

Data-driven Methods for Climate Change Modelling in Hydrology

A Use Case for Deep Learning in Rainfall-Runoff Simulation

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Matthias Eder, BSc.

Registration Number 01624856

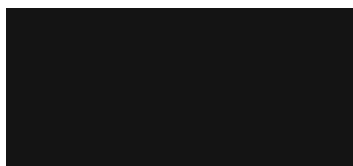
to the Faculty of Informatics

at the TU Wien

Advisor: Shashikant Shankar Ilager, M.Tech. PhD

Assistance: Univ.Prof.in Mag.a rer.soc.oec. Dr.in rer.soc.oec Ivona Brandić

Vienna, 23rd January, 2024



Matthias Eder



Shashikant Shankar Ilager



Datengesteuerte Methoden für Klimawandelmodellierung in der Hydrologie

Ein Anwendungsfall für Deep Learning in der Niederschlags-Abfluss Simulation

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Matthias Eder, BSc.

Matrikelnummer 01624856


an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Shashikant Shankar Ilager, M.Tech. PhD

Mitwirkung: Univ.Prof.in Mag.a rer.soc.oec. Dr.in rer.soc.oec Ivona Brandić

Wien, 23. Jänner 2024


Matthias Eder
Shashikant Shankar Ilager

Erklärung zur Verfassung der Arbeit

Matthias Eder, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 23. Jänner 2024

A black rectangular box redacting the signature of Matthias Eder.

Matthias Eder

Danksagung

Zuallererst möchte ich meinem Betreuer Shashikant Ilager und meiner Supervisorin Ivona Brandić danken. Shashikants Feedback war immer sehr wertvoll und fundiert und hat mich oft aus gedanklichen Sackgassen wieder in die richtige Richtung für meine Arbeit gelenkt. Ich bedanke mich für die gute Zusammenarbeit und die wertvollen Erkenntnisse, die ich daraus gewinnen konnte.

An Lisa, vielen Dank, dass du immer für mich da bist und mir wie in allen Lebenslagen auch bei dieser Arbeit geholfen hast. Das Leben als Student wäre ohne dich nicht vorstellbar gewesen. Die Besuche mit Moritz oder gemeinsame Spaziergänge haben mich oft auf andere Gedanken gebracht und mir gezeigt, dass es auch Wichtigeres gibt als ein Studium.

An Mama und Papa, Danke, dass ihr mich immer unterstützt und mir das Studium ermöglicht habt. Ihr habt mich ermutigt, auch schwierige Entscheidungen zu treffen und meinen Weg sowohl in akademischer als auch persönlicher Sicht zu gehen. Ohne euch wäre ich nicht, wer ich bin - Danke für Alles.

An Laura, du hast mich vom Anfang bis zum Ende dieser Arbeit begleitet, hast dir meine stundenlangen Beschwerden angehört und hattest trotzdem immer einen guten Rat für mich. Ich danke dir von ganzem Herzen für deine Liebe, deine Unterstützung, deine Geduld, dein Interesse, deine Motivation und die vielen Erinnerungen, dass ich das Ganze schon schaffen werde.

An meine Großeltern, ich wünschte, ihr hättet mich alle in diesem Kapitel begleiten können. Vielen Dank für die Zeit, die wir gemeinsam verbracht haben, euer Interesse, die gemeinsamen Mittagessen und spannenden Schachpartien und dass ihr immer für mich da seid.

An Markus und Max, Danke für das monatliche Jour Fixe und die vielen Gespräche, sei es in unserem zweiten Wohnzimmer, auf Discord, im Kaffeehaus oder beim Kebabstand. Ich bin froh, dass ich euch in meinem Leben habe.

An alle hier nicht genannten Freund*innen und Wegbegleiter*innen, ich bin dankbar für jede Hilfe, jeden guten Rat und jede gemeinsame Aktivität. Ich bin sehr froh über eure Unterstützung und freue mich auf die nächsten Kapitel mit euch.

Kurzfassung

Der Klimawandel hat neue Herausforderungen für die hydrologische Modellierung mit sich gebracht, da extreme Ereignisse wie Überschwemmungen, Dürren oder Hitzewellen immer häufiger auftreten. Dadurch wird die Robustheit herkömmlicher, prozessbasierter hydrologischer Modelle in Frage gestellt. Insbesondere die Niederschlags-Abfluss-Simulation ist ein zentraler Anwendungsfall für Modelle, die das Abflussverhalten bei Niederschlagsereignissen in einem Wassereinzugsgebiet erklären sollen. Diese Arbeit evaluiert und vergleicht die Robustheit und Genauigkeit eines modernen datengesteuerten Deep Learning (DL)-Ansatzes im Bereich der Large-Sample-Hydrologie (LSH), wo sein Aufkommen zu einer Neudefinition der Anforderungen geführt hat, und demonstriert seine Fähigkeit, verborgene Beziehungen in komplexen hydrologischen Prozessen aufzudecken. Gängige LSH-Datensätze werden verglichen, der Datensatz LamaH-CE, der 479 Einzugsgebiete in Mitteleuropa abdeckt, wird analysiert, und es werden Schritte zum Pre-Processing der Daten eingesetzt, um domänenspezifische Probleme zu behandeln, wie z.B. die Imputation fehlender Stromabflussdaten und die Erfassung von Anomalien. Eine Trendanalyse zeigt einen allgemeinen Erwärmungstrend von $\bar{T} + 1,53^\circ\text{C}$ während des 39-jährigen Untersuchungszeitraums.

In dieser Arbeit werden drei Arten von Modellen verglichen: das konzeptionelle prozessgesteuerte Modell HBVEdu, das gradientenbasierte Machine-Learning-Modell XGBoost und das moderne Deep-Learning-Modell EA-LSTM. Um die Robustheit der Modelle unter wechselnden Klimabedingungen zu bewerten, wird ein Differential-Split-Sample-Test-Ansatz angewandt. Dabei werden vier Referenzzeiträume eingesetzt, die extreme Temperatur- und Niederschlagsschwankungen repräsentieren, sowie ein längerer Bezugszeitraum zum Vergleich mit konventionellen Datensplitting-Verfahren.

Das DL-Modell übertrifft sowohl die prozessgesteuerten als auch die ML-Modelle in allen klimatischen Referenzperioden und im Bezugszeitraum deutlich. Das EA-LSTM-Modell zeigt eine kompetitive und robuste Leistung mit einem durchschnittlichen NSE von 0,73486. Im Vergleich dazu übertrifft das XGBoost-Modell das physikalisch basierte HBVEdu-Modell mit einem mittleren NSE von 0,56306 bzw. 0,48528. Eine Analyse des ML-Modells zeigt jedoch, dass es empfindlich auf Schwankungen in den Daten reagiert. Bemerkenswert ist, dass es keinen signifikanten Unterschied in der Modelleistung zwischen den klimatischen Referenzperioden und der Basisperiode gibt. Dies deutet darauf hin, dass Modelle, die für kurze Zeiträume mit extremen klimatischen Bedingungen trainiert

wurden, nicht schlechter abschneiden als solche, die für lange Zeiträume ohne solche Bedingungen trainiert wurden, in denen die Daten willkürlich aufgeteilt wurden. Darüber hinaus wurde kein signifikanter Unterschied bei der Imputation von Daten mit Random-Forest-Regressionsmodellen im Vergleich zur Verwendung des einzugsgebietsspezifischen Medianwerts festgestellt.

Abstract

Climate change has introduced new challenges to the domain of hydrological modelling due to the increasing frequency of extreme events, such as floods, droughts or heat waves. Thus, the robustness of traditional, process-based hydrology models is called into question. Rainfall-runoff in particular is a key application of hydrological models aiming to explain the discharge response to precipitation events in a watershed. This thesis evaluates and compares the robustness and accuracy of a state-of-the-art data-driven Deep Learning (DL) approach in the field of Large-Sample Hydrology (LSH), where its emergence has led to a redefinition of requirements, and demonstrates its power to uncover hidden relationships in complex hydrological processes. Prevalent LSH datasets are compared, the LamaH dataset covering 479 catchments in Central Europe is analysed thoroughly, and pre-processing steps are employed to address domain-specific issues, such as imputation of missing streamflow records and anomaly detection. A trend analysis highlights an overall warming trend of $+1.53^{\circ}\text{C}$ over the 39-year study period.

Three types of models are compared in this work: the conceptual process-driven model HBVEdu, the gradient-based Machine Learning model XGBoost, and the state-of-the-art DL model EA-LSTM. To evaluate the robustness of models under transient climatic conditions, a differential split-sample testing approach is employed. This involves four reference periods that represent extreme temperature and precipitation variations, as well as a longer baseline period for comparison to conventional data splitting methods.

Our experimental study has revealed many key insights and results. The DL model significantly outperforms both the process-driven and ML models in all climatic reference periods and the baseline. The EA-LSTM model demonstrates competitive and robust performance with a mean Nash-Sutcliffe Efficiency (NSE) of 0.73486. In comparison, the XGBoost model outperforms the physics-driven HBVEdu model with a mean NSE of 0.56306 and 0.48528, respectively. However, an assessment of the ML model reveals that it is strongly underfitting and sensitive to fluctuations and noise in the data. Notably, there is no significant difference in model performance between the climatic reference periods and the baseline period. This suggests that models trained on short periods with extreme climatic conditions do not perform worse than those trained on long periods without such conditions, where the data is arbitrarily split. Furthermore, there is no significant difference observed when imputing data with Random Forest regression models compared to using the catchment-specific median value.

Contents

Kurzfassung	ix
Abstract	xi
Contents	xiii
1 Introduction	1
1.1 Problem Statement	2
1.2 Aim of the Work	3
1.3 Outline	5
2 Modelling in Hydrology	7
2.1 Processes and Concepts in the Hydrological Cycle	9
2.2 Rainfall-Runoff Modelling	18
2.3 Hydrological Model Types	19
2.4 Hydrological Modelling in the Era of Climate Change	28
3 Large-Sample Hydrology and Data Overview	31
3.1 Large-Sample Hydrology	31
3.2 Description of the Study Area	42
3.3 Data Description	44
3.4 Data Analysis	46
4 Methodology	61
4.1 Data Preparation	61
4.2 Model Architectures and Implementation	69
4.3 Model Evaluation	75
4.4 Experiment Design	79
5 Results and Discussion	83
5.1 Experimental Setup	83
5.2 Test Results	85
5.3 Validation Results	91
5.4 Discussion	95
	xiii

6 Conclusion	107
6.1 Research Results	107
6.2 Limitations and Future Work	109
A Static Catchment Attributes	111
B Hyperparameter Settings	113
B.1 Estimators for Handling Missing Data and Outliers	113
B.2 Settings and Configurations for Models	114
C Algorithms	117
D Test Statistics	121
D.1 Model Performance	121
D.2 DSST Periods	122
D.3 Imputation Methods	123
List of Figures	125
List of Tables	129
List of Algorithms	131
Acronyms	133
Bibliography	137

CHAPTER 1

Introduction

Hydrological modelling plays a pivotal role in understanding and predicting the highly complex and non-linear dynamics of water systems. Forecasts are essential for effective water resource management and environmental planning. Climate change has introduced new challenges and demands for numerous domains. As a result, it is crucial to evaluate the robustness and effectiveness of state-of-the-art hydrological models, specifically focusing on process-driven, data-driven, and hybrid models. While Data Science (DS), in particular techniques from Machine Learning (ML), has contributed immensely to improve the accuracy of hydrological models, it is now necessary to assess their performance and reliability when they are applied to problems under transient conditions. This work aims to investigate the most prominent difficulties within the field of hydrology that have gained importance due to shifting climatic conditions. Further, commonly used large-sample datasets and different types of models based on physical laws, Deep Learning (DL) as well as hybrid models are compared and assessed in order to address the evolving requirements posed by climate change on hydrological modelling outside the typical calibration range [ODO20].

Varying conditions have appeared in climate records only in the recent past, making it difficult to predict or simulate e.g. extreme events such as flash floods, droughts, heat waves, or severe storms far into the past or future. Meteorological and geophysical factors that favour these extremes may not be represented in the period used as training data for hydrological models, or these factors may be difficult to represent physically [KJK⁺22].

One of the key tasks in hydrology is rainfall-runoff modelling, and this will be the main focus of this work. Rainfall-runoff is a broad concept describing the movement of water within a watershed or basin following precipitation events. Simulations strive to explain the process by which moisture (rain or snow) falling onto land surface is transformed into runoff and eventually flows into streams, rivers, or other water bodies. These simulations are necessary for flood prediction, water resource and quality management and hydro power plant optimisation [Bev12].

While process-driven, physical models have been prevalent in hydrology for the better part of the last century [Nas57, SB22, LLWB94, BFMC73, HW23] leading to a better understanding of the complex physical processes involved, data-driven models have emerged in the recent past consistently outperforming their analytical counterparts [KYG⁺21, HKK⁺21, KKB⁺18, ZKBFH21]. These models are able to capture important relationships in highly complex processes. Current research now applies both of these model types as hybrid models in various forms of coupling to leverage the advantages and mitigate limitations of either type [LHB⁺23].

1.1 Problem Statement

Numerous approaches to rainfall-runoff modelling exist, each coupled with advantages and limitations. In the past, the research community has traditionally relied on numerical, process-driven models based on physical laws to predict in a simplified representation of the hydrological cycle. Whilst these models have contributed to the field by providing accurate predictions for various hydrological applications, they suffer from inherent biases, are often highly complex, and require significant domain knowledge and understanding of the study area to produce valuable results [SB22, LLWB94, BFMC73].

Recently, there has been a shift in prevalent hydrological modelling with the emergence of ML and DL as powerful data-driven alternative model types. Especially Artificial Neural Network (ANN)-based models have achieved great success due to their ability to learn complex and potentially hidden relationships in the hydro-meteorological input variables directly from the data.

Furthermore, the DS pipeline as proposed by Biswas et al. includes various stages and processes for data acquisition, exploration, analysis, and preparation, for model training and evaluation, and for interpretation and communication of their results [BWR22]. Adhering to these theory-guided principles has the potential to significantly improve hydrological modelling, as they underlie the powerful interdisciplinary ML paradigm. Yet, pre-processing large scale datasets is a resource-intensive and often empirical process that can possibly introduce bias or uncertainty into the data, rendering predictions inaccurate or sensitive to extreme events [KKG⁺22].

However, as the impacts of climate change have intensified around the world in recent years, it is now necessary to assess the robustness and predictive power of different model paradigms under transient conditions, which describe the increase of variability over certain periods of time, typically associated with conditions that deviate significantly between training and calibration periods [CAP⁺12]. It is particularly beneficial to evaluate model performance on data from vulnerable regions where the effect of climate change is suspected to be strongly present in the data [EEA23]. Several state-of-the-art datasets and collections exist in the domain of Large-Sample Hydrology (LSH), which cover different study areas and have varying limitations. It is of interest to evaluate the requirements of large-sample datasets in the context of climate change modelling in hydrology.

1.2 Aim of the Work

Based on the problem statement, this thesis aims to address the following research questions (RQs):

- RQ 1** What are the challenges and opportunities associated with utilising state-of-the-art hydrological datasets for Machine Learning¹-based modelling, particularly in the context of data engineering and pre-processing?
- RQ 1.1** What are the state-of-the-art datasets in the domain and how do they compare?
- RQ 1.2** What are the current requirements for large-sample datasets in the domain, with a focus on the future considerations for models within the Machine Learning paradigm?
- RQ 1.3** What are the necessary pre-processing steps to analyse the quality of hydrological data and to prepare the data for hydrological experiments?

Hydrological datasets, typically encompassing several decades of meteorological data, have traditionally been used to fulfil the requirements of process-based models over the years. Nonetheless, utilising these invaluable datasets for modern ML applications requires extensive data pre-processing to evaluate and improve data quality, thereby making it suitable for ML-ready formats. This data engineering process is frequently resource-intensive and incurs substantial costs. Recognising the potential of cutting-edge hydrological datasets for Machine Learning-based modelling, research question **RQ 1** aims to methodically examine the difficulties and opportunities associated with the use of such datasets in the context of data engineering and pre-processing in the domain of hydrology. Specifically, the research objective is to understand the range of state-of-the-art datasets available, conduct a comparative analysis, and identify current and future needs for large-sample datasets in the field. Additionally, we explore the key pre-processing stages required to guarantee the quality and readiness of hydrological data for machine learning experiments. The motivation of this research question is driven by the demand of the research community to enhance the utilisation of current hydrological datasets with the emergence of highly complex Deep Learning models, assimilating these

¹Note that the thesis distinguishes between the terms ML and DL. According to Jakhar and Kaur's definitions, ML involves computational methods that enable machines to learn from data without explicit programming. It entails training with provided data and algorithms, enabling machines to make decisions based on processed information. ML is dynamic and capable of self-modification when exposed to additional data. DL is a subset of ML that utilises models which imitate the structure of organic neural networks in the brain, known as ANN. DL enhances the abilities of conventional ML methods by incorporating multiple layers in a deep architecture (hence the term "deep" in DL) to extract hidden patterns and representations from data [JK20]. While both ML and DL encompass distinct algorithms and models, the two terms are often used synonymously. Since ML is the superset, this term is used for the definition of the research questions.

data sources into robust and effective hydrological models [ADAG⁺20].

RQ 2 What are the necessary steps to employ Machine Learning and Deep Learning models for rainfall-runoff modelling in hydrology, and how can their modelling pipeline be compared to that of traditional models?

RQ 2.1 What state-of-the-art Deep Learning models are suited to model rainfall-runoff in hydrology, as identified from the current literature?

RQ 2.2 Which methods and metrics are appropriate to evaluate and compare the performance of these models?

In the field of hydrology, this work conducts a thorough investigation into the use of Deep Learning models for rainfall-runoff modelling. The main objective of research question **RQ 2** is to outline the required stages of applying DL models for this task, with a comparison drawn with the modelling pipelines of conventional models. The current literature will be reviewed and the most appropriate DL architectures to improve hydrological predictions will be identified. To this end, appropriate methods and metrics will be selected to thoroughly evaluate and compare the performance of these models. The primary objective is to improve the understanding of the necessary steps and factors involved in the applications of DL models in hydrology, with a focus on domain-specific evaluation.

RQ 3 In the domain of hydrology, how do state-of-the-art Machine Learning and Deep Learning models compare to traditional physics-based models when applied to the task of rainfall-runoff modelling?

RQ 3.1 Is there a difference in model performance with respect to common evaluation metrics among the modelling paradigms?

RQ 3.2 What strategies can be employed to design experiments for rainfall-runoff modelling to observe the impacts of climate change, and how do models trained using these specific strategies compare with each other?

RQ 3.3 Do variations in model performance arise from employing different data pre-processing methods?

RQ 3.4 What are the insights and conclusions that can be drawn from the results of the modelling experiments?

A major goal of this work is to comparatively evaluate the efficacy of modern models from the domains of ML and DL in contrast to traditional process-based, physics-informed models for the task of rainfall-runoff modelling in hydrology. **RQ 3** addresses the differences in performance between modelling paradigms and evaluates their complexity in terms of runtimes and parameters. Methodologies for designing experiments that can measure the effects of climate change on rainfall-runoff

modelling are investigated, and models that were trained using these strategies are compared. This research question aims to provide useful insights by evaluating the impact of different techniques for pre-processing data devised in **RQ 1.3** on model performance and determining hydrological or meteorological variables that can serve as reliable indicators for reference periods associated with climatic conditions. Ultimately, the research aims to derive significant conclusions from the modelling experiments, improving the comprehension of the effectiveness of various modelling techniques in hydrology.

1.3 Outline

The structure of this thesis is as follows: First, Chapter 2 introduces the most important concepts and processes that are part of the hydrological cycle and provide necessary domain-specific knowledge. Particularly, the task of rainfall-runoff modelling is described and a state-of-the-art analysis of hydrological model types is presented. This chapter further includes an introduction to the requirements and best practise approaches to hydrological modelling in the era of climate change. Chapter 3 contains an introduction to the field of LSH as well as literature research and assessment of prevalent datasets in the domain. Furthermore, the study area and the variables covered by the selected dataset are described. The chapter concludes by an in-depth exploration and analysis of the data, including a trend analysis with respect to the impacts of climate change in Central Europe. Then, Chapter 4 presents the methodological aspects of this work. First, the applied steps to pre-process the static and dynamic data are stated. The architectures and implementation details of the three selected candidate models from the process-based and data-driven modelling approaches are discussed. Evaluation metrics and guidelines are presented before finally describing the experimental design with regard to the various input sets resulting from differential split-sample testing to investigate the robustness of models trained on transient climatic conditions.

The setup and implementation of model training, as well as the specifications of the computing environment are given in Chapter 5. The centre of this chapter, however, is the presentation of model results on the test and validation sets, respectively, for each of the four climatic reference periods as well as the baseline period. The results are visualised and analysed in detail. A thorough discussion of the implications of the modelling results and the findings from the literature review and data analysis undertaken as part of this thesis for the research questions concludes the chapter 5. Lastly, Chapter 6 presents the answers to the research questions addressed in this thesis. An analysis of the limitations of the research and experiments, as well as an overview of future work and open questions in the discussed field conclude this work.

Modelling in Hydrology

Hydrological modelling is a key discipline within the field of hydrology. The aim is to construct and apply mathematical representations that model the behaviour of water within the Earth's hydrological cycle. Hydrological processes are highly complex, involve many physical variables and are characterised by considerable spatial and temporal variability. Fundamental processes and variables such as precipitation, evapotranspiration, runoff, and groundwater flow drive the hydrological cycle. Models utilise numerical representations of these processes to mimic the complex interactions within catchments. While the behaviour of water is often seen as a simple cycle of precipitation, evaporation, and condensation, in reality, the underlying processes of the hydrologic cycle are much more complex and interrelated, requiring sophisticated modelling approaches to achieve an accurate representation. By assimilating observational data from heterogeneous sources and meteorological inputs, hydrological models provide important insights into water quality and availability, the prediction of runoff, floods and droughts, and the impact of anthropogenic activities on water resources. Hydrological modelling is an important tool to facilitate decision-making in water management and climate research based on physical laws, thus promoting a holistic understanding of the water cycle [Nat19].

Extensive research effort has been put into deriving models that represent hydrologic processes resulting in significantly advanced understanding of the factors that drive the behaviour of water on Earth. Initial conceptual models largely aimed at simplifying the representations of hydrologic processes to provide first insights into the very basic mechanisms of the water cycle such as groundwater storage or streamflow. A major theory proposed in early research was the *unit hydrograph* [Doo59]. A hydrograph refers to the graph showing the discharge (or rate of flow) versus time at a specific point in a watershed. These graphs typically allow the correlation of precipitation events and the change in runoff over time. The theory of the unit hydrograph is based on the idea that the hydrograph resulting from one unit of excess rainfall in a catchment will have the same shape and only needs to be scaled for different amounts of rainfall. It offers a

straightforward method of predicting the runoff response of a watershed to any particular precipitation event.

With ever increasing computational power and availability of data, the development of process-based physical and analytical models accelerated. These models became the most widely used and powerful representations of hydrological processes and are still extensively employed in research and real-world applications today. At the core of the process-based models are complex mathematical equations for various hydrological components that are derived directly from physical laws, establishing a high degree of physical realism for diverse hydrological and climatic conditions. The components rely on an array of input parameters explaining characteristics of the specific catchment, topography, vegetation, soil, precipitation, and other hydro-geological factors. Calibration is required to tune the parameters and achieve an accurate representation of the system at hand. These models are easy to interpret, give good results on a coarse scale and are well researched. However, the high computational cost and systematic bias associated with this type of hydrologic model led to further development of alternative approaches [GGJP20, GRA⁺22].

In recent years, hydrologic modelling has undergone a paradigm shift towards data-driven models based on advances in DL. The ability of ANNs to capture complex hidden relationships in the data, and the excellent performance of Long Short-Term Memory (LSTM) models in time series prediction have led to DL becoming the leading approach for hydrologic modelling. While these models are highly adaptable and computationally efficient, they are also considered black-box models because they lack explainability and interpretability and are generally not bound by physical laws. Recently, however, numerous extensions to well-studied ANN models have been proposed to overcome inherent problems of DL or hydrological modelling in general [KKH⁺19, KJK⁺22, OEAF21].

The National Oceanic and Atmospheric Administration (NOAA) offers a visualisation of the hydrological cycle, shown in Figure 2.1, which includes all the main variables. It illustrates the fundamental processes that drive the movement and transformation of water through various topographies at different phases. Although the water cycle is a universal process that can be applied to any catchment, the *uniqueness of place* described by Beven requires regionalised modelling due to the individual characteristics at the catchment scale. Different expressions of topography, vegetation, soil, human modification complicate the extrapolation of knowledge to catchments where no data is available [Bev00]. This issue is closely related to the problem of Prediction in Ungauged Basins (PUB), which is considered one of the major challenges in hydrology [HSB⁺13].

Runoff is measured in only a fraction of catchments worldwide. In stream systems, gauges must be established to measure water levels and flow, which can be converted into runoff, which represents the volume of water flowing through a given cross-section per unit of time. However, as there are very few gauges available in catchments worldwide, there is generally little information on runoff. This data is urgently needed for managing water availability and quality, and for forecasting floods and droughts. PUB refers to this exact problem, that is, the need to simulate runoff in catchments without any means of

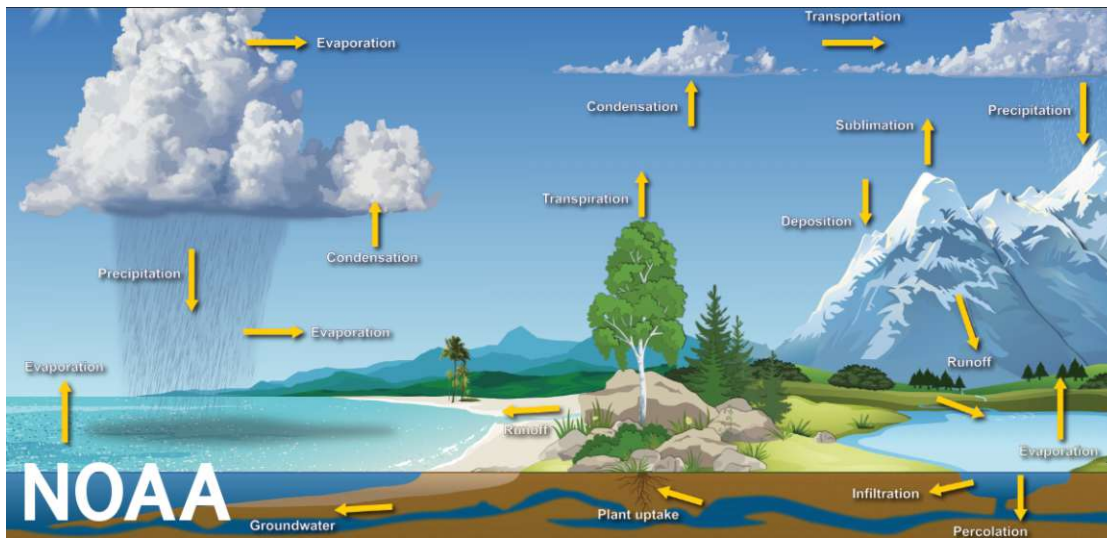


Figure 2.1: Conceptualisation of the water cycle visualising key processes and variables by the NOAA. Image credit: Dennis Cain/NWS [Nat19].

measurement by using information from gauged catchments or alternative approaches [BSW⁺13].

This chapter includes a thorough definition of the variables and processes involved in the hydrologic cycle. Subsequently, these concepts are placed in the context of the field of rainfall-runoff modelling, which is the focus of this work. Understanding the interplay of processes is an integral step before a novel model can be developed to solve the most common problems in hydrological modelling. An overview of the state of the art of different physical, data-driven and hybrid model types is given, with a focus on models based on DL to provide comprehensive information that will be incorporated into the design of a new model in section 4. This is followed by a thorough description of the most common datasets and collections in the field. The compilation of models and datasets in this section also forms the basis for baseline model and data selection for the experiments in this work. This chapter concludes with an overview of the factors influencing hydrological modelling due to climate change impacts and a review of the current state of research in this field.

2.1 Processes and Concepts in the Hydrological Cycle

In the field of hydrological modelling, catchments form the very basis of analysis, as all hydrological processes are inherently linked to these essential geographical units. In order to explain the movement, distribution and transformation of water in the Earth's hydrological cycle, it is essential to understand the processes in the catchment areas.

A *catchment* is typically defined as a specific geographic area that is delineated by natural

boundaries, such as ridges, hills, or mountains. All surface water of a catchment flows and converges into a common channel, ultimately forming rivers or creeks. A catchment thus represents a self-contained hydrological unit, where processes such as precipitation, runoff, and groundwater interplay within a coherent closed system. Catchments exhibit a high degree of spatial variability and can differ in many characteristics, making them difficult to classify and compare. In many cases, catchments consist of smaller areas, called *subcatchments*, which in turn are themselves separated by ridges or other boundaries. The term catchment is often used synonymously with the terms (*drainage*) *basin* or *watershed* [Dep21]. Sivakumar et al. emphasise the need to categorise catchments according to their variability, complexity and internal interconnections, but there is no unified classification system in current research [SSBK15]. A *hydrologic unit* typically refers to a combination of regionally related catchments. Several catchments drained by a river system and its tributaries form a single hydro-geologically interconnected unit.

A *water year* or *hydrologic year* describes the period of time that covers the natural course of the hydrologic seasons with a duration of twelve months. The beginning of the period is usually marked by the season of soil moisture recharge followed by the season of maximum runoff or groundwater recharge, and ends with the season of maximum evapotranspiration or soil moisture utilisation. In the Northern Hemisphere, this period typically occurs between 1 October and 30 September, and in the Southern Hemisphere from 1 July to 30 June [AMS23]. The reason why this period differs from the calendar year is that most precipitation usually occurs in autumn and winter. The rainfall at the end of the calendar year does not affect streamflow until spring. Thus, the water year usually begins and ends with the wet season. The purpose of a water year is to establish a consistent and standardised cycle that can be used for comparison in hydrological modelling, especially in regions with distinct wet and dry seasons.

2.1.1 Precipitation

The hydrologic cycle describes the movement of water around the planet and consists of various components. *Precipitation* (usually denoted as P) is the central process in the cycle and serves as the primary input of water into the system, driving its distribution throughout the other processes. Depending on regional atmospheric conditions such as temperature and air pressure, precipitation can occur in different forms, e.g. as (freezing) rain, snow, hail, sleet or drizzle. Water vapour in the atmosphere cools and condenses to form clouds. Once water droplets and ice crystals within the clouds are sufficiently heavy and condense on nuclei (so called *condensation nuclei*), such as dust or smoke particles, they fall to the surface. This occurs only if the velocity of the droplet exceeds the cloud updraft speed. In summary, three conditions need to be met for precipitation to form: cooling of the atmosphere, condensation of water droplets onto nuclei, and growth of the droplets. The amount of precipitation is determined by static and dynamic factors. Static influences are, for example, topography, altitude, aspect and slope and do not vary between rainfall events. Dynamic factors are mainly changes in weather conditions. While static influences usually predominate on a smaller scale, static and dynamic factors

can interact on a continental scale and strongly affect the distribution of precipitation [Dav08, USG23].

Although measuring precipitation on a broader scale is relatively easy, it proves difficult to measure all types of precipitation accurately and they show considerable spatio-temporal variability within a catchment. Precipitation is quantified as a vertical depth of liquid water. The amount of precipitation is usually measured in millimetres as a depth rather than in volume units such as litres or cubic metres. This measure represents the depth of water that would accumulate at the surface if all the rain remained at its point of impact. Snow can also be represented as depth, but in hydrological modelling it is usually more useful to express snow as water depth equivalent, i.e. water depth if the snow melts. This accounts for the fact that snow occupies about 90% more volume than liquid water. Snow Water Equivalent (SWE) refers to the corresponding amount of liquid water that is stored in the snow cover. It represents the water column that would be generated if the entire snowpack were to melt instantaneously and is defined as the product of the depth and density of the snowpack. The variable notation for this is SWE . Rainfall intensity and storm duration are other rainfall measures of interest. It is important to note that due to the extremely high variability of precipitation at the catchment level, all measurement methods should rather be considered *estimation* methods [Dav08].

Once precipitation reaches the surface, it can follow various paths. Water can be captured by vegetation or other topographical features, triggering processes such as evaporation or transpiration that return the water into the atmosphere. Some of the water can directly add to the runoff in the respective catchment area, thus changing and shaping streams and rivers. Other parts of the precipitation can be converted into soil moisture and further stored as part of the groundwater by infiltrating into the soil.

2.1.2 Evaporation

The process of *evaporation* refers to the transformation of liquid water into its gaseous state and its subsequent diffusion back into the atmosphere. This process takes place when there is enough energy (i.e. heat) to break the covalent bond within a water molecule and the atmosphere is sufficiently dry to take up the water vapour. The influence of evaporation on the hydrologic cycle varies regionally, seasonally and climatically. The presence of water (or lack thereof) defines distinctions in the evaporation process. Evaporation over a body of water, e.g. a lake or an ocean, is defined as *open water evaporation* (often denoted as E_o) and accounts for the majority of evaporation. *Potential evaporation* (PE) refers to the maximum possible evaporation that occurs above the land surface. The underlying assumption of PE is that the water supply is unrestricted and the whole surface is completely covered by water. *Actual evaporation* (E_t) defines the actual evaporation that occurs and is equal to PE for very wet conditions, e.g. directly after a heavy rain. E_t accounts for the limited availability of water at the surface. Evaporation above land occurs as E_t or *transpiration* from vegetation in the process of photosynthesis and respiration controlled by the available water in the soil, or a combination of both, usually referred to as *evapotranspiration*. Evapotranspiration and actual evaporation

can be considered equivalent processes for hydrologic modelling and are generally both denoted by E_t [Dav08].

The primary energy source driving evaporation is radiation from the sun. The sum of all forms of heat fluxes at the surface is described as *net radiation* (Q^*) and can be described by the following equation:

$$Q^* = Q_S \pm Q_L \pm Q_G \quad (2.1)$$

Sensible heat (Q_S) is energy that can be measured and felt as warmth. Latent heat (Q_L) is energy absorbed or released during a phase change of water. This can be a negative flux when energy is absorbed (i.e. liquid to gas) or positive in the opposite direction (i.e. gas to liquid). Soil heat flux (Q_G) is energy released from the soil and can usually be disregarded. Other energy sources that are important for the evaporation process are anthropogenic energy (e.g. through heating or industrial efforts) and advective energy, i.e. energy transported from far away to the surface where evaporation takes place (e.g. after a cyclone) [Dav08].

Evaporation as a key process in the hydrologic cycle is heavily dependent on the availability of water. In open water evaporation, the water is taken directly from a lake, river lake, river, pond or other body of water. While water can evaporate directly near the soil surface, the process becomes more complex when the water is stored deeper in the soil. A moisture gradient attracts water from deeper layers, and it must simultaneously overcome gravity and the forces exerted by soil capillaries. Vegetation can also help water reach the surface through osmosis in the root systems. For water vapour to diffuse into the atmosphere, it must not yet be saturated. The amount of water stored in a specific portion of air depends on temperature and pressure and *relative humidity* describes the saturation of the atmosphere with water vapour. *Atmospheric mixing* expresses the quality of the diffusion of a parcel of air with the surrounding atmosphere and is indicated by wind speeds above the evaporation surface [Dav08].

Evaporation and evapotranspiration are very difficult to measure and are therefore often only estimated or only indicators such as temperature, pressure, soil moisture content, and wind speed at different altitudes are given. The reverse process of evaporation is *condensation*, which describes the transferal of water vapour into liquid water. Saturated air needs to be cooled down sufficiently for condensation to occur. The processes *sublimation* and *deposition* describe the transition from the solid to the gaseous state of matter and vice versa without an liquid intermediate stage. In hydrology, sublimation is considered as the conversion of ice into water vapour; deposition is the opposite process [Dav08].

2.1.3 Water Storage

The hydrological cycle various stages where water is stored at least temporary, including soil moisture, groundwater, snow, ice, lakes and reservoirs. The term *water storage* is an

essential component in the hydrologic processes and is usually denoted as S . However, since storage itself is not static but a dynamic movement where inflow and outflow do not necessarily coincide in a given period of time, it is more appropriate to consider the *change in storage* or ΔS . On Earth, water is mainly stored in the form of snow and ice on the polar ice caps and makes up the majority of fresh water. All water stored below the surface is considered groundwater. Water above the water table, however, is considered unsaturated and is usually referred to as soil water. Water below the water table is saturated and considered groundwater, although the two types continuously mix vertically and horizontally through infiltration and flow, respectively [Dav08].

Soil is a conglomerate of water, air and solid particles, such as minerals or organic matter. The rate of *infiltration* determines the amount of water seeping into the soil over a period of time and depends on the content of water in the soil (i.e. its saturation) and the ability of the soil to transmit water. An important property of soil is *porosity*, which is the fraction of pore space in the volume of soil. While technically the pores can be completely filled with water, corresponding to the maximum volumetric water content (i.e. porosity), in practice the volumetric soil moisture rarely reaches the porosity value. When it does, gravity causes rapid drainage through the profile, reducing the moisture content below the porosity value. *Field capacity*, on the other hand, represents the stable saturation point after rapid drainage [Dav08].

Water stored in the saturated zone below the water table can only escape through transpiration, but not evaporation. Groundwater is constantly in motion, be it as part of large fossil water reserves or underground river systems in limestone. During dry seasons, groundwater is essential for maintaining streamflows. Similar to most other processes in hydrology, the storage of water in the soil or ground is difficult to measure, especially at a significant spatial scale [Dav08].

2.1.4 Runoff and Streamflow

The term *runoff* broadly describes the movement of water in a channelised stream on or below the surface at various velocities and is denoted as Q . The water always moves in a channelled form towards the ocean as soon as it reaches a river, which is then referred to as *streamflow*. Streamflow is expressed as *discharge*, i.e. the volumetric flow rate of water in a stream over a period of time in m^3/s . In rainfall-runoff modelling, streamflow is frequently also expressed in mm per day and scaled by catchment area to facilitate comparisons with rainfall data, which is expressed in terms of depth. In this context, streamflow in mm per day represents the equivalent depth of water that would have to fall over the catchment to produce the observed flow. Streamflow (or its average) plotted in a continuous record over time is called a *hydrograph*. This is a simple but significant visualisation and source of information for hydrological modelling.

A hydrograph commonly exhibits multiple peaks interspersed with intervals of steady flow. These peaks are referred to as *peakflow* or *stormflow* and occur after substantial precipitation events. The intervening stable periods of low amid peaks are called *baseflow*.

The shape of a hydrograph is determined by catchment characteristics such as size, slope, shape, vegetation cover and type, antecedent soil moisture and urbanisation, as well as by the storm, such as intensity and duration of precipitation, and other physical characteristics. A storm hydrograph consists of a *rising limb* leading to peakflow, driven by precipitation falling directly onto the stream, and a *recession limb* characterised by a slow decline in discharge until baseflow is restored. Figure 2.2 illustrates a typical hydrograph as well as a storm hydrograph.

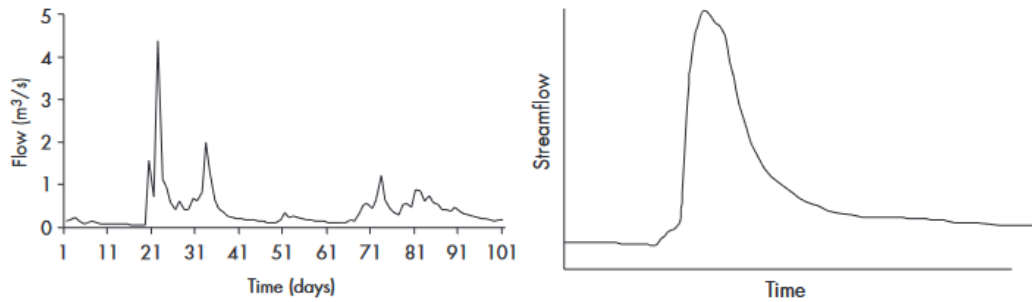


Figure 2.2: Left: A typical hydrograph with flow values in m^3/s plotted against time in days. Peakflow and baseflow are easily distinguishable. Right: A typical storm hydrograph with a rising limb, peakflow and a recession limb. Taken from [Dav08].

The exact mechanisms of runoff are highly complex and have been subject of the most substantial research efforts in the field of hydrology since the inception of this science. On a high level, runoff can be expressed as a combination of three processes at the hillslope scale: *overland flow* (Q_o), *throughflow* (Q_t) and *groundwater flow* (Q_G). Their respective relative influence on streamflow depends on the characteristics of the catchment at hand as well as the properties of precipitation during a storm event. It is important to note that rainfall is the driving factor for any type of runoff [Dav08].

The prevailing mechanisms of Q_o vary depending on the climate zone in which the catchment is located. For humid areas, *saturated overland flow* is dominant, which describes the rising of the water table until it reaches the surface due to a combination of infiltration and throughflow. The resulting overland flow is thus driven by return flow, meaning water that was already stored in the soil and resurfaces, and by rainfall striking areas that are already saturated. Naturally, the water table is closest to the surface and is most likely to rise above it in the most saturated areas, which are near the banks of a catchment or at the base of a slope. The concept of the *variable source area* describes that the saturated partial response area of a catchment that contributes most significantly to peakflow is dynamic in space and time during any storm event. This concept is an essential part of explaining stormflow. In semi-arid to arid climates, however, the predominant mechanism is *infiltration excess overland flow*, which occurs whenever the intensity of rain exceeds the infiltration capacity of the soil. This results in thin layers of water running over the ground and is one of the main causes of flash floods. This type of overland flow is especially common in agricultural or highly urbanised areas

where the ground is sealed resulting in hydrophobic soil [Dav08].

Below the surface, water in the soil or in the fully saturated zone that was stored there prior to storm events also contributes to the storm hydrograph in the form of throughflow or groundwater flow. Throughflow or Q_t refers to the movement of water through the unsaturated zone either within the soil matrix or along preferential flow paths. Once the soil is saturated, water movement is not limited to vertical flow, and a declining water table on a hillside can cause water to flow downslope. However, the velocity of water moving through a saturated soil matrix is fairly slow. To contribute to storm-related runoff, throughflow must occur through mechanisms beyond the soil matrix. One theory to explain the rapid movement of water from the subsurface into the stream is translational flow, which compares the process to a piston where pressure at the top of the soil column leads to a release of water at the bottom, creating a pressure wave. Groundwater also contributes to stormwater runoff by raising the water table in the immediate vicinity of a stream. A small increase in water can quickly change the soil moisture from an unsaturated to a saturated state causing a ridge in the groundwater. This so-called capillary fringe effect begins to take place well before any throughflow occurs [Dav08].

Groundwater is the predominant source of baseflow, which is created in particular by the slow infiltration of water from the groundwater into a stream. In fact, streams or lakes are usually formed in areas where the groundwater table breaks the surface [Dav08].

When the water reaches a stream, it flows through a network of channels leading to the main river. The amount of water present, the gradient of the channel, and the flow resistance in the channel bed control the flow rate of the water. Topographical features such as rocks, sand ridges or vegetation influence the water velocity. Channel networks exhibit a high degree of spatio-temporal variability [Dav08].

Measuring streamflow is referred to as *hydrometry* and is performed either instantaneously or continuously. While models in hydrology can simulate almost all processes in the water cycle, runoff is usually considered the outcome of most relevance. Runoff and precipitation are the two key processes in the specific problem of *rainfall-runoff simulation*. This research area is the core of this work and is explained in detail in section 2.2.

Floods

A *flood* is typically associated with the inundation of land adjacent to a river that occurs due to unusually large runoff or intrusion of seawater. However, flood sources can also extend to lakes or the ocean, and there is no clear definition of the amount of excess water required to cause a flood. Floods usually are assessed on the basis of the extent of damage caused, which leads to an assessment based on cost rather than hydrological criteria. By far the most important cause of river flooding is abnormal rainfall that exceeds a river's capacity and pushes it over its banks. Factors that influence the intensity of floods include antecedent soil moisture, deforestation, urbanisation, river channel modification, land drainage and climate change. The increase in heavy precipitation events due to climate

change leads to an elevated probability of flooding [Dav08]. Blöschl et al. find regional patterns of increasing and decreasing river flood discharges in Europe over a 50-year period that are indicative of climate-related changes. These flood trends are roughly consistent with climate projections for the future, highlighting the need to incorporate climate change assessments into river flood management [BHV⁺19].

Streamflow Analysis

The evaluation of continuous runoff records for the analysis of streamflow is an essential task of hydrology. For instance, separating baseflow from peakflow in a hydrograph is a supposedly simple, but in reality imprecise and a subjective exercise. To standardise the analysis, the concept of the *unit hydrograph* was introduced, which assumes that the shape of the hydrograph of a storm is determined by the physical properties of the catchment. These properties are constant over time, so that an averaged hydrograph for a given storm size is able to predict other storm events. In essence, the idea behind this concept is that identical rainfall on a catchment with the same antecedent conditions should yield identical hydrographs. By encompassing temporal variation of discharge, insights beyond peakflow and baseflow can be gained. The unit hydrograph then allows to directly derive the amount of hourly runoff per *mm* of effective rainfall for a storm event, which is equivalent to overland flow. The process to construct a unit hydrograph from a storm is complex [Dav08].

Another important tool is the frequency analysis of a specific flow, which is best expressed in the *flow duration curve*. This plot shows how often a certain flow is exceeded and typically uses daily mean flows for longer periods as input data. The flow values are transformed by the natural logarithm and plotted against the percentage of the cumulative frequency of a flow. The resulting line in the diagram is indicative of the tendency of a catchment to experience frequent high or low flows (i.e. its variability) or towards relative stability [Dav08].

Flood frequency analysis is pivotal technique in hydrology. The analysis primarily focuses on peakflow and uses annual maxima in streamflow values as input data. As floods often occur in the wet periods in response to heavy rainfall or in spring/summer in response to high melt, it is imperative to use the water year as the reference period when acquiring annual peakflow maximum data. The conventional calendar year would lead to cutoffs during flood events. The initial step in flood frequency analysis is to create (relative) frequency histograms and derive probability distributions. Three interrelated terms are of interest that need to be calculated. The *probability of exceedence* $P(X)$ describes the likelihood that a flow is greater than or equal to a value X . The *relative frequency* $F(X)$ describes the likelihood of a flow to be less than X . The *average recurrence interval* $T(X)$ describes the chance of exceedence once every T years [Dav08].

Computer models in hydrology are often considered black boxes which try to achieve an accurate simulation of a relationship in the data by employing numerical methods. A prototypical process to be modelled is the relationship between annual rainfall and runoff,

which can be represented as a linear regression line in a simplistic way. The term *black box* is used because all variables of the hydrological cycle that are relevant to precipitation are used as input, however, the resulting model only applies to the chosen time scale or the climatic and geologic conditions of the used catchment [Dav08]. More sophisticated models achieve the capability to generalise to other spatio-temporal resolutions and catchment properties. In section 2.3, historical and state-of-the-art models that use numerical methods in computer simulations in hydrology are described in detail.

2.1.5 Water Balance

Now that the key processes of precipitation, runoff, evaporation and (change in) storage that drive the hydrological cycle have been defined, the water balance of a system can be expressed in an equation. The *general water balance equation* is as follows:

$$P \pm Q \pm E_t \pm \Delta S = 0 \quad (2.2)$$

The plus-minus sign is appropriate because each term can be considered either a loss or a gain for the system, depending on the perspective. Precipitation, for example, can be considered a loss of water from the atmosphere, but also a gain of water at the surface. Normally, all terms are considered a gain for the surface, so only evaporation is regarded as a loss. The equation expresses that all inflows and outflows to a water system are equivalent with the addition of a change of water storage over time accounting for the principles of continuity regarding mass and energy. Thus, the hydrological cycle can be considered a closed system where all mass (i.e. water) or energy is preserved. Equation 2.2 represents the fundamental theory of hydrology and is the target of most modelling in the field [Dav08].

2.1.6 Water Quality

Groundwater often acts as a pressure wave response to recharge with rainwater. The water entering the stream is not the same water that originally infiltrated and triggered that reaction. Consequently, the water entering the stream may be several years older and its quality may not be affected by current changes in land use. As part of a holistic hydrological analysis, it is imperative to consider not only water availability and water behaviour, but also water quality. It is evident that the quality of drinking water in a stream is most strongly influenced by pollution, but the natural occurrence of suspended matter also plays an a major role. The amount of sediment is determined by the velocity of the water. The slowing down of the water, e.g. by damming or relocating the stream bed, directly leads to sediment deposition, which in turn reduces the capacity of the reservoir downstream of the source of the slowdown [Dav08].

In the case of human-induced sources of water pollution, a distinction can be made between diffuse sources, i.e. sources distributed over a large area without a precise point of origin, and point sources, i.e. specific locations that cause damage. Diffuse sources can be fertilisers or pesticides in excess, point sources can be sewage pipelines. Furthermore,

three types of pollutants can be identified: *toxic compounds*, *oxygen balance affecting compounds*, and *suspended solids*. Streams often carry large amounts of human waste. Through degradation, dilution and dispersion, rivers can mitigate the effect of waste on water quality. The power of these three processes is determined by the temperature, pH value, the amount of water in the stream and the mixing potential [Dav08].

There are many physical and chemical parameters that influence the quality of the water to a greater or lesser extent. These include temperature, dissolved and suspended solids, electrical conductivity, turbidity (i.e. “cloudiness”), pH, dissolved oxygen, trace organics and biochemical oxygen demand. Contents of nitrate, phosphate, heavy metals and chlorine are also of interest [Dav08].

2.2 Rainfall-Runoff Modelling

As presented in the previous section, the hydrological cycle is highly complex and consists of many interrelated processes. Runoff represents the link between precipitation and streamflow and constitutes the principal outcome of all modelling approaches in the field of hydrology. The first attempts to simulate the extent of runoff after rainfall events using regression techniques date back to 1850 and were presented by Mulvaney [Mul50]. Over the years, approaches to modelling this relationship have become more sophisticated and have been constantly refined by gradually integrating physical laws into the mathematical frameworks. Spatial and temporal variability and physical properties as well as boundary conditions of catchments were progressively incorporated into the models, leading to increased accuracy. Advances in computing power and the availability of high-resolution data have greatly accelerated the development of model accuracy [KKB⁺18].

Rainfall-runoff modelling is especially concerned with surface runoff, also known as overland flow (Q_o) as described in Section 2.1.4. This occurs when rain fall to the surface without infiltrating the soil and instead flows over the land surface, ultimately joining surface waters such as rivers, lakes, or reservoirs. Surface runoff plays a vital role in maintaining the balance of the hydrological cycle by regulating excess precipitation and affecting the inflow into stream systems. Simulations of rainfall and runoff are a crucial tool for monitoring water availability and quality, predicting floods, assessing ecological relationships, and conducting research in general. Surface runoff is a significant factor in the dispersion and transport of pollutants and therefore an important instrument for effective water resource management [SKP⁺18].

Hydrological process interactions are non-linear and dependent on the current specific state of the catchment system, which represents the system’s memory. Kratzert et al. describe a mathematical formulation of the state-space approach. The authors suggest that the state S of the system at a specific point in time t is dependent on the input I_t , the state at the previous time S_{t-1} , and a set of additional parameters Θ_i . Hence, the new state of the system can therefore be represented as a function of the aforementioned components:

$$S_t = f(I_t, S_{t-1}, \Theta_i) \quad (2.3)$$

The discharge at time t is affected by the current system conditions and the meteorological events that happened in previous time intervals. The output runoff Q_t of a rainfall-runoff model can then be described as:

$$Q_t = g(I_t, S_t, \Theta_j) \quad (2.4)$$

where $g(\cdot)$ is the mapping function connecting system states, inputs and output [KHK⁺19].

Beven classifies runoff models based on their structure as empirical, conceptual, and physical models, and by the spatial interpretation of the model's catchment area as lumped, semi-distributed, and distributed models [Bev12]. These classifications provide a basis for the model types presented in section 2.3.

2.3 Hydrological Model Types

Sitterson et al. categorise hydrological models into three types: empirical, conceptual, and physical. These types increase in complexity and required domain understanding as listed. Empirical, also referred to as data-driven, models employ non-linear relationships to inputs and outputs, but do not utilise physics-informed knowledge of the catchments and typically only allow for a single output variable. Methods from ML and ANNs belong to the empirical rainfall-runoff model type. Conceptual models rely on simplified equations of physical hydrological processes, requiring significantly more parameters to calibrate. These models are simple in structure and their calibration is a straightforward optimisation process. Physical models directly incorporate laws of physics, such as conservation of mass and energy, momentum, and kinematics, into the modelling of hydrological processes. Spatio-temporal variability is explicitly accounted for and model parameters are directly connected to physical catchments. However, physical models are highly complex in nature and require large amounts of data to calibrate [SKP⁺18].

With the emergence of methods from DL, such as ANN architectures, as the leading concept for interdisciplinary modelling across many domains, the categorisation of hydrological models requires reconsideration. Mohammadi et al. combine conceptual and physical model types into a single category: the Process-driven model (PDM). This type aims to simulate all processes of the hydrological cycle and incorporates a certain amount of physics-informed knowledge into the simplified architecture of a model. Advantages and limitations of both physical and conceptual model types apply to PDMs, however, they are more diverse in structure and cover an array of different popular models. On the other hand, the authors propose a second model type, the Data-driven model (DDM), which relates to the empirical model type. Historical time series data is used to predict the runoff behaviour on unseen data. This type does not require significant understanding of the domain, physical laws or hydrological processes. Rather, black-box models leveraging the potential of large amounts of data are applied without the need to express highly complex hydrological cycle as simplified equations. These models can learn relationships between hydrological concepts and catchments directly from the data, and are capable of uncovering hidden relations. DL methods exploit the non-linearity of streamflow as

they are not restricted to linear modelling [MMCD21, SKP⁺18, HKK⁺21]. Kraft et al. introduce the terminology for a third model type, the Hybrid hydrological model (H2M), which combines the PDM and DDM in a hybrid approach. The authors motivate the hybridisation by developing a hydrological modelling architecture “that exploits the data adaptivity of neural networks for representing uncertain processes within a model structure based on physical principles (e.g., mass conservation)” [KJK⁺22].

The naming and classification of hydrological model types varies in the literature, and there is no standard definition to categorise models, given the increasing importance of DL models. In order to use consistent terminology, the three model types described above, namely PDMs, DDMs and H2Ms, are used throughout this work [LLH⁺18].

An abundance of hydrological models exists for numerous applications and problems. This section focuses on models specifically designed for rainfall-runoff simulation. In the following, several state-of-the-art representative model architectures are discussed and compared for each of the three model types.

2.3.1 Process-Driven Models

PDMs are physical, analytical simulations of rainfall-runoff processes. They typically suffer from large uncertainties associated with hydrological processes due to their high degree of complexity and the need for model parameterisation and calibration. The high computational cost and systematic bias in results are other disadvantages. However, these models are well researched and calibrated, represent physical processes realistically, generalise well, are easy to interpret and yield good results on a coarse scale [GGJP20, GRA⁺22]. PDMs are classified into (i) lumped models applied to a single region using spatially averaged characteristics, (ii) semi-distributed models where basins are broken down into sub-basins and runoff volumes are accumulated downstream to estimate the output at the outlet, and (iii) fully distributed models representing processes at a high resolution by utilising grids, sub-basins, flowplanes or triangulated irregular networks [KYG⁺21]. Fully distributed models typically provide the most detailed predictions.

The Nash Model, which is also referred to as the “Linear Cascade”, considers a catchment as a sequence of linear reservoirs. Here, the output from the upstream reservoir is directly transmitted as the input to the downstream reservoir forming an arrangement in the form of a cascade. The Nash model has been used as the basis for numerous rainfall-runoff models and, despite its publication in 1957, still provides an important mathematical basis for describing surface runoff [Nas57].

The Hydrologiska Byråns Vattenbalansavdelning (HBV) model is a lumped to semi-distributed conceptual hydrology model developed in the 1970s in Sweden to analyse snow accumulation and melt, soil moisture, and runoff response and has since gained large popularity across the world. Its advantages lie in the simplistic representation of processes, good performance, and prevalence in industry and research, especially in Scandinavia. The model has been applied in about 100 different countries. The number of parameters and forcing input requirements (typically T and P) are relatively low. The

HBV model can be tested via an easy-to-use software tool called “HBV-light” that is freely available¹. There are numerous extensions and modifications available for this model [SKP⁺18, SB22, AH10].

Seibert and Bergström identify three distinguishable physical components in the structure of the model: snow accumulation and melting, the accounting of soil moisture and the response to runoff, which contains groundwater dynamics. It consists of four distinct routines: snow, soil moisture, response and routing. Catchments are divided into rather large grid cells of typically 1 km², each treated as a single unit with aggregated inputs and outputs. The principal use of HBV is in rainfall-runoff modelling, although research examining the impact of climate change on water resources is increasingly significant. A critical concern in this context is the transferability of the model results under climatically transient conditions. Differential split-sample tests have revealed that transferring calibrated parameters from climatically differing reference to testing periods can lead to significant uncertainty. Further challenges arise from the potential impacts of climate change on catchments, which may change the vegetation and soil over time. This issue calls for modifications to model parameters rather than holding them as constant factors [SB22].

Continuous Semi-distributed Runoff (COSERO) is a conceptual semi-distributed model originally developed for runoff forecasting in alpine catchments in Austria and is now an important cornerstone in hydrological research, applied to different climatic zones and spatio-temporal resolutions. The model makes use of a vast array of attributes including soil water storage, snow accumulation/melting, glacier melting, evapotranspiration, etc. It requires a time series for precipitation, air temperature, and potential evapotranspiration (PET) as input. It makes use of a vast array of attributes including soil water storage, snow accumulation/melting, glacier melting, and evapotranspiration. Notably, COSERO is used by Klingler et al. as the baseline model in the study presenting the LamaH dataset, which is selected as basis for the rainfall-runoff experiments in the course of this work. The authors selected this model as baseline because its performance has been tested in varying climatic conditions and spatio-temporal resolutions [KSH21]. However, a major disadvantage is the lack of open-source implementations of COSERO.

The Sacramento Soil Moisture Accounting Model (SAC-SMA) model is a lumped, continuous, conceptual soil moisture accounting model from the US and simulates movement of water through a watershed. It incorporates key processes such as precipitation, snow accumulation and melt, temperature, and potential evapotranspiration as inputs and provides soil moisture, evapotranspiration, and runoff as outputs. In the conceptualisation of the model, basins are divided into lower and upper zones with regard to certain depths. The distribution of moisture as well as free water components are parameterised separately for both zones. Furthermore, the model is capable of expressing effects of frozen ground as part of the rainfall-runoff process, which is a unique aspect. River and water supply forecasting and estimation of hydrological extremes as well as basin-specific

¹Source: <https://www.geo.uzh.ch/en/units/h2k/Services/HBV-Model.html>

climate change are key applications of the SAC-SMA model. Several years of data can be used for model calibration. The model is especially popular in the United States, where the National Weather Service of the NOAA relies on its predictions. The model was originally developed by Burnash et al. in 1973 in Fortran and is publicly accessible [BFMC73]. There are modern implementations of the model, e.g. as *R* packages ².

The Soil and Water Assessment Tool (SWAT) is a physically-based, long-term continuous and semi-distributed model that simulates the hydrological cycle at the watershed scale at a daily resolution. The model can be considered a hydrological transport model aiming to describe the movement of sediment, pollutants, and water in general across a network of basins. It features numerous components to represent weather, land use, soil properties, vegetation, and climate to simulate processes such as evapotranspiration, infiltration, runoff, and nutrient transport. The model has been found to be robust, and produces accurate predictions for water resource management at a global scale [HW23].

Five critical types of input data are necessary to use the model: weather, topography, soil, land use, and land management. These diverse types of data typically exhibit high degrees of spatio-temporal variability, and their collection can be a difficult process. Due to its design for the landscape of the United States, application of the model to other areas requires resource intensive data pre-processing. For this reason, many countries have devised custom extensions explicitly tailored to the prevalent conditions. It is mostly used for water resource and quality management, but can also quantify the impact of land management on river basins. SWAT is one of the most popular hydrological models and frequently finds application in the domain of agricultural modelling. It can be tested via a command-line tool available for Windows and Linux on the official website³ [HW23].

The Variable Infiltration Capacity (VIC) model, a semi-distributed physical hydrological model operating at the macro-scale, has been developed since the 1990s by Liang et al. at the University of Washington and has found widespread use in a variety of applications [LLWB94]. VIC is used for tasks such as hydrological dataset construction, trend analysis, data assimilation, forecasting, coupled climate modelling and climate change impact assessment. The model assumes the land surface to be at a large scale with uniform grid cells exceeding 1 km. Sub-daily meteorological variables P , T , W , wind speed, and atmospheric as well as vapour pressure are required as input. Notably, water can only enter the system from the atmosphere and there is no exchange of moisture with the soil. Furthermore, exchange of water between catchments is also not modelled [SKP⁺18].

The recently developed VIC-5 represents a significant advancement, with a revised source code available via a public GitHub repository⁴. This encourages collaboration and enhances the model's versatility for modern hydrological modelling applications. The code is distributed under the permissive MIT license and is written in C for Linux/Unix platforms with an experimental Python driver also available. VIC-5 is equipped with a robust testing infrastructure to ensure reliability and reproducibility [HNB⁺18].

²Source: <https://github.com/tanerumit/sacsmaR>

³Source: <https://swat.tamu.edu/>

⁴Source: <https://github.com/UW-Hydro/VIC/>

2.3.2 Data-Driven Models

DDMs typically encompass models from the domains of Machine Learning, and, more specifically in recent years, Deep Learning. Notably, these models usually do not incorporate a foundation on physical laws or aim to represent hydrological processes as simplified parameterised equations by design. Much rather, DDMs leverage the potential of the data itself to uncover complex patterns and relationships inherent in hydrological systems. By harnessing large datasets containing information on hydro-meteorological variables, these models excel at capturing non-linearities and spatio-temporal relationships. The strength of DDMs lies in their ability to adapt and learn directly from observational data, without the need to explicitly formulate physical processes. This adaptability makes DDMs particularly valuable for hydrological applications in diverse and dynamic environments, where the underlying processes may be influenced by a multitude of factors. However, it is important to note that the interpretability of these models can be a challenge, as the learned relationships do not always agree with established hydrological theories, and the lack of transparency in the calculated weights and produced outputs can be a major drawback. Despite this, DDMs represent a promising frontier in hydrological modelling, offering new insights and predictive capabilities in the face of evolving data landscapes and changing climate conditions [KKH⁺19].

Deep Learning (DL) techniques can effectively represent complex physical processes and discover hidden, long-lasting relationships in the data and output. These models are highly adaptable, computationally efficient, and can be easily calibrated. The problem of exploding/vanishing gradients in Recurrent Neural Network (RNN) is usually mitigated by using LSTM models. DDMs require large amounts of training data, and the training process is very resource intensive to obtain reliable predictions. Due to the size of these models, they are considered black-box models. Also, DDMs tend to struggle to generalise outside their calibration range and their lack of physics-informed components in their architecture can lead to implausible outputs [KKH⁺19, LHB⁺23, KJK⁺22, OEAF21].

There is a discrepancy in terminology between traditional hydrological modelling and modern, data-driven approaches. The optimisation of a set of model parameters across a pre-defined number of iteration steps in order to represent the whole period of data, and the evaluation of the performance based on objective metrics is referred to as the process of “calibration” in traditional hydrology. In ML, this process is called “training”. More specifically, the iterations are referred to as epochs in DL, where model architectures process data in subsets (i.e. batches). Otherwise the process of finding ideal model parameters or weights is similar for both approaches [KKB⁺18].

The problems posed in hydrological modelling are typically complex, non-linear and data-intensive. These characteristics are inherent to ML methods. Their data-driven, computationally efficient design allows them to excel at learning patterns directly from observational data, which leads to increased adaptability, flexibility, and robustness. ML methods are capable of producing accurate results in time series forecasting. Spatio-temporal dependencies of input variables can be efficiently processed and represented.

ML techniques have been applied to various use cases in hydrology in the literature. For example, El-Haddad et al. employ four different ML methods to delineate flood-prone zones in Egypt. Among boosted regression tree, multivariate discriminant analysis, a general linear model, and functional data analysis, the latter method exhibits superior performance [EHYP⁺21]. Rahmati et al. compare the three popular ML methods random forests, support-vector machines, and k-nearest neighbour (kNN) regression to predict the levels of nitrate in the groundwater, thereby analysing water quality. The authors find that the kNN and Random Forest (RF) models produce good results with respect to predictive power and uncertainty [RCF⁺19]. Furthermore, Kabir et al. present a case study with an evaluation of three ML-based methods for multi-step ahead streamflow forecasting at an hourly time scale for several river systems in the United Kingdom. The results indicate that a support-vector regression model is accurate up to two hours but experiences a gradual decline in performance beyond that, and that a wavelet-ANN model is characterised by higher system non-linearity [KPP20].

Furthermore, Zounemat-Kermani et al. provide a comprehensive meta-analysis of the employment of ensemble learning methods to hydrological problems, such as rainfall-runoff, flooding, or water quality. The authors find that novel boosting techniques, such as AdaBoost and XGBoost, have been applied extensively and successfully in recent years and have shown promising results. Significant improvements of efficiency and accuracy are reported in studies employing such models. Furthermore, bootstrap and bagging techniques, such as an RF model, are popular approaches in rainfall-runoff modelling [ZKBFH21].

Mosaffa et al. cite the high variability in spatial and temporal scales inherent to hydrological modelling and issues associated with over- and underfitting arising from the lack of quality and volume of the training data, uncertainty, and missing records as obstacles in the application of ML methods [MSM⁺22].

Excursion: Deep Learning for Time Series Analysis

In time series modelling, the input space is augmented by the time dimension. Consequently, the input sequence x takes the form $x = [x_1, x_2, \dots, x_T]$ for T time steps, with each x_t ($1 \leq t \leq T$) comprising D input variables following the description of the feature space. As a result, the training data space (or input space) shall be $N \times T \times D$.

One of the primary drawbacks of RNNs is their limited capacity to remember sequences beyond ten time steps. In hydrological modelling, many processes are subject to major time lags between causes and results. For instance, accurately representing the relationship between precipitation, storage processes in groundwater, snow, or glaciers and subsequent discharge may take months or even years within a catchment's memory. This issue is especially relevant to hydrological modelling, since most datasets provide the time series input at a resolution of the daily average at a minimum. Consequently, an RNN can only utilise the previous ten days of information to forecast the runoff for the subsequent day, presenting a significant constraint [BSF94, KKB⁺18].

The LSTM model is a special type of RNN introduced by Hochreiter and Schmidhuber in 1997 [HS97]. This network incorporates so-called *memory cells* that are capable of storing

information over long time intervals. Moreover, LSTMs are not subject to the issue of vanishing or exploding gradients as compared to other types of RNNs due to their more complex and carefully designed internal structure. Thus, LSTMs are an appropriate option for modelling dynamic systems, such as watersheds, due to the similarity of memory cells to state vectors in conventional dynamic system models.

The forward pass of an input sequence $x = [x_1, x_2, \dots, x_T]$ for T time steps is presented in Equation 2.5. Each vector x_t includes the input features at time t ($1 < t < T$), resulting in a coherent and systematic flow of information.

$$\begin{aligned}
 i[t] &= \sigma(W_i x[t] + U_i h[t-1] + b_i) \\
 f[t] &= \sigma(W_f x[t] + U_f h[t-1] + b_f) \\
 o[t] &= \sigma(W_o x[t] + U_o h[t-1] + b_o) \\
 g[t] &= \tanh(W_g x[t] + U_g h[t-1] + b_g) \\
 h[t] &= o[t] \odot \tanh c[t] \\
 c[t] &= f[t] \odot c[t-1] + i[t] \odot g[t]
 \end{aligned} \tag{2.5}$$

Here, the parameters W , U , and b are learned for each gate (denoted by the subscript). Initially, all cell states are set to zero vectors. $\sigma(\cdot)$ designate the sigmoid function, $\tanh(\cdot)$ the hyperbolic tangent function. In this model, the $c[t]$ cell states represent the over-arching memory of the architecture. These states can be modified by the respective gates: $f[t]$ deletes states, $i[t]$ and $g[t]$ update states and introduce new information to the system. Then, $o[t]$ determines which information stored in the cell states is emitted as output [KKS⁺19].

ANN-based architectures have experienced widespread recognition in the field of hydrology since the 1990s [Dan91]. DL models have been found to be highly accurate for the task of rainfall-runoff modelling. Previously, the majority of published studies have applied feed-forward neural network architectures to this task, but in recent years, models with memory-like components based on RNNs, such as LSTM models, have been able to perform well in the domain [KHK⁺19].

Wang et al. used a dilated causal convolutional neural network to predict water levels. Their results showed the increased performance of the model compared to a multilayer perceptron and SVM models [WLC⁺19]. Sun et al. test the performance of graph neural networks and report that they are robust and computationally efficient. The performance is at least similar compared to an LSTM baseline model [SJMC21]. Zou et al. experiment with a combination of auto-regressive RNNs and a novel, enhanced RNN called ResLSTM alongside other techniques in a multi-step-ahead flood probability prediction model [ZWLL23].

Kratzert et al. state that the LSTM architecture is especially suitable for this task as the evolution of states can be modelled explicitly through time and mapped to a given output. This allows for a direct comparison to the general definition of rainfall-runoff modelling given in Equations 2.3 and 2.4: the system states defined in the formulations can be compared to the memory cell states of the LSTM architecture. The parameters can be translated to the learnable network weights [KHK⁺19]. Notable examples of DDMs with

an LSTM-based architecture include an early LSTM model to simulate rainfall-runoff proposed by Hu et al. in 2018 [HWL⁺18]. Sahoo et al. use an LSTM for low-flow time series forecasting; Zhang et al. apply an LSTM as well as an SVR model for reservoir operation simulation and later show an improved attention-based LSTM for urban flood forecasting [SJSK19, ZLP⁺18, ZQM⁺23]. The latter model features a specific attention mechanism with double-time sliding windows and a weighted mean square error loss function to address the issues associated with high temporal variability in urban flood prediction. The authors apply the model at a very high temporal resolution (at the minute-scale) and report superior performance underlined by R^2 scores exceeding 0.85. Furthermore, Ouma et al. compare an LSTM with a wavelet-ANN for spatio-temporal prediction of rainfall-runoff time series in data-scarce basins [OCW22]. Anshuka et al. propose an LSTM-based model for spatio-temporal hydrological extreme forecasting [ACB⁺22]. Kim et al. present a case study comparing two PDMs to two DDMs (an ANN and a LSTM). They report the competitive performance of the data-driven approaches and emphasise the high accuracy. However, the authors find that DL models are only capable of producing accurate results if they are provided with a sufficiently high amount of data due to the lack of physical water routing information implemented as part of the models' architectures [KYG⁺21].

In the field of rainfall-runoff modelling utilising memory-based neural network models, Frederik Kratzert, Daniel Klotz, and Grey Nearing are among the leading researchers, frequently presenting cutting-edge model architectures that extend LSTMs to address various issues related to this specific modelling task. After the presentation of an initial LSTM to model rainfall-runoff in 2018, the authors also propose an entity-aware LSTM, which incorporates static information on catchments into the architecture. Kratzert et al. further present several case studies in the field of hydrological modelling with DL [KKB⁺18, KKHH18, KKH⁺19, KKS⁺19].

Furthermore, the authors are involved in the development of a mass-conserving LSTM architecture by Hoedt et al. in 2019. The model architecture integrates physical conservation laws, which govern the re-distribution of quantities in a system. The authors are able to show that their novel model competes with an ensemble of LSTMs and outperforms several popular PDMs, such as HBV and VIC. Interestingly, the authors show that the mass-conserving LSTM is capable of learning to track snow in memory cells without requiring snow data as input for training [HKK⁺21]. Beyond that, Kratzert et al. are responsible for the publication and maintenance of the `NeuralHydrology` open-source Python package for DL-based hydrological modelling, as well as the state-of-the-art `Caravan` large-sample data collection [KHK⁺19, KGNK22, KNA⁺23]. These publications are significant to the development of an inter-disciplinary research effort that integrates the capabilities of DL into the domain of hydrology.

2.3.3 Hybrid Hydrology Models

Hybrid models combine physical and data-driven models into a single end-to-end simulation pipeline [Raz21]. Due to reduced computational cost, they can be trained more

efficiently to explore both long and short-term forecasts on localised scales [GRA⁺22]. By leveraging ML practises and physics-driven processes, the resulting model can reduce uncertainty and bias while at the same time improving explainability and consistency [KJK⁺22]. Lian et al. classify hybrid modelling into (a) surrogate modelling, (b) one-way coupling, and (c) modular coupling [LHB⁺23]. Okkan et al. introduce a fourth type, (d) nested hybridisation [OEAF21].

Noori et al. use a coupled hybrid model using a process-based watershed model and an ANN for water quality prediction. The resulting model achieves optimised calibration and validation processes [NKI20]. Lian et al. apply the modular coupling approach in that they use an RF as a sub-model to represent the evapotranspiration simulation for streamflow estimation in the PDMs XAJ and SWAT, respectively, and achieve improved accuracy [LHB⁺23]. Kraft et al. present a hybrid framework based on a dynamic Neural Network simulating time-varying coefficients that are fed to a simple hydrograph. The model is capable of simulating the dynamics of snow, soil moisture, and fluxes and storage in groundwater [KJK⁺22]. Mohammadi et al. present an extensive evaluation of two PDMs and seven H2Ms that are based on Adaptive Neuro-Fuzzy Inference Systems (ANFIS) and Support Vector Machines (SVM), and conclude that AI-based hybrid models generally lead to more accurate streamflow estimates [MMCD21].

Ren et al. propose a hybrid model integrating a Bayesian Neural Network and an SVM with the process-based model HBV to improve streamflow prediction in alpine regions. In this setup, HBV provides the snow/glacier-melt runoffs as input to the ML techniques [RYH⁺18]. Gharbia et al. introduce a hybrid geophysical tool called GEO-CWB consisting of GIS-based algorithms that parameterise the result of catchment-dynamic water balance for climate and land-use changes. The underlying models are based on physics, statistics and Machine Learning and allow for localised point forecasts in the long and short term [GGJP20]. In 2022, the authors follow up with an array of proposed surrogate models for their tool to simulate water flow and level simulations. They experiment with (wavelet)-ANNs and -SVMs and achieve reduced computational cost [GRA⁺22].

Dong et al. propose a hybrid framework model where a high-resolution PDM is coupled with XGBoost and an ANN as well as a calibration-free conceptual scheme for data-scarce reservoirs. The gradient boosting technique outperforms the ANN, and the hybrid framework shows improved performance in reconstructing daily streamflow of the investigated basin [DGC⁺23].

The nested hybrid modelling approach is introduced by Okkan et al. for rainfall-runoff simulation. Outputs from conceptual PDMs are directed to ANN and SVMs. The embedded ML part is inactive until the rainfall-runoff calibration in the conceptual model is completed. The nested model outperforms standalone models as well as coupled model variants in mean and high flows [OEAF21].

2.4 Hydrological Modelling in the Era of Climate Change

In recent years, awareness of non-stationarity in climate and hydrology has continually increased. The interest of the research community now gradually shifts towards the assessment of model performance in climatic conditions that extend beyond their training data and calibration range. The ongoing drastic changes in climate pose a dilemma for the reliability of land surface models in simulating future climate scenarios, regardless of their satisfactory performance under known conditions [ODO20]. Blöschl et al. emphasise the significance of studying the effects of environmental modifications and the associated climate change in their compilation of 23 open problems in hydrology that were identified by the research community in 2018. The importance of evaluating hydrological models under contrasting conditions compared to their calibration reference is explicitly mentioned by the authors [BBC⁺19].

Modelling experiments have now turned their focus on evaluating the ability of models to extrapolate under non-stationary conditions. The most common method used in simulations to assess this capacity in models is Differential Split-Sample Testing (DSST). This method employs a reference period, such as wet and cold seasons, during calibration. The models are then assessed using a period of different climatic conditions, such as dry and warm seasons, to evaluate the robustness of their predictions [ODO20]. The ultimate goal of a model undergoing DSST is to exhibit robust transferability of calibrated parameters to validation data, whilst displaying minimal sensitivity to environmental and climatic conditions [NdMSD22].

Previous research has largely neglected the context of changing conditions in the input data when analysing performance, accuracy, flexibility and robustness of state-of-the-art models. While complex models may accurately depict hydrological processes, they are prone to being over-parameterised, despite applying stricter physical limitations and incorporating considerably more data regarding the system in question when compared to physically-based conceptual models [Bev89, ODO20]. However, initial studies that employ DSST have found no significant difference between complex and simple models with regard to the number of parameters used [CAP⁺12].

Sungmin O et al. examine how varying hydro-climatic conditions impact model performance with consideration of model complexity. The authors compare two PDMs, a highly complex physically-based model (HTESSEL) that considers land surface and a conceptual model (SWBM) of medium complexity focused on hydrological processes, respectively, and a rather simple black-box DDM, in this case an LSTM, which generates a runoff time series. All models are calibrated using only streamflow data from 161 catchments in 11 European countries from 1984 to 2007. Catchments are classified into humid, moderate, and arid hydro-climatic regimes according to their aridity index. According to the authors, the aridity index, commonly known as the ratio of atmospheric water supply to its demand, is defined as “ratio between mean net radiation and respective unit-scaled precipitation during the entire 24 years” [ODO20].

The simulation setup employed in this study starts with a selection of catchments based on

model suitability and data quality over the entire period. The PDMs are then calibrated for each catchment individually and the LSTM uses all catchments simultaneously, while both approaches only use reference periods of the wettest and driest years according to annual mean precipitation at this stage, respectively. Then, the models are assessed according to their performance in the remaining 23 years to illustrate the difference under different climatic conditions. The authors coin the increasingly drier years *wet2dry* and the increasingly wetter years *dry2wet* [ODO20].

Precipitation is used as the driving variable to define the climatic reference periods as it has been shown to be most influential compared to temperature or potential evapotranspiration [CAP⁺12]. The single-year calibration run is repeated iteratively until the model reaches an equilibrium. The authors highlight that process-based models derive an advantage from their foundation in the laws of physics to model the various conditions, whereas data-driven models acquire this knowledge solely from the input data. However, for the LSTM the differential split-sampling approach needs to be configured in such a way that more training data is available at the point of calibration to compensate for the lack of physical knowledge. The authors conduct extra experiments using an additional LSTM, trained using a randomly selected year as a reference period. This approach enables a greater diversity of hydro-climatic information to be represented, allowing the relationships between the hydrological variables to be learned from all observations [ODO20]. Coron et al. present a 10-year sliding window approach in their study to examine the effect of contrasting climate conditions on conceptual rainfall-runoff models. They find that calibration using especially dry or wet period might lead to overestimation of simulated runoff [CAP⁺12].

The evaluation is performed in the same way as the calibration, but using the remaining years as input data. Transient climatic conditions are represented in the evaluation data, enabling the assessment of robustness under conditions that change progressively. This setup is a promising basis for further experiments. However, Ji et al. acknowledge that experiments utilising DSST should include a validation period separate from calibration and evaluation (known as training and testing, respectively, in the context of DL) to gain a comprehensive understanding of the model's performance under diverse climatic conditions [JML⁺23].

The authors find that the robustness of model performance in changing conditions improves gradually as the models incorporate more physical principles as constraints. The decline in model performance with increasing differences from reference conditions is attributed to temporal shifts between hydro-climatic patterns that cannot be effectively characterised by static model parameters. The challenge can be addressed by the authors through the inclusion of training data from several catchments simultaneously, which does not require an escalation in training data. It is important to note that there is a fundamental difference between process-based models and data-driven models. PDMs are constructed using explicit knowledge about the climatic conditions and hydrological behaviours, while DDMs are black-box systems by their very nature and learn all relationships from the input data, potentially lacking vital a-priori information.

Nonetheless, these issues can be addressed in DL techniques. The challenge can be conquered by adopting regional, rather than local, calibration (refer to section 4.4.1), and by including not only strictly constrained reference periods, but also integrating all relevant conditions into the model [ODO20]. These findings are confirmed by de Moura et al. in a similar experimental setup [NdMSD22].

O et al. present a vital study focused on the impact of climate change on hydrological modelling. Hydrological models are crucial for assessing climate impacts on systems, but they can yield significant predictive errors in the face of changing climates. This issue is particularly pronounced in regions expected to undergo substantial hydroclimatic changes. The comparison of PDMs of varying complexity with state-of-the-art approaches from DL and the use of DSST provides an important overview of the state of research in this regard. However, the study also leaves room for improvement and further experimentation to address pivotal issues. For instance, the period of training data stops in 2007 and thus may not represent the drastic climatic conditions experienced in the recent past. The calibration period of only a single year should be extended in further experiments. Furthermore, the catchment and forcing data contains only few variables in comparison to already discussed state-of-the-art datasets in LSH. The here employed DDM uses a very simplistic architecture and the quality and the insights gained from the output of experiments could be much improved by using an augmented LSTM model. For instance, a physics-informed setup could compensate for the missing knowledge about physical laws in the experiments by O et al. Additionally, the inclusion of static catchment attributes can greatly enhance the available information. Sophisticated hyperparameter tuning approaches could also lead to improved transferability. It is crucial to also observe the complexity of models in addition to their mere performance [ODO20, NdMSD22].

Large-Sample Hydrology and Data Overview

3.1 Large-Sample Hydrology

Hydro-climatic variables such as precipitation, temperature, humidity or wind as well as geo-ecological properties such as soil, land use and land cover, anthropogenic impacts or vegetation are the driving factors for hydrological processes such as streamflow generation. A key issue in hydrology is the collection of information about large amounts of catchments in variables that are capable of describing the complex and heterogeneous processes in hydrology and allow for accurate simulations in various spatio-temporal resolutions and subject to varying climatic conditions. Since data from individual catchments or river gauges are not able to explain the diversity of general hydrological behaviour, the available data must be combined to obtain large sample sizes that promote consistency and generalisability. Large-Sample Hydrology (LSH) is a sub-discipline of comparative hydrology and aims to address these challenges by establishing large-scale datasets that follow consistent formats and include a vast amount of samples from a diverse set of hydrological conditions in order to be able to simulate behaviours for catchments not included in these sets. Furthermore, modelling strategies are promoted that attain reliability, robustness, and realism, generalise well and are transposable, and where parameter estimation from data is facilitated [ADAG⁺20, GPB⁺14, KNA⁺23].

Gupta et al. define the ultimate goal of hydrological sciences “to achieve a degree of process understanding that enables construction of models that are capable of providing detailed and physically realistic simulations across a variety of different hydrologic environments, and at multiple spatial and temporal scales” [GPB⁺14].

Datasets in hydrology typically consist of time series data in key hydrological variables: streamflow, precipitation, temperature, potential evapotranspiration, and snow water

equivalent. Additionally, catchment attributes are given to provide further context of the covered basins and allow for more in-depth analyses and powerful models. Basic catchment identifiers such as the name of the gauging station, the coordinates and country, the catchment area and quality identifiers of gauging stations are also commonly present. Differences commonly exist among datasets in terms of the spatio-temporal resolution provided, the extent of catchments covered, the specific attributes of catchments offered, and the overall availability. Hydrological datasets are primarily distinguished by the inclusion of specific regions or climate zones. While only a few datasets or collections strive to offer global coverage, the majority are focused on continental, national, regional, or local scales, providing descriptive information at those respective climatically or geographically specific levels.

Addor et al. reviewed the progress of large-sample hydrology in a comprehensive study in 2019. The authors identified key limitations in the field: (i) the lack of common standards hindering basin comparability between datasets, (ii) the lack of metadata and uncertainty estimates impeding the assessment of data reliability, (iii) negligence in describing the extent of human impacts, (iv) infrequent adherence to the findable, accessible, interoperable and reusable (FAIR) principle [WDA⁺16]. Furthermore, the authors outlined guidelines and requirements for the generation of future LSH datasets: (i) providing basic data for each basin, (ii) using consistent naming for variables, (iii) relying on publicly available code for data processing, (iv) publishing uncertainty estimates for time series and catchment attributes, (v) incorporating anthropogenic descriptors, and adhering to the FAIR principle. According to this work, the central challenges going forward are to progressively move datasets to the cloud so that the increasing computational load can be mitigated and comparability of datasets is promoted. The authors also emphasise the importance of comprehensively describing anthropogenic factors to elucidate human impact on water systems [ADAG⁺20].

A major challenge in LSH is to generate publicly available datasets, since many important studies in hydrology are based on data that is not open to public inspection. Examples are the study by Blöschl et al. who investigated the impact of climate change on floods in European rivers where approximately only one third of the data is publicly available, and the regionalisation and calibration experiments performed by Beck et al. [BHV⁺19, BvDdR⁺16, BPL⁺20].

Excursion: Meteorological Forcings and Climate Reanalysis

Besides generating real-time data of the Earth's climate, it is also vital to produce accurate, consistent, holistic, long-term records of past climatic conditions. *Climate reanalysis* refers to the process of assimilating climate-related information from multiple sources, including satellite imagery, radar, buoys, historical weather observations, and topographic data into a comprehensive climate model using the laws of physics. The aim is to reconstruct the Earth's past atmospheric and land conditions spanning several decades so that historical climate conditions can be represented coherently.

Hydrological, ecosystem and land cover/land use models use gridded near-surface me-

teorological data, commonly referred to as *meteorological forcings*, as inputs for their simulations. The data typically contains information on the factors that drive atmospheric, surface, and water-specific conditions. Meteorological forcings can be considered inputs to a comprehensive climate reanalysis model and typically represent the time series data used for hydrological modelling.

In the field of hydrological modelling, several datasets are widely recognised and frequently used as input to models. Most notably, Global Land Data Assimilation System (GLDAS) is a major NASA reanalysis project that combines satellite and land-based observations into a comprehensive record on a global scale. The system incorporates a vast amount of observation-based data at high resolutions, ranging from 2.5° to 1 km. Furthermore, it is capable of delivering near-real-time results. The data provided by GLDAS are used for various current and planned applications for climate and weather forecasting as well as for the simulation of water resources, quality and cycles. Data on elevation, soil, vegetation, precipitation and radiation are compiled by high-quality data assimilation systems to optimally drive the models [RHJ⁺04].

The 5th Generation of European ReAnalysis (ERA5) is a popular global reanalysis project providing data from 1940 onward at an hourly temporal resolution and includes uncertainty estimates as well as daily and monthly aggregates. It is developed by the Copernicus Climate Change Service (C3S) at the European Centre for Medium-Range Weather Forecasts (ECMWF). The atmospheric, land, and ocean climate variables provided by ERA5 encompass the Earth on a 30-kilometre Gaussian grid. 137 levels are employed to resolve the atmosphere from the surface up to 80 kilometres in altitude. The spatial resolution for the reanalysis ranges from 0.25° to 0.5° while the uncertainty estimates are given at reduced spatial and temporal resolutions (0.5° to 1°) [MnSDAP⁺21].

The NCEP/NCAR reanalysis set is widely used and provides extensive near-surface meteorological information at a global scale. This dataset is freely available, continuously updated and developed by the National Centers for Environmental Prediction (NCEP) in collaboration with the National Center for Atmospheric Research (NCAR). The data is provided from 1948 onwards in six-hour intervals and consists of atmospheric quantities and parameters computed by numerical weather prediction. The global grid has a resolution of 2.5° for both latitude and longitude and uses 17 pressure levels for the atmosphere, which is coarser than e.g. ERA5 and therefore causes difficulties in regions with few observations. However, the global coverage and extensive historical records make it a valuable resource [KKK⁺96].

Based on NCEP/NCAR, Sheffield et al. produced the Princeton Global Forcing (PGF) dataset from 1948 to the present at 3-hour intervals at a resolution of 1° across the globe. This dataset is specifically designed for land surface hydrology models and incorporates several heterogeneous observation-based sets with the NCEP/NCAR reanalysis to refine the resolution and provide comprehensive information for land surface fluxes and conditions. Corrective measures are applied to account for biases in the reanalysis precipitation data using the observation-based data. PGF represents a globally consistent, long-term dataset of near-surface meteorological variables and is used widely across hydro-ecological models [SGW06].

3.1.1 MOPEX

The first LSH dataset resulted from the Model Parameter Estimation Experiment (MOPEX) and consists of data from 438 catchments in the contiguous United States [SCD06, KNA⁺23]. The goal of MOPEX was to develop approaches for the a-priori estimation of parameters that are applied in land surface parameterisation procedures in atmospheric and hydrological models. Various hydro-meteorological observations as well as attributes for catchments that represent varying hydro-climatic conditions are incorporated. The data are available at hourly and daily resolution from 1948 to 2003. Time series attributes encompass precipitation, temperature, runoff, and potential evaporation (P, T, Q and PE). In addition, the data set includes catchment attributes that describe the topography, land cover and soil, as well as climate indices. The meteorological data were derived from more than 16.000 weather stations across the country. The dataset is available for free from the website of the NOAA¹. This dataset is particularly relevant to the problem of PUB, but has lost some of its relevance as it has not been updated since 2003. It is still used as an important reference dataset in hydrological modelling [ADAG⁺20].

The related dataset Canadian Model Parameter Estimation Experiment (CANOPEX) features 698 catchments and focuses specifically on Canada. The format of the data is aligned with the parent project MOPEX, but it does not provide any catchment attributes. The data is also available for free for non-commercial applications². MOPEX and CANOPEX represent two of the most important datasets at the national scale [ABODB16, ADAG⁺20].

3.1.2 CAMELS

The Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) collection emphasises the need for consistency in data by applying the same preparation techniques and publishing the data in a consistent format. Several datasets have been published in this collection for specific countries, using the same data preparation and formatting techniques. These include data for the contiguous United States, Brazil, Chile, the United Kingdom, Australia, France, contiguous China and Switzerland, typically covering a few hundred catchments each. The datasets are usually suffixed with the country code. These datasets combine hydro-meteorological time series and static catchment attributes aggregated to polygons [ANMC17]. Additional datasets that comply with or build on the CAMELS standard have been proposed for specific catchment data relevant for spatially distributed hydrological modelling and information transfer to data-sparse regions, or containing dedicated atmospheric and stream water chemistry data [KC22, MFL⁺21, SPL⁺22]. According to Addor et al.'s assessment, the datasets contain information on streamflow, precipitation, temperature, and potential evapotranspiration making them one of the most comprehensive collections in LSH [ADAG⁺20].

¹https://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/

²<http://canopex.etsmtl.net/>

As most other prevalent datasets in hydrology, CAMELS covers the full upstream area of individual catchments, but does not describe interconnected river networks [KSH21].

The original dataset in this collection covers the contiguous United States in 671 catchments from 1980 to 2014 [NCS⁺15]. It is publicly available on the Geoscience Data Exchange platform³. Both time series data and catchment attributes are part of the collection, providing a comprehensive representation of the hydrological cycle. A major advantage of CAMELS is that it provides a detailed explanation of the approaches employed to obtain catchment attributes and a discussion of some of the caveats associated with the data sources. Furthermore, the code used to generate most attributes of the CAMELS collection are publicly available as open-source scripts⁴. This facilitates the reproducibility and comparability of data sets and improves the transparency of hydrological experiments.

3.1.3 GRDB and EWA

The Global Runoff Data Base (GRDB) and the European Water Archive (EWA) are both freely available large-scale data collections from the associated Global Runoff Data Centre (GRDC). GRDB is arguably the most common dataset used for streamflow research at a global scale⁵. The GRDB has been operational since 1988 and comprises information from more than 100 hydrological and meteorological services worldwide. The streamflow data can be acquired in daily or monthly aggregates and can be queried by station, country, time-period, region or sub-region, which refers to a hydrographical region representing all or parts of a river basin, or land that drains in total via coastal sections between defined sub-regions. For example, the Danube sub-region contains 186 stations in the respective area. In total, the collection contains 841 sub-regions and 10.702 stations from 159 countries across the globe. Naturally, some countries and regions are underrepresented in this collection. The period of available data differs per station, but the earliest data is from 1931 [ADAG⁺20].

The EWA can be considered the counterpart to GRDB at a continental scale with 3.700 river gauging stations in 29 countries across Europe. The GRDC now also hosts this collection, and some national hydrology services have opted to integrate their data from EWA into GRDB. However, EWA has been discontinued since 2014 and no new data can be expected to be added to this collection [ADAG⁺20].

3.1.4 GSIM

Global Streamflow Indices and Metadata Archive (GSIM) is an extension of the GRDB to accommodate access to rural station data and provides global monthly, seasonal, and annual streamflow indices for more than 35,000 catchments. The data can be acquired in

³<https://gdex.ucar.edu/dataset/camels.html>

⁴<https://github.com/naddor/camels>

⁵<https://www.bafg.de/GRDC>

daily, monthly, seasonal, and yearly resolution⁶ [DGLW18]. The collection incorporates 12 streamflow databases (seven national, five international) to compile daily streamflow time series. GSIM consists of three parts: (i) a catalogue with fundamental metadata for each time series, (ii) catchment boundaries outlining the area that contributes to each gauge, and (iii) catchment metadata from the 12 gridded databases. These three metadata products are highly relevant in LSH and hydrological modelling. The extensive metadata records, streamflow indices and catchment boundaries are used as input to other important datasets such as HYSETS [ABM⁺22].

3.1.5 E-HYPE

European Hydrological Predictions for the Environment (E-HYPE) is based on the semi-distributed, process-based hydrological model “HYPE” proposed by Lindström et al. in 2010 and incorporates data from open and freely available sources [LPR⁺10]. For instance, EWA has been integrated into E-HYPE. This model is calibrated on European data; when applied only for Europe it is called E-HYPE. Version 2.1 of this dataset includes information on more than 35.000 sub-basins across Europe with a median size of 214 km². The dataset includes 18 variables concerning meteorological, snow, soil, hydrological and nutrient concentration data. At the moment, the daily historical data from 1989 to the present day can be acquired for 4.000€ from the Swedish Meteorological and Hydrological Institute⁷ [DAA16, KAHW17]. This is a rare example of an LSH dataset where the acquisition or usage of the data is associated with cost.

3.1.6 HYSETS

The collection Hydrometeorological Sandbox - École de technologie supérieure (HYSETS) covers 14.425 watersheds in North America (Canada, Mexico, contiguous United States) in the period from 1950 to 2018. This dataset is one of the most comprehensive collections of hydrologically and meteorologically important data on a continental scale. Daily precipitation aggregates are compiled from seven data sources, discharge time series from one source per country, snow water equivalent is taken from ERA5 and SNODAS, and catchment characteristics are provided from an additional source with information on watershed area, elevation slope, land use, soil properties. The authors plan to update the database with the emergence of new datasets. The code to generate attributes is available upon request to the authors. A major advantage of this collection is the incorporation of a large set of catchment attributes. HYSETS is provided by the Open Science Framework repository⁸ [ABM⁺22].

⁶<https://doi.pangaea.de/10.1594/PANGAEA.887477>

⁷<https://hypeweb.smhi.se/water-services/data-delivery-services/standard-historical-e-hype/>

⁸<https://osf.io/rpc3w/>

3.1.7 LamaH

Large-Sample Data for Hydrology and Environmental Sciences for Central Europe (LamaH)⁹ consists of 859 gauged catchments in an area of 170.000 km^2 in nine countries in Central Europe. It applies a very similar structure to the data compared to the CAMELS collection. More than 60 attributes were compiled in this dataset that describe catchments in detail with information on (river) topography, climate indices, land cover. LamaH is one of the first datasets specifically designed for the purpose of LSH. The data is available from 1981 to 2017 and the time series data is forced using the ERA5-Land forcings. The authors explicitly refer to the suggestions brought up by Addor in their landmark LSH meta-analysis paper from 2019 [ADAG⁺20, KSH21]. LamaH is freely available in a versioned state with a permissive license¹⁰.

A novelty of the LamaH dataset is the inclusion of basin delineations describing inter-catchment areas of neighbouring gauging stations. This additional topographic classification allows for the simulation of local runoff generation in a river network. Furthermore, the temporal resolution of daily and hourly hydrometeorological time series data is a distinctive feature of LamaH. Few other datasets provide data at hourly resolution, although data of this granularity is critical to the reliability of results from simulations of processes that involve changing patterns over the course of a day [KSH21].

LamaH was incorporated into the Caravan collection to meet the standards of LSH in that the data should be consistent with and its continental, region-specific data should augment other state-of-the-art datasets. Central Europe is a climatically volatile region where the effects of climate change are clearly perceptible. For this reason as well as the good coverage of high-quality gauges, the highly granular temporal resolution, the high amount of catchment characteristics and the novel basin delineation make LamaH a fitting candidate for rainfall-runoff experiments in this work. Chapter 3 includes a thorough description of the study area, the variables at hand and the further course of data handling.

3.1.8 Caravan

Caravan is the first major global data collection specifically developed for LSH and was introduced by Kratzert et al. in 2023. It is a collection of consistent, region-specific datasets that conform to a standardised format, are publicly available in a versioned state¹¹ with a permissive licence, and can be accessed and extended by the community. Caravan currently combines HYSETS, CAMELS, and LamaH and contains observations from 6830 basins in 14 countries on four continents. The described catchments cover almost all of the 18 climate zones represented in Global Environmental Stratification (GEnS) (arctic, extremely cold and arid regions are not yet available at the time of this work) [MBJ⁺13]. The collection provides daily data spanning four decades from 1981 to

⁹Please note that the additional suffix “CE” for “Central Europe” is omitted in this work.

¹⁰<https://zenodo.org/record/5153305>

¹¹<https://zenodo.org/record/7944025>

2020. Basins between 93 and 2000 km^2 were selected. Meteorological forcing data was applied from ERA5-Land [MnSDAP⁺21]. The authors accompany the data collection with an extensive open-source code repository that facilitates the addition of further LSH datasets, contains the scripts used to derive all the current data and serves as a community hub for issues, extensions and contributions¹².

Seven region-specific datasets are combined in Caravan and provided as separate but combinable sets in the following distribution of basins:

- 4621 basins from HYSETS
- 482 basins from CAMELS-US
- 479 basins from LamaH-CE
- 408 basins from CAMELS-GB
- 376 basins from CAMELS-BR
- 314 basins from CAMELS-CL
- 150 basins from CAMELS-AUS

Figure 3.1 shows the distribution of catchments provided in Caravan across the globe as well as their distribution over the GEnS climate zones. It is obvious that the majority of catchments are taken from the HYSETS dataset which covers North America. Most catchments lie in cold and mesic, cool and moist, warm and mesic, cold and dry, and hot and dry areas.

Requirements for dataset selection were (i) the inclusion of catchment boundaries for each streamflow gauge, and (ii) a permissive license to allow redistribution. In Section 3.2, the study area and the variables of LamaH as the candidate LSH dataset for experiments in this work is described. Here, the version included in *Caravan* will be used to comply with the aim for consistency across hydrological experiments.

The authors account for the vision for LSH proposed by Addor et al. in 2019 by providing data that is standardised at the global scale, publicly available with an open license and extensible by open-source software as well as ready for the cloud [ADAG⁺20, KNA⁺23].

3.1.9 FutureStreams

FutureStreams is a representative of a different approach to datasets in LSH. The authors present a projection of future streamflow and water temperature estimates for varying climatic conditions up to the year 2099 including past data going back to 1976. Four different greenhouse gas emission scenarios are included for climate comparison. The

¹²<https://github.com/kratzert/Caravan/>

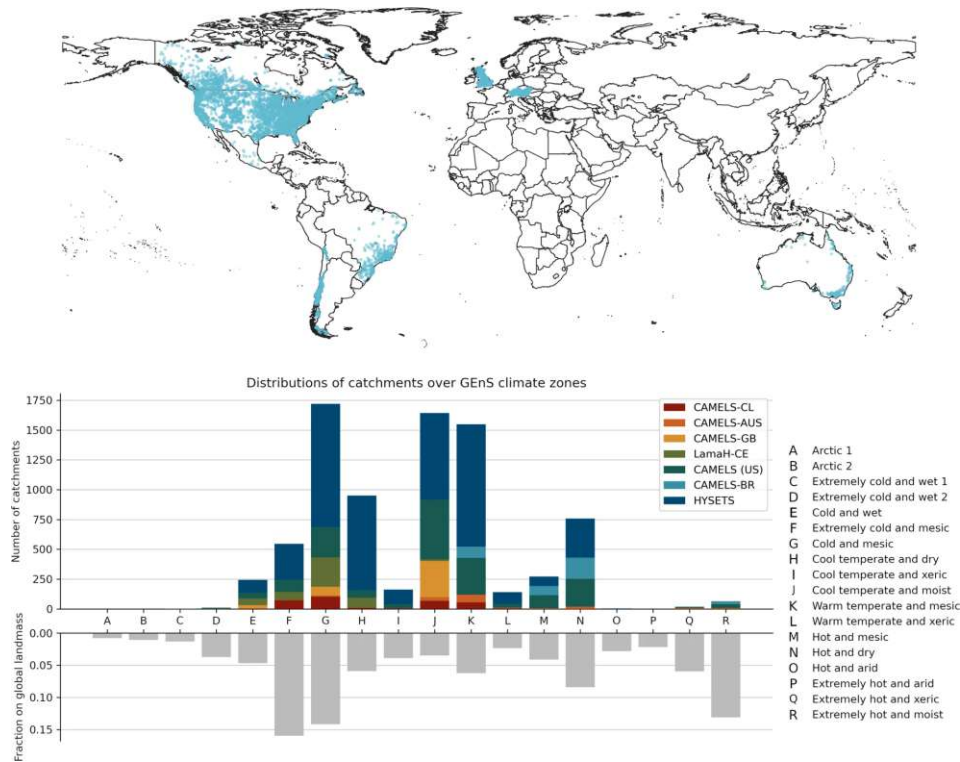


Figure 3.1: Top: Global distribution of catchments included in Caravan. Bottom: Distribution of catchments among the GEnS climate zones (the bottom part shows the fraction of a particular climate zone on the total land mass) [KNA⁺23].

streamflow and water temperature data is provided in a weekly aggregation. The authors use a spatial resolution of 5 arc minutes to generate data on a global scale. Additionally, ecologically and bioclimatically relevant indicators are given that are derived from streamflow and water temperature records. This dataset was developed specifically as a tool to simulate threats to freshwater systems arising from climate change, focusing on water temperature and biodiversity at a larger scale. The hydrological model PCR-GLOBWB coupled to the dynamical water temperature model DynWat were used to generate the future projections in this dataset. The authors provide a repository containing the code to generate all derived variables, and the dataset is freely retrievable from Utrecht University¹³ [BWB⁺22].

3.1.10 Other Notable Datasets

FLUXNET2015 is a dataset at the ecosystem level specifically concerned with the exchange of CO_2 , water and energy between the biosphere and the atmosphere. The data is provided for 212 individually operated site across the globe and processed in a

¹³<https://public.yoda.uu.nl/geo/UU01/T7TVTQ.html>

standardised way. The period of offered data varies per site with most sites available for more than 20 years up to 2014. FLUXNET2015 introduces supplementary data products, including gap-filled time series, estimations of ecosystem respiration and photosynthetic uptake, uncertainty assessments, and measurement metadata. This dataset has already been used for various applications such as ecophysiology and remote sensing studies as well as for the development of ecosystem and Earth system models. The data as well as the standardised pre-processing pipeline can be retrieved under the Creative Commons license¹⁴ [PTC⁺20].

3.1.11 Summary

Table 3.1 summarises the characteristics of the most common state-of-the-art datasets and data collections in LSH. The majority of these datasets is specifically concerned with streamflow records emphasising the importance of this information in hydrological modelling. The variability in the temporal and spatial resolution, the national, continental or global scale, the number of catchments covered, the variables provided and the attributes used to describe the catchments is high across all investigated datasets and collections. In summary, the key suggestions proposed by Addor et al. for Large-Sample Hydrology (LSH) have yet to be implemented at a large scale. The gradual integration of these proposals into existing, ubiquitously used, state-of-the-art LSH datasets is particularly important as a supplement to the development of new datasets such as *Caravan*, which already adhere to these standards. This should allow existing large-scale applications that depend on known data sets to seamlessly integrate the latest data standards and adopt requirements for availability, transparency, reproducibility and generalisability. *Caravan* currently represents the most recent and comprehensive data collection in the field. The large number of globally consistent catchment attributes, the standardised format of hydrometeorological time series with a highly granular temporal resolution provided by high quality monitoring stations in climatically volatile regions, and the approach of providing data in an extensible, available and open source manner makes it a cornerstone of LSH on which future datasets should be built.

The comprehensive state-of-the-art research on LSH in Section 3.1, and the general quality of datasets specifically developed and used for hydrological modelling with physically based models and DL leads to important conclusions concerning the appropriate dataset to use in the modelling experiments conducted in this work.

The dataset of choice is LamaH by Klingler et al. in the version contained in the *Caravan* collection [KSH21, KNA⁺23]. *Caravan* is used in version 1.2 released on 17 May 2023. The dataset is accessible on Zenodo in a citable form that is frequently updated and versioned¹⁵. Major advantages of the *Caravan* collection are that it is open-source, consistent and extensible by design. As a data engineer working with large-sample datasets in the domain of hydrology, these characteristics are crucial to create

¹⁴<https://fluxnet.org/data/fluxnet2015-dataset/>

¹⁵Source: <https://zenodo.org/records/7944025>

Dataset	Catchments	Location	Hydrological variables	Catchment attributes
MOPEX	438	United States	P, T, PE, Q	Topography, land cover, soil, climatic indices
CANOPEX CAMELS	698 few hundred each	Canada US, BRA, CHN, CHL, GBR, AUS, FRA, SUI	P, T, PE, Q P, T, PE, Q	/ Topography, climatic indices, hydrological signatures, land cover, soil, geology, water use
GRDB	10.702	Global	Q	/
EWA	3.731	Europe	Q	/
GSIM	35.002	Global	Q	Topography, land cover, geology, irrigation, human population, soil
E-HYPE	35.447	Europe	P, T, Q, SWE	Topography, hydrological signatures, land cover, geology, soil, climatic indices
HYSETS	14.425	North America	P, T, Q, SWE	Topography, land use, soil
LamaH	859	Central Europe	P, T, PE, Q	Topography, climatic indices, land cover, vegetation, soil, geology, anthropogenic impact
Caravan	3.830	Global	P, T, PE, Q	Topography, climatic indices, land cover, vegetation, soil, geology, anthropogenic impact
FutureStreams	5' global grid	Global (future projections)	T (Water), Q	/

Table 3.1: Summary of the predominant datasets in Large-Sample Hydrology (LSH) [ADAG⁺20].

interpretable, reproducible, and transferable model experiments and results. *Caravan* is a landmark data collection that accounts for these requirements and provides high-quality, state-of-the-art datasets in a consistent format, covering very diverse areas across the world over reasonably long periods of time. This allows for comprehensive and informative experiments, including climate change research in the field of hydrology.

The choice of LamaH is motivated by the fact that the effects of climate change are drastically evident in this region. According to the European Environment Agency, there is a stark increase in extreme weather events in Europe due to human-induced climate change. A decrease in summer rainfall and severe weather events, such as heavy precipitation, river floods, droughts and fire, are to be expected for Central Europe. At the same time, a reduction in snowfall is predicted. This makes Central Europe a very interesting area to study. The organisation further states that the availability of high-quality data is essential in the assessment of how climate change will affect Europe, conforming to the requirements of LSH [EEA23].

It should be noted that there are differences between the originally published LamaH dataset and the version used in *Caravan*. These differences are due to the constraints imposed by the *Caravan* collection on all its datasets to ensure consistency.

3.2 Description of the Study Area

LamaH features a total area of approximately 170,000 km² across nine countries in Central Europe: Austria, Germany, Czech Republic, Switzerland, Slovakia, Italy, Liechtenstein, Slovenia and Hungary. The main focus of the domain of coverage lies on the upper Danube area close to the Austrian-Slovakian border, its tributaries and various other catchment areas and adjacent upstream areas in Austria and its bordering countries. The Danube is the most prominent river and the catchments of its major tributaries divide the study area into 18 distinct river regions. These regions are depicted Figure 3.2, which also shows the runoff gauges along with their elevations. All of these catchment areas except for numbers 1 and 11 belong to the greater Danube catchment area. The water labelled as “Danube B” originates from regions outside the project area in Hungary or Croatia. The first river region includes the upper catchments of the Rhine up to Lake Constance. The eleventh region covers the Austrian catchment area of the Vltava (Moldau), which is the largest tributary of the Elbe [KSH21].

The largest river regions represented in LamaH are the Danube, Inn, Morava, Drava, Mur, Rhine, Salzach and Enns.

At present, the time period provided in the LamaH edition of *Caravan* spans from 2 January 1981 to 31 December 2020. The dataset offers information collected from 479 gauging stations from five countries: 307 from Austria, 120 from Germany, 35 from the Czech Republic, 16 from Switzerland and 1 from Liechtenstein. The difference to the 882 gauges in the originally published version of LamaH (refer to [KSH21]) is due to the limitation of *Caravan* to only include catchments between 100 and 2,000 km²

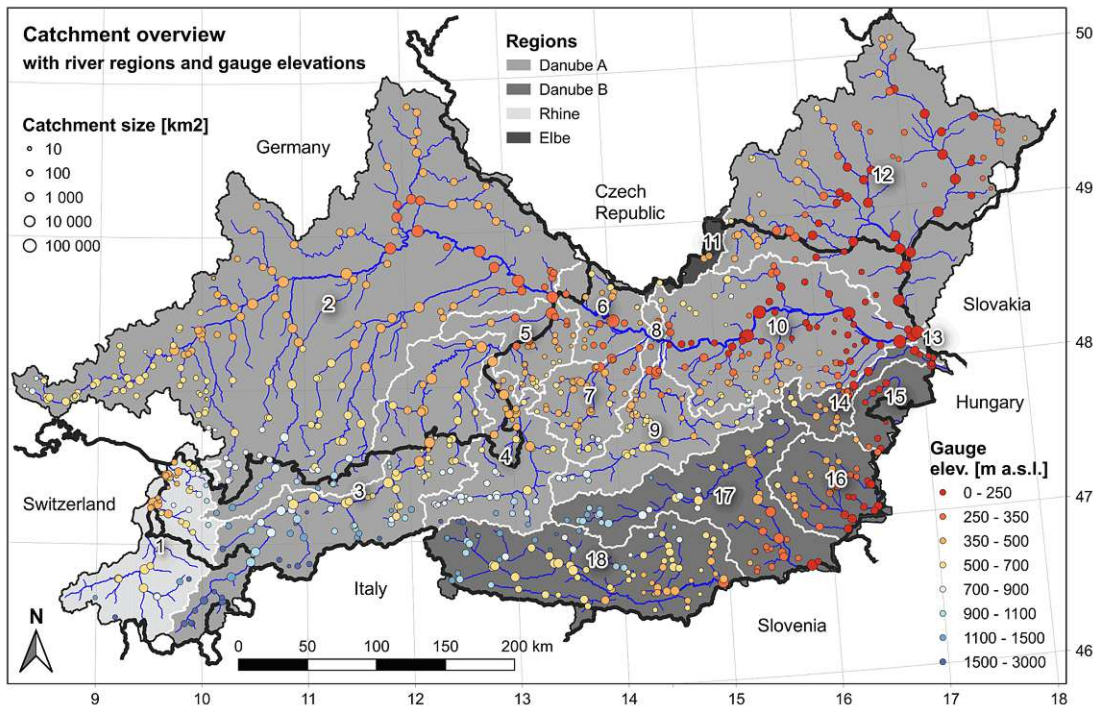


Figure 3.2: Overview of the domain of coverage of LamaH (in the original version). The discharge gauges are represented on the map as circles, with the size indicating the size of the catchment area and the colour indicating the elevation of the station. Numbers denote the 18 distinct river regions. Taken from [KSH21].

[KNA⁺23]. The mean catchment area is 452 km². Figure 3.3 illustrates the distribution of catchment areas by country, with the exception of Liechtenstein due to its sample size of one. Austrian catchments tend to be smaller in size and exhibit many outliers while those in the Czech Republic are larger and have a wide range of variability.

The average elevation is 975 metres above sea-level. The highest measuring point, situated at 2,605 metres, is the Ötztaler Ache, a tributary of the Inn in Tyrol, and the lowest, at 190 metres, is Rußbach in Lower Austria. This results in a maximum difference in altitude of 2,415 metres. The mean precipitation per day is around 3.4 mm. The highest recorded rainfall amount in a single day is 132.64 mm, which occurred in August of 2002 at the river Krems, a tributary of the Traun in Upper Austria, during a major flooding event that only occurs once every 50 to 100 years and affected considerable areas of Austria [FMoAM23]. The maximum amount of streamflow recorded was 112.1 mm at the stream Ostrach, a tributary of the Danube, in Baden-Württemberg, Germany, in May of 1999. This record also coincided with a major flooding event. On average, the gauging stations measure 1.78 mm of streamflow per day. The highest snow depth water equivalent (*SWE*) recorded was 2,62 metres also at Ötztaler Ache. The highest mean wind speeds recorded on a daily basis at a height of 10 meters are 8.92 m/s in an easterly

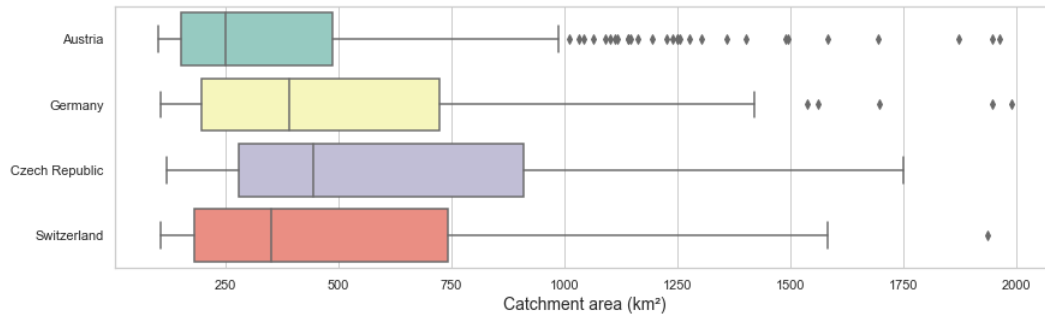


Figure 3.3: Distribution of catchment areas by country (excluding Liechtenstein).

direction, 6.48 m/s in a westerly direction, 7.76 m/s in a northerly direction, and 8.29 m/s in a southerly direction. The mean wind velocity, however, is negligibly low. The average temperature throughout the entire study region is 6.11 °C, ranging from -30.51 °C to +30.83 °C. The temperature variation of more than 60 °C indicates the extremely diverse climatic conditions prevailing in the study area.

3.3 Data Description

The open-source state-of-the-art LSH dataset *Caravan* includes both meteorological and hydrological time series attributes as well as a large number of static catchment attributes that allow for a holistic representation of the heterogeneous processes in hydrology [KNA⁺23]. Because of the diverse set of standardised variables for different climates available in this collection, it was chosen to conduct experiments in this work in the hope of providing initial results on which future research can be based. *Caravan* offers 14 distinct time series attributes, 56 distinct catchment attributes, and ten (also static) climate indices. The total number of aggregated static features is 206.

3.3.1 Time Series Attributes

Meteorological forcing data for the time series attributes are derived from the ERA5-Land reanalysis dataset [MnSDAP⁺21]. The authors of *Caravan* name global coverage, spatial consistency, sub-daily resolution, availability in the cloud, and the permissive license as reasons why this dataset was chosen. In addition to the meteorological forcing data, attributes for soil moisture and snow states are also taken from ERA5-Land. These hydrological reference model states are in themselves already modelled as they are originally taken from different reanalyses [MnSDAP⁺21, KNA⁺23]. It is important to note that streamflow data is given in mm per day and normalised by catchment area. In total, *Caravan* presents 14 distinct time series attributes (nine from meteorological forcings, five from model states). As all variables except precipitation and potential evaporation are offered aggregated by daily minimum, maximum, and mean, the total

Attribute	Description	Daily agg.	Unit
total_precipitation	Precipitation	sum	mm/day
potential_evaporation	Potential evaporation	sum	mm/day
temperature_2m	Air temperature	min/max/mean	°C
streamflow	Observed streamflow	min/max/mean	mm/day
M surface_net_solar_radiation	Shortwave radiation	min/max/mean	Wm^{-2}
surface_net_thermal_radiation	Surface net thermal radiation	min/max/mean	Wm^{-2}
surface_pressure	Surface pressure	min/max/mean	kPa
u_component_of_wind_10m	Eastward wind component	min/max/mean	m/s
v_component_of_wind_10m	Northward wind component	min/max/mean	m/s
snow_depth_water_equivalent	Snow water equivalent	min/max/mean	mm
volumetric_soil_water_layer_1	Soil water volume 0-7 cm	min/max/mean	m^3/m^3
H volumetric_soil_water_layer_2	Soil water volume 7-28 cm	min/max/mean	m^3/m^3
volumetric_soil_water_layer_3	Soil water volume 28-100 cm	min/max/mean	m^3/m^3
volumetric_soil_water_layer_4	Soil water volume 100-289 cm	min/max/mean	m^3/m^3

Table 3.2: Description of the time series attributes derived from ERA5-Land that are included in the Caravan dataset; largely taken from [KNA⁺23].

number of time series features is 39. Combining all 479 catchments for the available time period from 1981 to 2020 results in a total of 6,997,711 samples.

Table 3.2 shows the two groups of time series attributes included in the Caravan collection: meteorological forcings and hydrological model states. The groups are denoted as *M* for the meteorological forcings attributes and *H* for the hydrological reference model states. Each observation of the time series is given in the local time of the respective basin. The variables are computed as the area-weighted spatial average with a spatial resolution of around 9 km.

3.3.2 Static Catchments Attributes

The static, catchment-specific variables are primarily derived from HydroATLAS, a standardised database that contains descriptive hydro-ecological attributes for catchments around the world at high spatial resolution utilising polygons of sub-basins and river reach lines. The authors chose HydroATLAS due to its coverage of globally distributed catchments and permissive license. The variables are divided into six groups making up a total of 56 distinct attributes: hydrology (10), physiography (3), climate (9), land cover (16), soils and geology (8), and anthropogenic (8). Some of the attributes are given in several aggregations such as the monthly or annual mean to provide more information. The sharpest spatial resolution is used for the derived HydroATLAS attributes in Caravan (level 12 polygons). The combination of attributes and aggregation types results in a total of 196 variables [LLOD⁺19, KNA⁺23].

Additionally, the authors of Caravan provide further climate indices that were derived from the ERA5-Land time series data. These additional 10 distinct attributes contain information on precipitation including trends of high/low precipitation days, poten-

tial evaporation, aridity, moisture, and a seasonality index expressing changes in the water/energy budget [MnSDAP⁺21, KNA⁺23].

In combination, these 212 static attributes (including descriptive information and metadata) are ideally suited to be fed into a Deep Learning model as additional static input, as Kratzert et al. suggest [KKS⁺19]. Appendix A includes detailed information on the static catchment attributes from *Caravan* alongside their variable names, types of aggregation and units. Table A.2 summarises the 56 distinct static catchment attributes from HydroATLAS alongside the types of aggregation and unit. The six groups are denoted as *H* for hydrology, *P* for physiography, *LC* for land cover, *S&G* for soils and geology, and *A* for anthropogenic. Table A.1 shows the ten climate indices that are derived from the ERA5-Land time series.

3.3.3 Metadata

The metadata of the covered catchments include the latitude and longitude coordinates of the gauge as well as its name and the country. In addition, the catchment area is stated in km². Each basin is described by a unique identifier `gauge_id`, which is present for both the static and time series attributes to allow for consistent linking to the basin.

3.4 Data Analysis

The study area presented in the *Caravan* version of LamaH covers six of the 18 GEnS climate zones. Overall, the prevailing climate in the examined area of Central Europe is cold to varying degrees and varies between mesic, wet, moist and dry. Figure 3.4 shows the distribution of the GEnS zones as well as their area among the catchments in the domain. The vast majority of the area is made up by the *cold and mesic* zone with 59.6%. The second most common climate zone is *cool temperate and dry*, which spans large parts of the northern, north-eastern and eastern parts of the study area and covers 21.3%. The *extremely cold and wet* zone only applies to a single high-altitude catchment in Tyrol, Austria. Only two distinct terrestrial biomes are present in the examined area. The northern part of the area belongs to *temperate broadleaf and mixed forests* and the southern part to *temperate conifer forests*; both biomes account for about 50% each [DOJ⁺17].

The *Budyko* curve depicted in Figure 3.5 provides important insights into the climatic budget of a catchment [Bud74]. The scatter plot displays the aridity index, which is the potential evapotranspiration (*PE*) divided by precipitation (*P*), on the x-axis and the evaporative index, i.e. the ratio of actual evapotranspiration (*E_t*) to precipitation (*P*), on the y-axis. Furthermore, the catchments are categorised according to the mean elevation above sea level; colour provides an indication of this. In addition, the size of the catchment is displayed proportionally in km². These two indices have a clear polynomial correlation. There is also a distinct separation between high and low altitude catchments. All the catchment areas at higher altitudes (above 1000m above sea level)

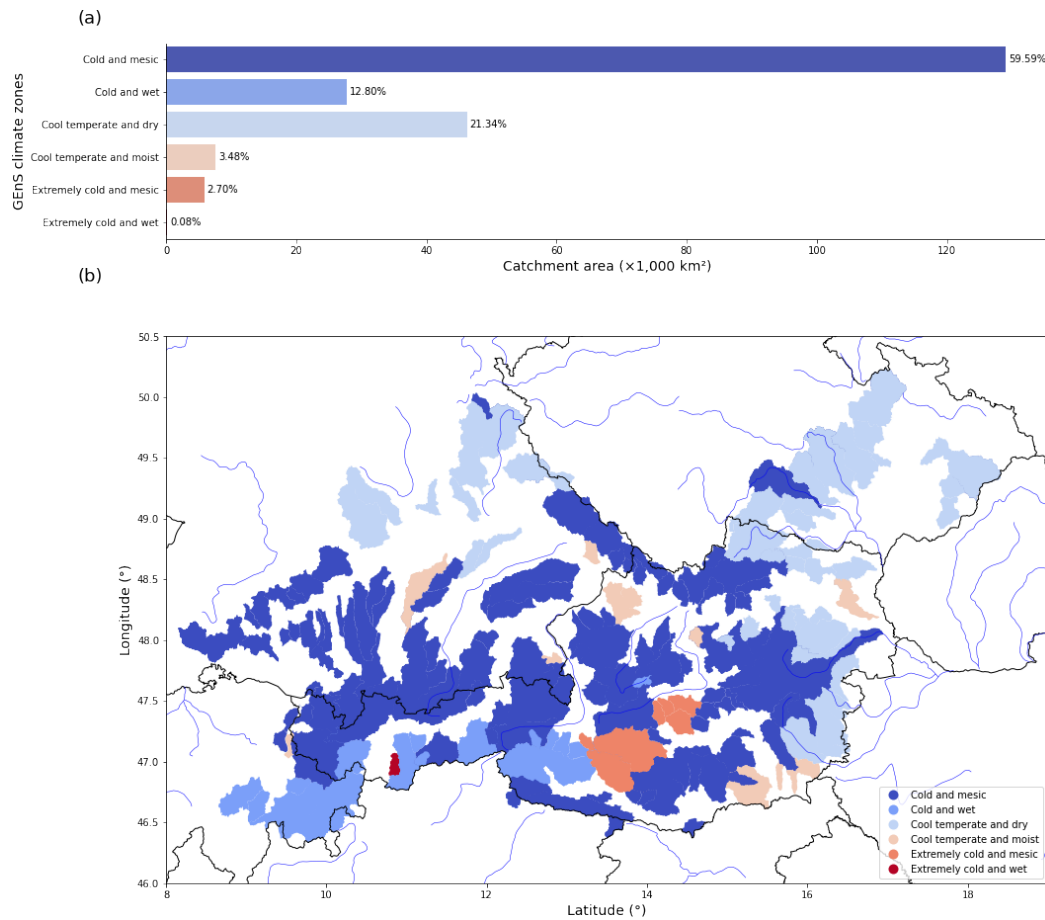


Figure 3.4: (a) Distribution of the catchment areas across the GEnS climate zones including the percentage of the overall catchment area covered in LamaH. (b) Geographical distribution of the GEnS climate zones across the study area.

have a low index of aridity as well as a low index of evaporation. This analysis suggests that catchments located at higher elevations generally have a greater water supply than that lost through evapotranspiration due to their location in wetter climates. In contrast, catchments at lower elevations are drier and limited by water availability. No correlation between size and elevation of the catchments can be found.

The actual *Budyko* curve is derived from the formula stated in equation 3.1.

$$Budyko = \sqrt{\Phi_P \tanh\left(\frac{1}{\Phi_P}\right) (1 - \exp(-\Phi_P))} \quad (3.1)$$

with Aridity index $\Phi_P = \frac{PE}{P}$

3. LARGE-SAMPLE HYDROLOGY AND DATA OVERVIEW

All of the available catchments lie above the *Budyko* curve indicating that the water systems of the study area tend to experience higher E_t relative to precipitation and PE . In other words, the catchments are effectively utilising a significant amount of the available moisture in the system for evapotranspiration. At the same time, almost all catchments have an aridity index smaller than 1.0, which indicates that the systems receive more precipitation relative to their PE . Therefore, the catchments are located in regions where more moisture is present than is lost through evapotranspiration, even though they consume a significant amount of available water. This suggests a humid to sub-humid climate, which aligns with the predominantly mesic to wet GEnS climate zones represented in the study area. There is notable variability in the local climate that may support the hydrological conditions within the region covered by LamaH.

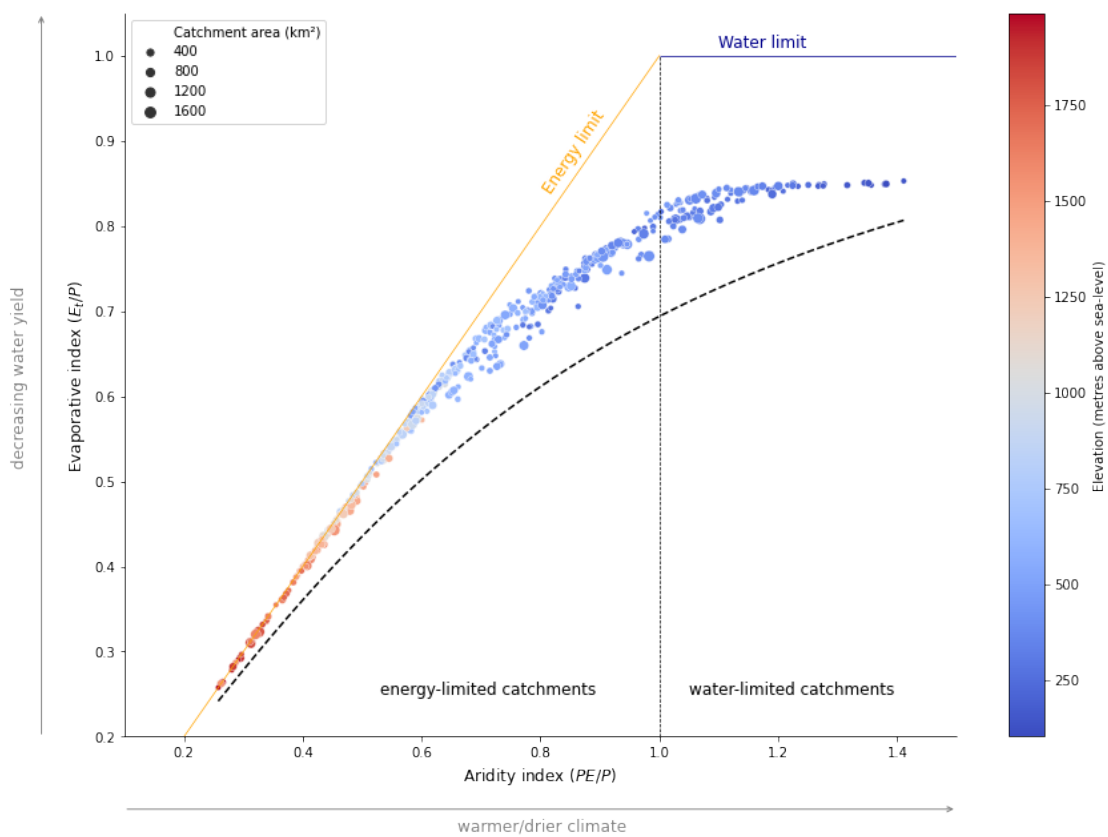


Figure 3.5: The Budyko curve is displayed for all 479 catchments in this study. It plots the aridity index (PE/P) against the evaporative index (E_t/P), with point size proportional to the catchment area, and point colour indicating the mean average elevation of the catchments. The Budyko curve is represented by the dashed line [Bud74].

The observation that most catchments have an aridity index below 1.0, but are consistently plotted above the *Budyko* curve, suggests an interesting hydrological pattern. The catchments appear to use water efficiently, meeting both evaporation and transpiration

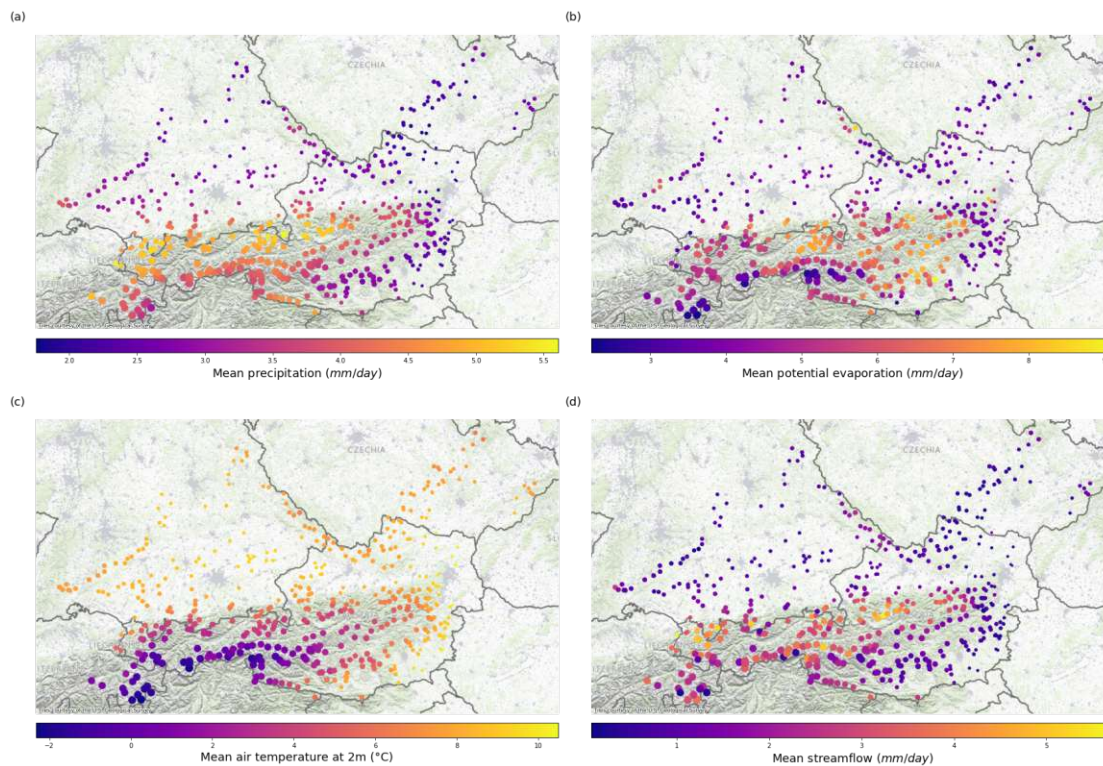


Figure 3.6: Geographical distribution of key hydrological measures, including precipitation (P), potential evaporation (PE), air temperature (T) and streamflow (Q), across the study area. The base layer of the map reflects the topography of the domain. The colour of the points represents the intensity of the respective measure, while their size corresponds to the catchment elevation. Four metrics are presented: **(a)** average daily precipitation (mm/day), **(b)** average daily potential evaporation (mm/day), **(c)** average daily air temperature calculated at 2 meters ($^{\circ}\text{C}$), and **(d)** average daily streamflow (mm/day) normalised with respect to the catchment area.

needs adequately. This hydrological efficiency may be due to vegetation, soil characteristics, or topography in the study area that promotes water retention and utilisation. Water management in the area appears to be relatively sustainable. Another conclusion that can be drawn from the diagram is that the runoff is rather low because most of the moisture is absorbed by the soil, used up by the vegetation or lost through evapotranspiration. This interpretation is to be confirmed by the experiment in this study. Furthermore, most of the catchments are plotted to the left of the energy-water boundary (shown as the dashed black vertical line in Figure 3.5). This indicates that most of the catchments are limited by energy rather than water, further supporting the interpretation of efficient water use in the system.

Figure 3.6 displays the geographical distribution of the key hydrological metrics P , PE , T and Q across the LamaH study area. The size of the points corresponds to the elevation

of the respective catchment, while their colour indicates the intensity of the respective measure. The numbers are compiled as daily averages over the whole study period. In general, the observations line up with the geography and topography of the domain of coverage. The predominant physical feature of the area comprises the Alps, specifically the Eastern, Central, and Limestone Alps located in Austria. These are accompanied by the foothills of the Alps. The considerable differences in altitude and temperature in the study area are mainly due to these characteristics. Other significant geographic and topographic features of the study region include the Danube and Rhine rivers, the Vienna Basin and the Pannonian Lowlands, the Schwarzwald in southern Germany, the Czech Highlands and Bohemian Massif, along with the Moravian Karst, Lake Constance, Swiss Plateaus, and the foothills of the Jura Mountains.

There are clear trends in the investigated metrics. As demonstrated in Figure 3.6a, catchments situated at higher altitudes receive considerably more rainfall than those in low-lying areas. The mean air temperature naturally aligns with the topography and elevation, with higher temperatures at lower altitudes and much lower temperatures at higher altitudes in the Alps, as shown in Figure 3.6c. As anticipated, there exist positive and negative linear relationships between both precipitation as well as temperature and elevation, respectively.

Figure 3.6b reveals a more varied pattern for streamflow than for the other metrics. PE is generally low for regions at high altitude as well as the lowlands and higher for mid-elevation catchments. The areas experiencing the highest amounts of PE ($PE > 7\text{mm/day}$, i.e. the 90% quantile) are situated in the Central Eastern Alps (specifically the Lavanttal Alps in Carinthia), the North Tyrol Limestone Alps and the Bayerischer Wald; they are located at moderate altitudes with a mean elevation of 1,148 metres. The geographical distribution of streamflow values is heterogeneous across the domain.

A key finding from Figure 3.6 is that many regions located at high altitudes with high precipitation values at the same time experience low evaporation. This is in agreement with the observation that water seems to be used efficiently in most systems of the study area and catchments are limited by energy. Figure 3.7a provides more insight into this matter by visualising the difference between P and PE . Catchments in the Schwarzwald, the Western Rhaetian Alps in Italy and Switzerland, Vorarlberg, Kaunertal, Ötztaler Alps through Osttirol and Gailtaler Alps in Carinthia, along with a few catchments in the Salzkammergut in Salzburg and Upper Austria, show water surplus whereby the system retains more moisture than lost through evaporation. These catchments are typically at higher altitudes, of smaller size and experience far less energy from radiation compared to precipitation through rain or snow. They are thus limited by energy to a high degree. The majority of catchments in the remaining study area show small negative values, signifying insignificant water limitations.

However, significant differences exist between these regions and catchments in Central Austria, the Northern Limestone Alps, the Vienna Basin, and certain basins in the Schwarzwald and Bayerischer Wald in Southern Germany. These areas exhibit signs

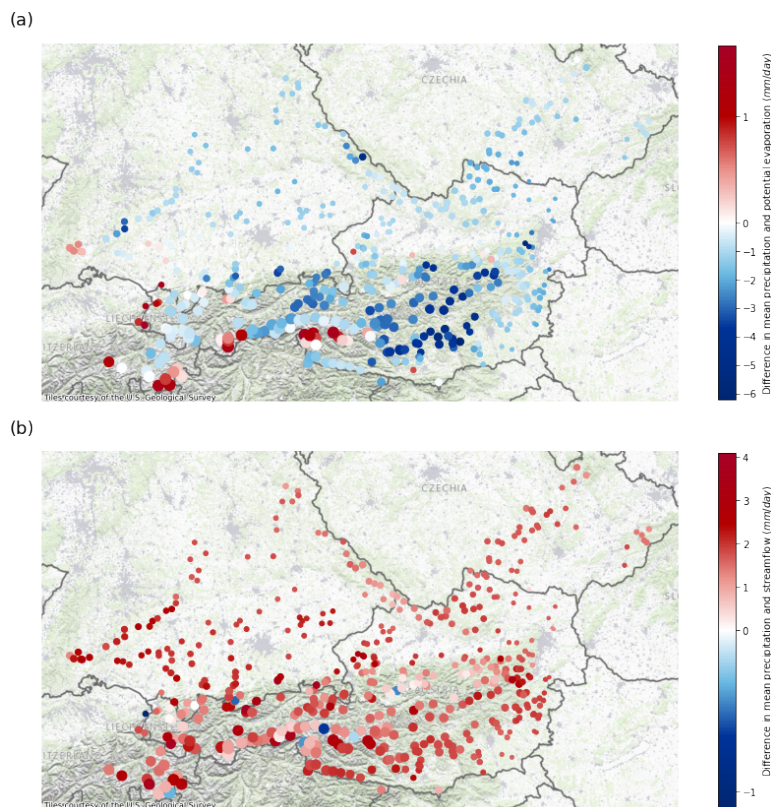


Figure 3.7: (a) Difference in mean daily precipitation and potential evaporation in mm/day. (b) Difference in mean daily precipitation and streamflow in mm/day.

of water scarcity. The mountain ranges Niedere Tauern as well as the Lavanttaler and Gurktaler Alps in Austria experience especially drastic differences in precipitation and potential evaporation, which is confirmed by the high aridity index in these regions.

Figure 3.7b displays the difference in P and Q . The dominant pattern is that more precipitation is measured than runoff. This is to be expected due to the principles of the hydrological cycle as water is typically subject to losses through processes like evaporation, transpiration, infiltration into the soil, percolation into groundwater, which contribute to baseflow. Water storage mechanisms further influence the timing of runoff. Streamflow typically operates with reduced water quantities due to losses and a delayed timing. Two anomalies experiencing greater streamflow than precipitation are situated in the Rhine Valley in eastern Switzerland and the Hohe Tauern in Salzburg and Osttirol. Notably, these anomalies are encompassed by watersheds with greater Q than P .

Throughout the study region, the average disparities in rainfall (P) and potential evapotranspiration (PE) as well as rainfall (P) and streamflow (Q) are moderately small, with values of -1.34 mm/day and 1.67 mm/day, respectively. Although varying topography results in noticeable regional and local differences in these relationships, these

findings confirm the generally efficient use of water in the system as a whole.

Caravan provides a seasonality attribute that shows the extent to which a catchment experiences variations in its water-energy-budget over the course of a year. The feature is represented by an index range of 0 to 2. The region generally undergoes a mild transition from drier to wetter seasons, with an average index of 1.13. The catchments located in the northernmost part of Germany (Bayerischer Wald) and the Czech Republic experience the most marked changes from arid to wet. The water and energy budget see the fewest changes in the Vienna Basin, as well as the Gurktaler and Lavanttaler Alps located in southern Austria.

Between 1993 and 2009, the human footprint index increased by 3.37%, signifying the progress of anthropogenic impact on the study region. The greatest increase took place in the southern and south-western parts of the domain, with southern Tyrol and eastern Switzerland experiencing the most significant gains.

Analysis of Exemplary Catchments

Time series data is the core of the provided information in the *Caravan* collection. The hydrological data spans 39 years at daily temporal resolution in 14 attributes. The time series data, alongside the catchments' static attributes, are the main input to the hydrological models used in the experiments to predict the relationship between rainfall and runoff in further stages of this thesis. To examine and visualise these important data as part of a comprehensive data analysis process, it is useful to first select exemplary catchments that are capable of providing information that can be generalised to the entire study region.

To this end, the z-score ($z = \frac{X-\mu}{\sigma}$) is calculated for each variable of the static catchment attributes (see Section 3.3.2). The mean absolute z-score is then calculated for each catchment to indicate the tendency to deviate from the standard behaviour of the study area. The calculations exclude 23 columns that contain only zero values, which relate to variables concerning land cover, natural vegetation, and wetland extents. The catchments with the highest and lowest mean absolute z-scores are chosen as candidate catchments for subsequent analysis. A higher score suggests that the catchment records notably large or small values, resulting in the most deviance from the other examined regions. Conversely, a lower score indicates that the catchment is a representative candidate of the entire study area, with attributes closest to the overall mean values.

The catchment area with the greatest mean absolute z-score is located in Pontresina, specifically in the Berninabach region, covering an area of 106.8 km², with an elevation of 2,575 metres above sea-level in the Rhaetian Alps of eastern Switzerland. This catchment has the highest negative deviation in the variables describing E_t , with the month of May typically experiencing the most drastic deficit in actual evapotranspiration. On the other hand, the most significant positive deviation is observed in the attributes related to snow cover extent during the months of July and August (variables explaining spatial means of land cover and potential natural vegetation for specific classes are excluded). This

result is in accordance with the notably high-altitude position of the catchment, which is located just 30 metres below the point of maximum elevation in the research area. In most catchments there is no snow cover during the summer months. However, in this high altitude catchment in the Swiss Alps, where the average air temperature is below freezing ($-2.76\text{ }^{\circ}\text{C}$), there is a considerable amount of snowfall and the snow cover is maintained throughout the summer.

Although the catchment area with the lowest mean absolute z-score is the Schwarza at Loipersbach in the Lower Austrian Prealps, it cannot serve as a candidate catchment due to missing streamflow records between 1991 and 1993. Therefore, the catchment with the second lowest score is chosen, which is the Schwarza at Gloggnitz, which is only 14.5 km from Loipersbach. The basin covers an area of 469.5 km^2 and is located at an elevation of 954 metres. This catchment area aligns most closely with the average of those studied in the area.

Figure 3.8 displays the rainfall and streamflow records as monthly averages for the whole 39-year period in the data for both exemplary catchments. The high-deviation catchment Berninabach at Pontresina experiences far more precipitation and streamflow than the low-deviation catchment Schwarza at Gloggnitz. Although both catchments receive similar amounts of precipitation, with a consistent pattern of minima and maxima throughout the year, there is a noticeable anomaly in the highly deviating catchment, which has a much higher frequency of extreme peaks, exceeding 10 mm/day on several occasions. In terms of both amplitude and frequency, the peaks tend to be more constant for the low deviation catchment. The difference in magnitude for both catchments is especially drastic for the streamflow records. On average, the Berninabach experiences more than twice the flow of the Schwarza. Peakflow is significantly more pronounced, and the curve's shape is relatively homogeneous for the high-deviation catchment. The periods of baseflow are also clearly distinguishable. It is evident that there is a consistent hydrological pattern in each year for the Berninabach. When comparing the peaks of precipitation with those in the streamflow records, a clear time delay can be observed between intense rainfall events and the peakflow. Shape and intensity of streamflow for the Schwarza are heterogeneous and do not reveal any clear patterns. The high-deviation catchment experiences comparable levels of rainfall and streamflow, whereas the low-deviation catchment receives over twice the amount of precipitation in comparison to streamflow.

The hydrographs of both catchments, shown in Figure 3.9, reveals the drastic difference in streamflow records for both exemplary catchments. The hydrographs are obtained by calculating the mean streamflow for each day of the calendar year with reference to all 39 years of data (the hydrological year is not used here as it is easier to visually discern the usual pattern of the hydrograph throughout the normal calendar year). The Berninabach shows a typical hydrograph pattern. An initial period of baseflow during the winter, when most precipitation falls down as snow and ice and the temperature is constantly below freezing, is succeeded by gradually increasing streamflow with rising temperature and the onset of the springtime period of snowmelt. Between June and

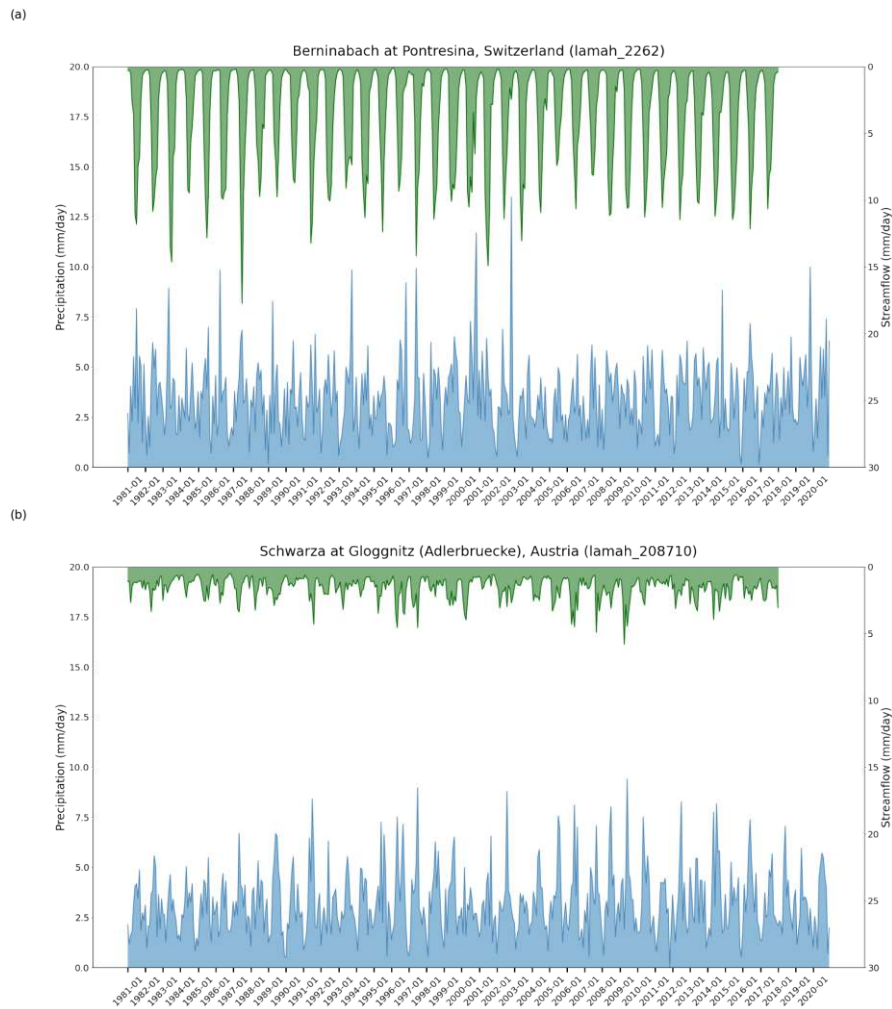


Figure 3.8: Rainfall and streamflow records over the whole period covered in *Caravan* for the two selected exemplary catchments. Daily records from each month were averaged and are displayed as a blue line for precipitation on the left y-axis and as a green inverted line for streamflow on the right y-axis. **(a)** The measurements for the high-deviation catchment Berninabach at Pontresina, Switzerland. **(b)** The measurements for the low-deviation catchment Schwarza at Gloggnitz, Austria.

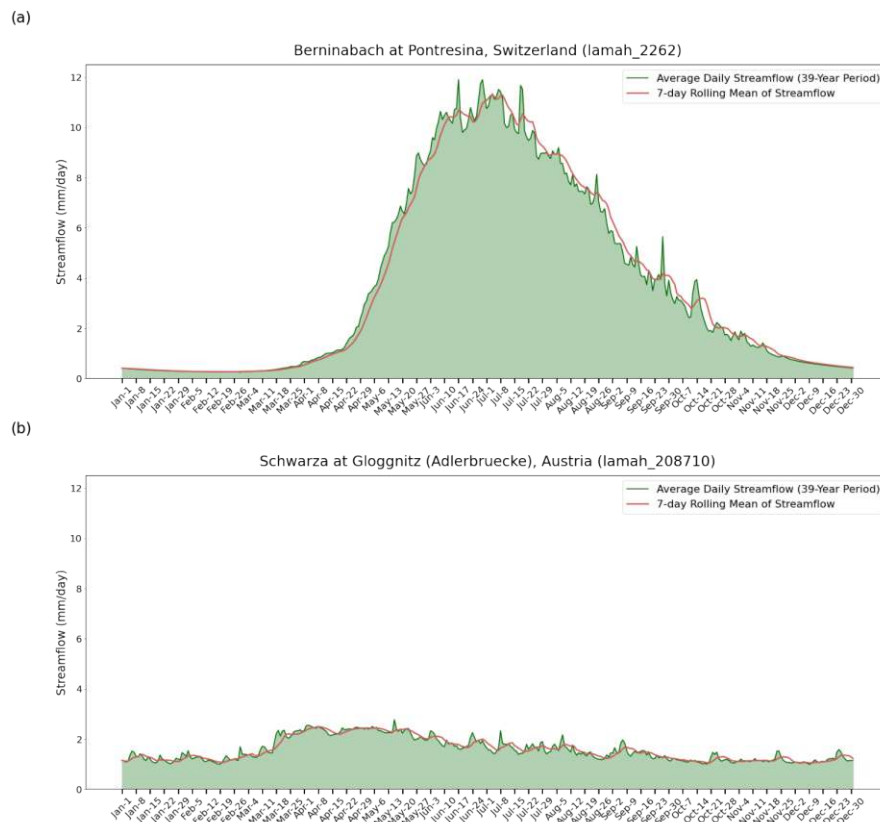


Figure 3.9: The hydrographs showing the mean daily streamflow records averaged over the entire 39-year period covered in *Caravan* for the two selected exemplary catchments. The 7-day rolling average of streamflow records is depicted by the red line. (a) The hydrograph for the high-deviation catchment Berninabach at Pontresina, Switzerland. (b) The hydrograph for the low-deviation catchment Schwarza at Gloggnitz.

July, peakflow is reached and afterwards the measurements continue to decrease with declining temperature in autumn. The characteristic pattern of the hydrograph as shown in Figure 2.2 becomes even clearer when the 7-day rolling average (shown as the red line) is considered. It is easy to distinguish between peak flow and base flow.

The hydrograph for the Schwarza catchment, however, differs significantly. Besides the overall lower streamflow magnitude, no distinct seasonal variations can be observed and there is only a slight increase in the spring compared to the rest of the year. Distinguishing baseflow from peakflow is not possible. It is rare for there to be negligible variation in streamflow volume throughout a year.

The substantial differences in rainfall and streamflow between the two analysed exemplary catchments are likely due to their greatly contrasting topographical features. The presence of large volumes of snow during winter (but also throughout the year) in the significantly higher elevated Rhaetian Alps leads to a pronounced snowmelt effect during spring,

which is not present for the Schwarza in the Lower Austrian Prealps. Glacial meltwater can significantly influence streamflow patterns in alpine catchments. Additionally, the consistently colder temperatures in the Rhaetian Alps likely contribute to the prolonged winter baseflow and gradual snowmelt. While precipitation records are similar for both catchments, distribution, magnitude, form and timing vary, which can impact streamflow patterns. Moreover, the different geological composition of the area the catchments are located in may affect the storage and release of water to a significant degree.

Trend Analysis

To investigate changes in the main hydrological variables and draw conclusions about the presence and impact of climate change on the study area, the time series data can be subject to trend analysis. Figure 3.10 shows the de-seasonalised trend in the average yearly temperature for all 479 catchments across the whole study period. The observation of trends in highly seasonal time series data is facilitated by the use of de-seasonalised data, which contributes to the objectivity of the analysis. This is achieved by subtracting the average value of the respective month from each observation. Equation 3.2 shows the calculation of the de-seasonalised time series d_t by subtracting the average temperature \bar{T} in a specific period from the original time series y_t . The periodicity function $m(T, d)$ takes an integer time index T and a periodicity d and calculates a filtered version of y_t . Only those observations that are the same month as input T are included determined by the modulus of T and the index t .

$$d_t = y_t - \bar{T} [m(t, 12)] \quad (3.2)$$

where $m(T, d) = y_t \{ \text{mod}(T, d) == \text{mod}(t, d) \}$

The trend of the de-seasonalised temperature data is clearly increasing. This is confirmed by analysing the seasonal data in the bottom four sub-plots of Figure 3.10. While the temperature in winter does not increase significantly and rather shows non-periodic maxima and minima, the other three seasons experience a clear trend of warming over the years. The observed increase is especially drastic in spring and summer. Therefore, the observations of warming are in accordance with the de-seasonalised data.

Table 3.3 displays differences in key statistical metrics for hydrological variables, which include temperature (T), streamflow (Q), precipitation (P), snow water equivalent (SWE) and potential evaporation (PE), between the first five-year period (1981-1986) and the last five-year period (2013-2018) for 479 catchments in the study area. The reference date was 1st January. Despite the data covering three more years, streamflow records are only available up to 2018. This analysis reveals noteworthy trends. The mean temperature in the region increased significantly by 1.53°C, demonstrating a clear trend of warming. Over time, there was an average decrease of -0.213 mm in streamflow which suggests a decline in water flow. On average, precipitation increased slightly by 0.124 mm, potentially indicating a minor rise in rainfall, while the snowpack displayed

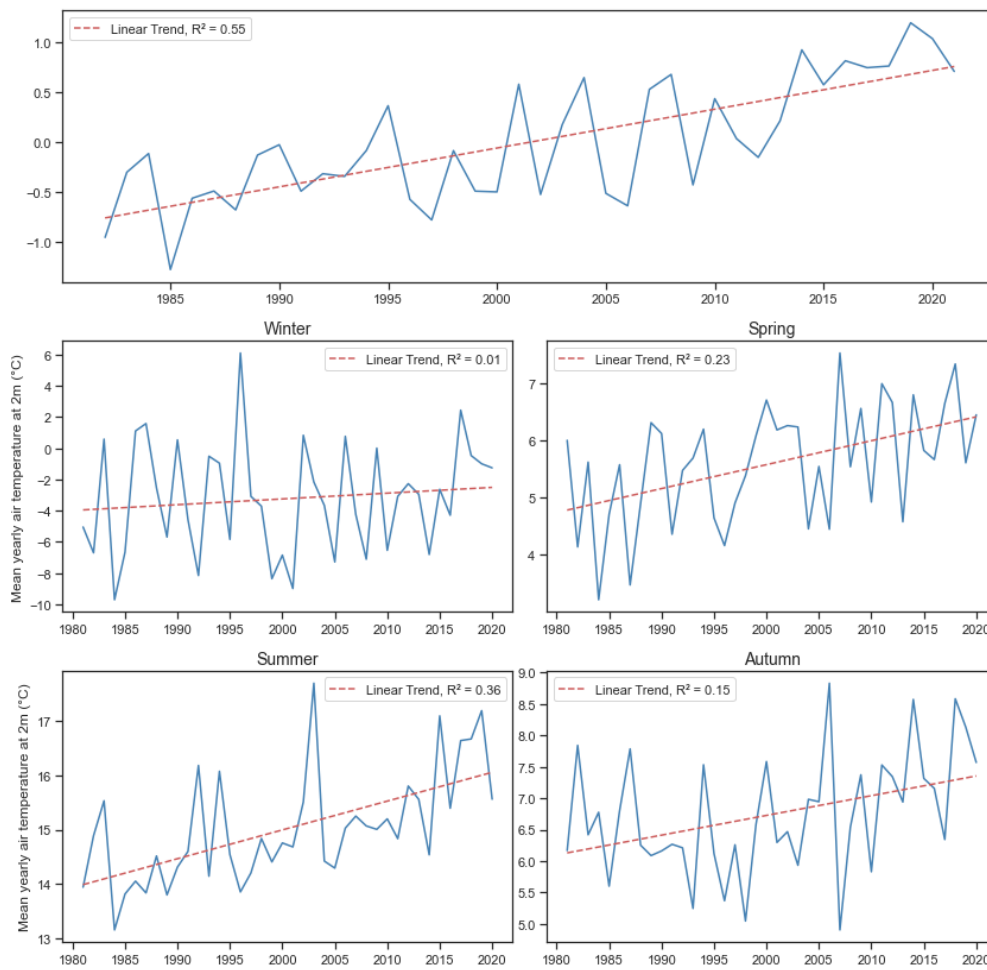


Figure 3.10: Top: De-seasonalised mean yearly air temperature for all catchments. Bottom: Seasonal mean yearly air temperature trends. Trends are indicated by red dashed linear regression lines.

a significant decrease of -33.943 mm, indicating a decrease in snow accumulation. PE increased by 0.423 mm, potentially linked to the rise in temperature.

Other trends were evident in the median values. The median temperature rose by 0.942°C , further suggesting a warming trend in the region. There was a decrease in median streamflow by -0.224 mm, while median precipitation increased by 0.049 mm, suggesting a possible rise in median rainfall and a simultaneous decrease in streamflow. The median value of SWE showed a significant decline of -15.841 mm, also indicating a reduction in snowpack. Meanwhile, the median value of potential evaporation also increased.

While the standard deviation of temperature and streamflow decreased, indicating less variability in these variables, an exception is precipitation, where the standard deviation

Metric	T (°C)	Q (mm)	P (mm)	SWE (mm)	PE (mm)
Mean	1.530	-0.213	0.124	-33.943	0.423
Median	0.942	-0.224	0.049	-15.841	0.279
Standard Deviation	-0.443	-0.163	0.066	-34.946	0.377
Minimum	1.600	0.006	0.000	0.000	0.032
Maximum	-0.313	-3.642	-14.364	-107.222	1.578
25th Percentile	1.916	-0.075	0.014	-0.979	0.021
75th Percentile	0.748	-0.400	0.159	-80.759	0.421

Table 3.3: Difference in key statistical metrics for hydrological variables (T , Q , P , SWE , PE) between the first five years (1981-1986) and the last five years (2013 - 2018) with complete records for 479 catchments in the study area.

increased, indicating more variability in rainfall patterns. SWE showed a significant decrease in standard deviation, suggesting less variability in snow cover, likely influenced by decreasing snow depth. Furthermore, the standard deviation of PE increased, reflecting increased variability in evaporation that may be associated with changing temperature patterns. Collectively, these trends highlight changes in the variability of key hydrological parameters in Central Europe over the period analysed. This analysis provides valuable insights into the changing hydrological conditions in Central Europe, with implications for water resource management, flood prediction, and climate monitoring.

In order to statistically confirm the observed trends, an unmodified Mann-Kendall test is performed for the variables P , Q , and T [Man45, HM19]. Figure 3.11 displays the number of catchments with statistically significant ($\alpha = 0.05$) trends based on monthly averaged data. In summary, the majority of catchments show no trends in these variables. Confirming the results of the exploratory trend analysis, precipitation and temperature show a slightly increasing trend, with the latter variable showing a stronger tendency. The trend analysis for streamflow is much more mixed, with 91 catchments showing a decreasing trend and 58 showing an increasing trend. Interestingly, the 65 catchments exhibiting an increasing trend in temperature are all located in the Austrian Alps. They are scattered across Carinthia, Salzburg and (East) Tyrol.

A major conclusion from this trend analysis is that, as a data engineer, it would be beneficial to have access to longer periods of complete data. Although the dataset at hand is a pioneering example in LSH and provides high-quality data spanning 39 years, naturally, it is important to note that climate trends take time to develop and may not yet be discernible in the data. It is desirable to have access to high spatio-temporal resolution data going back to the 19th century to study the impact of the industrial revolution and the large subsequent increase in greenhouse gas emissions on the climate. However, this can only be achieved as part of a major research effort, with particular emphasis on the development of high-quality climate reanalysis techniques and databases.

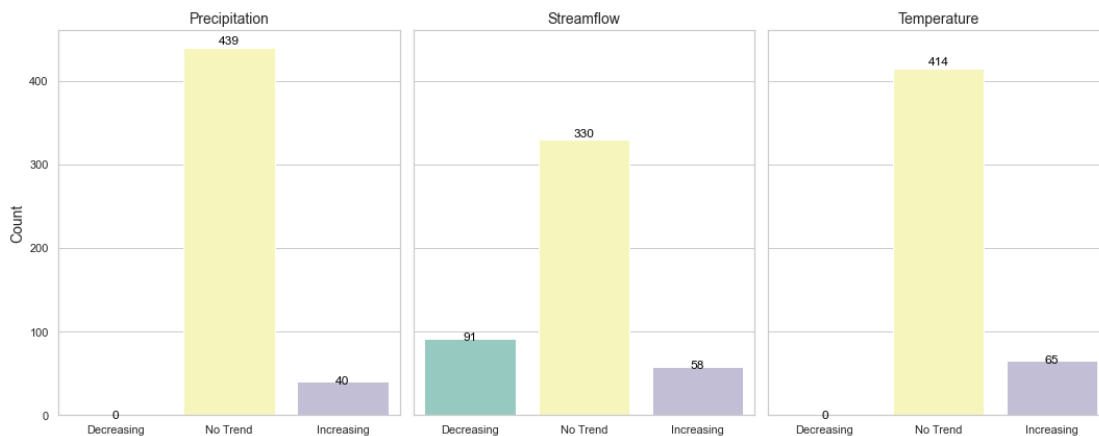


Figure 3.11: Number of statistically significant ($\alpha = 0.05$) trends in mean air temperature, precipitation, and streamflow for the 479 catchments in the study area over a 39-year period. The data was averaged on a monthly basis and analysed using an unmodified Mann-Kendall test.

Summary

The large number of available variables presents significant potential for conducting further analyses into other aspects of the data. For instance, an investigation into the implications of the geological composition of the study area on hydrological metrics or exploring the effects of different types of vegetation on water use could be explored.

The study area in the *Caravan* version of LamaH is located in Central Europe, with a focus on the greater Danube region as well as the Eastern Alps of Austria, Switzerland, Germany and Italy. This region is characterised by a remarkable climatic diversity, encompassing six GEnS climate zones ranging from cold to mesic, humid, wet and dry conditions. In particular, the cold and mesic zone dominates, covering 59.6% of the area. Within this diverse landscape, two primary terrestrial biomes emerge: temperate broadleaf and mixed forests in the northern region, and temperate coniferous forests in the southern part. These unique climatic and ecological characteristics make this area an ideal candidate for hydrological experiments, particularly for assessing the potential impacts of climate change on the relationship between rainfall and runoff.

The considerable hydrological efficiency of all the catchments in the study area is one of the key findings of the analysis. Actual evapotranspiration (E_t) consistently exceeds precipitation (P), indicating that the catchments are efficient users of available water. Furthermore, most catchments have an aridity index below 1.0, indicating a humid to sub-humid climate. This observation suggests that catchments in the region are generally limited by energy rather than water availability, underlining the influence of topography and climate on hydrological processes.

Exemplary catchments were selected on the basis of their deviations from the mean

absolute z-scores in order to examine the hydrological patterns within this region more closely. The hydrographs of these selected catchments, the Berninabach near Pontresina and the Schwarza near Gloggnitz, show striking differences. The Berninabach shows distinct seasonal variations in discharge, characterised by baseflow in winter and a pronounced snowmelt effect in spring. In contrast, the Schwarza has a much lower flow with no discernible seasonal pattern. These significant differences in precipitation and streamflow can be attributed to distinct topographical features, including elevation, snowfall patterns and geological composition. This stark contrast between the sample catchments highlights the key role of topography in shaping hydrological patterns within this diverse study area.

The trend analysis showed interesting patterns. The analysis of de-seasonalised annual temperature data shows a clear warming trend across all catchments, particularly in spring and summer. Similarly, important statistical metrics for hydrological variables, including temperature, streamflow, precipitation, snow cover and potential evaporation, indicate significant changes. Temperature has increased considerably, while streamflow and snow cover have decreased. The variability in temperature and streamflow has decreased over time, whereas precipitation variability has increased. The analysis underlines the importance of longer data records to capture climate trends, and emphasises the need for extensive spatio-temporal datasets to fully investigate the impacts of climate change.

Methodology

4.1 Data Preparation

Further investigation of the available data is required before it can serve as input to any hydrological model. Data preparation is a crucial preliminary step in hydrological modelling, regardless of the modelling approach employed. The presence of missing data, outliers, and inconsistencies may produce a noticeable effect on the performance of the model and the precision of its predictions. Its main purpose is to guarantee the reliability and precision of the model's results. By analysing the patterns of missing values and subsequently handling them accordingly, the model possesses complete information to operate with, which thereby reduces potential biases. Identifying and addressing outliers is crucial in preventing these extreme values from disproportionately influencing model results and promoting interpretability. Moreover, it is essential to encode and scale data to harmonise variables with different units and scales, facilitating fair comparisons and effective parameter estimation. Furthermore, DL models commonly require input data to be of a specific format with correct scaling applied. Meticulous data preparation is essential for improving the quality and robustness of hydrological models, regardless of whether they are conceptual, physical, or utilising advanced DL techniques [ZGSG18, LR02].

4.1.1 Initial Analysis and Processing Steps

An initial analysis reveals that the static catchment attributes from HydroATLAS and ERA5 as well as the descriptive information and metadata are complete and do not contain any missing values in the 212 available variables. Since these values are taken from well-maintained re-analysis sources, they can be considered to be reliable and do not require outlier detection.

The time series data contain missing values only for the variable describing streamflow; all other attributes are complete for all catchments throughout the period under consideration.

The authors of *Caravan* recognise the potential occurrence of missing data in streamflow records, and attribute this to the fact that the availability of data depends on the local, regional or national organisations that collect the data from the respective gauges [KNA⁺23]. 18.71% of the streamflow records are missing, equivalent to 0.46% of the entire time series data.

Missing streamflow affects all 479 catchments. The missing records are typically distributed in sequences of varying length with an average duration of 29 days. Conversely, the average length of sequences with successive streamflow values is only 27 days.

No streamflow data is available for the last three years of the study period (2 January 2018 to 31 December 2020) in any of the catchments. Therefore, it is impossible to perform data imputation for this period. As a result, the time series data has to be truncated from 2 January 2018, with the last three years being excluded. This step reduces the data by 7.5%. The reason of missingness for this pattern is Missing Not At Random (MNAR) as the missing records depend on the date of the observation [LR02].

Remarkably, although the catchments in the Czech Republic represent only 10.2% of the area covered by the data, their missing streamflow records account for 39.8% of all missing values. In addition, streamflow data for the Czech Republic are only available from 2004 onwards. Several catchments are missing data for later years as well. As a result, Czech records are considered unreliable and cannot be imputed. The reconstruction of the data would be too complex and beyond the scope of this work. The only viable option is to exclude the 35 catchments from the Czech Republic due to the severely limited streamflow records. This step further reduces the data by 7.3%. The records before 2004 are MNAR. The reason for the remaining missing records cannot be attributed to any systematic relationship, and hence, it is assumed to be Missing Completely At Random (MCAR) [LR02].

After implementing the data truncation and catchment exclusion steps for the observed patterns, there is still a significant proportion of missing values in the remaining streamflow data. The assumption that the missing data is MCAR is supported by Little's MCAR Test [Lit88]. Simply removing rows with missing values is not possible as it is necessary to preserve continuous sequences of time series data. Therefore, Multiple Imputation (MI) is employed in Section 4.1.3.

Various catchments still have excessive amounts of missing streamflow, which introduces significant bias into the data when imputation is used. While Madley-Dowd et al. found that it is possible to produce unbiased model results through imputation with data containing proportions of missing values as high as 90%, it is still advisable to define a reasonable cut-off [MDH19]. Thus, 12 catchments exceeding 75% of MCAR streamflow data are excluded, reducing the data by a further 2.7%. These three steps preserve 83.43% of the original data in 432 of the 479 catchments and leave the rest for MI.

This analysis results in three initial steps to prune the data and account for missing values in the streamflow variable before imputation:

1. **Truncate**

The time series is truncated from 2 January 2018, thus shortening it by three years.

2. **Exclude**

All 35 catchments from the Czech Republic are excluded from the time series data and the static attributes.

3. **Cut-off**

12 catchments exceeding 75% missing values are cut from the time series data and the static attributes.

4.1.2 Data Splitting

With the data now in a pruned, but yet unmodified state, the data can now be split to prepare for DSST. Sungmin O et al. defined climatically motivated reference periods for model training to assess the robustness of rainfall-runoff models under changing conditions in their landmark study from 2020. They chose the driest and wettest year on average of the study period for calibration and used the remaining 24 years of data as the evaluation period. The authors recognised that these very short training periods should be extended in further research. Additionally, a wider range of climatic conditions could be utilised as reference periods [ODO20].

Building on the first-order results of O et al., their approach is adopted and extended in this work. The choice of reference periods is modified to include longer periods with more diverse conditions. The aim is to adapt the reference periods by extending the duration and incorporating precipitation as a climatic condition. Therefore, four climatic reference periods are to be chosen: *hot2cold* and *cold2hot* for temperature, *wet2dry* and *dry2wet* for precipitation. The naming aligns with the initial proposal of Sungmin O et al. [ODO20].

The approach to find these periods is as follows: First, the time series data is searched for consecutive periods of high or low average values for either temperature or precipitation. High periods are determined by days exceeding the 66th percentile of the respective variable and low periods are defined by days that fall below the 33rd percentile. To increase the length of these periods, there may be periods of up to five days that do not meet the percentile threshold requirement. These periods are then aggregated by a sliding window of four years to find clusters of periods within the sliding window. The resulting aggregations are then analysed by the mean of the temperature or precipitation, respectively, as well as the percentage of days that meet the percentile threshold criteria within the total duration of the period, and also the deviation from the overall mean of the variable of interest. Then, the most fitting period (reasonable length, high deviation, high percentage of threshold days) is chosen for each of the four reference periods.

A detailed description of the algorithms to find and aggregate consecutive high or low periods are given as pseudo code in Algorithm C.1 and C.2. Table 4.1 displays the characteristics of the four reference periods.

	Temperature $\bar{X} = 5.87 \text{ }^\circ\text{C}$		Precipitation $\bar{X} = 3.48 \text{ mm}$	
	hot2cold	cold2hot	wet2dry	dry2wet
Start Date	2013-04-25	1984-11-14	1997-05-20	1989-08-04
End Date	2016-10-07	1988-04-07	2001-04-27	1993-01-26
Threshold	10.33 $^\circ\text{C}$	1.54 $^\circ\text{C}$	3.65 mm	0.69 mm
% of Threshold Days	39.89 %	46.21 %	45.76 %	41.31 %
No. of Days (Train)	1,261	1,240	1,438	1,271
No. of Days (Test)	11,886	11,907	11,709	11,876
Mean	7.84 $^\circ\text{C}$	4.18 $^\circ\text{C}$	3.75 mm	3.18 mm
Δ from \bar{X}	+1.97 $^\circ\text{C}$	-1.69 $^\circ\text{C}$	+0.27 mm	-0.30 mm

Table 4.1: Definition of the climatic periods for Differential Split-Sample Testing.

Start Date	1982-01-01
End Date	2000-01-01
No. of Days (Train)	6,575
No. of Days (Test)	6,573
\bar{T}	5.54 $^\circ\text{C}$
Δ from \bar{T}	-0.33 $^\circ\text{C}$
\bar{P}	3.43 mm
Δ from \bar{P}	-0.05 mm

Table 4.2: Definition of the baseline reference period.

Additionally, a baseline reference period is defined to evaluate the differences in model performance for the experiments with climatic reference periods. The reference period encompasses 50% of the available study period (18 years) starting from one year after the beginning of the records in the time series. The remaining years are used as an evaluation period similar to the experiments based on DSST. The offset of one year is chosen due to a warm start period appended to the chosen DL models prior to the actual start date. The baseline period characteristics are defined in Table 4.2. The length of the baseline period’s training set has been deliberately chosen to be considerably longer than the duration of the climatic reference periods. The rationale behind this approach is to emulate state-of-the-art experiments that typically use training periods of similar length [KKS⁺19, KKG⁺22]. While it is likely that extreme periods of temperature or precipitation will become longer as a result of climate change, these durations will always be shorter than an unaffected baseline period that is not subject to any constraints. Therefore, it is of interest to compare the performance of the reference periods with a model that has been trained on a much longer period of data.

This approach results in five reference periods for model calibration/training with the respectively remaining periods of the data used as the evaluation period. The last year

of the data (1st January 2017 to 1st January 2018) is used as the validation period for all reference periods. Therefore, ten subsets are left for further processing. It is vital to perform the subsequent data processing steps separately on each of these subsets to avoid data leaking from the training to test sets. This prevents the introduction of information about the data distribution and thus potentially significant bias into the models [PVG⁺23].

4.1.3 Missing Data

MI aims to “fill in missing values by generating plausible numbers derived from distributions of and relationships among observed variables in the data set” [LSA15]. This approach creates multiple datasets by using different estimators or random states, thus making it possible to analyse how the subsequent modelling results vary due to the inherent uncertainty caused by the missing values.

Two imputation techniques are applied in this work. Univariate imputation is implemented by using the per-catchment median. Multivariate imputation is implemented by using RF regression. The Python ML package `scikit-learn` is utilised to implement imputation [PVG⁺11]. The RF estimator was tuned based on empirical experiments and the resulting hyperparameter settings are given in Section B.1. Both strategies are explained in more detail below:

1. Per-Catchment Median

Univariate imputation using the Per-Catchment Median (PCM) is a naive but robust strategy. This approach takes advantage of the historical pattern of the catchment and provides a reasonable approximation of the missing data. Utilising the median values specific to each catchment is a first step in increasing the specificity and precision of the imputed value, and could be improved in further research by using the grouped seasonal median per catchment or other time-dependent constraints. Since the distribution of streamflow records is skewed, the median is a better estimator compared to the mean [HHR⁺20, HRS21].

2. Random Forest Regression

The meta estimator `ensemble.RandomForestRegressor`¹ fits a number of regression decision trees on bootstrapped sub-samples of the data and uses the average of the tree predictions as the imputed value. The advantages of this non-parametric estimator are its usage of the ensemble technique and robustness to outliers. However, it is computationally intensive and prone to overfitting. This approach provides numerous parameters that can be tuned.

There is a special case for the imputation using the catchment-specific median. Due to the specific data splits it may occur that a column contains only missing value, thus

¹Source: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#sklearn.ensemble.RandomForestRegressor>

leading to the grouped median for the respective catchment also being NaN. This is accounted for by imputing the overall column-specific median instead.

The two different imputation strategies are employed to handle the remaining missing values and allow for uncertainty estimation as part of the model evaluation. Combining the variations from splitting the data into baseline and climatic reference periods, a total of ten unique train/test subsets emerge as input to models.

Within the field of geospatial science, there exist state-of-the-art alternative techniques for data imputation, such as Arithmetic Averaging Nearer Stations Based On Conditions (AVwC), Normal Ratio Method With Respect To Distance (NRM_D) or Inverse Distance Weighting (IDW) [CPN16]. Extensive research has been done to analyse flow imputation in hydrological data. PDMs such as the SWAT can be used to substitute values, as well as advanced methods from DL like ANNs or Self-Organising Maps, as shown by [KBL⁺15]. Further frequently applied methods in the domain include Principal Component Analysis and Two-Directional Exponential Smoothing [HHR⁺20]. However, the selected strategies are chosen due to their ease of implementation in fulfilling the scope of the thesis, providing an overview of available approaches, and accounting for uncertainty due to missing values and their imputation. This selection was made considering the complexity of advanced methods.

4.1.4 Outliers

The detection and handling of anomalies in a dataset is a crucial task in data engineering. Especially domains where the data originates from sensors, gauging stations or human input, such as geospatial sciences, are strongly affected by the presence of outliers. Possible causes of these errors may include faulty sensors, human error, or errors in data engineering, such as inaccurate climate reanalysis. Similar to the management of missing data, the detection of outliers has been the focus of extensive research efforts and various domain-specific state-of-the-art techniques have been proposed. Besides simpler statistical methods, such as the z-score or the exponentially weighted moving average, and models from ML and DL, such as Support-Vector Machines or Isolation Forests, also ANN-based DL approaches have been applied extensively. The latter methods have largely been used to seamlessly detect and impute anomalous values, for instance, using a bi-directional RNN to capture temporal information. Notably, sliding window algorithms have been used in combination with ML-based data imputation for identifying outliers in hydrological time series data [KCM⁺21].

In this work, Isolation Forests are used to detect outliers in time series data after missing values. This ensemble technique uses a group of Decision Trees to efficiently isolate observations by arbitrarily selecting a splitting value within the range of the minimum and maximum values of the selected variable. The algorithm then averages the path lengths from the root to the terminal node for each tree, which corresponds to the number of necessary splittings to isolate the observation. The inverse is used as a measure of anomaly for a given observation. Thus, samples with shorter average path lengths are

considered easier to isolate and more likely to be outliers [LTZ12, PVG⁺11]. The chosen implementation of the algorithm is `ensemble.IsolationForest`². A separate model is fitted for each catchment and attribute to detect outliers, which are then flagged for later imputation. The model's hyperparameters were tuned empirically and the settings are listed in Table B.2.

The selected substitution technique for detected outliers is imputation by the PCM (see strategy 1 of missing value imputation). To avoid potential data leaking from the missing data imputation performed prior to the outlier imputation, the median values per catchment used here are sampled from the original data (after step 3) per subset split. While all methods discussed for dealing with missing values could be applied for outlier imputation, this would result in too many input datasets for experiments and subsequent uncertainty analysis also considering train/test splits, which is well beyond the scope of this work. Therefore, the simple strategy to impute the catchment-specific median suffices here, which is also proven to be a robust measure [KCM⁺21].

Of the 39 meteorological variables in the time series data, the only one with no detected outliers is the attribute describing the minimum net solar radiation at the surface, because it is always zero by nature. The contamination parameter of the Isolation Forest model used is set to a very low value of 0.004 to preserve natural anomalies and reduce noise at the same time. The mean proportion of outliers over all columns and strategies is approximately 0.25% and the mean anomaly score is 0.273. After outlier imputation, the standard deviation is changed by an average of -1.08%. There are no rows containing only outliers.

4.1.5 Feature Selection

Due to the large number of static attributes available (212) resulting from aggregations, it is recommended to decrease feature dimension prior to providing them as additional inputs for experiments. Monthly aggregate variables are unlikely to be relevant in explaining the feature space and may even be redundant. Furthermore, a high number of similar features can contribute to overfitting. Therefore, feature selection is performed to find a more compact representation of the available information. Since the static attributes are based on the catchments without a direct link to the time series data and the target variable of observed streamflow, an unsupervised algorithm needs to be chosen [Has23]. Only numerical attributes with a non-zero standard deviation are subject to feature selection since normalisation would fail otherwise. This leaves 186 attributes for selection.

To this end, the widely used and efficient Regularised Self-Representation (RSR)-based unsupervised feature selection algorithm proposed by Zhu et al. is applied to the static catchment attributes (the metadata such as gauge name, country, etc. are omitted from the feature selection and later appended to the resulting feature set). This method

²Source: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html#sklearn.ensemble.IsolationForest>

leverages the self-representation property of high-dimensional data in that a feature can be expressed effectively by a linear combination of its so-called relevant features. Furthermore, $L_{2,1}$ norm-group sparsity regularisation is employed on the feature weights matrix W to increase robustness [ZZZ⁺15]. Equation 4.1 shows the minimisation problem of the feature weights matrix:

$$\hat{W} = \arg \min_W \|X - XW\|_{2,1} + \lambda \|W\|_{2,1} \quad (4.1)$$

where $\|\cdot\|_{2,1}$ denotes $L_{2,1}$ norm-group sparsity regularisation
and λ is a positive constant

The chosen implementation of this model is FRUFS³, which utilises supervised algorithms such as XGBoost to calculate feature importance [Has23]. The highest-ranking 30% of the static catchment attributes are chosen to serve as additional input. This results in a total of 59 features (including the three numerical metadata attributes).

For the time series data, only the variable `surface_net_solar_radiation_min` is removed since it is always zero.

4.1.6 Scaling

Scaling data is a crucial pre-processing step, particularly for ML and DL models. This step is necessary to achieve a well-balanced initialisation of the weights and to avoid vanishing/exploding gradients, thus increasing convergence speed and promoting generalisation and interpretability. Various scaling methods are available, including standardisation, robust scaling, and min-max-scaling [LBOM12].

The `NeuralHydrology` Python library, which is used for the DL experiments in this thesis, standardises the input data by default before feeding it to the model. Equation 4.2 calculates the z-score separately for each feature to achieve a mean of zero and a standard deviation of one. To benefit from this standard behaviour of the library, the data for all models is standardised.

$$z = \frac{X - \mu}{\sigma} \quad (4.2)$$

Naturally, to prevent data leakage from the distributions of the respective subsets, the scaling is applied individually to all subsets. The static catchment attributes are also standardised.

4.1.7 State of the Data after Preparation

After applying the described pre-processing steps to split the data into training/test/validation sets according to four DSST periods and an additional reference period, remove missing

³Source: <https://github.com/atif-hassan/FRUFS>

values and handle them using two different approaches, to detect and impute outliers for each of the strategies, to standardise the data, and to select the most important features, the data is in a very different state compared to its raw form.

In total, 47 catchments were completely removed from both the time series and the static catchment data due them containing too large proportions of missing values. The time series data was also shortened because the most recent three years are missing streamflow records. These steps reduced the time series data by 16.5%. It now contains 5,838,048 observations across 36 years for 432 catchments that have their respective gauging stations located in four different countries (Austria, Germany, Switzerland, Liechtenstein).

Two approaches were used to address missing values: imputing them with the catchment-specific median and using the ensemble ML regression algorithm RF for imputation. However, each of these techniques introduces some level of uncertainty to the data, which requires further analysis. These two techniques were applied to each data split which yields a total of ten sets as input to modelling experiments and further analyses.

Outlier detection was performed using the ML algorithm Isolation Forest and then imputed using the catchment-specific median. Finally, the data was scaled by applying the z-score. The results of feature selection lead to discarding 62 static attributes due to them having a standard deviation of zero. In the end, 59 attributes were chosen to serve as additional input to the models where applicable. A single feature from the time series data was also dropped.

Figure 4.1 shows a visualisation of the complete data processing pipeline. Reproducibility is guaranteed by using fixed random seeds for each step of the pipeline (see Appendix B).

4.2 Model Architectures and Implementation

While the input data of all three model types was subject to the same pre-processing steps as described in Section 4.1, the models have different requirements with regard to input features and formats. Therefore, the exact input data differs for all three model types.

4.2.1 Process-based Model: HBVEdu

The widely applied conceptual model HBV is used in an educational version called HBVEdu proposed by Aghakouchak and Emad in 2010 [AH10]. This PDM operates on a per-catchment basis. Therefore, all catchments are subject to calibration separately and local models are built in contrast to the regional ML and DL models (see Section 4.4.1).

HBVEdu is a simplified, spatially-lumped model that treats a catchment as a single unit and disregards spatial variations. The four modules snowmelt and snow accumulation, soil moisture and precipitation, evapotranspiration, and the runoff response are represented in the educational version and can be seen in Figure 4.2. Precipitation is considered to be either rain or snowfall depending on the temperature of the respective day, which serves

4. METHODOLOGY

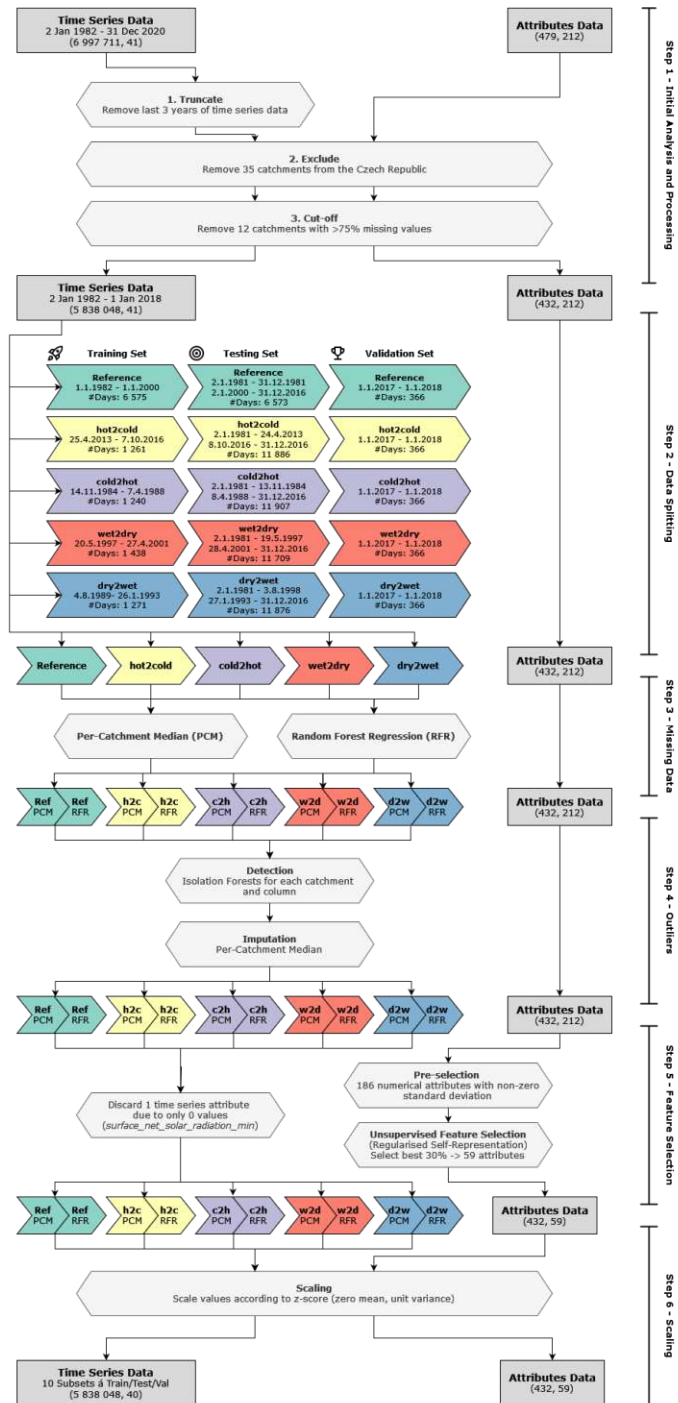


Figure 4.1: The data preparation pipeline for the raw LamaH time series and static attribute data.

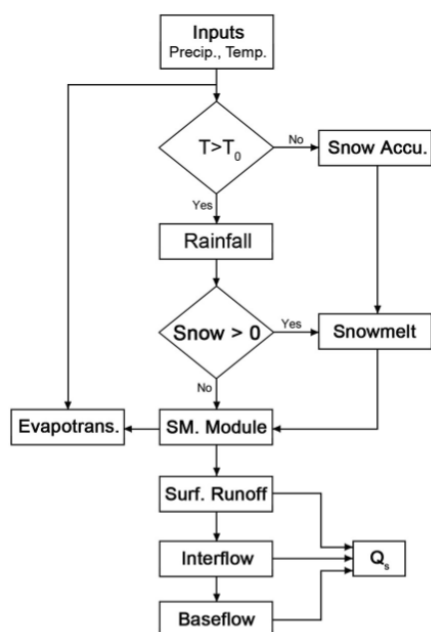


Figure 4.2: The components and processes of HBVEdu [AH10].

as a threshold. The resulting precipitation is then evaluated in the soil moisture module, where the effective input of water contributing to surface runoff is calculated. The remaining rainfall is considered to contribute to soil moisture storage, which evaporates as long as there is sufficient water content in the subsurface. The output is the discharge runoff at the watershed outlet, consisting of surface runoff, interflow, and baseflow. The model parameters (such as the threshold temperature, etc.) are adjusted during calibration (i.e. model training). The inputs to the HBVEdu model and their origins from the pre-processed data are stated in Table 4.3.

HBVEdu does not support static catchment attributes as additional input. The remaining time series features are discarded.

The model is implemented utilising the Python library RRMPG in version 0.0.1, which provides an out-of-the-box implementation of HBVEdu⁴. Internally, the rainfall-runoff model is treated as a multivariate function and the differential evolution method is used to find the global optimum of this problem [SP97]. The optimisation criterion is set to the Root Mean Squared Error (RMSE). To ensure reproducibility, the random seed is set to 2,609. The maximum number of allowed iterations during optimisation is increased from 1,000 to 3,000 to account for convergence issues.

It should be noted that the decision for PDM selection was made between HBVEdu and VIC-5. While the available implementation of VIC is much more advanced and

⁴Source: <https://github.com/kratzert/RRMPG>

Name	Input	Engineered feature
qobs	Target variable streamflow	Variable streamflow
temp	Mean temperature	Variable temperature_2m_mean
prec	Summed precipitation	Variable total_precipitation_sum
month	Month of each time step	Engineered from the time step date
PE_m	Long-term mean monthly PE	Re-sampled var. potential_evaporation
T_m	Long-term mean monthly T	Re-sampled var. temperature_2m_mean

Table 4.3: Description of how the inputs for the PDM HBVEdu are engineered.

comprehensive, its design primarily for macro-scale modeling and development in C were influential factors in favor of selecting HBVEdu. The study area in LamaH primarily encompasses small catchments, making the gridded approach with large cells in VIC less suitable for the specific characteristics of the area.

4.2.2 Machine Learning Model: XGBoost

XGBoost (eXtreme Gradient Boosting) is a state-of-the-art open-source ML framework in the field of ensemble learning. Built as an extension of the gradient tree boosting algorithm, it sequentially constructs decision trees with each tree improving the errors of its predecessors. The method was specifically designed to be used on large, complex data sets and achieves a significantly lower runtime than other popular ML techniques. XGBoost is scalable, computationally efficient, flexible, and supports distributed computing. The framework has received significant recognition for its success in Data Science and Machine Learning contests on Kaggle and has now become an essential component of research in various domains. Since XGBoost utilises L_1 - and L_2 -regularisation, the chance of overfitting is reduced drastically [CG16].

XGBoost’s capability to learn complex relationships within vast and diverse data through ensemble techniques positions it as an ideal solution for rainfall-runoff modelling. The incorporation of regularisation is a key characteristic that is expected to result in effective generalisation across the set of heterogeneous catchments in the diverse study area at hand. The scalability of the algorithm as well as its ability to be parallelised should allow for a low runtime even with a large amount of input data. As such, this work employs XGBoost as the representative ML model in the experiments.

Since XGBoost only accepts numerical input variables, the timestamp of a sample has to be broken down into five numerical components: day of the week/month/year, number of the month, and the year. The remaining time series variables are used as-is with the exception of the catchment ID. As with the process-based model, it is not possible to use the static attributes as input in this case either.

XGBoost is implemented using its Python library in version 2.0.2⁵. A regional model us-

⁵Source: <https://github.com/dmlc/xgboost>

ing all catchments at once is built, thus interpreting the data as a multivariate regression problem with the target variable being streamflow. To find the best model settings, a grid search using 3-fold cross-validation is performed. The parameter ranges as well as the optimal settings are stated in Table B.3. The fixed parameters are: `tree_method` set to `hist` (uses a faster histogram optimised approximate greedy algorithm for tree construction), `booster` set to `gbtree` (uses gradient-based tree boosting), `random_state` set to 2,609 (ensures reproducibility). In total, 11,664 candidate parameter combinations are applied in the grid search resulting in 34,992 fits with cross-validation. The split parameter combinations are scored based on a custom implementation of the NSE. The hyperparameter search took approximately 18.7 hours and was carried out once before model training. The optimal parameter set was then utilised throughout all DSST periods.

4.2.3 Deep Learning Model: EA-LSTM

Kratzert et al. introduced an adaptation to a classic LSTM neural network called Entity Aware - LSTM (EA-LSTM) that incorporates a set of static attributes to facilitate the learning of similarities between catchments. The authors showed in their initial study that this approach can outperform locally and regionally calibrated process-based models and contributes to model interpretability with regard to the way catchment-specific behaviours and inter-catchment relationships are learnt. The EA-LSTM is capable to efficiently model both temporal and spatial relationships in a single framework [KKS⁺19].

An additional attention layer aims to leverage the potential of the large volume of static attributes that introduces valuable information about each catchment. The underlying concept of the model architecture is that dynamic time series inputs should be processed conditionally based on the static attributes of the respective catchment. The authors propose an adaptation to the standard LSTM architecture given in Equation 2.5. Above all, the input gate $i[t]$ does not depend on the time step anymore and can thus be denoted simply as i , which incorporates the static inputs x_s . The dynamic time series attributes are denoted as $x_d[t]$ at time step t . Thus, the adapted notation of the LSTM gates are presented in Equation 4.3.

$$\begin{aligned}
 i &= \sigma(W_i x_s + b_i) \\
 f[t] &= \sigma(W_f x_d[t] + U_f h[t-1] + b_f) \\
 o[t] &= \sigma(W_o x_d[t] + U_o h[t-1] + b_o) \\
 g[t] &= \tanh(W_g x_d[t] + U_g h[t-1] + b_g) \\
 h[t] &= o[t] \odot \tanh c[t] \\
 c[t] &= f[t] \odot c[t-1] + i \odot g[t]
 \end{aligned} \tag{4.3}$$

In this architecture, both inputs are processed separately with different responsibilities. x_s controls which parts of the network should be activated for a specific catchment via the modified input gate i . The information is managed in the memory through the recurrent

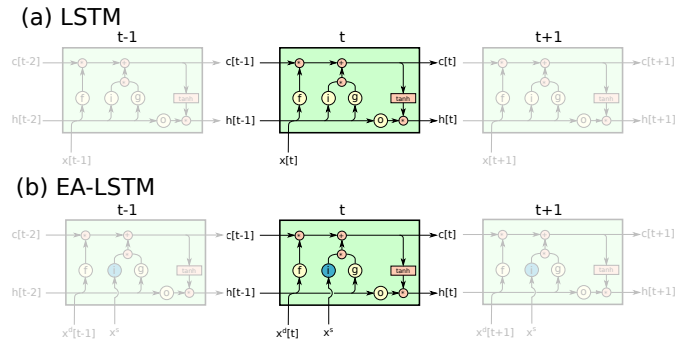


Figure 4.3: Differences between the architectures of the classic LSTM and the EA-LSTM model of Kratzert et al. [KKS⁺19].

inputs and x_d ($g[t]$ is responsible for writing, $f[t]$ for deleting, and $o[t]$ for output at time step t). The dynamic and static inputs are both used in an unchanged state after the data preparation steps. The EA-LSTM architecture, shown in Figure 4.3 with reference to the conventional LSTM design, is therefore able to distinguish between similar types of rainfall-runoff behaviours that differ between catchments. The static input gate i can determine which components of the network should be activated for a particular catchment through an embedding layer of real values that is available after training. The authors declare that this embedding layer enables the model to share information between catchments based on similarities that are based on, for instance, geological or anthropological attributes while other parts of the network are disabled due to differences in other characteristics [KKS⁺19].

The authors of the EA-LSTM are part of the AI for Earth Science group at the Institute for Machine Learning, Johannes Kepler University in Linz, who develop a popular open-source Python library called `NeuralHydrology`⁶, which is specifically designed for Deep Learning in hydrology. The library is used in this work to implement the EA-LSTM model out-of-the-box in version 1.9.0.

DL models from `NeuralHydrology` can be configured via a YAML file with an array of arguments. The applied values for the EA-LSTM model configuration mostly conform to those presented in the initial paper by Kratzert et al. [KKS⁺19] with a few adaptations following a high-degree, empirical hyperparameter tuning phase. Finding ideal parameters for neural network models is a complex and computationally costly process especially for large volumes of data and thus considered out-of-scope for this work. The relevant model settings, shown in Table B.4, are the same for all input sets (except for data-specific paths).

Following the initial experiment results from Kratzert et al., Gaussian noise $\mathcal{N}(0, \sigma)$ is added to the data during training. This noise affects the already standardised data so that σ is not influenced by the relative magnitude of the variable. The authors found

⁶<https://github.com/neuralhydrology/neuralhydrology>

that adding noise can contribute to increasing the robustness of a model [KKS⁺19]. In the experiment setup of this work, the standard deviation is set to 0.005.

4.3 Model Evaluation

4.3.1 Evaluation Metrics

In order to evaluate the performance of the various models on the data, six different metrics will be applied: Mean Absolute Error (MAE), RMSE, Coefficient of Determination (R^2), NSE, Kling-Gupta Efficiency (KGE), and Percent Bias (P_{BIAS}). The model results are analysed and compared using these metrics in detail in Section 5. Table 4.4 provides an overview of the formulae for the various metrics, their value ranges, optimal value, and the respective references. Additionally, reported runtimes for model training, evaluation, and validation offer insight into model complexity.

Metric definition	Value range	Optimum	Reference
$MAE = \frac{1}{n} \sum_t^n \hat{y}_t - y_t $	$[0, \infty)$	0	[WM05]
$RMSE = \sqrt{\frac{1}{n} \sum_t^n (\hat{y}_t - y_t)^2}$	$[0, \infty)$	0	[WM05]
$R^2 = \left(\frac{\sum_t^n (y_t - \mu_y)(\hat{y}_t - \mu_{\hat{y}})}{\sqrt{\sum_t^n (y_t - \mu_y)^2 \sum_t^n (\hat{y}_t - \mu_{\hat{y}})^2}} \right)^2$	$[0, 1]$	1	[MGPD15]
$NSE = 1 - \frac{\sum_t^n (\hat{y}_t - y_t)^2}{\sum_t^n (y_t - \mu_y)^2}$	$(-\infty, 1]$	1	[NS70]
$KGE = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}$	$(-\infty, 1]$	1	[GKYM09]
with $\alpha = \text{variability ratio} = \frac{\sigma_{\hat{y}}}{\sigma_y}$,			
$\beta = \text{bias ratio} = \frac{\mu_{\hat{y}}}{\mu_y}$,			
$r = \text{correlation coefficient}$.			
$P_{BIAS} = 100 \frac{\sum_t^n \hat{y}_t - y_t}{\sum_t^n y_t}$	$(-\infty, \infty)$	0	[GSY99]

Table 4.4: Overview of the evaluation metrics used to analyse and compare model performance.

The observed target values are denoted by y , the simulated values obtained from a model by \hat{y} . The term n refers to the forecast horizon (total number of fitted points or time steps in the time series); μ and σ are the mean and standard deviation of a sample. Pearson's linear correlation coefficient between observed and simulated values is denoted

by the term r and defined as

$$r = \frac{\sum_t^n (y_t - \mu_y)(\hat{y}_t - \mu_{\hat{y}})}{\sqrt{\sum_t^n (y_t - \mu_y)^2 \sum_t^n (\hat{y}_t - \mu_{\hat{y}})^2}} \quad (4.4)$$

with values ranging from -1 to 1 . A value of exactly 1 implies that a linear relationship between y and \hat{y} describes the relationship perfectly. A value of 0 indicates no linear relationship.

Mean Absolute Error

The MAE is a commonly used forecasting error measure in time series analysis. It describes the average absolute difference between simulated and observed values and is defined by the following equation:

$$\text{MAE}(y, \hat{y}) = \frac{1}{n} \sum_t^n |\hat{y}_t - y_t| \quad (4.5)$$

Values range from 0 to infinity with the ideal value being 0 . Here, each error contributes to the metric in proportion to the absolute error and does not overvalue or undervalue larger or smaller errors. This metric is scale-dependent and has the same unit as the underlying data. It is relatively robust against outliers as well as easy to explain and interpret [HK06].

Root Mean Squared Error

The RMSE has been used extensively in statistical modelling and is defined as

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_t^n (\hat{y}_t - y_t)^2} \quad (4.6)$$

where the range of values and the ideal value correspond to the MAE. Another common characteristic is that the output unit of both metrics is the same as the unit of the reported values (m^3/s in case of streamflow or rainfall-runoff simulation) further contributing to the interpretability of these measures. Each error contributes to the score in proportion to its square, thus causing the metric to exaggerate the influence of large errors which is an advantageous characteristic for model performance comparison. A consequence is that the RMSE is more sensitive to outliers compared to the MAE [HK06].

Coefficient of Determination

The metric R^2 is simply Pearson's correlation coefficient, which is presented in equation 4.4, squared and thus scaled to the interval from 0 to 1 . It is commonly utilised in hydrological modelling studies as standard metrics for evaluating model performance and

is widely recognised as a benchmark. The R^2 assesses the combined dispersion of the observed and predicted series in comparison to the dispersion of each series individually. It can be interpreted as the proportion of the observed dispersion that is accounted for by the prediction. The score can be overly influenced by extreme values, as highlighted by Krause et al. [KBB05].

Nash-Sutcliffe Efficiency

Nash and Sutcliffe proposed the NSE in 1970 [NS70]. This metric is most prominently applied in hydrology for model calibration and evaluation and can be understood as a normalised variant of the Mean Squared Error (MSE). The NSE is dimensionless and scaled onto the interval of $(-\infty, 1]$ with a value close to 1 considered ideal. This metric can be interpreted as a skill score implicitly comparing the prediction of a model to the mean of the observations with regard to the sum of squared errors. A value of 0 indicates that the model's predictive power is equivalent to the mean and a negative value implies that the mean is a better predictor than the model.

$$\text{NSE}(y, \hat{y}) = 1 - \frac{\sum_t^n (\hat{y}_t - y_t)^2}{\sum_t^n (y_t - \mu_y)^2} \quad (4.7)$$

The NSE has been extensively used to report results in hydrologic modelling and is most prevalent in literature alongside KGE. It is suitable for a wide range of target variables including streamflow and non-linear transformations as well as hydrological signatures such as flood or drought distribution. Interpretability, simplicity, the implicit comparison to a baseline, and the emphasis on errors are advantageous characteristics of this metric [MMA⁺23].

The mean of observations as the baseline model of the NSE is often criticised as being too naïve as it leads to an overestimation of the predictive power of the model, especially for strongly seasonal variables that are common in hydrology [MMA⁺23]. Kling and Gupta identify three components by decomposing the NSE: the linear correlation coefficient r , the normalised bias by the standard deviation of observed values β_n , and a measure of relative variability in the simulated and observed values α [GKYM09].

$$\begin{aligned} \text{NSE}(y, \hat{y}) &= 2\alpha r - \alpha^2 - \beta_n^2 \\ \text{with } \alpha &= \frac{\sigma_{\hat{y}}}{\sigma_y} \text{ and } \beta_n = \frac{\mu_{\hat{y}} - \mu_y}{\sigma_y} \end{aligned} \quad (4.8)$$

This decomposition turns the criterion into a multi-objective optimisation problem to find a balance between ideal values for the three components ($r = 1, \alpha = 1, \beta_n = 0$). However, due to α appearing twice in the decomposition and the bias being scaled by σ_y , the trade-off becomes unpredictable and models tend to be selected where bias might be underrepresented and variability underestimated [GKYM09]. Therefore, the NSE must always be used together with other metrics and put into context when used as benchmarking metrics [CVL⁺21].

Kling-Gupta Efficiency

The KGE is also commonly used in hydrological modelling and is motivated by the decomposition presented in equation 4.8. It aims to overcome the mentioned shortcomings by separating the three components variability ratio, bias, and correlation. Consequently, the multi-objective optimisation becomes more balanced, reducing the severity of the tendency to underestimate flow variability [GKYM09, Liu20].

$$\text{KGE}(y, \hat{y}) = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (4.9)$$

with $\beta = \frac{\mu_{\hat{y}}}{\mu_y}$

A notable difference to the decomposition of the NSE is that the bias term β is not normalised by the variability of the observed values but rather the ratio of both means. This improves the hydrologic interpretability and the ideal values for r , α and β all lie at 0. However, Knoben et al. demonstrate that values below 0 do not necessarily suggest poorer performance compared to the mean of the observations. As a result, the KGE cannot be directly compared to the NSE even though both metrics are frequently reported together in literature [KFW19]. Liu argues that the KGE still poses the problem of underestimation of peak flow or overestimation of low flow [Liu20, CVL⁺21].

P_{BIAS}

Another commonly utilised metric is the percentage long-term bias P_{BIAS}, indicating if the predicted data is likely to be larger or smaller than the observed data. Positive values express an underestimation bias, negative values overestimation bias, and the ideal value is 0 [GSY99].

$$\text{P}_{\text{BIAS}}(y, \hat{y}) = 100 \frac{\sum_t^n \hat{y}_t - y_t}{\sum_t^n y_t} \quad (4.10)$$

It serves several purposes: (1) assessing the model's ability to replicate average magnitudes in the desired output response, (2) facilitating continuous long-term simulations, (3) broad acceptance and robustness, as reflected in extensive reported values, (4) assisting in identifying average biases in model simulations (over- or underprediction), and (5) accommodating measurement uncertainties. Nonetheless, it is also recommended to consistently report P_{BIAS} alongside other evaluation criteria since the resulting values alone might be misleading if the model exhibits both overestimation and underestimation tendencies [MGPD15].

4.3.2 Evaluation Guideline

Moriasi et al. recommend to report NSE, P_{BIAS}, RMSE, and R² (in combination with the slope and gradient of the regression line) among others to get a holistic picture of

model performance. To provide additional information and to account for potential shortcomings of these criteria, KGE and MAE will also be used for evaluation. The MAE serves as an additional measure to the RMSE that is similarly easy to interpret and not restricted to hydrological modelling [MGPD15].

While it is common practice to define the threshold of 0 for model performance for NSE and KGE and to interpret values above 0 to indicate improvements upon the observed mean flow, Knoben et al. point out that this threshold is actually incorrect for the KGE, as it does not possess an inherent mathematical benchmark [KFW19]. Clark et al. comment on the state of the usage of performance metrics in hydrologic modelling and emphasise that it is constrained by sampling uncertainty which should be quantified in reports. Furthermore, they point out that statistical estimates should be improved by (1) calculating metrics separately per reference period, (2) always putting performance criteria into context and using strict, purpose-specific benchmarks, (3) considering limitations of performance metrics and critically analysing a model according to these shortcomings, and (4) using additional metrics to just NSE and KGE [CVL⁺21]. This thesis aims to take these recommendations into account when the model evaluation results are reported, analysed and discussed in chapters 5 and ??.

Metric	Performance Evaluation Criteria			
	Very Good	Good	Satisfactory	Unsatisfactory
MAE	$MAE < 0.1\sigma_y$	$0.1\sigma_y \leq MAE < 0.2\sigma_y$	$0.2\sigma_y \leq MAE < 0.5\sigma_y$	$MAE \geq 0.5\sigma_y$
RMSE	$RMSE < 0.1\sigma_y^2$	$0.1\sigma_y^2 \leq RMSE < 0.2\sigma_y^2$	$0.2\sigma_y^2 \leq RMSE < 0.5\sigma_y^2$	$MAE \geq 0.5\sigma_y^2$
R^2	$R^2 > 0.85$	$0.75 < R^2 \leq 0.85$	$0.60 < R^2 \leq 0.75$	$R^2 \leq 0.60$
NSE	$NSE > 0.80$	$0.70 < NSE \leq 0.80$	$0.50 < NSE \leq 0.70$	$NSE \leq 0.50$
KGE	$KGE > 0.80$	$0.60 < KGE \leq 0.80$	$0.40 < KGE \leq 0.60$	$KGE \leq 0.40$
P_{BIAS}	$P_{BIAS} < \pm 5$	$\pm 5 \leq P_{BIAS} < \pm 10$	$\pm 10 \leq P_{BIAS} < \pm 15$	$P_{BIAS} \geq \pm 15$

Table 4.5: Guideline for the evaluation of hydrological models with the six performance criteria, partly inspired by [MGPD15].

In order to keep in line with the literature, model performance will be classified in a four-step ordinal scale according to Moriasi et al. [MGPD15]. Table 4.5 shows the classification of values for the performance metrics into this scale. Since this thesis uses an array of state-of-the-art metrics from literature, the value ranges are compiled from various sources. It is important to note that there is no inherent or commonly agreed on benchmark or classification for the KGE. Frequently, the threshold of 0 is used for both NSE and KGE to mark desirable model performance, however, these two metrics should not be directly compared as stated above. MAE and RMSE make use of fractions of the standard deviation of the observed values for the classification into the scale.

4.4 Experiment Design

A cornerstone of the here presented DSST experiment design is that the setup of train and test sets for each of the five presented input sets differ intentionally. This is to simulate

a real-world setting of climate research in hydrology, where traditional experiments are mostly carried out on a single input set. Typically, data splits are chosen at random or at empirical percentage values with respect to the amount of data samples available. For the domain of climate research, this design may limit the meaningfulness and robustness of model results since potentially predominant climatic conditions in the train/test sets are unclear and a result of the possibly random data split.

With these limitations in mind, the approach of Differential Split-Sample Testing was chosen to emphasise climatic conditions and resulting differences in model performance. A further opportunity was explored by leveraging non-continuity between train and test splits, meaning that either the train or test set may be interrupted depending on the relevant climatic condition of the set. Additionally, in ML-based time series experiments it is common practice to use earlier periods as train sets and later periods as test sets. In this work, all test sets apart from the reference set are non-continuous, and often times use later periods as train sets, thereby breaking with experimental convention in order to align with the paradigm of DSST.

A drastic difference in the input set is the far larger number of input samples in the baseline reference set compared to the four DSST sets. However, this deliberate variation is adopted in order to contribute to one of the central research objective of this thesis (see **RQ 3.2**), that is, to find out whether models trained on data demonstrating a climatically intense period with respect to a meteorological variable can be compared with traditional modelling setups. Inter-model comparability is achieved by reporting and analysing performance on a validation set that is the same for all splits.

One representative model is chosen from each of the three modelling paradigms that are examined in this work: HBVEdu as a PDM, XGBoost as a more traditional ML model, and EA-LSTM as a DL model. The models are evaluated based on the metrics presented in Section 4.3 with respect to the DSST periods as well as the different methods used for the imputation of missing values during data pre-processing in Section 4.1.

4.4.1 Regionalisation

A major open research question in hydrology is the so-called “regional modelling problem”, which explores the methodology of utilising a model or a set of models to generate hydrological simulations that maintain spatial continuity over extensive geographic areas, encompassing regions ranging from regional to continental or even global scales and to extrapolate hydrologic information between spaces and scales. Most state-of-the-art hydrology models primarily employ strategies that calibrate on individual basins. However, incorporating the diverse characteristics of different catchments, including ecological, geological, and topographical factors, into a model capable of generalising well to local, regional, and global models remains an important task. It is essential to learn and encode these characteristics in order to capture the heterogeneous hydrological behaviours across catchments [KKS⁺19].

Razavi and Coulibaly identify *model-dependent* and *model-independent* approaches for regional modelling [RC13]. The difference here is that the first approach relies on a pre-defined (usually process-based) hydrological model to derive parameters for simulations from the given data, thus aiming to simulate the complex behaviours in hydrological processes merely on observable catchment characteristics. Model-independent regionalisation, however, aims to extract information directly from the available data removing the necessity to gain prior knowledge of the hydrological processes. In model-dependent regionalisation, the challenge of equifinality arises from the complex probability distribution of model parameters, which is influenced by the interdependencies between these parameters [KKS⁺19].

Model-dependent regionalisation has gained significant interest within the hydrological community, resulting in a wide range of approaches that are currently available. Successful strategies include the use of a conceptual model calibrated for more than 1700 catchments globally as a reference library to identify similar catchments in a parameterised simulation ensemble for new catchments [BvDdR⁺16]. Another notable example is the study presented by Prieto et al., where hydrological signatures are regionalised by a RF regression model and then used to calibrate a rainfall-runoff model [PLVK⁺19].

Model-independent regionalisation make use of ancillary data and meteorological inputs to directly learn the mapping to streamflow or other flux types. The influence of catchment characteristics and other additional data should then allow to distinguish different catchment response modes. Although hydrological modelling commonly achieves higher accuracy by calibrating a model on a single catchment, data-driven approaches have demonstrated the advantages of utilising diverse training data from multiple heterogeneous sites. These approaches leverage the ability to transfer knowledge across basins, leading to improved results in hydrological modelling [KKS⁺19]. Kratzert et al. show that their regional catchment-agnostic LSTM model is capable of outperforming the SAC-SMA, which was calibrated on a single catchment [KKB⁺18, KKHH18].

In this work, regionalisation is achieved based on the model type in question. While the PDM HBVEdu is calibrated separately for each catchment, thereby producing one optimised model per catchment, the ML and DL models are trained utilising a model-independent approach. Both build a single model for all catchments at once. Thus, regionalisation happens purely by utilising the available data, which helps to reveal the complex interactions between catchment attributes. This strategy is initially proposed by Kratzert et al. for the EA-LSTM model. The authors demonstrate the increased performance of this approach in comparison to locally and regionally calibrated benchmark models using model-dependent strategies [KKS⁺19].

Results and Discussion

In order to provide the basis for answering the main objective of this study, that is **RQ 3** and the comparative analysis of model performance, a series of rainfall-runoff modelling experiments are conducted. This chapter describes the design of the experiments and the training configurations of the models before presenting the evaluation results. In Section 5.1, the computational infrastructure and findings during model training are described. Section 5.2 presents the evaluation results of modelling experiments on the respective test sets. Finally, Section 5.3 contains the results of the models on the validation set as well as visualisations and a comparative analysis of these results.

5.1 Experimental Setup

The experiments are performed on the GPU server of the High Performance Computing (HPC) Research Group of the Institute of Information Systems Engineering at Vienna University of Technology. The system specifications are stated in Table 5.1.

5.1.1 HBVEdu

Training (calibration) and evaluation of the PDM runs are performed on the CPU as the utilised library (RRMPG) does not support GPU computing and the effort to add this to the implementation would have been significant. On average, the calibration took 973.45 seconds in 35.9 iterations and the mean RMSE was 1.0888. For calibration, each local model per catchment is initialised with a random set of parameters, which are then optimised until the optimisation criterion of the differential evolution algorithm is met. Few catchments were not calibrated successfully (3 (0.7%) on average: reference: 6, hot2cold: 3, cold2hot: 1, wet2dry: 5, dry2wet: 0). These catchments failed to converge and exceeded the maximum number of allowed iterations, which had already been increased to 3,000 from an original 1,000. There appears to be a correlation between

Specification	Details
Server Name	hpcgpu1
Manufacturer	HPC Research Group
CPUs	AMD EPYC 7452 32-Core (x2)
Number of CPU Cores	32 (x2)
Base Clock	2.35 GHz (x2)
GPU	NVIDIA Quadro RTX 8000
CUDA Cores	4,608
Tensor Cores	576
VRAM	48 GB GDDR6
Operating System	Ubuntu 20.04.6

Table 5.1: Specification of the hpcgpu1 server of the HPC Research Group at Vienna University of Technology.

the length of the training period and the number of catchments that fail to converge. As the number of catchments failing calibration is so negligible over all five periods, they are simply discarded and no strategy is put in place to facilitate convergence.

5.1.2 XGBoost

The experiments for the ML model are performed on the GPU in a parallelised way. For model training, only the runtime for training the single model with the best parameter set according to the hyperparameter tuning is reported since the grid search was performed only a single time for all input sets.

5.1.3 EA-LSTM

The neural networks are implemented using the `NeuralHydrology` Python library and they are trained for 30 epochs each. One third of all catchments (144 out of 432) are randomly selected for validation at every five epochs during the training run. The NSE is used as the loss metric. The average loss at every epoch is shown in Figure 5.1. While there is some initial variation between the imputation methods, by epoch three at the latest, this is balanced out. The points at which the learning rate is set can be clearly identified from the plot by the jumps in the loss values (at epochs 10 and 20). After epoch 20, the loss seems to converge for all five reference period models, indicating that shorter training runs may result in similar performance. The wet2dry and cold2hot models exhibit the lowest losses, whilst those from the dry2wet period show the highest losses. To ensure reproducibility, the training random seed is set to 2,609.

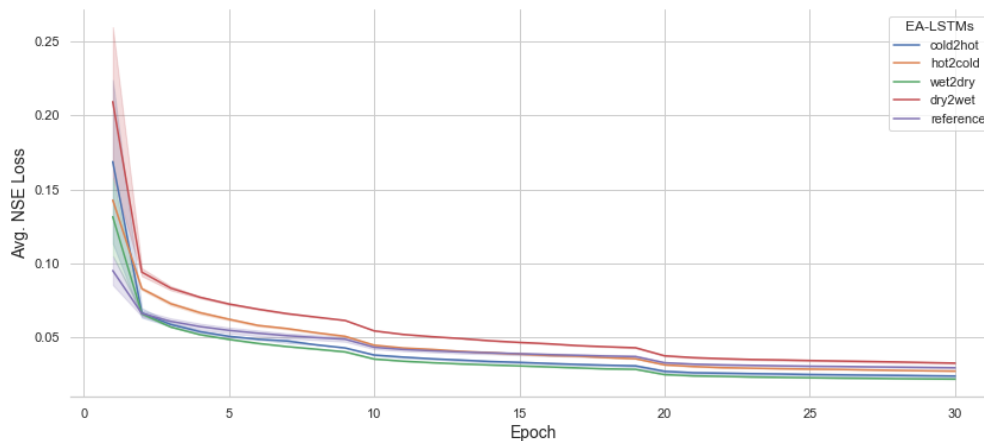


Figure 5.1: Average loss in NSE for the EA-LSTM training runs per reference period.

5.2 Test Results

5.2.1 reference

The reference period serves as a baseline data split and encompasses the earlier half of the data as the training and the later half as the test set. It uses by far the most samples during training, which suggests that this may lead to more accurate but also more overfitting results of the streamflow predictions. Furthermore, as models for this period are trained on data further in the past, where the effects of climate change on the study area may not have been as apparent as in more recent data, less robust performance on data covering climatically diverse conditions is to be expected. Evaluation results for all model combinations of this period are shown in Table 5.2.

	HBV		XGBoost		EA-LSTM	
	PCM	RF	PCM	RF	PCM	RF
MAE	0.7068	0.6986	0.7982	0.7397	0.5009	0.4778
RMSE	1.4149	1.3952	1.3647	1.3203	1.0752	1.0148
R ²	0.7574	0.7641	0.7674	0.7962	0.8664	0.8818
NSE	0.5556	0.5690	0.5763	0.6264	0.7491	0.7767
KGE	0.6191	0.6368	0.6921	0.7410	0.8109	0.8300
P _{BIAS}	16.3244	14.7541	-11.1658	-7.3564	4.7039	3.7570
Runtime (Train)	2,277.15 s	2193.26 s	168.44 s	159.14 s	70,476.34 s	69,190.06 s
Runtime (Test)	1.41 s	1.15 s	12.06 s	11.83 s	2,029.40 s	2,051.99 s

Table 5.2: Experiment results for the reference period on the respective test set.

With respect to the six evaluation metrics, the DL model EA-LSTM trained on the RF-imputed data achieves the best performance. The P_{BIAS} indicates a minor tendency to underestimate streamflow for EA-LSTM. HBVEdu shows a more pronounced proneness of underestimation. However, the XGBoost models clearly exhibit overestimation bias

with negative P_{BIAS} values. Overall, the EA-LSTM outperforms the other two models in all metrics, while XGBoost surpasses HBVEdu. This cautiously confirms the previous assumptions of the model results. Furthermore, the results of models trained on RF-imputed data are superior to models trained on PCM-imputed data across all model types and metrics. Regarding complexity (runtime), there is a clear difference between the model types. The ANN-based EA-LSTM is by far the most complex approach with training times of approximately 19.5 hours. In contrast, the HBVEdu models require approximately 37 minutes for training, whereas the XGBoost models can be fully trained in just 2.7 minutes.

5.2.2 hot2cold

The hot2cold set refers to the DSST period, for which the hydrological models are trained using data from a period that was $1.97\text{ }^{\circ}\text{C}$ warmer than the average (1,261 days). The used samples are taken from the years 2013 to 2016 for all catchments, which is in accordance to the assumption that the effects of climate change and the trend of warming examined in Section 3.4 are more pronounced in the more recent past. The training period is far shorter than that of the reference period. To the best of the author’s knowledge, the temperature has rarely been used as the driving reference variable in hydrological DSST experiments. Therefore, it is highly interesting to see how models trained on this input set and its counterpart cold2hot perform in contrast to the other DSST periods, which are driven by precipitation rather than temperature. The evaluation results are given in Table 5.3.

	HBV		XGBoost		EA-LSTM	
	PCM	RF	PCM	RF	PCM	RF
MAE	0.7336	0.7364	0.8957	0.8642	0.5205	0.5256
RMSE	1.4268	1.4185	1.4263	1.3934	1.0824	1.0604
R^2	0.7526	0.7599	0.7544	0.7663	0.8574	0.8656
NSE	0.5423	0.5529	0.5410	0.5673	0.7342	0.7479
KGE	0.6862	0.6839	0.6365	0.6552	0.8012	0.8130
P_{BIAS}	13.4619	14.8356	-19.2067	-16.3363	3.6598	4.1344
Runtime (Train)	655.55 s	650.74 s	44.28 s	46.54 s	12,988.04 s	13,217.60 s
Runtime (Test)	1.24 s	1.18 s	12.57 s	12.12 s	1,644.65 s	1,646.70 s

Table 5.3: Experiment results for the DSST period hot2cold on the respective test set.

In the context of the hot2cold (DSST) period, the EA-LSTM trained on PCM-imputed data again stands out as the top performer across the evaluation metrics. Notably, EA-LSTM achieves the best values for MAE, RMSE, and KGE, indicating superior accuracy and goodness of fit. Additionally, the model exhibits a modest positive P_{BIAS} , suggesting a slight tendency to overestimate streamflow. While the P_{BIAS} values for the HBVEdu models are similar in quality to those obtained from the reference period again exhibiting a slightly stronger tendency of underestimation than the EA-LSTMs, the XGBoost models are now far more prone to include overestimation bias in their runoff predicitions (P_{BIAS} decreasing from -11 to -19 and -7 to -16, respectively). Interestingly,

the EA-LSTM model trained on PCM-imputed data performs marginally better in terms of MAE and P_{BIAS} for this input set. A notable difference here is that HBVEdu actually outperforms XGBoost in terms of the KGE. This could be attributed to the fact that the KGE tends to underestimate peak flow to a lesser extent than the NSE. Since the XGBoost models, particularly those trained on the hot2cold set, are prone to overestimate flow, it is likely that the KGE is lower in this case.

In contrast, the XGBoost model trained on PCM-imputed data demonstrates the fastest training runtime, completing in just 44.28 seconds. The HBV models, although showing competitive performance, have longer training times, with EA-LSTM having the longest runtime, exceeding 3.5 hours. This underscores the trade-off between model complexity, training time, and predictive accuracy. The significantly shorter training times compared to the reference set are a consequence of the much smaller training sets. The results from the hot2cold period align with the overall trends observed in the reference period, reaffirming the effectiveness of EA-LSTM and the influence of imputation methods on model performance.

5.2.3 cold2hot

The cold2hot dataset acts as the counterpart to the hot2cold set and covers a period that was 1.69 °C colder than the average temperature of the study period under consideration. The training set covers a duration of 1,240 days in the years of 1984 to 1988. In contrast to the warmer reference period in the hot2cold set, the colder training period here is closer to the beginning of the study period, which again agrees with the general trend of warming. The evaluation results of the models for the cold2hot input set are stated in Table 5.4.

	HBV		XGBoost		EA-LSTM	
	PCM	RF	PCM	RF	PCM	RF
MAE	0.7519	0.7225	0.8647	0.8435	0.5765	0.5425
RMSE	1.4788	1.4346	1.4643	1.4303	1.1919	1.1141
R^2	0.7378	0.7558	0.7354	0.7580	0.8334	0.8569
NSE	0.5228	0.5557	0.5342	0.5603	0.6893	0.7312
KGE	0.6019	0.6320	0.6385	0.6791	0.7892	0.8312
P_{BIAS}	17.8697	15.1313	-8.9549	-11.9461	6.0481	1.3295
Runtime (Train)	624.42 s	596.68 s	46.46 s	44.45 s	12,878.03 s	12,887.18 s
Runtime (Test)	1.48 s	1.56 s	13.17 s	12.26 s	1,652.86 s	1,601.92 s

Table 5.4: Experiment results for the DSST period cold2hot on the respective test set.

The performance of models built on the cold2hot input set conforms to the already observed trend and order of the above reported DSST periods. The EA-LSTM trained on PCM-imputed data consistently outperforms the other models. These results consolidate the assumption of general model robustness of the ANN model EA-LSTM across various climatic conditions. It is particularly worth emphasising that the EA-LSTM models

exhibit P_{BIAS} values that are very close to zero (≈ 1.33 for RF), indicating near-optimal prediction of runoff.

The XGBoost models demonstrate competitive results, but they again show clear overestimation bias. On the other hand, the cold-to-hot transition poses challenges for HBV models, particularly when trained on PCM-imputed data, resulting in increased values of MAE, RMSE, and P_{BIAS} . The analysis of runtime indicates that XGBoost maintains the shortest training time, demonstrating its effectiveness while still performing reasonably well. The trends recognised in the period when temperatures transitioned from cold to hot correspond with those in the reference and hot2cold periods.

5.2.4 wet2dry

The wet2dry input set consists of training data collected over a 1,438-day period between 1997 and 2001 that experienced 0.27 mm more precipitation per day than the average daily P . In contrast to the hot2cold and cold2hot sets, both wet2dry and its counterpart dry2wet are driven by P instead of T . They correspond to the previous experiments with DSST setups to examine the robustness of models among varying climatic conditions by O et al. and Coron et al. [ODO20, CAP⁺12]. Since Coron et al. observed a tendency to overestimate runoff for calibration periods with higher precipitation levels, it is of interest to see whether their results can be reproduced in the experiments. The performance metrics of this DSST period are given in Table 5.5.

	HBV		XGBoost		EA-LSTM	
	PCM	RF	PCM	RF	PCM	RF
MAE	0.7035	0.6929	0.7982	0.7878	0.5098	0.4903
RMSE	1.3768	1.3614	1.3647	1.3422	1.0542	1.0219
R ²	0.7548	0.7662	0.7674	0.7795	0.8655	0.8760
NSE	0.5526	0.5687	0.5763	0.5955	0.7457	0.7640
KGE	0.6487	0.6531	0.6921	0.7034	0.8424	0.8549
P_{BIAS}	15.0231	15.8401	-11.1658	-11.3313	2.1391	2.8575
Runtime (Train)	881.19 s	889.30 s	52.11 s	48.21 s	15,283.87 s	15,604.86 s
Runtime (Test)	1.29 s	1.42 s	13.68 s	13.11 s	1,634.21	1,619.43 s

Table 5.5: Experiment results for the DSST period wet2dry on the respective test set.

Aligning with previously observed performance metrics, the EA-LSTM trained on PCM-imputed data maintains its superior adaptability, achieving optimal results on all key metrics. In accordance with the other reported results, the P_{BIAS} values imply a probable underestimation of runoff for the physics-based HBVEdu models. In contrast, XGBoost models exhibit a tendency to overestimate, which is in line with the results reported by Coron et al., whereas EA-LSTM maintains a well-balanced prediction [CAP⁺12]. Again, the HBV models, especially when trained on PCM-imputed data, show difficulties in adapting to the wet-to-dry transition. The performance of XGBoost models is slightly improved compared to previous results (e.g. NSE increased from 0.560 to 0.596 compared to cold2hot).

Interestingly, the training runtimes of the HBVEdu models exhibits an increase of about 84% compared to those of the dry2wet period, although the training period for wet2dry is only 167 days (13% increase) longer. The EA-LSTM models also show an increase in runtime, but to an extent (14%) that suggests a linear relationship with the increase in days of the training set. No increase of runtime can be observed for the XGBoost models. The consistency of wet2dry results with previously observed trends adds confidence to the robustness of the models in capturing complex hydrological dynamics.

5.2.5 dry2wet

In contrast to the wet2dry set, the 1,271-day dry2wet DSST period from 1989 to 1993 includes training data from a period when daily precipitation was 0.30 mm below average. Coron et al. found that drier calibration periods lead to underestimation bias in runoff prediction [CAP⁺12]. The results of the dry2wet training and evaluation are stated in Table 5.6.

	HBV		XGBoost		EA-LSTM	
	PCM	RF	PCM	RF	PCM	RF
MAE	0.7549	0.7372	0.7738	0.7578	0.5549	0.5202
RMSE	1.4328	1.4140	1.3811	1.3560	1.1168	1.0645
R ²	0.7526	0.7629	0.7637	0.7762	0.8540	0.8692
NSE	0.5512	0.5676	0.5830	0.6023	0.7260	0.7536
KGE	0.6663	0.6813	0.6548	0.6833	0.8137	0.8258
P _{BIAS}	12.8554	12.3571	-0.1697	-1.6656	4.8194	4.5077
Runtime (Train)	491.09 s	472.37 s	49.19 s	43.50 s	13,178.41 s	13,303.06 s
Runtime (Test)	1.20 s	1.60 s	13.96 s	12.20 s	1,925.73 s	1,669.94 s

Table 5.6: Experiment results for the DSST period dry2wet on the respective test set.

The order of model performance for the dry2wet input set is again unchanged to the previously observed results: EA-LSTM clearly outperforms XGBoost and HBVEdu (NSE of 0.75 compared to 0.60 and 0.56, respectively). XGBoost again exhibits the shortest training runtimes. Notably, both XGBoost models achieve P_{BIAS} values very close to zero, even outperforming the values reported for EA-LSTM for cold2hot. Generally, these values suggest that the model can perfectly estimate (peak) flow. However, paired with the low KGE values, these results contradict those observed for hot2cold, where HBVEdu also matched the performance of XGBoost. It is interesting that similarly to the metrics reported for hot2cold, HBVEdu performs equally well (KGE) if not slightly better (MAE) compared to XGBoost. The results underscore the robustness and versatility of EA-LSTM across different hydrological periods, supporting its potential as a reliable model for capturing complex hydrological dynamics.

5.2.6 Summary of the Test Results

Although the test results for the baseline reference period showed the best performance, the difference in key metrics to the four DSST periods is not as significant as previously

assumed. The variations are apparent in Figure 5.2, which shows a box-plot of the NSE values for each period across all models. The reference period (Median $NSE \approx 0.62$) clearly outperforms the others with the two pairs driven by meteorological variables - hot2cold and cold2hot (Median $NSE \approx 0.56$), as well as wet2dry and dry2wet (Median $NSE \approx 0.59$) - each performing similarly among each other. Another difference in performance can be taken from the plot: models built on both DSST sets driven by P (wet2dry and dry2wet) seem to outperform those driven by T . This hypothesis is examined in Section 5.4.3.

	HBVEdu	XGBoost	EA-LSTM
MAE	0.7238	0.8073	0.5219
RMSE	1.4154	1.3811	1.0796
R^2	0.7564	0.7687	0.8626
NSE	0.5538	0.5805	0.7418
KGE	0.6509	0.6812	0.8212
P_{BIAS}	14.8453	-9.5122	3.7957
Runtime (Train)	973.17 s	70.23 s	24,900.74 s
Runtime (Test)	1.35 s	12.69 s	1,747.68 s

Table 5.7: Mean performance metrics for each model (trained on RF-imputed data) across all reference periods on the respective test sets.

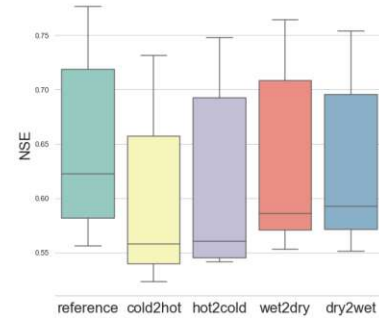


Figure 5.2: Boxplot of the distribution of NSE values for the DSST periods on the test sets.

Table 5.7 shows the mean metric results per hydrological model type. As observed for all examined train/test sets, the DL model EA-LSTM achieves superior performance compared to the PDM and the ML model measured across all six evaluation metrics. Particularly the values for the domain-specific metrics NSE and KGE can be considered “Good” and “Very Good”, respectively, with respect to the evaluation guideline proposed in Section 4.3.2. In terms of accuracy metrics, XGBoost consistently outperforms HBVEdu. Specifically, XGBoost demonstrates lower MAE and RMSE values, indicating superior precision in predicting streamflow. However, when it comes to metrics assessing goodness of fit, such as NSE) and KGE, HBVEdu tends to perform competitively with or even surpass XGBoost. This suggests that while XGBoost excels in precision, HBVEdu may exhibit comparable or better overall performance in capturing the variability of observed data. It is noteworthy that EA-LSTM seems to handle over- and underestimation bias well with values in the positive range close to zero, whereas HBVEdu clearly underestimates runoff with consistent values around 15. XGBoost has a unique pattern in P_{BIAS} that fluctuates in the negative range, indicating issues with the representation of overestimation and underestimation. It is to be examined how well the models actually generalise.

There is a clear difference in runtime between the three model types. The DL model is the most computationally expensive with an average training time of approximately 6.9 hours. XGBoost consistently achieves the shortest training times of only around 70 seconds, which is 355 times shorter than the mean training times of EA-LSTM. However, an intra-model comparison of runtimes does not reveal major differences among the

various input sets. The most striking outlier is the reference set described in Table 5.2 with a training time exceeding that of the models trained on DSST periods by several magnitudes for each model type. However, this can be attributed to the training set of the reference period consisting of more than five times as many samples compared to the other periods. Furthermore, the test set of the reference period contains only about half as many samples as the other test sets, explaining the much smaller difference in test runtimes among the reference period and the DSST sets.

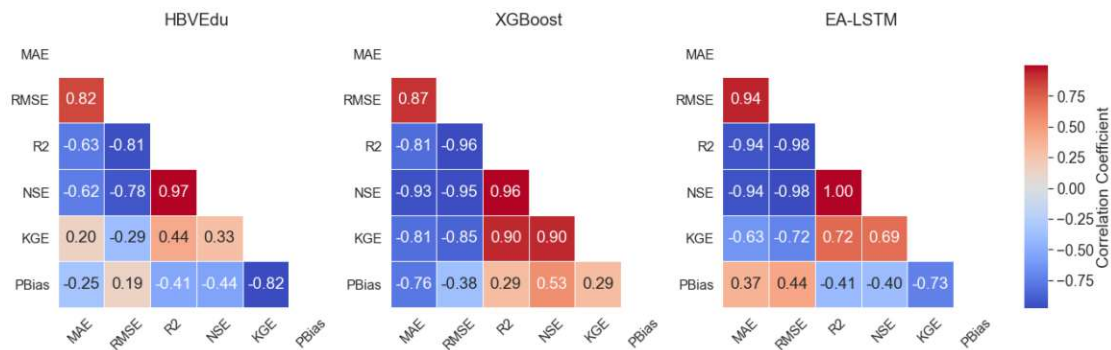


Figure 5.3: Correlation heat-maps depicting the relationships among key performance metrics for the test set results among the three hydrological models (HBVEdu, XGBoost, EA-LSTM).

Figure 5.3 shows the correlation matrices for HBVEdu, XGBoost, and EA-LSTM. The patterns in the relationships between performance metrics reveal unique characteristics of each model. Naturally, there are strong positive correlations between MAE and RMSE for all models. Additionally, lower absolute errors are typically associated with higher coefficients of determination, NSE, and a balanced representation of overestimation and underestimation. Similarly, in the case of XGBoost, higher goodness-of-fit metrics and efficiency measures are associated with lower absolute errors. The strong negative correlations between P_{BIAS} and NSE as well as KGE highlight the interplay between balanced error representation and efficiency. It is important to note that the correlation between P_{BIAS} and MAE can be positive, even when models have higher absolute errors.

5.3 Validation Results

Since the test period differs for each of the DSST sets, it is of interest to report and analyse the performance metrics for validation period, which is the same for all models. The validation period consists of 366 days from 1st January 2017 until 1st January 2018. Therefore, the models are validated on the most recent year of data, where the hydrological and meteorological variability is expected to be at its highest across the study period due to the increasingly prevalent impacts of climate change [EEA23]. These results are thereby comparable not only across models, but also across all five climatic periods. As demonstrated empirically in Section 5.2, using RF regression for missing

5. RESULTS AND DISCUSSION

data imputation leads to better performance across all metrics. Therefore, model results are reported for these input sets for all five periods. The performance metrics are shown in Table 5.8.

Period	Model	MAE	RMSE	R ²	NSE	KGE	P _{BIAS}	Runtime
Ref	HBVEdu	0.7622	1.4121	0.7657	0.4779	0.4319	34.9765	0.97 s
	XGBoost	0.7169	1.2553	0.7803	0.6021	0.7276	-5.4138	5.41 s
	EA-LSTM	0.4741	0.9695	0.8754	0.7627	0.8082	7.1817	95.96 s
h2c	HBVEdu	0.7435	1.3943	0.7760	0.5116	0.4736	32.9673	0.98 s
	XGBoost	0.7389	1.2761	0.7677	0.5888	0.6727	-2.7179	3.98 s
	EA-LSTM	0.4758	0.9976	0.8698	0.7487	0.7777	10.3123	97.64 s
c2h	HBVEdu	0.7859	1.4502	0.7521	0.4667	0.4308	33.7535	1.29 s
	XGBoost	0.8978	1.4339	0.7055	0.4808	0.6147	-12.3145	4.03 s
	EA-LSTM	0.5619	1.0969	0.8369	0.6962	0.8026	3.5861	98.59 s
w2d	HBVEdu	0.7559	1.4089	0.7672	0.4813	0.4372	35.4326	1.32 s
	XGBoost	0.7809	1.3011	0.7641	0.5726	0.6767	-11.3840	4.16 s
	EA-LSTM	0.4817	1.0171	0.8635	0.7388	0.7836	9.8048	98.64 s
d2w	HBVEdu	0.7694	1.4227	0.7567	0.4889	0.4643	31.7442	1.29 s
	XGBoost	0.7433	1.3035	0.7558	0.5710	0.6646	-0.5933	3.98 s
	EA-LSTM	0.5053	1.0381	0.8550	0.7279	0.8064	5.5842	97.54 s

Table 5.8: Validation results for all DSST periods reported for the best performing models for each period.

The DL model EA-LSTM trained on the baseline reference period reaches the best results for all metrics except P_{BIAS} and the runtime. The errors are consistently lower compared to the other models and climatic periods. Figure 5.4 shows bar plots for key metrics RMSE, NSE, and P_{BIAS} across all periods for each model type. The visualised results emphasise the superiority of the EA-LSTM across all experiments. A notable variation in results is that the EA-LSTM reports marginally worse results in metrics RMSE and NSE for the cold2hot period in comparison to the other periods. EA-LSTMs trained on the DSST periods show competitive performance compared to the baseline. Overall, the DL model trained on the hot2cold most closely matches the performance of the baseline.

The superiority in performance of the DL model is clearly apparent across all metrics, except for the P_{BIAS} values. Here, the ML model XGBoost achieves the best value for the dry2wet set (-0.59), which indicates almost optimal estimation of flow. Generally, the overall order of performance observed for the test results extends to the validation. A noteworthy observation is that the P_{BIAS} validation values of the HBVEdu models are significantly higher compared to the test results (15 to 33), which points to a drastic underestimation of flow. This further indicates that the locally calibrated PDMs overfit to the training data and struggle to generalise well to new data. Model robustness is adversely affected.

The Cumulative Distribution Function (CDF) plot for NSE and RMSE for the three model types in Figure 5.5 sheds light on the distribution of these key metrics across all

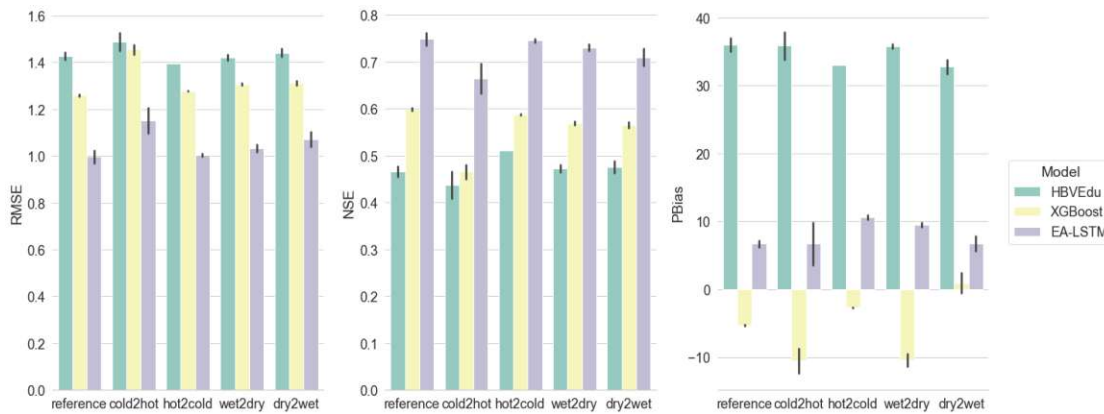


Figure 5.4: Key evaluation metrics from the validation results of the models for all reference periods.

individual catchments. XGBoost and EA-LSTM models trained on the cold2hot show consistently better distributions for both metrics. For NSE values, there is significant variability in value distribution in the lower percentile of the value range, except for HBVEdu. The PDM shows very similar metric distribution across the whole value ranges, especially for the RMSE. The only difference here is that the hot2cold HBVEdu models appear to have lower NSE values.

The relative change in metrics between validation and test results, stated in Table 5.9, underscores the decrease in performance of the PDM while both XGBoost and EA-LSTM achieve similar results with marginal deviations. Notably, the issue of underestimation of flow exhibited by EA-LSTM worsens for EA-LSTM with an increase of the P_{BIAS} value by $\approx 120\%$, which is at similar scale to the increase reported for the HBVEdu model. The ML and DL models exhibit comparable improvements in terms of accuracy metrics, but at the same time deterioration in terms of measures describing the goodness of the model fit. Regarding the relative change in performance for the reference periods, there is a noticeable improvement for models trained on the hot2cold input set. In contrast, the cold2hot models show deteriorated performance, aligning with previously observed results. The most drastic deviation is the increase in P_{BIAS} of 1,440 %. The KGE values decreased for all five sets.

As expected, the PDM has the shortest runtime for the validation runs with an average of 1.17 seconds, which is 3.7 times faster than XGBoost (4.32 s) and 83.5 times faster than EA-LSTM (97.66 s).

In summary, model robustness of the DL model EA-LSTM seems not to be significantly affected by shorter training periods driven by a high degree of climatic variability. The hydrological key metric NSE decreases by only 2.63% compared to the test results, indicating that the model is able to generalise well to data with very different conditions compared to the training period. XGBoost also performs reasonably well on the validation set and achieves competitive results. However, the physics-based model HBVEdu has

5. RESULTS AND DISCUSSION

	MAE	RMSE	R ²	NSE	KGE	P _{BIAS}
HBVEdu	+6.39 %	+0.92 %	+0.23 %	-13.77 %	-31.92 %	+131.59 %
XGBoost	-2.88 %	-3.98 %	-2.65 %	-4.63 %	-3.06 %	-33.33 %
EA-LSTM	-2.25 %	-2.97 %	-1.13 %	-2.63 %	-4.24 %	+119.88 %
reference	+1.93 %	-2.50 %	-0.85 %	-6.56 %	-10.87 %	+229.41 %
cold2hot	+6.50 %	+0.05 %	-3.22 %	-11.02 %	-13.73 %	+454.30 %
hot2cold	-7.90 %	-5.27 %	+0.90 %	-1.02 %	-10.60 %	+1,440.15 %
wet2dry	+2.41 %	+0.04 %	-1.11 %	-7.03 %	-14.19 %	+359.57 %
dry2wet	+0.14 %	-1.83 %	-1.70 %	-7.05 %	-11.65 %	+141.69 %

Table 5.9: Percentage change in mean performance metrics between validation and test results. Upper section: model types. Lower section: DSST periods.

significant issues to generalise to unseen data, adversely affecting its robustness with respect to climatic variability in the time series data.

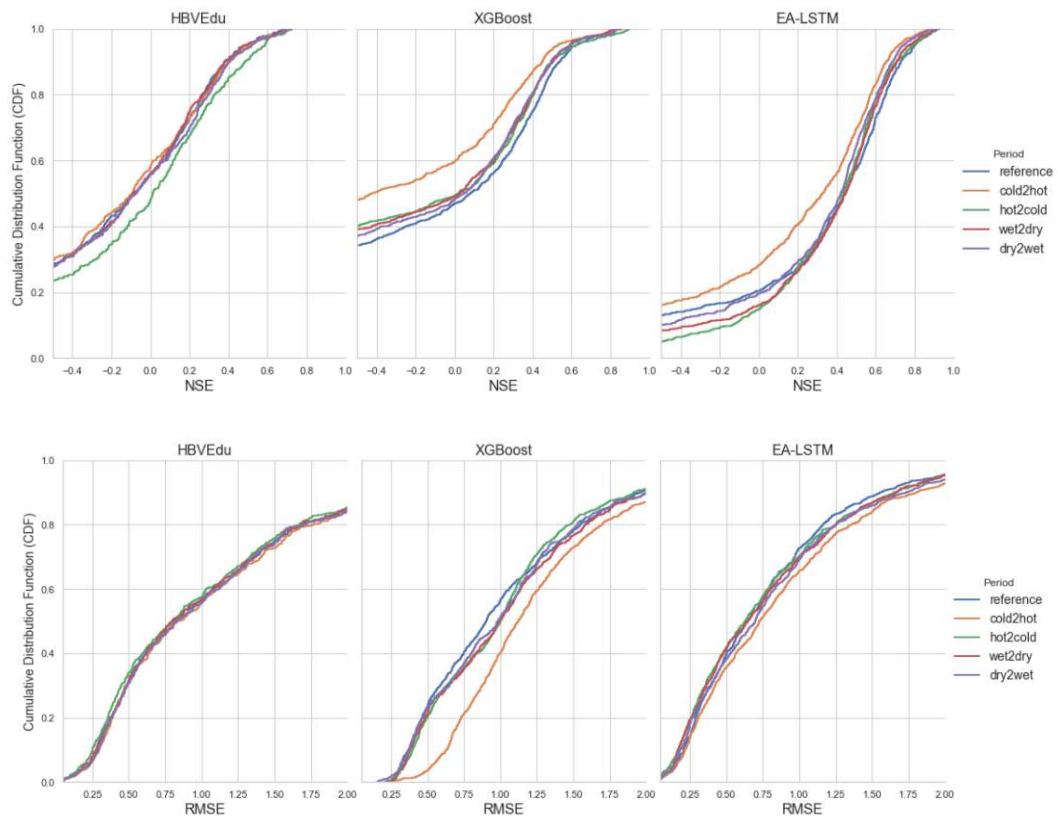


Figure 5.5: Cumulative distribution function of the NSE (top) and RMSE (bottom) values of HBVEdu, XGBoost, and EA-LSTM.

Figure 5.6 shows the observed and simulated streamflow results for the validation period. The streamflow during the year 2017 is depicted for the high-deviation catchment at Berninabach at Pontresina, already examined in detail as part of the analysis of exemplary catchments in Section 3.4. The hydrographs are plotted for each of the five climatic reference periods. The analysis of this plot reveals the ability of the EA-LSTM models to fit closely to the curve, which is subject to strong temporal deviations. The models are generally capable of accurately representing peakflow and adhering to periods of extended baseflow, which is prevalent at the beginning and at the end of the validation period. Judging from the plot, the EA-LSTMs appear to prematurely detect local maxima. Furthermore, the reported tendency of underestimation is confirmed by the consistently lower amplitude of predicted peakflow. The only striking difference between reference periods is that the model trained on the cold2hot dataset seems to suffer from underestimation bias to the least degree and most accurately matches maxima, as well as minima. The DL approach suggests a high degree of robustness to climatic variability and a sufficient ability to generalise across all periods.

The picture is different for the ML model and the PDM. HBVEdu significantly underestimates flow and fails to capture important peaks. Notably, baseflow is underestimated and set to zero for the first 100 days with no variance at all. Flow variations during the summer months are not well fitted, with only the first peak at the beginning of June, and that too early by at least one week.

On the other hand, the XGBoost models appear to be drastically underfitting and suggest high variance. The predictions seem to be highly sensitive to fluctuations and are noisy throughout the validation period. Peakflow is not detected. Interestingly, given the P_{BIAS} values indicating a general overestimation bias in all experiments, the model actually underestimates periods of peakflow and overestimates only baseflow.

5.4 Discussion

Based on the literature review on hydrological modelling, the domain-specific datasets, the required steps in data analysis and pre-processing, the experimental design and the results of the rainfall-runoff modelling, the research objectives of this thesis can now be addressed and put into context.

5.4.1 Large-Sample Hydrology and Data Engineering

To address **RQ 1**, the current state-of-the-art in the field of LSH is examined. The progress of complex interdisciplinary modelling in hydrology, often rooted in different paradigms, such as Machine Learning, entails the provision of high-quality data collections that adhere to modern standards and meet new, previously overlooked requirements. Traditional hydrological datasets are often proprietary in nature, collected for a specific application or research without the use of consistent data formats, often in violation of the FAIR principle, and sometimes not publicly accessible either due

5. RESULTS AND DISCUSSION

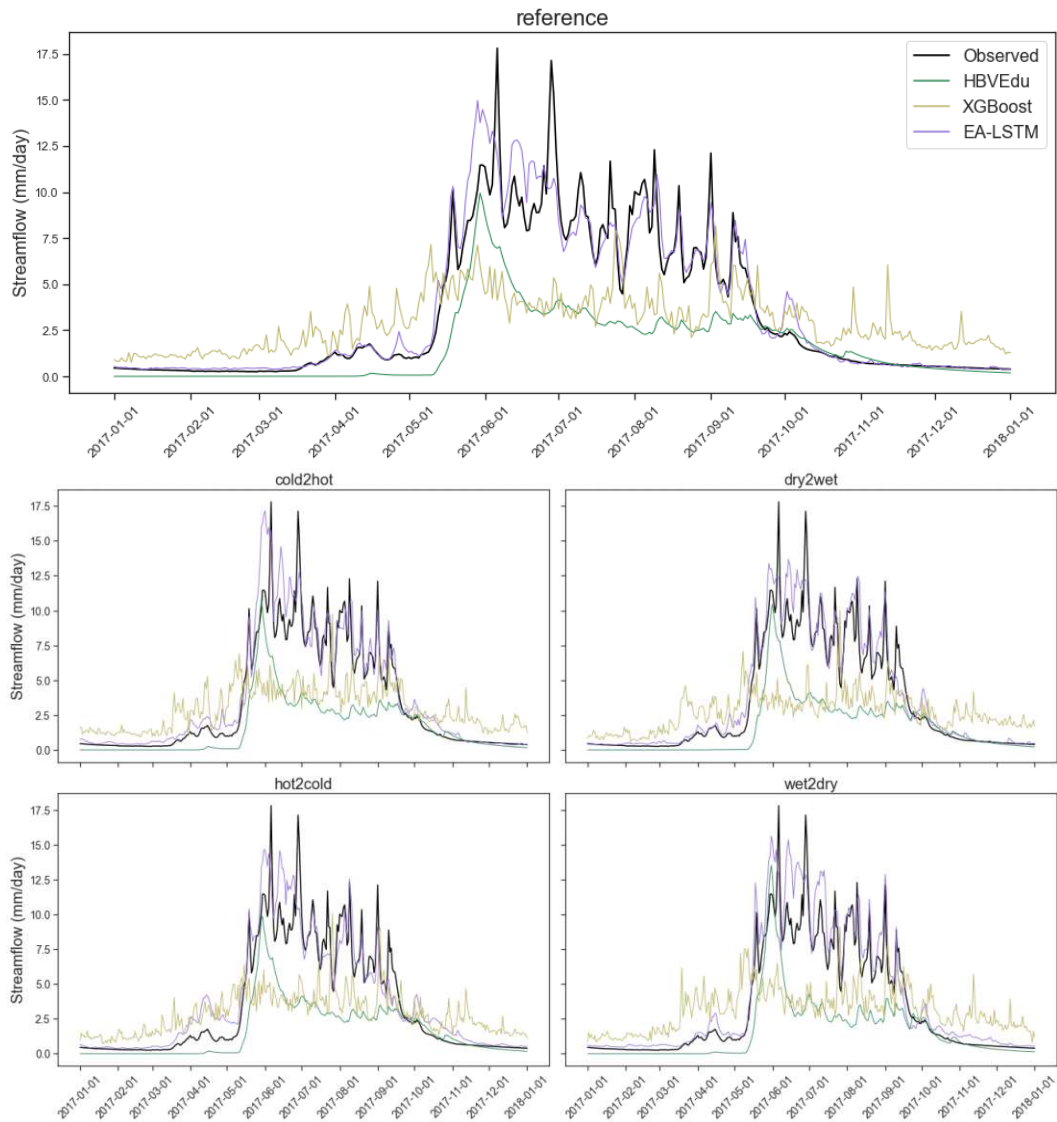


Figure 5.6: Observed and simulated streamflow results for the validation period of the high-deviation catchment located at Berninabach at Pontresina, Switzerland (lamah_2262). Values are plotted for each DSST period and each model.

to regulatory concerns or negligence. Furthermore, the available datasets differ significantly in their coverage, resolution and several key characteristics, such as the applied catchment delineation strategies and catchment differentiation approach, location and local/regional/national/continental/global scale, hydro-meteorological variables provided, meteorological forcings used, temporal-spatial resolution and catchment attributes available. Therefore, presented research results are often difficult to reproduce and compare, especially among different catchments. Blöschl et al. and Beck et al. conducted important

studies in hydrological modelling. However, the data is only partly available to the public, which means that the modelling results cannot be fully reproduced. This illustrates the issue of reproducibility in hydrological modelling [BHV⁺19, BPL⁺20].

The above-mentioned differences and the requirements of modern hydrological modelling lead to the need to establish consistent, comparable, extensible, and available open source data collections in the field of LSH, following the FAIR principle. A 2019 landmark study by Addor et al. presents an assessment of the state-of-the-art in LSH as well as key limitations, requirements, and opportunities in the domain. The authors notably mention the lack of common standards and metadata to facilitate comparability, missing estimates of anthropogenic impacts and violations of the FAIR principle as prevalent limitations. According to the authors, the most vital requirements for new datasets are the use of consistent data formats, providing data publicly available, open-source ways, reporting uncertainty estimates, and presenting human impact factors [ADAG⁺20].

With the limitations and suggestions presented by Addor et al. in mind, this work analyses and compares the most prevalent datasets used in both research and applications in the domain of LSH. There have been attempts towards establishing comparable, large-scale national data collections as early as 2006 with the development of MOPEX. This dataset already covered key requirements of LSH as it is publicly available and incorporates the most important hydro-meteorological variables at a high spatio-temporal resolution across 55 years. Further cornerstone data collections within the domain include CAMELS, EWA, GRDB, and GSIM [SCD06, NCS⁺15, DGLW18, ADAG⁺20].

The Caravan collection is the most recent addition to the domain of LSH and can be considered the most comprehensive and up-to-date collection of hydrological data yet. Kratzert et al. introduced this global data collection in 2023 specifically with the suggestions of Addor et al. in mind and took into account the emergence of DDMs in hydrological modelling based on their own experience with the use of ANNs in the domain [KNA⁺23, KKS⁺19]. The collection currently comprises the state-of-the-art hydrological datasets HYSETS, CAMELS (five subsets to date), and LamaH. Data is available as versioned packages from Zenodo, GitHub or the cloud platform Google Earth Engine with a permissive license, and covers several thousand highly diverse catchments across the world. The period covered by the time series is the same for each subset and currently covers 39 years from 1980. A major benefit of Caravan is its design as an open-source software, providing well-documented code to extend the data, generate catchment attributes, and reproduce experiments performed on the data. By including an extensive set of more than 200 static attributes for each catchment including anthropogenic influences on the basin and its surrounding area, the authors fulfil the requirement to report on human influence on catchments raised by Addor et al. for the first time [ADAG⁺20]. The original datasets from various sources are subject to standardised data processing and formatting procedures, with the scripts to perform these steps available as part of the code repository. This allows for consistent and comparable use of data no matter their origin across a highly heterogeneous set of catchments. By providing open-source interfaces and tutorials to extend the collection in a standardised

way not only from members of the research community and official hydrological services, but from any individual who has access to high-quality data, the authors account for the needs of large-scale data collections and potentially facilitate a faster growth of hydro-meteorological datasets in previously underrepresented areas, following the FAIR principle. The collection does not provide estimates of uncertainty related to the data. In summary, Caravan can be considered a milestone in LSH, representing an open platform for the hydrology community. The issue of reproducibility in hydrological modelling is addressed by providing the data collection as open source by design, in a versioned, extensible and frequently updated publicly accessible manner. Therefore, Caravan serves as a good candidate data collection in the domain of LSH.

The application of data-driven models in rainfall-runoff modelling has changed the requirements for DS methodologies in hydrology. Key findings from experiments with different modelling paradigms highlight the importance of having data in a consistent, non-proprietary and model-agnostic format. This minimises the need for time-consuming data wrangling prior to processing or modelling. Specific needs of DS researchers require simple interfaces such as Caravan that provide high quality, standardised data collections.

Access to public, freely available data platforms and open-source modelling software, such as the Python package `NeuralHydrology`, is crucial for successful DDM experiments in LSH. `NeuralHydrology` streamlines the data engineering process, supports different DDM architectures, and allows customisation through a comprehensive configuration file. However, there is a need to improve and consolidate existing open-source software solutions for PDM experiments. The current state of PDM software is characterised by inconsistency and incompatibility, making inter- and intra-model comparability a challenge. Despite the issues, the Python package `RRMPG` was chosen for its relatively well-designed API and ease of use, although it has limitations such as lack of support for popular model architectures and GPU computation. Future steps could include integrating successful models into comprehensive platforms such as `NeuralHydrology` for efficient hydrological modelling across different paradigms.

Data exploration and pre-processing are exceptionally resource-intensive and complex tasks in hydrological modelling. The steps undertaken to prepare the data for modelling experiments in this work include initial explorative analysis and cleaning, splitting into training/test/validation sets with respect to baseline and climatic reference periods to allow for Differential Split-Sample Testing (DSST) later on, missing data imputation, outlier detection and imputation, feature selection, and data standardisation. Each of these steps requires comprehensive analysis and research to find suitable strategies and methods that can be applied to the data and study area at hand. Particularly in the field of hydrology, it is necessary to incorporate domain-specific knowledge into these decisions. Sharma et al. confirm that, generally, “the relative effects of preprocessing and postprocessing depend strongly on the forecasting system (e.g., forcing, hydrological model, statistical processing technique), and conditions (e.g., lead time, study area, season), underscoring the research need to rigorously verify and benchmark new forecasting systems that incorporate statistical processing” [SSR⁺18].

Domain-specific issues, such as missing streamflow data in large-scale collections, are a major challenge for researchers. Tencaliec et al. state that technical or maintenance issues, damaged gauging stations, e.g. during flood events, and the complex tasks of long hydro-metric data production and management can lead to intervals of missing data in streamflow records. The authors conclude that this entails information loss and incorrect interpretation of the data or unreliable analysis and research communication [TFPM15]. Few of the state-of-the-art, yet highly sophisticated methods are easily accessible to the modelling community as they are typically not implemented as (open-source) software. Custom implementation of such techniques can significantly increase the workload of pre-processing the data. Therefore, it is easier to resort to well-known non-proprietary imputation methods from data-driven domains. However, it would be highly beneficial to leverage the potential of the domain-informed methods introduced in the literature. The same issue applies to the detection and management of outliers. This preparation step includes the additional requirement of safely discarding anomalies while at the same time keeping naturally occurring phenomena in the data.

Performing Multiple Imputation (MI) can be computationally expensive, and does not scale well with increasing number of samples and employed estimators. In the presented work, imputation was necessary for the extrapolation of more than 700,000 samples with missing streamflow records, as well as for several ten-thousand samples marked as outliers. This led to high computational load and thus long runtimes for pre-processing, which was aggravated due to the five separate input sets each having to undergo data preparation separately to avoid data leakage. Furthermore, since MI produces distinct datasets for each imputation strategy or random seed, each resulting set must undergo subsequent pre-processing steps separately as well, further increasing the workload. Naturally, all resulting sets must then be input to the hydrological models, which can be infeasible due to the computational cost of some models (e.g. ANNs). Adhering to these principles of data-driven modelling guarantees scientific soundness of the reported results, but also leads to exploding computational load, vast numbers of hydrological models to evaluate and compare, and to imprecise analyses and conclusions. Uncertainty estimates are necessary and part of good scientific practise, but they are costly and resource-intensive. An important step in consolidating the pre- and post-processing strategies in hydrological modelling is to facilitate the implementation of imputation and subsequent uncertainty analysis to get a reproducible, comparable understanding of the robustness of models. However, uncertainty has to be accounted for at a large scale in rainfall-runoff modelling as a model is only a simplified representation of physical hydrological processes, and uncertainty is inherent to the target variable streamflow due to its instability and proneness to fluctuations. Every model is therefore uncertain to some extent and every step of the modelling process can introduce more imbalance. Modelling in this domain must acknowledge the significant issue of uncertainty [SKP⁺18, MMCD21].

Furthermore, the strategies employed to split the data in the context of Differential Split-Sample Testing (DSST) for climate change modelling are not yet represented comprehensively in literature and appear to be an open issue in current research. For this

work, custom algorithms (see Algorithms C.1 and C.2) are devised to find and aggregate consecutive periods of significant deviations in key hydrological variables (P and T). It would be useful to rely on a scientifically agreed upon framework for DSST in hydrology and design reference periods accordingly [CAP⁺12].

5.4.2 Hydrological Modelling

As shown in Chapter 5, the EA-LSTM by Kratzert et al. performs well for modelling the runoff of the LamaH study area in Central Europe. The model design incorporates catchment-specific information into the architecture. Time series data is processed conditionally based on the catchment it belongs to and can thus tailor the predictions to produce a single universal forecasting system for all basins in the study area. The model is capable of activating parts of its network based on the processed catchment as shown in Figure 4.3, which contributes to the high degree of flexibility and accuracy.

In general, the set up, configuration, implementation, and evaluation of the EA-LSTM model using the `NeuralHydrology` package by Kratzert et al. is a straightforward process [KGNK22]. There are only few model-specific pre-processing steps required to prepare the data. One proprietary step is to remove features with a standard deviation of zero. However, this step is beneficial for all types of models since such features would be redundant as input. A further necessary step is to create and fill the configuration file, which at times requires data in very specific formats. For instance, basin identification numbers and start and end dates of non-continuous training, test, and validation periods must be in separate files of pre-defined formats. Furthermore, the configuration of the LSTM architecture, such as the number of neurons per fully-connected layer, learning rates at specific epochs, the activation function of the embedded network, or the applied dropout require significant effort to be fine-tuned to the available data. However, default architectures can already result in good performance. For the experiments in this work, the parameters were tuned to the values presented in the original publication of the EA-LSTM with only little adaptation. This architecture leads to well-fitting predictions, which are reasonably sensitive to peakflow, robust to transient climatic conditions, and superior to a PDM and a traditional ML model (see Section 5.4.3).

The process to set up the DL model used in this work does not significantly differ from the configuration of a hyperparameter-rich ML model, such as XGBoost. Tuning the parameters of such models is a highly complex process, typically performed in a grid or random search with pre-defined value ranges. The computational cost associated with a hyperparameter search can be a major limiting factor in achieving accurate model results. For example, the assessment of 11,664 parameter combinations in a cross-validated grid search resulted in a runtime of almost 19 hours for a single training set. Performing this search for all sets separately and for more diverse value ranges would have been infeasible for this work.

The flexibility, versatility, and robustness of LSTM models marks them as suitable architectures for countless inter-disciplinary tasks, one of them being rainfall-runoff

modelling. The possible volume and dimension of the input data and the relationships represented by these models, as well as their ability to use internal memory structures and conditional processing of information, qualify them for extensive use in such tasks. The accessibility of LSTM models is improved in comparison to PDMs as they do not require domain knowledge and rather consider highly complex problems from the Earth Sciences as multi-variate classification or regression tasks.

In the case of rainfall-runoff modelling, the most important step to apply the EA-LSTM is the thorough pre-processing of data, which, however, applies to all modelling tasks just the same. The model-specific configuration and set up of the programming environment negligible steps in comparison to the data preparation, which is partly due to the well-documented source code of the used library `NeuralHydrology`. However, while the use of this library significantly reduced the workload of the ANN-based experiments in this work, designing a similar neural network architecture using state-of-the-art software packages such as `TensorFlow` or `PyTorch` would have been relatively straightforward. Nevertheless, it is essential to have access to a modern GPU at the workstation level in order to leverage their computational power for numerical calculations. Relying on a CPU without using GPU acceleration for neural network modelling is not feasible for large-scale input datasets, which are prevalent in most geo-science domains, such as hydrology. The experiments performed in this work would not have been possible without access to the GPU server of the High Performance Computing Lab at Vienna University of Technology. The runtimes reported for EA-LSTM training and evaluation exceed those of the PDM and the ML models by several magnitudes. Yet, the improvement in accuracy seems to be worth the computational cost.

In summary, the EA-LSTM model can be considered a state-of-the-art approach to hydrological modelling, achieving great predictive power and featuring simple configuration as well as high degrees of flexibility and versatility. While PDMs have been prevalent in the past, the success of DDMs have now made them the state-of-the-art approach to modelling in the domain. In future research, the performance and robustness of hybrid model that combine process- as well as data-driven architectures should be tested and analysed. Models such as the mass-conserving LSTM have already uncovered notable hidden relationships in input data, which would not have been discovered without the incorporation of physical laws into the model architecture [HKK⁺21]. The potential of such models could further contribute to leveraging the power of large amounts of data and producing accurate results in rainfall-runoff modelling, further increasing the importance and reliability of forecasts for other applications such as flood predictions. Additionally, the robustness of LSTM-based models to varying climatic conditions is evaluated in this work and is found to be significant. However, more research into DSST experiments with different types of models is needed to better understand potential shifts in predictive power and accuracy as the impacts of climate change increase in vulnerable regions.

While the architectures, configurations, training and calibration procedures, and computational characteristics of process-driven and data-driven rainfall-runoff models differ significantly between model types, overall they are very similar in what they attempt

to represent. An array of samples in various hydro-meteorological input variables with a focus on precipitation is processed to explain the single outcome variable of runoff, also referred to as streamflow, over time. Therefore, the evaluation of such models does not necessarily differ depending on the model type or paradigm. In general, the error between observed and predicted values needs to be calculated to assess the accuracy of the model.

A first step in the comprehensive evaluation of the models in light of the effects of climate change was the decision to use a Differential Split-Sample Testing (DSST) approach. This test setup allows to evaluate a model's capability to extrapolate under non-stationary conditions, such as transient climatic conditions. Robustness is a key characteristic of models subject to high degrees of variability in the input data. In the compilation of open problems in the domain of hydrology, Blöschl et al. name the assessment of model robustness under contrasting climatic conditions as one of the key issues [BBC⁺19]. The experiments and their evaluation are therefore designed to make a step in the direction of comprehensive, climate-resilient, modelling. O et al. and de Moura et al. highlight the importance of further analyses in the field of DSST-based modelling utilising LSTMs [ODO20, NdMSD22].

Furthermore, six different metrics have been applied to evaluate the performance of the three hydrological model types employed in the experiments presented in this work. This accounts for the suggestion to incorporate various metrics in model evaluation rather than relying on the domain-specific (e.g. the NSE) to gain a comprehensive understanding of the advantages and limitations of a model [MGPD15, CVL⁺21].

The application of the ordinal evaluation guideline presented in Table 4.5 allows for a holistic categorisation of model performance in the domain of rainfall-runoff modelling. In contrast to limiting the evaluation to a narrow comparison of predictions achieved in this study alone, domain-wide standards are consulted to gain insight into the overall competitiveness of the model results. Opening up the scope of evaluation is an important step in recognising the effective robustness of models and in removing experimentation biases. Table 5.10 presents the classification of model results on the validation set into the proposed evaluation guideline. Overall, the EA-LSTM reports the best score with an average of 2.2 in the four-point scale and is the only model to achieve the grade *Good* with respect to domain-wide model results. In contrast, the PDM and the ML-based DDM perform worse, both being graded as *Satisfactory*. However, there is a difference in average scores with XGBoost achieving slightly better results (2.6 compared to 3.2). These results are in line with the state-of-the-art research, which revealed the generally superior performance of DL models in the domain (see Section 2.3).

5.4.3 Model Comparison

Building on the experiment results presented in this chapter, this section presents an analysis of hypotheses that are formulated based on the research questions addressing differences in model performances (**RQ 3**). The evaluation metrics and guidelines

	HBVEdu	XGBoost	EA-LSTM
MAE	Satisfactory (3.0)	Satisfactory (3.0)	Satisfactory (3.0)
RMSE	Satisfactory (3.0)	Satisfactory (3.0)	Satisfactory (3.0)
R ²	Good (2.2)	Good (2.3)	Very Good (1.3)
NSE	Unsatisfactory (3.8)	Satisfactory (3.2)	Good (2.3)
KGE	Satisfactory (3.1)	Good (2.1)	Good (1.6)
P _{BIAS}	Unsatisfactory (4.0)	Good (1.8)	Good (2.1)
Average	Satisfactory (3.2)	Satisfactory (2.6)	Good (2.2)

Table 5.10: Validation performance evaluation of all model types according to the criteria presented in Table 4.5.

presented in Chapter 4 are applied and assessed in statistical significance tests in order to investigate systemic variations in performance based on certain experiment characteristics, such as DSST reference periods, imputation methods or model types. Therefore, this section is a direct application of the outcome of **RQ 3.1** discussed in the previous section.

Figure 5.7 provides a motivation for further spatial analyses of model results. The RMSE values of modelling results from the EA-LSTM trained on the hot2cold split are visualised across the study area. The size of the points is equivalent to the relative catchment elevation. It is apparent that catchments in non-alpine, lowland areas achieve lower errors. Basins at high elevations experience higher errors. This figure showcases the potential of in-depth evaluations and hypotheses tests of the model results. However, the assessment of the results is limited to the questions raised in **RQ 3** and its sub-questions to fulfil the scope of this work.

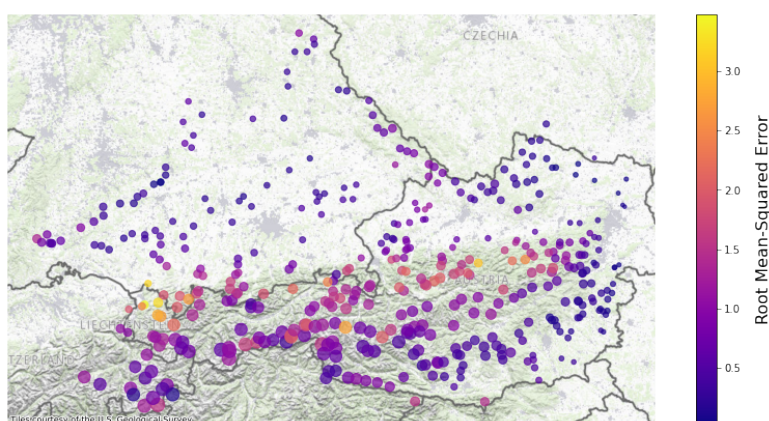


Figure 5.7: The catchment-specific RMSE values of the EA-LSTM model trained on the hot2cold period across the study area of LamaH.

Difference in Model Performance

The results indicate that there are clear differences in model performance with respect to the recorded metrics among the three model types HBVEdu, XGBoost, and EA-LSTM. The distribution of key performance metrics among groups and reference periods, depicted in Figure 5.4, suggests that the EA-LSTM significantly outperforms the other two models. Furthermore, the graph indicates that the ML model XGBoost also achieves better results than the PDM HBVEdu throughout the experiments, although to a lesser extent.

The Hypotheses 5.1 are therefore formulated to answer research question **RQ 3.1**.

H_0 : A process-based model (HBVEdu), a Machine Learning model (XGBoost), and a Deep Learning model (EA-LSTM) perform equally well. (5.1)

H_1 : There exists a statistically significant difference in model performance among the three model types.

Examining these hypotheses contributes to the main objective of this thesis, covered by **RQ 3**. To address this research question, the three model types are subject to the non-parametric Wilcoxon signed-rank tests in order to investigate the difference in locations of two related populations using paired samples for each evaluation metric, i.e. differences in performance [Wil92]. The test results are stated in Table D.1.

In summary, it is statistically sound to reject the H_0 and accept the H_1 . The only insignificant differences in model performance are between XGBoost and HBVEdu for metrics MAE and R^2 with adjusted p-values of 0.9 and 0.735, respectively. All other model-metric combinations prove the assumptions that there are statistically significant differences in model performance among the three model types. Therefore, the answer to **RQ 3.1** is that the selected DL model outperforms the other models, and the ML model also shows better performance in comparison to the traditional physics-based model.

Difference in DSST Periods

In their 2012 study analysing the ability of hydrological models to extrapolate under different climatic conditions, Coron et al. found estimation biases in models trained on periods where significant parameter transfer occurs. In fact, the researchers observed a tendency to overestimate runoff for calibration periods similar to the wet2dry periods presented here, and a tendency to underestimate runoff for a dry2wet calibration period [CAP⁺12].

The results reported for the calibration and validation periods in Sections 5.2 and 5.3 largely suggest comparative model performance across all five periods and three model types. While Table 5.9 points to an overall decrease in performance, there are no striking differences for a specific model type or period that would indicate significant deterioration.

In fact, Figure 5.5 and suggest that, at times, periods cold2hot and hot2cold even show marginal improvements for the EA-LSTM models. Table 5.11 illustrates the relative changes of the four climatically-driven DSST periods in comparison to the baseline reference period. Means are shown for all models with relative changes in brackets, except for the reference period where the standard deviation is shown next to the mean. This table illustrates that there are no drastic differences in performances for models trained on varying climatic conditions. The DSST periods suggest competitive performance and robustness for all model types.

Period	RMSE	NSE	KGE	PBias
reference	1.2123 (\pm 0.18)	0.6142 (\pm 0.12)	0.6559 (\pm 0.16)	12.2481 (\pm 16.87)
cold2hot	1.3270 (+9.46 %)	0.5479 (-10.79 %)	0.6160 (-6.08 %)	8.3417 (-31.89 %)
hot2cold	1.2227 (+0.86 %)	0.6164 (+0.35 %)	0.6413 (-2.22 %)	13.5206 (+10.39 %)
wet2dry	1.2424 (+2.48 %)	0.5975 (-2.71 %)	0.6325 (-3.57 %)	11.2844 (-7.87 %)
dry2wet	1.2548 (+3.50 %)	0.5959 (-2.98 %)	0.6451 (-1.65 %)	12.245 (-0.03 %)

Table 5.11: Mean performance metrics and relative changes compared to the reference period for all four DSST periods. For the reference period, the standard deviation is shown in brackets next to the mean value.

Based on these assumptions, Hypotheses 5.2 are formulated in order to answer **RQ 3.2**.

H_0 : There is no significant difference in the performance of models trained on periods characterised by extreme climatic conditions compared to models trained on a traditional baseline period. (5.2)

H_1 : Models trained on climatically-drive periods of data differ significantly in performance compared to those trained on a baseline period.

To address these hypotheses, non-parametric Kruskal-Wallis tests are performed to determine whether there are statistically significant differences between the medians of the groups [KW52]. The validation results are grouped by training periods and then carried out for each model and metric. Since the training periods differ for each of the DSST periods, each group can be considered independent. Therefore, five groups (reference, cold2hot, hot2cold, wet2dry, dry2wet) are input to each Kruskal-Wallis test per model and metric, resulting in 18 tests. The results are shown in Table D.2.

The test results confirm that there is no statistical difference in model performance among the split-sample periods. Thus, H_0 can be accepted for all model types and metrics. This is a key result of this study as it indicates that any model trained on a traditional train/test-split period with no regard of climatic trends and much larger amounts of training samples does not outperform models trained on short periods exhibiting strong climatic conditions for a specific variable. A further finding is that models are generally

robust to transient climatic conditions. However, the experiment results presented for the test and validation sets suggest a higher degree of robustness for the EA-LSTM.

Difference in Pre-processing Methods

Judging from the performance criteria reported for all model combinations of both the test and validation periods, it seems as that the models built on data with missing values imputed by a Random Forest show consistently better performance compared to using the catchment-specific median. Since the three different modelling approaches represent distinct approaches, the effect of pre-processing must be considered and compared among the model types separately.

Therefore, the intuition is that there is a statistical difference in performance for PDM and models from ML as well as DL that were built on data with missing values imputed using either Random Forests or the catchment-specific median. This is reflected in the Hypotheses 5.3 where based on the aforementioned intuition, H_0 should be rejected and H_1 accepted.

$$\begin{aligned} H_0 &: \text{Models based on different imputation methods of missing values} \\ &\quad \text{perform equally well with respect to the six evaluation metrics.} \\ H_1 &: \text{There exists a statistically significant difference in performance} \\ &\quad \text{between models based on different imputation methods.} \end{aligned} \tag{5.3}$$

Based on the distribution of metric differences for the distinct groups of models and imputation methods, either an independent Student's t-test or a Mann-Whitney U-test is performed to find out whether there is a difference in performance. The test results are shown in Table D.3.

For a large part of the results, the H_0 cannot be rejected. This indicates that there is no significant difference in performance between the two imputation methods. This is largely in line with the reported results: While the performance of the models trained on the RF-imputed sets is slightly superior to the PCM in most cases, this is not always the case and the predominant pattern is that both methods result in very similar, only marginally different performance. Most notably, the P_{BIAS} value closest to the optimum (-0.1697) is achieved with the PCM method. This finding indicates that a robust catchment-specific imputation value is as good a measure as a complex ensemble regression technique. The mean p-value is 0.39. The only statistical differences in performance can be detected for the EA-LSTM in metric KGE ($p = 0.045$) for the validation results. This may be due to the proneness of underestimating peakflow inherent to this metric.

Conclusion

The extensive literature review in the field of hydrological modelling and the comprehensive experiments, evaluations, and analyses of how ML and, more importantly, DL models can be applied to the task of rainfall-runoff modelling in the context of increasing climatic variability to provide crucial insights into model performance, the impact of transient climatic conditions, and the process of data engineering in the domain. This concluding chapter now summarises and emphasises the outcomes of the presented research and experiments. Finally, limitations of this work and recommendations for future research are stated.

6.1 Research Results

The state-of-the-art review reveals that datasets in the domain of LSH have significant shortcomings, including proprietary nature, inconsistent formats, low quality or unreliability in the data, and limited accessibility, thus hindering reproducibility of research results. The extensive research highlights the need for consistency and extensibility of data collections, and design as open-source software as the key requirements for large-sample datasets. Caravan, which includes long-term time series and catchment attributes for several popular datasets in a consistent format, is identified as the most comprehensive data collection. Its design as open-source software positions Caravan as a milestone in LSH. It addresses the recommendations of Addor et al.'s landmark study on the state of the field [ADAG⁺20]. To facilitate the application of DL models for domain-specific applications such as rainfall-runoff modelling, the use of open-source software, such as `NeuralHydrology`, is encouraged. Furthermore, it is necessary to improve and consolidate existing libraries for PDMs in order to promote a comprehensive software ecosystem in LSH that can be adapted to various modelling paradigms. Addressing domain- or model-specific challenges as part of the DS pipeline can be resource-intensive and infeasible for data engineers and modellers. Moreover, the modelling of climate

change is highly sensitive to data perturbations caused by transient climatic conditions. However, the steps involved are complex and require a thorough understanding of the data and domain. Undertaken measures to modify the data need to be guided by theory.

Hydrological modelling has traditionally relied on Process-driven model (PDM) approaches based on physical laws. However, alternative approaches have emerged as suitable candidates due to the limitations of these methods. DDM architectures, such as DL models, are now considered powerful tools and their effectiveness, exemplified by the EA-LSTM architecture in this work, highlights their ability to capture complex spatio-temporal patterns in hydrological processes. The flexibility and increasing accessibility of DL models contribute to their widespread acclaim. H2M approaches represent a new modelling paradigm that combines the strengths of both process- and data-driven models. The exploration of such models, e.g. the mass-conserving LSTM constitutes a move towards integrating physical laws with data-driven approaches for improved accuracy in hydrological forecasts [KKS⁺19, HKK⁺21]. The EA-LSTM model is identified as a suitable and effective architecture due to the inherent memory structure and the conditional processing of time series data based on catchment-specific data, which contributes to its robustness. However, key issues in the application of this model are its computational requirements, particularly the reliance on GPU acceleration, which raise practical concerns. The evaluation framework presented in this work, including theory-guided pre-processing, a DSST-based experimental setup, and a set of domain-specific and general error metrics, provides a step towards analysing the robustness of hydrological models in the face of changing climatic conditions.

Three models are employed to investigate the performance of different model paradigms in rainfall-runoff modelling: the PDM HBVEdu and two DDMs, the traditional ML model XGBoost and the DL model EA-LSTM. The EA-LSTM is statistically superior (Mean NSE = 0.73486) to the other models, while XGBoost (Mean NSE = 0.56306) outperforms the conventional physics-based model HBVEdu (Mean NSE = 0.48528). While the XGBoost model appears to produce satisfactory runoff predictions, it in fact suffers from a high sensitivity to fluctuations and is strongly underfitting. The evaluation of the models across climatic reference periods in the DSST setting suggests that there is no significant difference in the performance of models trained on shorter, more specific periods with transient climatic conditions compared to traditional baseline reference periods where data are arbitrarily split without regard to climatic variation. The EA-LSTM shows a high degree of robustness to extreme conditions, which indicates that the DL model excels in adapting to and capturing the complex relationships in hydrological processes. Although models constructed using data with missing values imputed by RF regression generally demonstrate better performance compared to the catchment-specific median, statistical tests do not reveal significant differences between the imputation methods. In conclusion, this work offers valuable insights into hydrological modelling. It demonstrates the superiority of the EA-LSTM model and thus the DL paradigm, the robustness under varying climatic conditions, and the reliability of different pre-processing methods as well as the possible uncertainty introduced by them.

6.2 Limitations and Future Work

It is important to note that static catchment attributes, which constitute a cornerstone of the success reported for DL models in this work, are not, in fact, static at all; they are much rather a long-term approximation to describe a system that is very much dynamic in nature. For instance, greater temporal granularity of catchment attributes can be reflected in dynamic embedding layers, adjusted at various stages of the time series modelling in the LSTM architecture. Furthermore, the DSST approach has limitations that can be overcome by exploring alternative hydro-meteorological driver variables beyond P and T , or by combining multiple attributes. Additionally, there is significant potential to enhance the chosen pre-processing strategies at every stage of the DS pipeline. Domain-specific methods should be integrated into the detection and imputation of outliers, for instance. As previously mentioned, imputation inevitably introduces some level of uncertainty into the data and therefore also the predictions. The lack of analysis of the resulting uncertainty is a major limitation of this work. Experimenting with different architectures for all discussed paradigms is highly beneficial to acquire a thorough understanding of the advantages and limitations of the model types.

A major development in weather forecasting was the publication of FourCastNet (Fourier Forecasting Neural Network) by Pathak et al. in 2022. The authors propose a highly efficient and inexpensive, purely data-driven ensemble forecasting system operating at the global scale. This system outperforms state-of-the-art numerical weather prediction models, but with drastically reduced power consumption by a factor of 12,000 and, most importantly, a 45,000-fold reduction in runtime [PSH⁺22]. The application of this system to the task of rainfall-runoff bears the potential of uncovering and predicting the effects of climate change at an unprecedented level. Combining such large-scale ensemble systems with state-of-the-art DL model architectures could provide a promising approach to further increase efficiency and effectiveness of hydro-meteorological predictions.

Future work in the domain should address the issues of uncertainty and complexity. Parameter uncertainty and model complexity as well as the analysis of system and cell states of neural networks carry the potential of providing valuable insights into the applicability, effectiveness, and robustness of models. The available mixture density networks provided as part of the `NeuralHydrology` library can be leveraged to account for the suggestion by Addor et al. to report uncertainty alongside model results [ADAG⁺20]. Although DL models have contributed immensely to the domain of LSH and have set new standards for rainfall-runoff model accuracy, the paradigm of hybrid hydrology models is quickly emerging as a promising approach. Combining the strengths of physical process-based models, which are capable of accurately representing natural processes such as the hydrological cycle, and data-driven approaches, which can leverage the potential of large-sample and high-dimensional datasets to reveal hidden relationships in hydro-meteorological variables, is a promising pathway for future modelling experiments. Finally, more effort should be put into the development of open-source software for hydrological modelling. The current state of PDM availability must be enhanced in the future to provide modellers with the means of convenient out-of-the-box model comparison across paradigms.

Static Catchment Attributes

Attribute	Description	Unit
p_mean	Mean daily precipitation (P)	mm/day
pet_mean	Mean daily potential evaporation (PE)	mm/day
aridity	Aridity index, ratio PE_{mean}/P_{mean}	/
frac_snow	Fraction of precipitation falling as snow	/
moisture_index	Mean annual moisture index in range $[-1, 1]$, where -1 indicates water-limited conditions and 1 energy-limited conditions	/
seasonality	Moisture index in range $[0, 2]$, where 0 indicates no changes in the water/energy budget during the year, 2 indicates a change from arid to humid	/
high_prec_freq	Frequency of high precipitation days, i.e. days where $P \geq 5 * P_{mean}$	/
high_prec_dur	Average duration of high precipitation events, i.e. number of consecutive days where $P \geq 5 * P_{mean}$	days
low_prec_freq	Frequency of low precipitation days, i.e. days where $P < 1$ mm/day	/
low_prec_dur	Average duration of low precipitation events, i.e. number of consecutive days where $P < 1$ mm/day	days

Table A.1: Description of the climate indices derived from ERA5-Land time series that are included in the Caravan dataset; largely taken from [KNA⁺23].

Note that the land cover extent attribute `glc_pc_sse` from group LC in table A.2 is incorrectly called `gla_pc_sse` in Caravan. This bug is corrected by renaming the attribute during data preparation.

A. STATIC CATCHMENT ATTRIBUTES

Group	Attribute	Description	Aggregation	Unit
H	dis_m3_p	Natural discharge	annual min/max/mean	m ³ s ⁻¹
	run_mm_syr	Land surface runoff	spatial mean of sub-basin runoff	mm
	inu_pc_s	Inundation extent	annual min/mean, long-term max	%
	lka_pc_sse	Limnicity - percent lake area	spatial extent	%
	lkv_mc_usu	Lake volume	at reach pour point	10 ⁶ m ³
	rev_mc_usu	Reservoir volume	at reach pour point	10 ⁶ m ³
	dor_pc_pva	Degree of regulation	index at reach pour point	/
	ria_ha_usu	River area	at reach pour point	ha
	ria_ha_usu	River volume	at reach pour point	10 ³ m ³
P	gwt_cm_sav	Groundwater table depth	spatial mean	cm
	ele_mt_s	Elevation above sea level	spatial min/max/mean	m
	slp_dg_sav	Terrain slope	spatial mean	°(x10)
	sgr_dk_sav	Stream gradient	mean of reach segments	dm/km
C	clz_cl_smj	Climate zones	spatial majority	n = 18
	cls_cl_smj	Climate strata	spatial majority	n = 125
	tmp_dc_s	Air temperature	monthly mean, annual min/max/mean	°C(x10)
	pre_mm_s	Precipitation	monthly and annual mean	mm
	pet_mm_s	Potential evapotranspiration	monthly and annual mean	mm
	aet_mm_s	Actual evapotranspiration	monthly and annual mean	mm
	ari_ix_sav	Global aridity	spatial mean	index (x10)
	cmi_ix_s	Climate moisture index	monthly and annual mean	index (x10)
LC	snw_pc_s	Snow cover extent	monthly mean, annual max/mean	% cover
	glc_cl_smj	Land cover classes	spatial majority	n = 22
	glc_pc_s	Land cover extent	spatial mean	%
	pnv_cl_smj	Pot. natural vegetation classes	spatial majority	n = 15
	pnv_pc_s	Pot. natural vegetation extent	spatial mean	%
	wet_cl_smjs	Wetland classes	spatial mean	n = 12
	wet_pc_s	Wetland extent	spatial mean	% & group
	for_pc_sse	Forest cover extent	spatial mean	%
	crp_pc_sse	Cropland extent	spatial mean	%
	pst_pc_sse	Pasture extent	spatial mean	%
	ire_pc_sse	Irrigated area extent (equipped)	spatial mean	%
	prm_pc_sse	Permafrost extent	spatial mean	%
	pac_pc_sse	Protected area extent	spatial mean	%
	tbi_cl_smj	Terrestrial biomes	spatial majority	n = 14
tec_cl_smj	Terrestrial ecoregions	spatial majority	n = 846	
fmh_cl_smj	Freshwater major habitat types	spatial majority	n = 13	
fec_cl_smj	Freshwater ecoregions	spatial majority	n = 426	
S&G	cly_pc_sav	Clay fraction in soil	spatial mean	%
	slt_pc_sav	Silt fraction in soil	spatial mean	%
	snd_pc_sav	Sand fraction in soil	spatial mean	%
	soc_th_sav	Organic carbon content in soil	spatial mean	t/ha
	swc_pc_s	Soil water content	monthly mean, annual mean	%
	lit_cl_smj	Lithological classes	spatial mean	n = 16
	kar_pc_sse	Karst area extent	spatial mean	%
ero_kh_sav	Soil erosion	spatial mean	kg/ha/yr	
A	pop_ct_usu	Population count	at reach pour point	x1000
	ppd_pk_sav	Population density	spatial mean	people/km ²
	urb_pc_sse	Urban extent	spatial mean	%
	nli_ix_sav	Nighttime lights	spatial mean	index (x100)
	rdd_mk_sav	Road density	monthly mean, annual mean	m/km ²
	hft_ix_s	Human footprint	spatial mean for 1993 & 2009	index (x100)
	gdp_ud_sav	Gross domestic product	spatial mean	USD
	hdi_ix_sav	Human development index	spatial mean	index (x1000)

Table A.2: Description of the static catchment attributes derived from HydroATLAS that are included in the Caravan dataset; largely taken from [KNA⁺23].

Hyperparameter Settings

B.1 Estimators for Handling Missing Data and Outliers

Parameter	Description	Value
random_state	Random seed	1,996
n_estimators	Number of estimators to create	7
max_depth	Maximum depth of the tree	10
bootstrap	Fit trees on random subsets with replacement	True
max_samples	Number of samples to draw during bootstrapping	0.5
n_jobs	Number of processors to use (-1 = all)	-1

Table B.1: Settings for the Random Forest estimator used to impute missing streamflow values during data preparation.

Parameter	Description	Value
random_state	Random seed	2,609
n_estimators	Number of estimators to create	7
contamination	Proportion of outliers in the data set	0.004
max_features	Proportion of features to draw from the data set	0.5
bootstrap	Fit trees on random subsets with replacement	True
n_jobs	Number of processors to use (-1 = all)	-1

Table B.2: Settings for the Isolation Forest model used to detect outliers during data preparation.

B.2 Settings and Configurations for Models

Parameter	Value range	Optimal setting
max_depth	8, 9, 10	10
n_estimators	1,000, 1,500, 2,000	2000
learning_rate	0.01, 0.05, 0.1	0.05
colsample_bytree	0.8, 0.9, 1.0	0.9
subsample	0.8, 0.9, 1.0	0.8
alpha	0, 0.5, 1, 2	2
lambda	0, 0.1, 0.5, 1	0.5
gamma	0.1, 0.2, 0.5	0.1

Table B.3: Parameter value ranges for the XGBoost model and optimal settings.

Parameter	Description	Value
seed	Random seed	2,609
validate_every	Validation frequency in epochs	5
validate_n_random_basins	No. of random basins for validation	2
metrics	Metrics to calculate during validation	NSE, KGE, RMSE
model	Model type	ea-lstm
head	Prediction head	regression
output_activation	Activation of regression output	linear
statics_embedding:type	Emb. net. type for static inputs	fc
statics_embedding:hiddens	Number of neurons per FC layer	30, 20, 64
statics_embedding:activation	Activation function of emb. net.	tanh
statics_embedding:dropout	Dropout applied to emb. net.	0.0
dynamics_embedding:type	Emb. net. type for dynamic inputs	fc
dynamics_embedding:hiddens	Number of neurons per FC layer	30, 20, 64
dynamics_embedding:activation	Activation function of emb. net.	tanh
dynamics_embedding:dropout	Dropout applied to emb. net.	0.0
hidden_size	No. of cell states of the LSTM	256
initial_forget_bias	Init. forget gate bias	3
output_dropout	Droupout applied to LSTM output	0.4
optimizer	Optimisation algorithm	Adam
loss	Loss function	NSE
batch_size	Mini-batch size for training	256
epochs	Number of training epochs	30
target_noise_std	Added σ of gaussian noise to labels	0.005
clip_gradient_norm	Clipped norm of gradients during training	1
predict_last_n	Which time step to predict loss for	1
seq_length	Length of the input sequence	365
use_basin_id_encoding	Use basin ID as a static input	True
learning_rate	Learning rates at epochs	0: 0.001, 10: 0.0005, 20: 0.0001

Table B.4: General configuration of the EA-LSTM DL model.

APPENDIX **C**

Algorithms

Algorithm C.1: Find Consecutive High/Low Periods**Input:***data*: Time series data with daily mean values*column*: Name of the column to analyze*percentile*: Percentile threshold (66 for high periods, 33 for low periods)*is_high*: True for high periods, False for low periods*min_period_length*: Minimum length of consecutive periods*days_without_threshold*: Days allowed without meeting the threshold**Output:***sorted_periods*: DataFrame with sorted consecutive periods

```

1  periods, current_period  $\leftarrow$   $\emptyset$ 
2  days_below_threshold  $\leftarrow$  0
3  for each row in data do
4      if (is_high and data[column]  $\geq$  percentile) or ( $\neg$ is_high and
      data[column]  $\leq$  percentile) then
5          if  $\nexists$  current_period then
6              Start a new current_period
7          else
8              Extend the current_period
9              days_below_threshold  $\leftarrow$  0
10         end
11     else
12         if  $\exists$  current_period then
13             if days_below_threshold  $\leq$  days_without_threshold then
14                 Extend the current_period
15                 days_below_threshold += 1
16             else
17                 if length(current_period)  $\geq$  min_period_length then
18                     Close the current_period
19                     periods  $\leftarrow$  periods + current_period
20                 end
21                 current_period  $\leftarrow$   $\emptyset$ 
22             end
23         end
24     end
25 end
26 sorted_periods  $\leftarrow$  sort(periods)
27 return sorted_periods

```

Algorithm C.2: Aggregate Consecutive High/Low Periods

Input:

data: Time series data with daily mean values
periods: Result from Algorithm C.1
column: Name of the column to analyse
is_high: True for high periods, False for low periods
num_years: Number of years for sliding window

Output:

sorted_periods: Sorted DataFrame with aggregated periods

```

1 aggregated_periods, current_start_date, current_end_date ← ∅
2 total_num_days ← 0
3 mean_for_column ← mean(data[column])
4 Sort periods by 'start_date'
5 for each row in periods do
6   start_date ← row['start_date']
7   end_date ← row['end_date']
8   num_days ← row['num_days']
9   if current_start_date is ∅ then
10    current_start_date ← start_date
11    current_end_date ← end_date
12    total_num_days ← total_num_days + num_days
13  end
14  else
15    time_span_years ← days(end_date - current_start_date)/365
16    if time_span_years ≤ num_years then
17      current_end_date ← end_date
18      total_num_days ← total_num_days + num_days
19    end
20    else
21      new_aggregation ← new aggregation from the current values
22      aggregated_periods ← aggregated_periods + new_aggregation
23      current_start_date ← start_date
24      current_end_date ← end_date
25      total_num_days ← num_days
26    end
27  end
28 end
29 if current_start_date is not ∅ then
30   new_aggregation ← new aggregation from the current values
31   aggregated_periods ← aggregated_periods + new_aggregation
32 end
33 Sort sorted_periods_df by 'mean' in descending order if is_high is True
34 return sorted_periods_df
  
```

Test Statistics

D.1 Model Performance

Metric	Model Comparison	W-statistic	p-value	H_0
MAE	HBVEdu vs XGBoost	26.0	0.921875	Accept
	HBVEdu vs EA-LSTM	0.0	0.001953	Reject
	XGBoost vs EA-LSTM	0.0	0.001953	Reject
RMSE	HBVEdu vs XGBoost	0.0	0.001953	Reject
	HBVEdu vs EA-LSTM	0.0	0.001953	Reject
	XGBoost vs EA-LSTM	0.0	0.001953	Reject
R ²	HBVEdu vs XGBoost	20.0	0.492188	Accept
	HBVEdu vs EA-LSTM	0.0	0.001953	Reject
	XGBoost vs EA-LSTM	0.0	0.001953	Reject
NSE	HBVEdu vs XGBoost	0.0	0.001953	Reject
	HBVEdu vs EA-LSTM	0.0	0.001953	Reject
	XGBoost vs EA-LSTM	0.0	0.001953	Reject
KGE	HBVEdu vs XGBoost	0.0	0.001953	Reject
	HBVEdu vs EA-LSTM	0.0	0.001953	Reject
	XGBoost vs EA-LSTM	0.0	0.001953	Reject
P _{BIAS}	HBVEdu vs XGBoost	0.0	0.001953	Reject
	HBVEdu vs EA-LSTM	0.0	0.001953	Reject
	XGBoost vs EA-LSTM	0.0	0.001953	Reject

Table D.1: Wilcoxon signed-rank test results ($\alpha = 0.05$) for model comparisons on evaluation metrics for Hypothesis 5.1.

D.2 DSST Periods

Model	Metric	H-statistic	p-value	H_0
HBVEdu	MAE	7.527273	0.110513	Accept
	RMSE	6.872727	0.142769	Accept
	R ²	7.963636	0.092919	Accept
	NSE	5.345455	0.253652	Accept
	KGE	7.418182	0.115372	Accept
	P _{BIAS}	5.563636	0.234192	Accept
XGBoost	MAE	8.727273	0.068290	Accept
	RMSE	8.400000	0.077977	Accept
	R ²	8.727273	0.068290	Accept
	NSE	8.400000	0.077977	Accept
	KGE	8.727273	0.068290	Accept
	P _{BIAS}	8.290909	0.081485	Accept
EA-LSTM	MAE	7.309091	0.120428	Accept
	RMSE	6.872727	0.142769	Accept
	R ²	6.872727	0.142769	Accept
	NSE	6.872727	0.142769	Accept
	KGE	4.145455	0.386678	Accept
	P _{BIAS}	6.218182	0.183436	Accept

Table D.2: Kruskal-Wallis test results ($\alpha = 0.05$) for DSST period comparisons on evaluation metrics for Hypothesis 5.2.

D.3 Imputation Methods

Model	Metric	Test	Test Statistic	p-value	H_0
HBVEdu	MAE	Independent t-test	-1.202546	0.263534	Accept
	RMSE	Independent t-test	-1.427566	0.191264	Accept
	R^2	Independent t-test	1.223412	0.255984	Accept
	NSE	Independent t-test	1.403142	0.198175	Accept
	KGE	Independent t-test	1.220500	0.257027	Accept
	P_{BIAS}	Independent t-test	-1.559898	0.157404	Accept
XGBoost	MAE	Independent t-test	-0.166832	0.871642	Accept
	RMSE	Independent t-test	-0.340564	0.742203	Accept
	R^2	Mann-Whitney U test	15.000000	0.676103	Accept
	NSE	Independent t-test	0.342733	0.740630	Accept
	KGE	Independent t-test	0.650784	0.533428	Accept
	P_{BIAS}	Independent t-test	-0.547746	0.598812	Accept
EA-LSTM	MAE	Independent t-test	-0.911325	0.388759	Accept
	RMSE	Independent t-test	-1.317229	0.224237	Accept
	R^2	Independent t-test	1.256947	0.244228	Accept
	NSE	Independent t-test	1.305485	0.228019	Accept
	KGE	Independent t-test	2.377000	0.044755	Reject
	P_{BIAS}	Independent t-test	-0.981665	0.355017	Accept

Table D.3: Independent t-test and Mann-Whitney U-test results ($\alpha = 0.05$) for Hypothesis 5.3.

List of Figures

2.1	Conceptualisation of the water cycle visualising key processes and variables by the NOAA. Image credit: Dennis Cain/NWS [Nat19].	9
2.2	Left: A typical hydrograph with flow values in m^3/s plotted against time in days. Peakflow and baseflow are easily distinguishable. Right: A typical storm hydrograph with a rising limb, peakflow and a recession limb. Taken from [Dav08].	14
3.1	Top: Global distribution of catchments included in Caravan. Bottom: Distribution of catchments among the GEnS climate zones (the bottom part shows the fraction of a particular climate zone on the total land mass) [KNA ⁺ 23].	39
3.2	Overview of the domain of coverage of LamaH (in the original version). The discharge gauges are represented on the map as circles, with the size indicating the size of the catchment area and the colour indicating the elevation of the station. Numbers denote the 18 distinct river regions. Taken from [KSH21].	43
3.3	Distribution of catchment areas by country (excluding Liechtenstein). . .	44
3.4	(a) Distribution of the catchment areas across the GEnS climate zones including the percentage of the overall catchment area covered in LamaH. (b) Geographical distribution of the GEnS climate zones across the study area.	47
3.5	The Budyko curve is displayed for all 479 catchments in this study. It plots the aridity index (PE/P) against the evaporative index (E_t/P), with point size proportional to the catchment area, and point colour indicating the mean average elevation of the catchments. The Budyko curve is represented by the dashed line [Bud74].	48
3.6	Geographical distribution of key hydrological measures, including precipitation (P), potential evaporation (PE), air temperature (T) and streamflow (Q), across the study area. The base layer of the map reflects the topography of the domain. The colour of the points represents the intensity of the respective measure, while their size corresponds to the catchment elevation. Four metrics are presented: (a) average daily precipitation (mm/day), (b) average daily potential evaporation (mm/day), (c) average daily air temperature calculated at 2 meters ($^{\circ}C$), and (d) average daily streamflow (mm/day) normalised with respect to the catchment area.	49
		125

3.7	(a) Difference in mean daily precipitation and potential evaporation in mm/day. (b) Difference in mean daily precipitation and streamflow in mm/day.	51
3.8	Rainfall and streamflow records over the whole period covered in <i>Caravan</i> for the two selected exemplary catchments. Daily records from each month were averaged and are displayed as a blue line for precipitation on the left y-axis and as a green inverted line for streamflow on the right y-axis. (a) The measurements for the high-deviation catchment Berninabach at Pontresina, Switzerland. (b) The measurements for the low-deviation catchment Schwarza at Gloggnitz, Austria.	54
3.9	The hydrographs showing the mean daily streamflow records averaged over the entire 39-year period covered in <i>Caravan</i> for the two selected exemplary catchments. The 7-day rolling average of streamflow records is depicted by the red line. (a) The hydrograph for the high-deviation catchment Berninabach at Pontresina, Switzerland. (b) The hydrograph for the low-deviation catchment Schwarza at Gloggnitz.	55
3.10	Top: De-seasonalised mean yearly air temperature for all catchments. Bottom: Seasonal mean yearly air temperature trends. Trends are indicated by red dashed linear regression lines.	57
3.11	Number of statistically significant ($\alpha = 0.05$) trends in mean air temperature, precipitation, and streamflow for the 479 catchments in the study area over a 39-year period. The data was averaged on a monthly basis and analysed using an unmodified Mann-Kendall test.	59
4.1	The data preparation pipeline for the raw LamaH time series and static attribute data.	70
4.2	The components and processes of HBVEdu [AH10].	71
4.3	Differences between the architectures of the classic LSTM and the EA-LSTM model of Kratzert et al. [KKS ⁺ 19].	74
5.1	Average loss in NSE for the EA-LSTM training runs per reference period.	85
5.2	Boxplot of the distribution of NSE values for the DSST periods on the test sets.	90
5.3	Correlation heat-maps depicting the relationships among key performance metrics for the test set results among the three hydrological models (HBVEdu, XGBoost, EA-LSTM).	91
5.4	Key evaluation metrics from the validation results of the models for all reference periods.	93
5.5	Cumulative distribution function of the NSE (top) and RMSE (bottom) values of HBVEdu, XGBoost, and EA-LSTM.	94
5.6	Observed and simulated streamflow results for the validation period of the high-deviation catchment located at Berninabach at Pontresina, Switzerland (lamah_2262). Values are plotted for each DSST period and each model.	96

5.7 The catchment-specific RMSE values of the EA-LSTM model trained on the hot2cold period across the study area of LamaH. 103

List of Tables

3.1	Summary of the predominant datasets in Large-Sample Hydrology (LSH) [ADAG ⁺ 20].	41
3.2	Description of the time series attributes derived from ERA5-Land that are included in the Caravan dataset; largely taken from [KNA ⁺ 23].	45
3.3	Difference in key statistical metrics for hydrological variables (T , Q , P , SWE , PE) between the first five years (1981-1986) and the last five years (2013 - 2018) with complete records for 479 catchments in the study area.	58
4.1	Definition of the climatic periods for Differential Split-Sample Testing. . .	64
4.2	Definition of the baseline reference period.	64
4.3	Description of how the inputs for the PDM HBVEdu are engineered. . . .	72
4.4	Overview of the evaluation metrics used to analyse and compare model performance.	75
4.5	Guideline for the evaluation of hydrological models with the six performance criteria, partly inspired by [MGPD15].	79
5.1	Specification of the hpcgpu1 server of the HPC Research Group at Vienna University of Technology.	84
5.2	Experiment results for the reference period on the respective test set. . .	85
5.3	Experiment results for the DSST period hot2cold on the respective test set.	86
5.4	Experiment results for the DSST period cold2hot on the respective test set.	87
5.5	Experiment results for the DSST period wet2dry on the respective test set.	88
5.6	Experiment results for the DSST period dry2wet on the respective test set.	89
5.7	Mean performance metrics for each model (trained on RF-imputed data) across all reference periods on the respective test sets.	90
5.8	Validation results for all DSST periods reported for the best performing models for each period.	92
5.9	Percentage change in mean performance metrics between validation and test results. Upper section: model types. Lower section: DSST periods.	94
5.10	Validation performance evaluation of all model types according to the criteria presented in Table 4.5.	103
5.11	Mean performance metrics and relative changes compared to the reference period for all four DSST periods. For the reference period, the standard deviation is shown in brackets next to the mean value.	105
		129

A.1	Description of the climate indices derived from ERA5-Land time series that are included in the Caravan dataset; largely taken from [KNA ⁺ 23].	111
A.2	Description of the static catchment attributes derived from HydroATLAS that are included in the Caravan dataset; largely taken from [KNA ⁺ 23].	112
B.1	Settings for the Random Forest estimator used to impute missing streamflow values during data preparation.	113
B.2	Settings for the Isolation Forest model used to detect outliers during data preparation.	113
B.3	Parameter value ranges for the XGBoost model and optimal settings.	114
B.4	General configuration of the EA-LSTM DL model.	115
D.1	Wilcoxon signed-rank test results ($\alpha = 0.05$) for model comparisons on evaluation metrics for Hypothesis 5.1.	121
D.2	Kruskal-Wallis test results ($\alpha = 0.05$) for DSST period comparisons on evaluation metrics for Hypothesis 5.2.	122
D.3	Independent t-test and Mann-Whitney U-test results ($\alpha = 0.05$) for Hypothesis 5.3.	123

List of Algorithms

C.1 Find Consecutive High/Low Periods	118
C.2 Aggregate Consecutive High/Low Periods	119

Acronyms

- ANN** Artificial Neural Network. 2, 3, 8, 19, 24–27, 66, 86, 87, 97, 99, 101
- C3S** Copernicus Climate Change Service. 33
- CAMELS** Catchment Attributes and Meteorology for Large-sample Studies. 34, 35, 37, 38, 41, 97
- CANOPEX** Canadian Model Parameter Estimation Experiment. 34, 41
- CDF** Cumulative Distribution Function. 92
- COSERO** Continuous Semi-distributed Runoff. 21
- DDM** Data-driven model. 19, 20, 23, 25, 26, 28–30, 97, 98, 101, 102, 108
- DL** Deep Learning. 1–4, 8, 9, 19, 20, 23, 25, 26, 29, 30, 40, 61, 64, 66, 68, 69, 74, 80, 81, 85, 90, 92, 93, 95, 100, 102, 104, 106–109, 115, 130
- DS** Data Science. 1, 2, 98, 107, 109
- DSST** Differential Split-Sample Testing. 28–30, 63, 64, 68, 73, 79, 80, 86–92, 94, 96, 98–103, 105, 108, 109, 122, 126, 129, 130
- E-HYPE** European Hydrological Predictions for the Environment. 36, 41
- ECMWF** European Centre for Medium-Range Weather Forecasts. 33
- ERA5** 5th Generation of European ReAnalysis. 33, 36–38, 44–46, 61, 111, 129, 130
- EWA** European Water Archive. 35, 36, 41, 97
- FAIR** findable, accessible, interoperable and reusable. 32, 95, 97, 98
- GEnS** Global Environmental Stratification. 37–39, 46–48, 59, 125
- GLDAS** Global Land Data Assimilation System. 33

GRDB Global Runoff Data Base. 35, 41, 97

GRDC Global Runoff Data Centre. 35

GSIM Global Streamflow Indices and Metadata Archive. 35, 36, 41, 97

H2M Hybrid hydrological model. 20, 27, 108

HBV Hydrologiska Byråns Vattenbalansavdelning. 20, 21, 26, 27, 69

HYSETS Hydrometeorological Sandbox - École de technologie supérieure. 36–38, 41, 97

KGE Kling-Gupta Efficiency. 75, 77–79, 85–94, 103, 106, 121–123

LamaH Large-Sample Data for Hydrology and Environmental Sciences for Central Europe. 37, 38, 40–43, 46–49, 59, 70, 72, 97, 100, 103, 125–127

LSH Large-Sample Hydrology. 2, 5, 30–32, 34, 36–38, 40–42, 44, 58, 95, 97, 98, 107, 109, 129

LSTM Long Short-Term Memory. 8, 23–26, 28–30, 73, 74, 81, 100–102, 108, 109, 126

MAE Mean Absolute Error. 75, 76, 79, 85–92, 94, 103, 104, 121–123

MCAR Missing Completely At Random. 62

MI Multiple Imputation. 62, 65, 99

ML Machine Learning. 1–4, 19, 23, 24, 27, 65, 66, 68, 69, 72, 80, 81, 84, 90, 92, 93, 95, 100–102, 104, 106–108

MNAR Missing Not At Random. 62

MOPEX Model Parameter Estimation Experiment. 34, 41, 97

MSE Mean Squared Error. 77

NCAR National Center for Atmospheric Research. 33

NCEP National Centers for Environmental Prediction. 33

NOAA National Oceanic and Atmospheric Administration. 8, 9, 22, 34, 125

NSE Nash-Sutcliffe Efficiency. xi, 73, 75, 77–79, 84–94, 102, 103, 108, 121–123, 126

P_{BIAS} Percent Bias. 75, 78, 79, 85–95, 103, 106, 121–123

PCM Per-Catchment Median. 65, 67, 85–89, 106

PDM Process-driven model. 19, 20, 26–30, 66, 69, 71, 72, 80, 81, 83, 90, 92, 93, 95, 98, 100–102, 104, 106–109, 129

PGF Princeton Global Forcing. 33

PUB Prediction in Ungauged Basins. 8, 34

R² Coefficient of Determination. 75–79, 85–90, 92, 94, 103, 104, 121–123

RF Random Forest. 24, 27, 65, 69, 81, 85–91, 106, 108, 129

RMSE Root Mean Squared Error. 71, 75, 76, 78, 79, 83, 85–94, 103, 121–123, 126, 127

RNN Recurrent Neural Network. 23–25, 66

RSR Regularised Self-Representation. 67

SAC-SMA Sacramento Soil Moisture Accounting Model. 21, 22, 81

SWAT Soil and Water Assessment Tool. 22, 27, 66

SWE Snow Water Equivalent. 11

VIC Variable Infiltration Capacity. 22, 26, 71, 72

Bibliography

- [ABM⁺22] Richard Arsenault, François Brissette, Jean-Luc Martel, Magali Troin, Guillaume Lévesque, Jonathan Davidson-Chaput, Mariana C Gonzalez, Ali Ameli, and Annie Poulin. Hysets - a 14425 watershed hydrometeorological sandbox over north america, May 2022.
- [ABODB16] Richard Arsenault, Rachel Bazile, Camille Ouellet Dallaire, and François Brissette. CANOPEX: A Canadian hydrometeorological watershed database. *Hydrological Processes*, 30(15):2734–2736, 2016.
- [ACB⁺22] Anshuka Anshuka, Rohitash Chandra, Alexander Buzacott, D. Sanderson, and Floris Van Ogtrop. Spatiotemporal hydrological extreme forecasting framework using lstm deep learning model. *Stochastic Environmental Research and Risk Assessment*, 36:3467–3485, 03 2022.
- [ADAG⁺20] Nans Addor, Hong X. Do, Camila Alvarez-Garreton, Gemma Coxon, Keirnan Fowler, and Pablo A. Mendoza. Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal*, 65(5):712–725, 2020.
- [AH10] Amir Aghakouchak and Emad H. Habib. Application of a conceptual hydrologic model in teaching hydrologic processes. *International Journal of Engineering Education*, 26:963–973, 2010.
- [AMS23] American Meteorological Society AMS. Glossary of meteorology, 5 2023. Online; accessed 1. Aug. 2023. Available at: https://glossary.ametsoc.org/wiki/Water_year.
- [ANMC17] N. Addor, A. J. Newman, N. Mizukami, and M. P. Clark. The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, 21(10):5293–5313, 2017.
- [BBC⁺19] Günter Blöschl, Marc F.P. Bierkens, Antonio Chambel, Christophe Cudennec, Georgia Destouni, Aldo Fiori, James W. Kirchner, Jeffrey J. McDonnell, Hubert H.G. Savenije, Murugesu Sivapalan, Christine Stumpp, Elena Toth, Elena Volpi, Gemma Carr, Claire Lupton, Josè Salinas,

Borbála Széles, Alberto Viglione, Hafzullah Aksoy, Scott T. Allen, Anam Amin, Vazken Andréassian, Berit Arheimer, Santosh K. Aryal, Victor Baker, Earl Bardsley, Marlies H. Barendrecht, Alena Bartosova, Okke Batelaan, Wouter R. Berghuijs, Keith Beven, Theresa Blume, Thom Bogaard, Pablo Borges de Amorim, Michael E. Böttcher, Gilles Boulet, Korbinian Breinl, Mitja Brilly, Luca Brocca, Wouter Buytaert, Attilio Castellarin, Andrea Castelletti, Xiaohong Chen, Yangbo Chen, Yuanfang Chen, Peter Chiffard, Pierluigi Claps, Martyn P. Clark, Adrian L. Collins, Barry Croke, Annette Dathe, Paula C. David, Felipe P. J. de Barros, Gerrit de Rooij, Giuliano Di Baldassarre, Jessica M. Driscoll, Doris Duethmann, Ravindra Dwivedi, Ebru Eris, William H. Farmer, James Feiccabrino, Grant Ferguson, Ennio Ferrari, Stefano Ferraris, Benjamin Fersch, David Finger, Laura Foglia, Keirnan Fowler, Boris Gartsman, Simon Gascoin, Eric Gaume, Alexander Gelfan, Josie Geris, Shervan Gharari, Tom Gleeson, Miriam Glendell, Alena Gonzalez Bevacqua, María P. González-Dugo, Salvatore Grimaldi, A. B. Gupta, Björn Guse, Dawei Han, David Hannah, Adrian Harpold, Stefan Haun, Kate Heal, Kay Helfricht, Mathew Herrnegger, Matthew Hipsey, Hana Hlaváčiková, Clara Hohmann, Ladislav Holko, Christopher Hopkinson, Markus Hrachowitz, Tissa H. Illangasekare, Azhar Inam, Camyla Innocente, Erkan Istanbuluoglu, Ben Jarihani, Zahra Kalantari, Andis Kalvans, Sonu Khanal, Sina Khatami, Jens Kiesel, Mike Kirkby, Wouter Knoben, Krzysztof Kochanek, Silvia Kohnová, Alla Kolechkina, Stefan Krause, David Kreamer, Heidi Kreibich, Harald Kunstmann, Holger Lange, Margarida L. R. Liberato, Eric Lindquist, Timothy Link, Junguo Liu, Daniel Peter Loucks, Charles Luce, Gil Mahé, Olga Makarieva, Julien Malard, Shamshagul Mashtayeva, Shreedhar Maskey, Josep Mas-Pla, Maria Mavrova-Guirguinova, Maurizio Mazzoleni, Sebastian Mernild, Bruce Dudley Mistear, Alberto Montanari, Hannes Müller-Thomy, Alireza Nabizadeh, Fernando Nardi, Christopher Neale, Nataliia Nesterova, Bakhram Nurtaev, Vincent O. Odongo, Subhabrata Panda, Saket Pande, Zhonghe Pang, Georgia Papacharalampous, Charles Perrin, Laurent Pfister, Rafael Pimentel, María J. Polo, David Post, Cristina Prieto Sierra, Maria-Helena Ramos, Maik Renner, José Eduardo Reynolds, Elena Ridolfi, Riccardo Rigon, Monica Riva, David E. Robertson, Renzo Rosso, Tirthankar Roy, João H.M. Sá, Gianfausto Salvadori, Mel Sandells, Bettina Schaeffli, Andreas Schumann, Anna Scolobig, Jan Seibert, Eric Servat, Mojtaba Shafiei, Ashish Sharma, Moussa Sidibe, Roy C. Sidle, Thomas Skaugen, Hugh Smith, Sabine M. Spiessl, Lina Stein, Ingelin Steinsland, Ulrich Strasser, Bob Su, Jan Szolgay, David Tarboton, Flavia Tauro, Guillaume Thirel, Fuqiang Tian, Rui Tong, Kamshat Tussupova, Hristos Tyralis, Remko Uijlenhoet, Rens van Beek, Ruud J. van der Ent, Martine van der Ploeg, Anne F. Van Loon, Ilja van Meerveld, Ronald van Nooijen, Pieter R. van Oel, Jean-Philippe Vidal, Jana von

- Freyberg, Sergiy Vorogushyn, Przemyslaw Wachniew, Andrew J. Wade, Philip Ward, Ida K. Westerberg, Christopher White, Eric F. Wood, Ross Woods, Zongxue Xu, Koray K. Yilmaz, and Yongqiang Zhang. Twenty-three unsolved problems in hydrology (uph) – a community perspective. *Hydrological Sciences Journal*, 64(10):1141–1158, 2019.
- [Bev89] Keith Beven. Changing ideas in hydrology — the case of physically-based models. *Journal of Hydrology*, 105(1):157–172, 1989.
- [Bev00] K. J. Beven. Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences*, 4(2):203–213, 2000.
- [Bev12] Keith Beven. *Down to Basics: Runoff Processes and the Modelling Process*, chapter 1, pages 1–23. John Wiley & Sons, Ltd, 2012.
- [BFMC73] R.J.C. Burnash, R.L. Ferral, R.A. McGuire, and Joint Federal-State River Forecast Center. *A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers*. U. S. Department of Commerce, National Weather Service, and State of California, Department of Water Resources, 1973.
- [BHV⁺19] Günter Blöschl, Julia Hall, Alberto Viglione, Rui A P Perdigão, Juraj Parajka, Bruno Merz, David Lun, Berit Arheimer, Giuseppe T Aronica, Ardian Bilibashi, Miloň Boháč, Ognjen Bonacci, Marco Borga, Ivan Čanjevac, Attilio Castellarin, Giovanni B Chirico, Pierluigi Claps, Natalia Frolova, Daniele Ganora, Liudmyla Gorbachova, Ali Gül, Jamie Hannaford, Shaun Harrigan, Maria Kireeva, Andrea Kiss, Thomas R Kjeldsen, Silvia Kohnová, Jarkko J Koskela, Ondrej Ledvinka, Neil Macdonald, Maria Mavrova-Guirguinova, Luis Mediero, Ralf Merz, Peter Molnar, Alberto Montanari, Conor Murphy, Marzena Osuch, Valeryia Ovcharuk, Ivan Radevski, José L Salinas, Eric Sauquet, Mojca Šraj, Jan Szolgay, Elena Volpi, Donna Wilson, Klodian Zaimi, and Nenad Živković. Changing climate both increases and decreases European river floods. *Nature*, 573(7772):108—111, September 2019.
- [BPL⁺20] Hylke E. Beck, Ming Pan, Peirong Lin, Jan Seibert, Albert I. J. M. van Dijk, and Eric F. Wood. Global Fully Distributed Parameter Regionalization Based on Observed Streamflow From 4,229 Headwater Catchments. *Journal of Geophysical Research: Atmospheres*, 125(17):e2019JD031485, 2020.
- [BSF94] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

- [BSW⁺13] G. Blöschl, M. Sivapalan, T. Wagener, A. Viglione, and H. H. G. Savenije. *Introduction*, page 1–10. Cambridge University Press, 2013.
- [Bud74] M. I. Budyko. *Climate and Life*. ISSN. Elsevier Science, 1974.
- [BvDdR⁺16] Hylke E. Beck, Albert I. J. M. van Dijk, Ad de Roo, Diego G. Miralles, Tim R. McVicar, Jaap Schellekens, and L. Adrian Bruijnzeel. Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, 52(5):3599–3622, 2016.
- [BWB⁺22] Joyce Bosmans, Niko Wanders, Marc Bierkens, Mark Huijbregts, Aafke Schipper, and Valerio Barbarossa. Futurestreams, a global dataset of future streamflow and water temperature. *Scientific Data*, 9, 06 2022.
- [BWR22] Sumon Biswas, Mohammad Wardat, and Hridesh Rajan. The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large. In *Proceedings of the 44th International Conference on Software Engineering, ICSE '22*, page 2091–2103, New York, NY, USA, 2022. Association for Computing Machinery.
- [CAP⁺12] L. Coron, V. Andréassian, C. Perrin, J. Lerat, J. Vaze, M. Bourqui, and F. Hendrickx. Crash testing hydrological models in contrasted climate conditions: An experiment on 216 australian catchments. *Water Resources Research*, 48(5), 2012.
- [CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.
- [CPN16] H.P.G.M Caldera, V. Piyathisse, and K D W Nandalal. A comparison of methods of estimating missing daily rainfall data. *Engineer: Journal of the Institution of Engineers, Sri Lanka*, 49:1, 11 2016.
- [CVL⁺21] Martyn P. Clark, Richard M. Vogel, Jonathan R. Lamontagne, Naoki Mizukami, Wouter J. M. Knoben, Guoqiang Tang, Shervan Gharari, Jim E. Freer, Paul H. Whitfield, Kevin R. Shook, and Simon Michael Papalexiou. The Abuse of Popular Performance Metrics in Hydrologic Modeling. *Water Resources Research*, 57(9):e2020WR029001, 2021.
- [DAA16] Chantal Donnelly, Jafet C.M. Andersson, and Berit Arheimer. Using flow signatures and catchment similarities to evaluate the e-hype multi-basin model across europe. *Hydrological Sciences Journal*, 61(2):255–273, 2016.
- [Dan91] TM Daniell. Neural networks. applications in hydrology and water resources engineering. In *National Conference Publication- Institute of Engineers. Australia*, 1991.

- [Dav08] Tim Davie. *Fundamentals of Hydrology*. Routledge Fundamentals of Physical Geography Series. Routledge, 2nd edition, 2008.
- [Dep21] Department of Environment and Science, Queensland. Hydrology – Catchment and subcatchment, 5 2021. Online; accessed 1. Aug. 2023. Available at: <https://wetlandinfo.des.qld.gov.au/wetlands/ecology/processes-systems/water/hydrology/landscape.html>.
- [DGC⁺23] Ningpeng Dong, Wenhai Guan, Jixue Cao, Yibo Zou, Mingxiang Yang, Jianhui Wei, Liang Chen, and Hao Wang. A hybrid hydrologic modelling framework with data-driven and conceptual reservoir operation schemes for reservoir impact assessment and predictions. *Journal of Hydrology*, 619:129246, 2023.
- [DGLW18] H. X. Do, L. Gudmundsson, M. Leonard, and S. Westra. The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata. *Earth System Science Data*, 10(2):765–785, 2018.
- [DOJ⁺17] Eric Dinerstein, David Olson, Anup Joshi, Carly Vynne, Neil D. Burgess, Eric Wikramanayake, Nathan Hahn, Suzanne Palminteri, Prashant Hedao, Reed Noss, Matt Hansen, Harvey Locke, Erle C Ellis, Benjamin Jones, Charles Victor Barber, Randy Hayes, Cyril Kormos, Vance Martin, Eileen Crist, Wes Sechrest, Lori Price, Jonathan E. M. Baillie, Don Weeden, Kierán Suckling, Crystal Davis, Nigel Sizer, Rebecca Moore, David Thau, Tanya Birch, Peter Potapov, Svetlana Turubanova, Alexandra Tyukavina, Nadia de Souza, Lilian Pintea, José C. Brito, Othman A. Llewellyn, Anthony G. Miller, Annette Patzelt, Shahina A. Ghazanfar, Jonathan Timberlake, Heinz Klöser, Yara Shennan-Farpón, Roeland Kindt, Jens-Peter Barnekow Lillesø, Paulo van Breugel, Lars Graudal, Maianna Voge, Khalaf F. Al-Shammari, and Muhammad Saleem. An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm. *BioScience*, 67(6):534–545, 04 2017.
- [Doo59] James C. I. Dooge. A General Theory of the Unit Hydrograph. *Journal of Geophysical Research*, 64(2):241–256, 2 1959.
- [EEA23] Europe’s changing climate hazards — an index-based interactive EEA report, February 2023. [Online; accessed 6. Nov. 2023. Available at: <https://www.eea.europa.eu/publications/europes-changing-climate-hazards-1>].
- [EHYP⁺21] Bosy A. El-Haddad, Ahmed M. Youssef, Hamid R. Pourghasemi, Biswa-jeet Pradhan, Abdel-Hamid El-Shater, and Mohamed H. El-Khashab. Flood susceptibility prediction using four machine learning techniques

- and comparison of their performance at Wadi Qena Basin, Egypt. *Natural Hazards: Journal of the International Society for the Prevention and Mitigation of Natural Hazards*, 105(1):83–114, January 2021.
- [FMoAM23] Regions Federal Ministry of Agriculture, Forestry and Water Management. Die Hochwasserereignisse im Jahre 2002 in Österreich, September 2023. [Online; accessed 15. Sep. 2023. Available at: <https://info.bml.gv.at/themen/wasser/wasser-oesterreich/hydrographie/chronik-besonderer-ereignisse/Hochwasser2002.html>].
- [GGJP20] Salem S. Gharbia, Laurence Gill, Paul Johnston, and Francesco Pilla. GEO-CWB: GIS-Based Algorithms for Parametrising the Responses of Catchment Dynamic Water Balance Regarding Climate and Land Use Changes. *Hydrology*, 7(3), 2020.
- [GKYM09] Hoshin V. Gupta, Harald Kling, Koray K. Yilmaz, and Guillermo F. Martinez. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1):80–91, 2009.
- [GPB⁺14] H. V. Gupta, C. Perrin, G. Blöschl, A. Montanari, R. Kumar, M. Clark, and V. Andréassian. Large-sample hydrology: a need to balance depth with breadth. *Hydrology and Earth System Sciences*, 18(2):463–477, 2014.
- [GRA⁺22] Salem Gharbia, Khurram Riaz, Iulia Anton, Gabor Makrai, Laurence Gill, Leo Creedon, Marion McAfee, Paul Johnston, and Francesco Pilla. Hybrid Data-Driven Models for Hydrological Simulation and Projection on the Catchment Scale. *Sustainability*, 14(7), 2022.
- [GSY99] Hoshin Gupta, Soroosh Sorooshian, and Patrice Yapo. Status of Automatic Calibration for Hydrologic Models: Comparison With Multilevel Expert Calibration. *Journal of Hydrologic Engineering - J HYDROL ENG*, 4, 04 1999.
- [Has23] How to perform Unsupervised Feature Selection using Supervised Algorithms, October 2023. [Online; accessed 30. Oct. 2023. Available at: <https://shorturl.at/gvzs9>].
- [HHR⁺20] Fatimah Bibi Hamzah, Firdaus Mohd Hamzah, Siti Fatin Mohd Razali, Othman Jaafar, and Norhayati Abdul Jamil. Imputation methods for recovering streamflow observation: A methodological review. *Cogent Environmental Science*, 6(1):1745133, 2020.
- [HHRS21] Fatimah Bibi Hamzah, Firdaus Mohd Hamzah, Siti Fatin Mohd Razali, and Hafiza Samad. A Comparison of Multiple Imputation Methods for Recovering Missing Data in Hydrological Studies. *Civil Engineering Journal*, 7(9):1608–1619, September 2021.

- [HK06] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.
- [HKK⁺21] P.J. Hoedt, F. Kratzert, D. Klotz, C. Halmich, M. Holzleitner, G.S. Nearing, S. Hochreiter, and G. Klambauer. MC-LSTM: Mass-Conserving LSTM. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4275–4286. PMLR, 18–24 Jul 2021.
- [HM19] Md. Hussain and Ishtiaq Mahmud. pymannkendall: a python package for non parametric mann kendall family of trend tests. *Journal of Open Source Software*, 4(39):1556, 7 2019.
- [HNB⁺18] J. J. Hamman, B. Nijssen, T. J. Bohn, D. R. Gergel, and Y. Mao. The variable infiltration capacity model version 5 (vic-5): infrastructure improvements for new applications and reproducibility. *Geoscientific Model Development*, 11(8):3481–3496, 2018.
- [HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, nov 1997.
- [HSB⁺13] M. Hrachowitz, H.H.G. Savenije, G. Blöschl, J.J. McDonnell, M. Sivalalan, J.W. Pomeroy, B. Arheimer, T. Blume, M.P. Clark, U. Ehret, F. Fenicia, J.E. Freer, A. Gelfan, H.V. Gupta, D.A. Hughes, R.W. Hut, A. Montanari, S. Pande, D. Tetzlaff, P.A. Troch, S. Uhlenbrook, T. Wagener, H.C. Winsemius, R.A. Woods, E. Zehe, and C. Cudennec. A decade of Predictions in Ungauged Basins (PUB) — a review. *Hydrological Sciences Journal*, 58(6):1198–1255, 2013.
- [HW23] Zhengzheng Hao and Desheng Wu. Data Preprocessing of Soil Attributes for Ecohydrological Applications Using SWAT Model at Xin’anjiang Upstream Watershed, China. *Ecohydrology and Hydrobiology*, 23(2):198–210, 2023.
- [HWL⁺18] Caihong Hu, Qiang Wu, Hui Li, Shengqi Jian, Nan Li, and Zhengzheng Lou. Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water*, 10(11), 2018.
- [JK20] D. Jakhar and I. Kaur. Artificial intelligence, machine learning and deep learning: definitions and differences. *Clinical and Experimental Dermatology*, 45(1):131–132, 01 2020.
- [JML⁺23] Hong Kang Ji, Majid Mirzaei, Sai Hin Lai, Adnan Dehghani, and Amin Dehghani. The robustness of conceptual rainfall-runoff modelling under climate variability – a review. *Journal of Hydrology*, 621:129666, 2023.

- [KAHW17] A. Kuentz, B. Arheimer, Y. Hundecha, and T. Wagener. Understanding hydrologic variability across Europe through catchment classification. *Hydrology and Earth System Sciences*, 21(6):2863–2879, 2017.
- [KBB05] P. Krause, D. P. Boyle, and F. Bäse. Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, 5:89–97, 2005.
- [KBL⁺15] Minjeong Kim, Sangsoo Baek, Mayzonee Ligaray, Jongcheol Pyo, Minji Park, and Kyung Hwa Cho. Comparative studies of different imputation methods for recovering streamflow observation. *Water*, 7(12):6847–6860, 2015.
- [KC22] Wouter Knoben and Martyn Clark. CAMELS-Spat: Catchment Data for Spatially Distributed Large-Sample Hydrology. In *EGU General Assembly Conference Abstracts*, EGU General Assembly Conference Abstracts, pages EGU22–6609, May 2022.
- [KCM⁺21] Lattawit Kulanuwat, Chantana Chantrapornchai, Montri Maleewong, Papis Wongchaisuwat, Supaluk Wimala, Kanoksri Sarinnapakorn, and Surajate Boonya-aroonnet. Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series. *Water*, 13:1862, 07 2021.
- [KFW19] W. J. M. Knoben, J. E. Freer, and R. A. Woods. Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, 23(10):4323–4331, 2019.
- [KGNK22] F. Kratzert, M. Gauch, G. Nearing, and D. Klotz. NeuralHydrology — A Python library for Deep Learning research in hydrology. *Journal of Open Source Software*, 7(71):4050, 2022.
- [KHK⁺19] F. Kratzert, M. Herrnegger, D. Klotz, S. Hochreiter, and G. Klambauer. *NeuralHydrology – Interpreting LSTMs in Hydrology*, pages 347–362. Springer International Publishing, 2019.
- [KJK⁺22] B. Kraft, M. Jung, M. Körner, S. Koirala, and M. Reichstein. Towards hybrid modeling of the global hydrological cycle. *Hydrology and Earth System Sciences*, 26(6):1579–1614, 2022.
- [KKB⁺18] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022, 2018.
- [KKG⁺22] D. Klotz, F. Kratzert, M. Gauch, A. Keefe Sampson, J. Brandstetter, G. Klambauer, S. Hochreiter, and G. Nearing. Uncertainty estimation with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, 26(6):1673–1693, 2022.

- [KKH⁺19] F. Kratzert, D. Klotz, M. Herrnegger, A.K. Sampson, S. Hochreiter, and G.S. Nearing. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, 55(12):11344–11354, 2019.
- [KKHH18] F. Kratzert, D. Klotz, M. Herrnegger, and S. Hochreiter. A glimpse into the Unobserved: Runoff simulation for ungauged catchments with LSTMs. In *Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on Neural Information Processing Systems (NeuRIPS 2018)*, 2018.
- [KKK⁺96] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, Roy Jenne, and Dennis Joseph. The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, 77(3):437 – 472, 1996.
- [KKS⁺19] F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23(12):5089–5110, 2019.
- [KNA⁺23] F. Kratzert, G. Nearing, N. Addor, T. Erickson, M. Gauch, O. Gilon, L. Gudmundsson, A. Hassidim, D. Klotz, S. Nevo, et al. Caravan-A global community dataset for large-sample hydrology. *Scientific Data*, 10(1):61, 2023.
- [KPP20] Syed Kabir, Sandhya Patidar, and Gareth Pender. Investigating capabilities of machine learning techniques in forecasting stream flow. *Proceedings of the Institution of Civil Engineers - Water Management*, 173(2):69–86, 2020.
- [KSH21] C. Klingler, K. Schulz, and M. Herrnegger. LamaH-CE: LARge-SaMple DATA for Hydrology and Environmental Sciences for Central Europe. *Earth System Science Data*, 13(9):4529–4565, 2021.
- [KW52] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.
- [KYG⁺21] Taereem Kim, Tiantian Yang, Shang Gao, Lujun Zhang, Ziyu Ding, Xin Wen, Jonathan J. Gourley, and Yang Hong. Can artificial intelligence and data-driven machine learning models match or even replace process-driven hydrologic models for streamflow simulation?: A case study of four watersheds with different hydro-climatic regions across the CONUS. *Journal of Hydrology*, 598:126423, 2021.

- [LBOM12] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. *Efficient BackProp*, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [LHB⁺23] Xie Lian, Xiaolong Hu, Jiang Bian, Liangsheng Shi, Lin Lin, and Yuanlai Cui. Enhancing streamflow estimation by integrating a data-driven evapotranspiration submodel into process-based hydrological models. *Journal of Hydrology*, 621:129603, 2023.
- [Lit88] Roderick J. A. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988.
- [Liu20] Dedi Liu. A rational performance criterion for hydrological model. *Journal of Hydrology*, 590:125488, 2020.
- [LLH⁺18] Zhongmin Liang, Yujie Li, Yiming Hu, Binquan Li, and Jun Wang. A data-driven SVR model for long-term runoff prediction and uncertainty analysis based on the Bayesian framework. *Theoretical and Applied Climatology*, 133(1-2):137–149, July 2018.
- [LLOD⁺19] Simon Linke, Bernhard Lehner, Camille Ouellet Dallaire, Joseph Ariwi, Günther Grill, Mira Anand, Penny Beames, Vicente Burchard Levine, Sally Maxwell, Hana Moidu, Florence Tan, and Michele Thieme. Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific Data*, 6:283, 12 2019.
- [LLWB94] Xu Liang, Dennis P. Lettenmaier, Eric F. Wood, and Stephen J. Burges. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research: Atmospheres*, 99(D7):14415–14428, 1994.
- [LPR⁺10] Göran Lindström, Charlotta Pers, Jörgen Rosberg, Johan Strömqvist, and Berit Arheimer. Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. *Hydrology Research*, 41(3-4):295–319, 04 2010.
- [LR02] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Ltd., Chichester, England, UK, August 2002.
- [LSA15] Peng Li, Elizabeth A. Stuart, and David B. Allison. Multiple Imputation: A Flexible Tool for Handling Missing Data. *JAMA*, 314(18):1966–1967, 11 2015.
- [LTZ12] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data*, 6(1), mar 2012.

- [Man45] Henry B. Mann. Nonparametric tests against trend. *Econometrica*, 13(3):245–259, 1945.
- [MBJ⁺13] Marc J. Metzger, Robert G. H. Bunce, Rob H. G. Jongman, Roger Sayre, Antonio Trabucco, and Robert Zomer. A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring. *Global Ecology and Biogeography*, 22(5):630–638, 2013.
- [MDH19] Paul Madley-Dowd, Rachael Hughes, Kate Tilling, and Jon Heron. The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110:63–73, 2019.
- [MFL⁺21] Kai Ma, Dapeng Feng, Kathryn Lawson, Wen-Ping Tsai, Chuan Liang, Xiaorong Huang, Ashutosh Sharma, and Chaopeng Shen. Transferring Hydrologic Data Across Continents – Leveraging Data-Rich Regions to Improve Hydrologic Prediction in Data-Sparse Regions. *Water Resources Research*, 57(5):e2020WR028600, 2021.
- [MGPD15] Daniel Moriasi, Margaret Gitau, Naresh Pai, and Prasad Daggupati. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. *Transactions of the ASABE (American Society of Agricultural and Biological Engineers)*, 58:1763–1785, 12 2015.
- [MMA⁺23] Thibault Mathevet, Nicolas Le Moine, Vazken Andréassian, Hoshin Gupta, and Ludovic Oudin. Multi-objective assessment of hydrological model performances using Nash–Sutcliffe and Kling–Gupta efficiencies on a worldwide large sample of watersheds. *Comptes Rendus. Géoscience*, 2023. Online first.
- [MMCD21] Babak Mohammadi, Roozbeh Moazenzadeh, Kevin Christian, and Zheng Duan. Improving streamflow simulation by combining hydrological process-driven and artificial intelligence-based models. *Environmental Science and Pollution Research*, 28:65752–65768, 12 2021.
- [MnSDAP⁺21] J. Muñoz Sabater, E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach, B. Martens, D. G. Miralles, M. Piles, N. J. Rodríguez-Fernández, E. Zsoter, C. Buontempo, and J.-N. Thépaut. Era5-land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383, 2021.
- [MSM⁺22] Hamidreza Mosaffa, Mojtaba Sadeghi, Iman Mallakpour, Mojtaba Naghdizadegan Jahromi, and Hamid Reza Pourghasemi. Chapter 43 - application of machine learning algorithms in hydrology. In Hamid Reza Pourghasemi, editor, *Computers in Earth and Environmental Sciences*, pages 585–591. Elsevier, 2022.

- [Mul50] T. J. Mulvaney. On the use of self-registering rain and flood gauges in making observations of the relations of rainfall and flood discharges in a given catchment. In *Proceedings of the Institution of Civil Engineers of Ireland*, volume 4, pages 18–31, 1850.
- [Nas57] J.E. Nash. The form of the instantaneous unit hydrograph, publication 42. *International Association Scientific Hydrology, Wallingford, England*, pages 114–112, 1957.
- [Nat19] National Oceanic and Atmospheric Administration (NOAA). Water cycle, 2 2019. Online; accessed 27. Jul. 2023. Available at: <https://www.noaa.gov/education/resource-collections/freshwater/water-cycle>.
- [NCS⁺15] A. J. Newman, M. P. Clark, K. Sampson, A. Wood, L. E. Hay, A. Bock, R. J. Viger, D. Blodgett, L. Brekke, J. R. Arnold, T. Hopson, and Q. Duan. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous usa: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, 19(1):209–223, 2015.
- [NdMSD22] Carolina Natel de Moura, Jan Seibert, and Daniel Henrique Marco Detzel. Evaluating the long short-term memory (LSTM) network for discharge prediction under changing climate conditions. *Hydrology Research*, 53(5):657–667, 04 2022.
- [NKI20] Navideh Noori, Latif Kalin, and Sabahattin Isik. Water quality prediction using swat-ann coupled approach. *Journal of Hydrology*, 590:125220, 2020.
- [NS70] J.E. Nash and J.V. Sutcliffe. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970.
- [OCW22] Yashon O. Ouma, Rodrick Cheruyot, and Alice N. Wachera. Rainfall and runoff time-series trend analysis using lstm recurrent neural network and wavelet neural network with satellite-based meteorological data: case study of nzoia hydrologic basin. *Complex and Intelligent Systems*, 8:213–236, 2022.
- [ODO20] Sungmin O, Emanuel Dutra, and Rene Orth. Robustness of process-based versus data-driven modeling in changing climatic conditions. *Journal of Hydrometeorology*, 21(9):1929 – 1944, 2020.
- [OEAF21] Umut Okkan, Zeynep Beril Ersoy, Ahmet Ali Kumanlioglu, and Okan Fistikoglu. Embedding machine learning techniques into a conceptual model to improve monthly runoff simulation: A nested hybrid rainfall-runoff modeling. *Journal of Hydrology*, 598:126433, 2021.

- [PLVK⁺19] Cristina Prieto, Nataliya Le Vine, Dmitri Kavetski, Eduardo García, and Raúl Medina. Flow Prediction in Ungauged Catchments Using Probabilistic Random Forests Regionalization and New Statistical Adequacy Tests. *Water Resources Research*, 55(5):4364–4392, 2019.
- [PSH⁺22] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, 2022.
- [PTC⁺20] Gilberto Pastorello, Carlo Trotta, Eleonora Canfora, Housen Chu, Danielle Christianson, You-Wei Cheah, Cristina Poindexter, Jiquan Chen, Abdelrahman Elbashandy, Marty Humphrey, Peter Isaac, Diego Polidori, Markus Reichstein, Alessio Ribeca, Catharine van Ingen, Nicolas Vuichard, Leiming Zhang, Brian Amiro, Christof Ammann, M. Altaf Arain, Jonas Ardö, Timothy Arkebauer, Stefan K. Arndt, Nicola Arriga, Marc Aubinet, Mika Aurela, Dennis Baldocchi, Alan Barr, Eric Beamesderfer, Luca Belelli Marchesini, Onil Bergeron, Jason Beringer, Christian Bernhofer, Daniel Berveiller, Dave Billesbach, Thomas Andrew Black, Peter D. Blanken, Gil Bohrer, Julia Boike, Paul V. Bolstad, Damien Bonal, Jean-Marc Bonnefond, David R. Bowling, Rosvel Bracho, Jason Brodeur, Christian Brümmer, Nina Buchmann, Benoit Burban, Sean P. Burns, Pauline Buysse, Peter Cale, Mauro Cavagna, Pierre Cellier, Shiping Chen, Isaac Chini, Torben R. Christensen, James Cleverly, Alessio Collalti, Claudia Consalvo, Bruce D. Cook, David Cook, Carole Coursolle, Edoardo Cremonese, Peter S. Curtis, Ettore D’Andrea, Humberto da Rocha, Xiaoqin Dai, Kenneth J. Davis, Bruno De Cinti, Agnes de Grandcourt, Anne De Ligne, Raimundo C. De Oliveira, Nicolas Delpierre, Ankur R. Desai, Carlos Marcelo Di Bella, Paul di Tommasi, Han Dolman, Francisco Domingo, Gang Dong, Sabina Dore, Pierpaolo Duce, Eric Dufrière, Allison Dunn, Jiří Dušek, Derek Eamus, Uwe Eichelmann, Hatim Abdalla M. ElKhidir, Werner Eugster, Cacilia M. Ewenz, Brent Ewers, Daniela Famulari, Silvano Fares, Iris Feigenwinter, Andrew Feitz, Rasmus Fensholt, Gianluca Filippa, Marc Fischer, John Frank, Marta Galvagno, Mana Gharun, Damiano Gianelle, Bert Gielen, Beniamino Gioli, Anatoly Gitelson, Ignacio Goded, Mathias Goeckede, Allen H. Goldstein, Christopher M. Gough, Michael L. Goulden, Alexander Graf, Anne Griebel, Carsten Gruening, Thomas Grünwald, Albin Hammerle, Shijie Han, Xingguo Han, Birger Ulf Hansen, Chad Hanson, Juha Hatakka, Yongtao He, Markus Hehn, Bernard Heinesch, Nina Hinko-Najera, Lukas Hörtnagl, Lindsay Hutley, Andreas Ibrom, Hiroki Ikawa, Marcin Jackowicz-Korczynski, Dalibor Janouš, Wilma Jans, Rachhpal

Jassal, Shicheng Jiang, Tomomichi Kato, Myroslava Khomik, Janina Klatt, Alexander Knohl, Sara Knox, Hideki Kobayashi, Georgia Koerber, Olaf Kolle, Yoshiko Kosugi, Ayumi Kotani, Andrew Kowalski, Bart Kruijt, Julia Kurbatova, Werner L. Kutsch, Hyojung Kwon, Samuli Launiainen, Tuomas Laurila, Bev Law, Ray Leuning, Yingnian Li, Michael Liddell, Jean-Marc Limousin, Marryanna Lion, Adam J. Liska, Annalea Lohila, Ana López-Ballesteros, Efrén López-Blanco, Benjamin Loubet, Denis Loustau, Antje Lucas-Moffat, Johannes Lüers, Siyan Ma, Craig Macfarlane, Vincenzo Magliulo, Regine Maier, Ivan Mammarella, Giovanni Manca, Barbara Marcolla, Hank A. Margolis, Serena Marras, William Massman, Mikhail Mastepanov, Roser Matamala, Jaclyn Hatala Matthes, Francesco Mazzenga, Harry McCaughey, Ian McHugh, Andrew M. S. McMillan, Lutz Merbold, Wayne Meyer, Tilden Meyers, Scott D. Miller, Stefano Minerbi, Uta Moderow, Russell K. Monson, Leonardo Montagnani, Caitlin E. Moore, Eddy Moors, Virginie Moreaux, Christine Moureaux, J. William Munger, Taro Nakai, Johan Neiryneck, Zoran Nestic, Giacomo Nicolini, Asko Noormets, Matthew Northwood, Marcelo Nosetto, Yann Nouvellon, Kimberly Novick, Walter Oechel, Jørgen Eivind Olesen, Jean-Marc Ourcival, Shirley A. Papuga, Frans-Jan Parmentier, Eugenie Paul-Limoges, Marian Pavelka, Matthias Peichl, Elise Pendall, Richard P. Phillips, Kim Pilegaard, Norbert Pirk, Gabriela Posse, Thomas Powell, Heiko Prasse, Suzanne M. Prober, Serge Rambal, Üllar Rannik, Naama Raz-Yaseef, Corinna Rebmann, David Reed, Victor Resco de Dios, Natalia Restrepo-Coupe, Borja R. Reverter, Marilyn Roland, Simone Sabbatini, Torsten Sachs, Scott R. Saleska, Enrique P. Sánchez-Cañete, Zulia M. Sanchez-Mejia, Hans Peter Schmid, Marius Schmidt, Karl Schneider, Frederik Schrader, Ivan Schroder, Russell L. Scott, Pavel Sedlák, Penélope Serrano-Ortíz, Changliang Shao, Peili Shi, Ivan Shironya, Lukas Siebicke, Ladislav Šigut, Richard Silberstein, Costantino Sirca, Donatella Spano, Rainer Steinbrecher, Robert M. Stevens, Cove Sturtevant, Andy Suyker, Torbern Tagesson, Satoru Takanashi, Yanhong Tang, Nigel Tapper, Jonathan Thom, Michele Tomassucci, Juha-Pekka Tuovinen, Shawn Urbanski, Riccardo Valentini, Michiel van der Molen, Eva van Gorsel, Ko van Huissteden, Andrej Varlagin, Joseph Verfaillie, Timo Vesala, Caroline Vincke, Domenico Vitale, Natalia Vygodskaya, Jeffrey P. Walker, Elizabeth Walter-Shea, Huimin Wang, Robin Weber, Sebastian Westermann, Christian Wille, Steven Wofsy, Georg Wohlfahrt, Sebastian Wolf, William Woodgate, Yuelin Li, Roberto Zampedri, Junhui Zhang, Guoyi Zhou, Donatella Zona, Deb Agarwal, Sebastien Biraud, Margaret Torn, and Dario Papale. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Sci. Data*, 7(225):1–27, July 2020.

[PVG+11]

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion,

O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

- [PVG⁺23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Common pitfalls and recommended practices - 10.2.1. Data leakage during pre-processing, 2023. [Online; accessed 8. Nov. 2023. Available at: https://scikit-learn.org/stable/common_pitfalls.html#data-leakage-during-pre-processing].
- [Raz21] Saman Razavi. Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling. *Environmental Modelling and Software*, 144:105159, 2021.
- [RC13] Tara Razavi and Paulin Coulibaly. Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods. *Journal of Hydrologic Engineering*, 18(8):958–975, 2013.
- [RCF⁺19] Omid Rahmati, Bahram Choubin, Abolhasan Fathabadi, Frederic Coulon, Elinaz Soltani, Himan Shahabi, Eisa Mollaefar, John Tiefenbacher, Sabrina Cipullo, Baharin Bin Ahmad, and Dieu Tien Bui. Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and unec methods. *Science of The Total Environment*, 688:855–866, 2019.
- [RHJ⁺04] M. Rodell, P. R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C.-J. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, J. K. Entin, J. P. Walker, D. Lohmann, and D. Toll. The Global Land Data Assimilation System. *Bulletin of the American Meteorological Society*, 85(3):381 – 394, 2004.
- [RYH⁺18] Weiwei Ren, Tao Yang, Ching-Sheng Huang, Chong yu Xu, and Quanxi Shao. Improving monthly streamflow prediction in alpine regions: integrating hbv model with bayesian neural network. *Stochastic Environmental Research and Risk Assessment*, 32:3381–3396, 2018.
- [SB22] J. Seibert and S. Bergström. A retrospective on hydrological catchment modelling based on half a century with the HBV model. *Hydrology and Earth System Sciences*, 26(5):1371–1388, 2022.
- [SCD06] J Schaake, S Cong, and Q Duan. U.S. MOPEX DATA SET. *IAHS Publication Series*, 307:9–28, 5 2006.

- [SGW06] Justin Sheffield, Gopi Goteti, and Eric F. Wood. Development of a 50-Year High-Resolution Global Dataset of Meteorological Forcings for Land Surface Modeling. *Journal of Climate*, 19(13):3088 – 3111, 2006.
- [SJMC21] A. Sun, Peishi Jiang, Maruti Mudunuru, and Xingyuan Chen. Explore spatio-temporal learning of large sample hydrology using graph neural networks. *Water Resources Research*, 57, 12 2021.
- [SJSK19] Bibhuti Bhusan Sahoo, Ramakar Jha, Anshuman Singh, and Deepak Kumar. Long short-term memory (lstm) recurrent neural network for low-flow hydrological time series forecasting. *Acta Geophysica*, 67(5):1471–1481, October 2019.
- [SKP⁺18] Jan Sitterson, Christopher D. Knightes, Rajbir Parmar, Kurt Wolfe, Brian Avant, and Muluken E. Muche. An Overview of Rainfall-Runoff Model Types. *International Congress on Environmental Modelling and Software*, 41, 2018.
- [SP97] Rainer Storn and Kenneth V. Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11:341–359, 1997.
- [SPL⁺22] G. Sterle, J. Perdrial, L. Li, T. Adler, K. Underwood, D. Rizzo, H. Wen, and A. Harpold. Camels-chem: Augmenting camels (catchment attributes and meteorology for large-sample studies) with atmospheric and stream water chemistry data. *Hydrology and Earth System Sciences Discussions*, 2022:1–23, 2022.
- [SSBK15] Bellie Sivakumar, Vijay P. Singh, Ronny Berndtsson, and Shakera K. Khan. Catchment Classification Framework in Hydrology: Challenges and Directions. *Journal of Hydrologic Engineering*, 20(1):A4014002, 2015.
- [SSR⁺18] S. Sharma, R. Siddique, S. Reed, P. Ahnert, P. Mendoza, and A. Mejia. Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system. *Hydrology and Earth System Sciences*, 22(3):1831–1849, 2018.
- [TFPM15] Patricia Tencaliec, Anne-Catherine Favre, Clémentine Prieur, and Thibault Mathevet. Reconstruction of missing daily streamflow data using dynamic regression models. *Water Resources Research*, 51(12):9447–9463, 2015.
- [USG23] U.S. Geological Survey USGS. Precipitation and the Water Cycle, 8 2023. Online; accessed 2. Aug. 2023. Available at: <https://www.usgs.gov/special-topics/water-science-school/science/precipitation-and-water-cycle>.

- [WDA⁺16] Mark Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gaby Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Olavo Bonino da Silva Santos, Philip Bourne, Jildau Bouwman, Anthony Brookes, Tim Clark, Merce Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris Evelo, Richard Finkers, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 03 2016.
- [Wil92] Frank Wilcoxon. *Individual Comparisons by Ranking Methods*, pages 196–202. Springer New York, New York, NY, 1992.
- [WLC⁺19] Jhih-Huang Wang, Gwo-Fong Lin, Ming-Jui Chang, I-Hang Huang, and Yu-Ren Chen. Real-time water-level forecasting using dilated causal convolutional neural networks. *Water Resources Management*, 33:3759 – 3780, 2019.
- [WM05] C. Willmott and K Matsuura. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Climate Research*, 30:79, 12 2005.
- [ZGSG18] Jinlin Zhu, Zhiqiang Ge, Zhihuan Song, and Furong Gao. Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data. *Annual Reviews in Control*, 46:107–133, 2018.
- [ZKBFH21] Mohammad Zounemat-Kermani, Okke Batelaan, Marzieh Fadaee, and Reinhard Hinkelmann. Ensemble machine learning paradigms in hydrology: A review. *Journal of Hydrology*, 598:126266, 2021.
- [ZLP⁺18] Di Zhang, Junqiang Lin, Qidong Peng, Dongsheng Wang, Tiantian Yang, Soroosh Sorooshian, Xuefei Liu, and Jiangbo Zhuang. Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. *Journal of Hydrology*, 565:720–736, 2018.
- [ZQM⁺23] Lin Zhang, Huapeng Qin, Junqi Mao, Xiaoyan Cao, and Guangtao Fu. High temporal resolution urban flood prediction using attention-based lstm models. *Journal of Hydrology*, 620:129499, 2023.
- [ZWLL23] Yongsong Zou, Jin Wang, Peng Lei, and Yi Li. A novel multi-step ahead forecasting model for flood based on time residual lstm. *Journal of Hydrology*, 620:129521, 2023.
- [ZZZ⁺15] Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Qinghua Hu, and Simon C. K. Shiu. Unsupervised feature selection by regularized self-representation. *Pattern Recognition*, 48(2):438–446, February 2015.