, [[2]] mean daily temperature, [[3]]daily poterntial
evapotranspiration;
        DISC,          # list(M catchments) - discharge
series (zoo) in m^3/s
        topology,      # list of M dataframes - each
dataframe: [,1] pixels touching the HU, [,2] weight, [,3] area in
m^2 of the intersection;
        Z_temp_pre,    # list of M dataframes - each
dataframe: Z_temp_pre[M]$Level = level of the HU for the
catchment; Z_temp_pre[M]$Z_temp = ID of the HU;
Z_temp_pre[M]$down = ID of downstream HU;
        COR,           # (numeric) Number of cores to use
in parallelization;
        me.fn,         # (character) model efficiency
function. Can be "KGE" for Kling-Gupta or "wNSE" for eighted Nash-
Sutcliffe
        ...){          # further arguments to be passed
#require(rpart)
# select parameter sets
N.cat <- nrow(X.cat) # number of CALIBRATED sub-catchments
N.grd <- nrow(X.grd) # number of TOTAL sub-catchments
if (class(Y) != 'PASS') {

    train.parameters <- Y
    options <- do.call(PASS.options, as.list(options))
    overall.eff.threshold <- rep(-999, options$nGroups)
    PASS.group <- list()
    PASS.save <- list()
} else {
    previousPASSout <- Y
    overall.eff.threshold <- sapply(previousPASSout$groups, function
(x) x$overall.eff)
    options <- do.call(PASS.options, as.list(options))
    options$nGroups <- length(previousPASSout$groups)
    train.parameters <- previousPASSout$train.parameters.updated
    PASS.group <- list()
    PASS.save <- previousPASSout
}

# parameter normalisation (to calculate distances in the regional
consistency algorithm)
train.parameters.norm <- train.parameters

```

```

for (i.cat in 1:N.cat) {
  dummy_max <- t(upper %*% t(rep(1,
nrow(train.parameters[[i.cat]])))
  dummy_min <- t(lower %*% t(rep(1,
nrow(train.parameters[[i.cat]])))
  train.parameters.norm[[i.cat]][,-1] <-
(train.parameters[[i.cat]][,-1] - dummy_min)/(dummy_max - dummy_min)
# transformation (the first column contains efficiencies)
}
dummy_max.cat <- t(upper %*% t(rep(1, N.cat)))
dummy_min.cat <- t(lower %*% t(rep(1, N.cat)))
dummy_max.grd <- t(upper %*% t(rep(1, N.grd)))
dummy_min.grd <- t(lower %*% t(rep(1, N.grd)))

cat('%%%% ----- START ----- %%%\n')
time0 <- Sys.time()
for (i.PASS in 1:options$maxLoops) {
  #if (i.PASS %% round(options$maxLoops/10) == 0) cat('      Loop',
i.PASS, 'out of', options$maxLoops, 'loops\n')
  cat(i.PASS)
  i.Group <- sample(options$nGroups, 1)
  if (options$sampling == 'random') {
    selected.parameters <- t(sapply(train.parameters, function (x)
x[sample(nrow(x), 1),])) # matrix with N rows (n. of catchments)
and M columns (n. of parameters)
    selected.parameters <- as.matrix(selected.parameters)
  } else if (options$sampling == 'optim') {
    selected.parameters <-
previousPASSout$groups[[i.Group]]$selected.parameters
    select.cat <- sample(N.cat,
round(N.cat*options$optim.subset.cat))
    selected.parameters[select.cat,] <-
t(sapply(train.parameters[select.cat], function (x)
x[sample(nrow(x), 1),]))
  }
  # parameter normalisation (to calculate distances in the
regional consistency algorithm)
  selected.parameters[, -1] <- (selected.parameters[, -1] -
dummy_min.cat)/(dummy_max.cat - dummy_min.cat) # transformation
  # Regional consistency algorithm:
  for (i.REG in 1:options$REGloops) {
    # regionalization
    regionalization <- .DT.app(par.in=selected.parameters[, -1],
catch_CDs=X.cat, grds_CDs=X.grd) # output = cat.par.pred and
grd.par.pred
    for (i.cat in 1:N.cat) {
      par.reg.cat <- regionalization$cat.par.pred[i.cat,]

```

```

        distances <- (train.parameters.norm[[i.cat]][, -1] -
t(par.reg.cat %*% t(rep(1,
nrow(train.parameters.norm[[i.cat]]))))^2
        selected.parameters[i.cat,] <-
train.parameters.norm[[i.cat]][which.min(apply(distances, 1, sum,
na.rm=TRUE)),]
    }
}
colnames(regionalization$cat.par.pred) <-
colnames(selected.parameters)[-1]
rownames(regionalization$cat.par.pred) <- rownames(X.cat)
colnames(regionalization$grd.par.pred) <-
colnames(selected.parameters)[-1]
rownames(regionalization$grd.par.pred) <- rownames(X.grd)
regionalization$cat.par.pred[regionalization$cat.par.pred > 1]
<- 1
regionalization$cat.par.pred[regionalization$cat.par.pred < 0]
<- 0
regionalization$grd.par.pred[regionalization$grd.par.pred > 1]
<- 1
regionalization$grd.par.pred[regionalization$grd.par.pred < 0]
<- 0
regionalization$cat.par.pred <-
regionalization$cat.par.pred*(dummy_max.cat - dummy_min.cat) +
dummy_min.cat
regionalization$grd.par.pred <-
regionalization$grd.par.pred*(dummy_max.grd - dummy_min.grd) +
dummy_min.grd
eff.lump <- rep(NA, N.cat)
eff.dist <- rep(NA, N.cat)

eff.lump <- model.eff.fn2(param =
regionalization$cat.par.pred, CI, DISC, topology, Z_temp_pre, me.fn =
me.fn, COR)

for (i.cat in 1:length(DISC)){
  qc <- which(names(train.parameters) == names(DISC)[i.cat])
  if (eff.lump[i.cat] >
options$proportion.max.eff.update*max(train.parameters[[qc]][, 1]))
{
  train.parameters[[qc]] <- rbind(train.parameters[[qc]],
c(eff.lump[i.cat], regionalization$cat.par.pred[qc,]))
}
}

param.grd <- regionalization$grd.par.pred
param.grd[is.na(param.grd)] <- -999

```

```

    eff.dist <- model.eff.fn(param = param.grd, GAR, CI, DISC,
topology, Z_temp_pre, me.fn = me.fn, COR)

    overall.eff.dist <- median(eff.dist)

    if (options$sampling == 'random') {
      i.Group <- which.min(overall.eff.threshold)
    } # in optim stay in the Group you are optimizing
    if (overall.eff.dist > overall.eff.threshold[i.Group]) {
      overall.eff.threshold[i.Group] <- overall.eff.dist
      PASS.group[['overall.eff']] <- overall.eff.dist
      PASS.group[['selected.parameters']] <- selected.parameters
      PASS.group[['regionalized.parameters']] <- regionalization
      PASS.group[['cat.eff.lump']] <- eff.lump
      PASS.group[['cat.eff.dist']] <- eff.dist
      PASS.save[['groups']] [[paste('Group', i.Group, sep=' ')] <-
PASS.group
    }
  } # end of for (i.PASS in 1:maxLoops)
  PASS.save[['train.parameters.updated']] <- train.parameters
  PASS.save[['PASS.options']] <- options
  cat('%%% ----- END ----- %%%\n')
  time1 <- Sys.time()
  print(time1 - time0)
  class(PASS.save) <- 'PASS'
  return(PASS.save)
}

```

F. ME Functions Examples

```

ME.TUWmodel <- function(param, GAR, CI, DISC, topology, Z_temp_pre,
me.fn, COR){

  ##### INITIALIZE SIMUL MATRIX #####
  G <- matrix(NA, ncol=length(GAR$geometry), nrow = dim(CI[[1]])[1])
  colnames(G) <- GAR$ZHYD
  timestamp()

  cat("----- Start Regional Simulation: Generative Module ----
-----", " \n", sep="")

  #### GENERATION ON ALL HUs ####

  current_cat <- sort(GAR$ZHYD, decreasing = FALSE)

```

```

cluster <- makeCluster(COR)
registerDoParallel(cluster)
series <- list()

series <- foreach(i = 1:length(current_cat), .packages =
c("TUWmodel")) %dopar% {

  ii <- current_cat[i]
  topo <- topology[[which(as.numeric(names(topology)) == ii)]]
  TUW <- TUWmodel(prec =
as.matrix(CI[[1]][,which(colnames(CI[[1]]) %in% topo$pxl)]),
          airt =
as.matrix(CI[[2]][,which(colnames(CI[[2]]) %in% topo$pxl)]),
          ep = as.matrix(CI[[3]][,which(colnames(CI[[3]])
%in% topo$pxl)]),
          area = topo$weights, param =
param[which(rownames(param) == ii),1:15])

  FC <- topo$zones*(10^-3)/86400

  flag_dummy <- FALSE
  if(dim(TUW$q0)[1] == 1){
    TUW <- t(TUW$qzones)*FC
    flag_dummy <- TRUE
  }
  if(flag_dummy == FALSE){
    conv <- c(1:dim(TUW$qzones)[2])
    TUW$qzones[,conv] <- TUW$qzones[,conv] %*%
diag(round(FC[conv],4))
    TUW <- apply(TUW$qzones,1,sum)
  }

  series[i] <- TUW

}
stopCluster(cluster)

for(s in 1:length(series)){
  G[,s] <- series[[s]]
}

cat("----- End Regional Simulation: Generative Module -----
-----", " \n", sep="")

#### ROUTING ON ALL HUs ####

cat("----- Start Regional Simulation: Routing Module -----
-----", " \n", sep="")

```

```

for(l_cat in 1:length(unique(GAR$Level))){
  cat("Level ",l_cat, " out of ",length(unique(GAR$Level)), "
\n",sep="")
  current_cat <- which(GAR$Level == l_cat)
  for(i in current_cat){
    Z_temp <- Z_temp_pre[[i]]
    pp <- max(Z_temp$Level)
    qualicat <- which(Z_temp$Level == pp)

    if(Z_temp$Level[qualicat] != 1){
      dummy <- Z_temp$Z_temp[which(Z_temp$down == qualicat)]
      if(is.null(dim(G[,which(colnames(G)%in%dummy)])) == TRUE){
        add <- G[,which(colnames(G)%in%dummy)]
        add[is.na(add)] <- 0
        G[,which(colnames(G) == Z_temp$Z_temp[qualicat])] <-
G[,which(colnames(G) == Z_temp$Z_temp[qualicat])] + add
      }
      if(is.null(dim(G[,which(colnames(G)%in%dummy)])) == FALSE){
        add <- apply(G[,which(colnames(G)%in%dummy)], 1, sum,
na.rm=TRUE)
        add[is.na(add)] <- 0
        G[,which(colnames(G) == Z_temp$Z_temp[qualicat])] <-
G[,which(colnames(G) == Z_temp$Z_temp[qualicat])] + add
      }
    }

    ##NASH
    N <- 1
    t <- seq(1:dim(G)[1])
    UH <- (1/((param[which(rownames(param) ==
Z_temp$Z_temp[qualicat]),16]^N)*factorial(N-1)))*(t^(N-1))*exp(-
(t/param[which(rownames(param) == Z_temp$Z_temp[qualicat]),16]))

    UH <- UH[1:1000]
    UH <- UH/sum(UH)

    G[,which(colnames(G) ==
Z_temp$Z_temp[qualicat])][is.na(G[,which(colnames(G) ==
Z_temp$Z_temp[qualicat])])] <- 0
    Q_out <- convolve(G[,which(colnames(G) ==
Z_temp$Z_temp[qualicat])], rev(UH), type='open')
    Q_out <- Q_out[1:length(G[,which(colnames(G) ==
Z_temp$Z_temp[qualicat])])]

    G[,which(colnames(G) == Z_temp$Z_temp[qualicat])] <- Q_out
  }
}

```

```

G <- zoo(G, time(CI[[1]][,1]))
timestamp()

ME <- vector()
for(i in 1:length(DISC)){
  simu <- G[,which(colnames(G) == names(DISC)[i])]
  simu <- window(simu, start=as.Date("2010-01-01"), end =
as.Date("2019-12-31"))
  obs_c <- DISC[[i]]
  dummy <- window(CI[[1]][,1], start=as.Date("2010-01-01"),
end=as.Date("2019-12-31"))
  obs_c <- window(obs_c, start=as.Date("2010-01-01"),
end=as.Date("2019-12-31"))
  obs_c <- apply(merge.zoo(obs_c,dummy),1,sum) - dummy

  if(me.fn == "KGE"){
    ME[i] <- KGE(simu,window(obs_c,start=as.Date("2010-01-01"),
end = as.Date("2019-12-31")))
  }
  if(me.fn == "wNSE"){
    ME[i] <- wNSE(simu,window(obs_c,start=as.Date("2010-01-01"),
end = as.Date("2019-12-31")))
  }
}

cat("----- End Regional Simulation: Routing Module -----
--", " \n", sep="")
timestamp()
return(ME)
}

ME.TUWmodel2 <- function(param, CI, DISC, topology, Z_temp_pre,
me.fn, COR){

  timestamp()
  cat("----- Start Local Simulation -----", " \n",
sep="")

  #n_cores <- detectCores()
  #cluster <- makeCluster(n_cores -2)
  cluster <- makeCluster(COR)
  registerDoParallel(cluster)

  ME <- list()

```

```

ME <- foreach(i = 1:dim(param)[1], .packages = c("TUWmodel",
"zoo", "hydroGOF"), .combine = "c") %dopar% {

  NOME <- names(DISC)[i]
  obs_c <- DISC[[i]]
  dummy <- window(CI[[1]][,1], start=as.Date("2000-01-01"),
end=as.Date("2020-12-31"))
  obs_c <- window(obs_c, start=as.Date("2000-01-01"),
end=as.Date("2020-12-31"))
  obs_c <- apply(merge.zoo(obs_c, dummy), 1, sum) - dummy

  Z_temp <- Z_temp_pre[[as.numeric(NOME)]]

  par_un <- as.numeric(param[which(rownames(param) == NOME),])

  inter <- matrix(NA, ncol = dim(Z_temp)[1], nrow =
length(CI[[1]][,1]))
  UH <- vector()
  t <- seq(1:5000)
  UH <- (1/((par_un[16])*factorial(0)))*(t^(0))*exp(-
(t/par_un[16]))
  UH <- UH[1:1000]
  UH <- UH/sum(UH)

  for(pp in 1:max(Z_temp$Level)){
    qualicat <- which(Z_temp$Level == pp)
    for(jj in 1:length(qualicat)){

      topo <- topology[[which(as.numeric(names(topology)) ==
Z_temp$Z_temp[qualicat[jj]])]]
      TUW <- TUWmodel(prec =
as.matrix(CI[[1]][,which(colnames(CI[[1]]) %in% topo$pxl)]),
airt =
as.matrix(CI[[2]][,which(colnames(CI[[2]]) %in% topo$pxl)]),
ep =
as.matrix(CI[[3]][,which(colnames(CI[[3]]) %in% topo$pxl)]),
area = topo$weights, param = par_un[1:15])

      FC <- topo$zones*(10^-3)/86400

      flag_dummy <- FALSE
      if(dim(TUW$qzones)[1] == 1){
        TUW <- t(TUW$qzones)*FC
        flag_dummy <- TRUE
      }

      if(flag_dummy == FALSE){
        conv <- c(1:dim(TUW$qzones)[2])

```

```

      TUV$qzones[,conv] <- TUV$qzones[,conv] %*%
diag(round(FC[conv],4))
      TUV <- apply(TUV$qzones,1,sum)
    }

    if(Z_temp$Level[qualicat[jj]] != 1){
      dummy <- which(Z_temp$down == qualicat[jj])
      if(is.null(dim(inter[,dummy])) == TRUE){
        add <- inter[,dummy]
        add[is.na(add)] <- 0
        TUV <- TUV + add
      }
      if(is.null(dim(inter[,dummy])) == FALSE){
        add <- apply(inter[,dummy], 1, sum, na.rm=TRUE)
        TUV <- TUV + add
      }
    }
    TUV[is.na(TUV)] <- 0
    Q_out <- convolve(TUV, rev(UH), type='open')
    Q_out <- Q_out[1:length(TUV)]
    inter[,qualicat[jj]] <- Q_out
    TUV[is.na(TUV)] <- 0
    Q_out <- convolve(TUV, rev(UH), type='open')
    Q_out <- Q_out[1:length(TUV)]
    inter[,qualicat[jj]] <- Q_out
  }
}

simu <- zoo(Q_out, time(CI[[1]]))
simu <- window(simu, start=as.Date("2010-01-01"),
end=as.Date("2019-12-31"))

if(me.fn == "KGE"){
  ME[[i]] <- KGE(simu, window(obs_c, start=as.Date("2010-01-
01"), end=as.Date("2019-12-31")))
}
if(me.fn == "wNSE"){
  ME[[i]] <- wNSE(simu, window(obs_c, start=as.Date("2010-01-
01"), end=as.Date("2019-12-31")))
}
}
cat("----- End Local Simulation -----", " \n", sep="")
stopCluster(cluster)
timestamp()
return(ME)
}

```

G. C-Method (calibration)

```
#####  
# Libraries  
#####  
  
library(zoo)  
library(VineCopula)  
  
#####  
# Load data (assumed to be in the working directory)  
#####  
  
load("Q_37_local.RData")  
load("Q_37_REG_all.RData")  
  
#####  
# Identify basins with complete observations  
#####  
  
L <- vector()  
  
for(i in 1:length(Q_37_local[[1]])){  
  L[i] <- sum(!is.na(Q_37_local$OBS[[i]]))  
}  
  
qualicat <- which(L == 3652)  
  
# remove problematic basin  
qualicat <- qualicat[-19]  
  
#####  
# Initialize output list  
#####  
  
SERIES <- vector("list", length(qualicat))  
names(SERIES) <- names(Q_37_local$SIM)[qualicat]  
  
# Loop over basins  
#####  
  
for(i in 1:length(qualicat)){  
  
  cat(i, " out of ", length(qualicat), "\n", sep = "")  
  
  iter <- 1000  
  
  ERR_RESLOG <- as.data.frame(matrix(NA, ncol = iter, nrow = 3652))
```

```

CAT <- qualicat[i]

obs <- Q_37_local$OBS[[CAT]]

sim <- Q_37_REG_all[[CAT]]
sim <- window(sim,
               start = as.Date("2010-01-01"),
               end   = as.Date("2019-12-31"))

#####
# Log residuals
#####

res_log <- log(sim) - log(obs)
res_log2 <- res_log

#####
# Monthly standardization
#####

mm <- sort(unique(format(time(res_log), "%m")))

M <- vector()
SD <- vector()

for(m in 1:length(mm)){

  sub <- res_log[format(time(res_log), "%m") == mm[m]]

  M[m] <- mean(sub, na.rm = TRUE)
  SD[m] <- sd(sub, na.rm = TRUE)

  res_log2[format(time(res_log2), "%m") == mm[m]] <-
    (res_log2[format(time(res_log2), "%m") == mm[m]] - M[m]) /
SD[m]
}

#####
# Copula selection
#####

COP2 <- BiCopSelect(
  pobs(as.numeric(res_log2)[-length(res_log2)]),
  pobs(as.numeric(res_log2)[-1]),
  indeptest = TRUE,

```

```

    familyset = 2
  )

#####
# Copula simulation
#####

cat("Inizio 1000 serie\n")

for(k in 1:iter){

  cat(k, " out of ", iter, "\n", sep = "")

  u2 <- pobs(as.numeric(res_log2))

  for(j in 1:(length(u2)-1)){

    u2[j+1] <- BiCopCondSim(
      1,
      cond.val = u2[j],
      cond.var = 1,
      COP2
    )
  }

  dummy2 <- as.numeric(
    quantile(as.numeric(res_log2),
      prob = u2,
      na.rm = TRUE)
  )

  ERR_RESLOG[,k] <- dummy2
}

#####
# Reintroduce seasonal statistics
#####

cat("Inizio reintroduzione DS\n")

for(k in 1:iter){

  cat(k, " out of ", iter, "\n", sep = "")

  dummy2 <- zoo(ERR_RESLOG[,k], time(res_log2))

  for(m in 1:length(mm)){

```

```

dummy2[format(time(dummy2), "%m") == mm[m]] <-
  (dummy2[format(time(dummy2), "%m") == mm[m]] * SD[m]) + M[m]
}

ERR_RESLOG[,k] <- as.numeric(dummy2)
}

#####
# Save results in list
#####

SERIES[[i]]$LOG <- ERR_RESLOG
}

save(SERIES, file = "Residual_series_copula.RData")

```

H. P-BLUECAT Method (BB7 case - calibration)

```

#####
# Libraries
#####

library(zoo)
library(VineCopula)
library(doParallel)
library(foreach)
library(progressr)

#####
# Functions
#####

fit_gpd_tail <- function(Q, threshold = NULL, auto_threshold = TRUE,
min_quantile = 0.80) {
  Q <- as.numeric(Q)
  Q <- Q[!is.na(Q)]
  if(length(Q) < 50) stop("Servono almeno 50 osservazioni per un fit
GPD sensato.")

  if(auto_threshold) {
    qs <- quantile(Q, seq(min_quantile, 0.98, by = 0.01))
    shape_vec <- scale_vec <- numeric(length(qs))
    for(i in seq_along(qs)) {
      thr <- qs[i]
      y <- Q[Q > thr] - thr
      if(length(y) < 20) next
    }
  }
}

```

```

fit <- try(gpd.fit(y, threshold = 0, show = FALSE), silent =
TRUE)
if(class(fit) == "try-error") {
  shape_vec[i] <- NA
  scale_vec[i] <- NA
} else {
  shape_vec[i] <- fit$mle[2]
  scale_vec[i] <- fit$mle[1]
}
}
idx <- which(!is.na(shape_vec))
dshape <- abs(diff(shape_vec[idx]))
best <- idx[which.min(dshape)]
u <- qs[best]
message(sprintf("Soglia selezionata automaticamente: u = %.3f",
u))
} else {
  if(is.null(threshold)) stop("Se auto_threshold=FALSE devi
fornire threshold.")
  u <- threshold
}

exceedances <- Q[Q > u] - u
if(length(exceedances) < 20) stop("Troppi pochi exceedances per
stimare la GPD.")
fit <- gpd.fit(exceedances, threshold = 0, show = FALSE)

return(list(
  threshold = u,
  n_exceed = length(exceedances),
  params = list(scale = fit$mle[1], shape = fit$mle[2]),
  fit = fit,
  exceedances = exceedances,
  Q = Q
))
}

gpd_quantile <- function(p_tail, sigma, xi) {
  if(abs(xi) < 1e-8) {
    y <- - sigma * log(1 - p_tail)
  } else {
    y <- (sigma / xi) * ((1 - p_tail)^(-xi) - 1)
  }
  return(y)
}

hybrid_cdf <- function(x, Q, u, sigma, xi, ecdf_fun = NULL) {
  if(is.null(ecdf_fun)) ecdf_fun <- ecdf(Q)
  F_u <- ecdf_fun(u)

```

```

res <- numeric(length(x))

idx_emp <- which(x <= u)
if(length(idx_emp) > 0) res[idx_emp] <- ecdf_fun(x[idx_emp])

idx_tail <- which(x > u)
if(length(idx_tail) > 0) {
  y <- x[idx_tail] - u
  if(abs(xi) < 1e-8) {
    G_y <- 1 - exp(- y / sigma)
  } else {
    G_y <- 1 - (1 + xi*y/sigma)^(-1/xi)
  }
  res[idx_tail] <- F_u + (1 - F_u) * G_y
}

return(res)
}

hybrid_quantile <- function(p, Q, u, sigma, xi, ecdf_fun = NULL) {
  if(is.null(ecdf_fun)) ecdf_fun <- ecdf(Q)
  F_u <- ecdf_fun(u)
  out <- numeric(length(p))

  idx_emp <- which(p <= F_u)
  if(length(idx_emp) > 0) {
    P_tot <- c(1:length(Q))/(length(Q)+1)
    for(j in 1:length(p[idx_emp])) {
      out[idx_emp[j]] <- sort(as.numeric(Q))[which.min(abs(P_tot -
p[idx_emp[j]]))]
    }
  }

  idx_tail <- which(p > F_u)
  if(length(idx_tail) > 0) {
    p_tail <- (p[idx_tail] - F_u)/(1 - F_u)
    y <- gpd_quantile(p_tail, sigma, xi)
    out[idx_tail] <- u + y
  }

  return(out)
}

hybrid_cdf_from_fit <- function(x, ris, ecdf_fun = NULL) {
  Q <- ris$Q
  u <- ris$threshold
  sigma <- ris$params$scale
  xi <- ris$params$shape
  if(is.null(ecdf_fun)) ecdf_fun <- ecdf(Q)

```

```

F_u <- ecdf_fun(u)

x_vec <- as.numeric(x)
out <- numeric(length(x_vec))

idx_emp <- which(x_vec <= u)
if(length(idx_emp) > 0) out[idx_emp] <- ecdf_fun(x_vec[idx_emp])

idx_tail <- which(x_vec > u)
if(length(idx_tail) > 0) {
  y <- x_vec[idx_tail] - u
  denom_term <- 1 + xi * y / sigma
  if(any(denom_term <= 0)) {
    warning("GPD non definita per alcuni x: saturiamo a 1")
    G_y <- numeric(length(y))
    valid <- denom_term > 0
    G_y[valid] <- 1 - denom_term[valid]^(-1/xi)
    G_y[!valid] <- 1
  } else {
    G_y <- 1 - denom_term^(-1/xi)
  }
  out[idx_tail] <- F_u + (1 - F_u) * G_y
}

return(out)
}

#####
# Load data (assumed in working directory)
#####

load("Q_37_local.RData")
load("Q_37_REG_all.RData")

#####
# Identify basins with complete observations
#####

L <- vector()
for(i in 1:length(Q_37_local[[1]])) L[i] <-
sum(!is.na(Q_37_local$OBS[[i]]))
qualicat <- which(L == 3652)
qualicat <- qualicat[-19]

#####
# Initialize output list
#####

ENS_COP_both_reg_all_par_tail <- vector("list", length(qualicat))

```

```

names(ENS_COP_both_reg_all_par_tail) <-
names(Q_37_REG_all)[qualicat]
count <- 1

#####
# Main loop over basins
#####

for(i in 1:length(qualicat)) {

  cat(i," out of ", length(qualicat),"\n", sep="")

  CAT <- qualicat[i]
  obs <- Q_37_local$OBS[[CAT]]
  sim <- Q_37_REG_all[[CAT]]
  sim <- window(sim, start = as.Date("2010-01-01"), end =
as.Date("2019-12-31"))

  Q <- as.numeric(obs)
  P_cut <- 0.97
  ris_obs <- fit_gpd_tail(Q, auto_threshold = FALSE, threshold =
quantile(Q, prob = P_cut))
  u <- ris_obs$threshold
  sigma <- ris_obs$params$scale
  xi <- ris_obs$params$shape
  ec_obs <- ecdf(Q)

  Q_s <- as.numeric(sim)
  P_cut_s <- 0.97
  ris_sim <- fit_gpd_tail(Q_s, auto_threshold = FALSE, threshold =
quantile(Q_s, prob = P_cut_s))
  u_s <- ris_sim$threshold
  sigma_s <- ris_sim$params$scale
  xi_s <- ris_sim$params$shape
  ec_sim <- ecdf(Q_s)

  u1 <- hybrid_cdf(Q, Q, u, sigma, xi, ec_obs)
  u2 <- hybrid_cdf(Q_s, Q_s, u_s, sigma_s, xi_s, ec_sim)
  COP <- BiCopSelect(u2, u1, familyset = 9, indeptest = TRUE,
rotations = FALSE)

  ncores <- parallel::detectCores() - 1
  cl <- makeCluster(ncores)
  registerDoParallel(cl)
  N <- 1000

  u1 <- hybrid_cdf(Q_s, Q_s, u_s, sigma_s, xi_s, ec_sim)

  handlers("txtprogressbar")

```

```

with_progress({
  p <- progressor(steps = length(u1))
  res <- foreach(
    i = 1:length(u1),
    .packages = "VineCopula",
    .export = c("hybrid_quantile", "obs", "u", "sigma", "xi",
"ec_obs")
  ) %dopar% {
    P <- BiCopCondSim(N, cond.val = u1[i], cond.var = 1, COP)
    sub2 <- hybrid_quantile(P, Q, u, sigma, xi, ec_obs)
    p()
    c(
      med      = quantile(sub2, 0.50, na.rm = TRUE),
      q10      = quantile(sub2, 0.10, na.rm = TRUE),
      q25      = quantile(sub2, 0.25, na.rm = TRUE),
      q75      = quantile(sub2, 0.75, na.rm = TRUE),
      q90      = quantile(sub2, 0.90, na.rm = TRUE),
      qpred    = ecdf(Q)(Q[i])
    )
  }
})

res_mat <- do.call(rbind, res)

temp2 <- list(
  med      = res_mat[,1],
  q10      = res_mat[,2],
  q25      = res_mat[,3],
  q75      = res_mat[,4],
  q90      = res_mat[,5],
  q_pred   = res_mat[,6]
)

ENS_COP_both_reg_all_par_tail[[count]] <- temp2
count <- count + 1
}

#####
# Save results in working directory
#####

save(ENS_COP_both_reg_all_par_tail, file =
"ENS_COP_both_reg_all_par_tail_plus_sim.RData")

```

