Hybrid Modelling in Hydrology Using a Neural Ordinary Differential Equations Approach

by Jonathan Schieren

Thesis
in partial fulfillment of
Master of Science
Civil Engineering

Assessment Committee
TU Delft:
Dr. Markus Hrachowitz
Dr. Elisa Ragno
Dr. Riccardo Taormina
Deltares:
Judith ter Maat
Martijn Visser

Department of Water Managment
Delft University of Technology
The Netherlands

March 25, 2024

Acknowledgements

Acknowledgements

There are many people I am grateful to for their support during this work. First and foremost, I would like to thank my Assessment Committee, not only for their feedback and advice, but also for their patience on this journey, which has taken longer than planned. I would also like to thank my partner Caroline for her continuous support. I would also like to thank all the researchers without whose work this thesis would not have been possible. This includes the provision of the following data sets that I used in the thesis:

I acknowledge the E-OBS dataset from the EU-FP6 project UERRA (https://www.uerra.eu) and the Copernicus Climate Change Service, and the data providers in the ECA&D project (https://www.ecad.eu)

Part of this work was generated using Copernicus Climate Change Service information [2023]

Maps were created using QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.org

Borders and country labels on maps are made with Natural Earth. Free vector and raster map data naturalearthdata.com.

Abstract

Conceptual models in hydrology are widely used, allow for easy interpretation and require little data. Machine learning models in hydrology often outperform conceptual models but lack the ease of interpretability, require large amounts of data and and do not obey physical laws. Hybrid approaches aiming to combine the advantages of both approaches are becoming more popular. A Neural Ordinary Differential Equations approach is introduced to combine a differential equation-based conceptual model with a neural network. Additionally, conceptual models and Long Short-term Memory (LSTM) models are used as benchmarks. The models are tested using the LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe (LamaH-CE) dataset as well as the E-OBS dataset. In many cases the hybrid models outperform the conceptual model. However, to further improve the performance of hybrid models more research is needed to make the models more computationally efficient and optimized training strategies are required to explore the full potential of the approach.

Contents

List of Figures

List of Tables

List of Abbreviations

ODE	Ordinary Differential Equation
NODE	Neural Ordinary Differential Equation
CAMELS	Catchment Attributes and MEteorologcy for Large-sample studies
DEM	Digital Elevation Model
EXP-Hydro	Exponential Bucket Hydrological Model
m.a.s.l.	meters above sea level
IVP	Initial Value Problem
FAO	Food and Agricultural Organization of the United Nations
PET	potential evapotranspiration
ECA&D	European Climate Assessment & Dataset
E-OBS	E-OBS Gridded Dataset
ERA5-Land	ERA5-Land Dataset
NWM	National Water Model
KGE	Kling-Gupte Efficiency
LamaH-CE	LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe
C3S	Copernicus Climate Change Service
MSE	Mean Squared Error
NSE	Nash-Sutfcliffe Efficiency
NWM	National Water Model
PSO	Particle Swarm Optimization
RK4	Runge-Kutta 4th-order
ECMWF	European Centre for Medium-Range Weather Forecasts
LSH	Large-sample Hydrology
LSTM	Long Short-term Memory
SI	Swarm Intelligence

1 Introduction

1.1 Motivation

Hydrological modelling plays an important role in many hydrology-related fields. Short- and long-term forecasts of streamflow are used for the management of water resource systems (Zealand et al., 1999; Bouaziz et al., 2021). Streamflow predictions are used for flood frequency estimation (Moretti and Montanari, 2008) and for predicting low-flows (Staudinger et al., 2011). Furthermore, hydrological modelling serves as a tool to improve the understanding of the interaction between storages and fluxes at catchment scale (Bouaziz et al., 2021).

Machine learning approaches to hydrological modelling such as LSTM models have shown great skill in rainfall-runoff modelling, often showing superior performance compared to conceptual models. However, the interpretability of machine learning models is more difficult compared to physics-based and conceptual models (Feng et al., 2020; Kratzert et al., 2018; Samek et al., 2019; Feng et al., 2023). Besides accurate streamflow predictions, hydrological system and process understanding is another common objective of hydrological modelling (Devia et al., 2015; Höge et al., 2022). Machine learning models such as LSTM models are fully data-driven or empirical. Hence, the translation from, in this context, meteorological inputs to streamflow is learned entirely from data and does not allow incorporation of prior existing hydrological knowledge (Höge et al., 2022; Devia et al., 2015).

Conceptual models are widely used in hydrological modelling and are semi data-driven or semi empirical. Simplified physical processes are implemented and parameters, which are often assumed as constant across the catchment area, are calibrated using streamflow observations (Merz and Blöschl, 2004; Devia et al., 2015). Conceptual models are easily interpretable and allow incorporation of prior hydrological knowledge.

In recent years different hybrid approaches, combining conceptual models and neural networks have been introduced, aiming to combine the strengths of both approaches (Höge et al., 2022; Feng et al., 2023; Rackauckas et al., 2020). Some hybrid approaches have displayed similar performance to state-of-the-art machine learning models such as LSTMs while additionally providing insight into physical variables and processes (Höge et al., 2022; Feng et al., 2022). Approaches chosen by Frame et al. (2021) and Jiang et al. (2020) apply the neural network component in a post-processing step while Höge et al. (2022), Tsai et al. (2021) and Feng et al. (2022) directly combine the conceptual component with the neural network component. The latter is assumed to result in better interpretability and physical significance (Feng et al., 2022).

1.2 Research Objective

The overarching goal of this work is to explore the utility of a Neural Ordinary Differential Equation (NODE) approach as a hybrid hydrological model. A study region with good data availability and sufficiently large diversity of catchment characteristics was selected to allow for model comparison across different types of catchments. The LamaH-CE dataset (Klingler et al., 2021), which is based on ERA5-Land dataset (Muñoz Sabater, 2019) used here. The ERA5-Land dataset is a derivative of the ERA5 dataset (Hersbach et al., 2020), which is a reanalysis dataset. The LamaH-CE dataset includes 859 catchments and the meteorological forcings are aggregated per catchment. A second dataset is used for comparison. The dataset is created based on the E-OBS dataset (Cornes et al., 2018) which contains ground-based meteorological observations and it is aggregated to the catchments.

A conceptual model is selected as base model which is then extended with a neural network. At the same time the conceptual model is used as a benchmark for the conceptual approach. The conceptual model EXP-Hydro (Patil and Stieglitz, 2014) is used as the model structure facilitates the incorporation of a neural network for the NODE approach. Additionally, a LSTM model is built in this work which serves as a benchmark for the machine learning approach.

The performance of the three types of models, conceptual models, hybrid models and LSTM models is compared. For the comparison of the models, the catchments are aggregated into clusters of hydrologically similar characteristics to allow for a comparison on different types of catchments. At last, strategies to improce the performance of the hybrid models are explored.

1.3 Research Questions

The main research question for this work is:

Can hybrid hydrological models, using a neural ordinary differential equations approach, improve performance of conceptual models?

To assess this, the following research questions are addressed:

SQ1: How does the performance of a hybrid hydrological model, based on a neural ordinary differential equations approach, compare to the performance of the underlying conceptual base model and to a LSTM model?

SQ2: How do the models (conceptual, hybrid, LSTM) perform on catchments with different catchment characteristics?

SQ3: What pathways can be explored to improve the performance of the hybrid models?

2 Theoretical Background

2.1 Hydrological Modelling

Hydrological models can be classified according to different criteria. They can be classified as deterministic or stochastic models, the spatial representation can be lumped, semi-distributed or fully distributed and the process description can be divided into physically-based, conceptual and data-driven or empirical (Singh, 2018; Peel and McMahon, 2020). Subsequently, a brief introduction is given.

2.1.1 Conceptual Hydrological Models

In the 1960s and 1970s several simple conceptual catchment runoff models were developed, the complex hydrological system was represented in a limited number storages and fluxes, partly due to computational constraints. Some of these simplistic models are still used today, for example the HBV (Hydrologiska Byråns Vattenbalansavdelning) model (Lindstrom et al., 1997; Seibert and Bergström, 2022). Further conceptual hydrological models include the GR4J model (Perrin et al., 2003) and EXP-HYDRO (Patil and Stieglitz, 2014). An extensive overview of hydrological models is provided in Peel and McMahon (2020). Advantages of conceptual models include the ease of interpretation, the ability to incorporate prior knowledge, prediction of unobserved variables as well as robust predictions in data-scarce scenarios (Höge et al., 2022; Feng et al., 2022).

Many conceptual models are composed of a combination of three key elements: (1) storages to represent the water that is stored and released, (2) lag functions to represent the temporal dynamics of fluxes and (3) junction elements to represent the splitting and merging of fluxes (Fenicia et al., 2011). One example for a simple conceptual model is EXP-Hydro introduced by Patil and Stieglitz (2014). The conceptual model is composed of two buckets, snow storage and water storage, it includes five mechanistic processes and six static parameters for calibration. A slightly more complex and widely applied conceptual model is the HBV model which includes two buckets for soil storage to represent fast and slow subsurface flow and accounts for interactions of the two (Lindstrom et al., 1997).

2.1.2 Machine Learning Hydrological Models

One approach to machine learning based modelling in hydrology has been presented by Kratzert et al. (2018) where the so called Long Short-Term Memory (LSTM) model, a special type of recurrent neural network is introduced for hydrological modelling. LSTMs have shown great skill in rainfall-runoff modelling and have been used as a benchmark for ther models in recent years (Kratzert et al., 2018; Feng et al., 2022; Höge et al., 2022). However, interpretability of LSTMs is limited as the relation of input variable (e.g. meteorological forcings) to the output variable (e.g. streamflow) has no physical basis and may not be easily interpretable (Feng et al., 2023).

2.1.3 Hybrid Hydrological Models

Hybrid hydrological models in this context can be grouped into two categories. Models that directly combine conceptual and neural network components and models that apply the neural network as a post-processing step. The latter can for example be used to mitigate systematic errors in hydrological models (Frame et al., 2021).

A direct combination of neural network component with a neural network has been introduced by Höge et al. (2022). Here a Neural Ordinary Differential Equations approach for hybrid hydrological modelling is described which builds upon the conceptual model EXP-Hydro by Patil and Stieglitz (2014). Two models are introduced, M50 and M100. The underlying conceptual model contains two model states (snow storage and water storage), five processes and six static parameters to be calibrated. The first model uses two small neural network to replace the evapotranspiration process and the discharge computation while in the second model replaces al five mechanistic processes of the conceptual models. Results show that particularly the second model performs as well as state-of-the-art deep learning models while maintaining the interpretability of the conceptual model (Höge et al., 2022).

In Feng et al. (2022) another approach that directly combines a neural network component with a conceptual model is introduced. The models, referred to as "differentiable, learnable, process-based models" build upon the conceptual HBV model and the regional parameterization of the model is replaced with a neural network and allows for either static or dynamic parameterization of the model. Also here the performance is comparable to state-of-the-art deep learning models, preserve interpretability and are able to impose physical constraints such as mass conservation (Feng et al., 2022).

One example for hybrid hydrological modeling with a post-processing approach is found in Frame et al. (2021) where the effect of using an LSTM model as post-processing step to the United States National Water Model (NWM), a large-scale hydrology simulator is assessed. Or, depending on the perspective, assessed the effect of using the output of the NWM as input to the LSTM, additionally or instead of the meteorological forcings. This was compared against a LSTM model that only takes the meteorological forcings as input. All three models additionally took static catchment characteristics as input (Frame et al., 2021). The models were compared on 531 out of the 671 basins using forcing and streamflow data as well as static catchment characteristics from the Catchment Attributes and MEteorologcy for Large-sample studies (CAMELS) dataset which covers the contiguous United States. The post-processing approach in which the LSTM received the outputs of the NWM as inputs (LSTM_PP), along with the static catchment attributes, resulted in improved NSE scores in 88% of the basins while it lead to a worse NSE score in 12% of the basins. For the post-processing approach in which the LSTM received both the outputs of the NWM and the meteorological forcings as inputs (LSTM_PPA), along with the static catchment characteristics, the NSE score showed improvements in 92% of the catchments and while it deteriorated in 8% of the catchments. The LSTM model which only received meteorological forcings (LSTM_A), along with the static catchment attributes as input outperformed the NWM in 89% of the cases w.r.t. NSE score while performance was worse in 11% of the basins. LSTM_PPA improved the NSE score and the peak timing error in the largest number of catchments while LSTM_A improved bias in the largest number of catchments. Overall the the authors concluded that the LSTM_PP and LSTM_PPA models showed significantly improved performance compared to the NWM. However, they did not consistently, nor significantly, outperform the LSTM_A model which did not take any information from the NWM as input (Frame et al., 2021).

Another example for a post-processing approach is introduced by Jiang et al. (2020) as "process-wrapped recurrent neural network (P-RNN)". Here the conceptual model EXP-Hydro (Patil and Stieglitz, 2014) is integrated into a deep learning framework. The underlying idea is that the conceptual model introduces physical knowledge and the data-driven component fills gaps in the physical knowledge implemented in the conceptual model.

2.2 Hydrological Concepts and Methods

In this section hydrological concepts that are used throughout this study are introduced.

2.2.1 Potential Evapotranspiration (ET_P) and Reference (Crop) Evapotranspiration (ET_0)

The concepts of potential evapotranspiration and reference evapotranspiration are closely related but do not describe the same concept. However, the terms are not always clearly distinguished. The concept of potential evapotranspiration (ET_P) is mostly used in climatology, meteorology and hydrogy while the concept of reference evapotranspiration (ET_0) is mostly used in ecology and agriculture (Xiang et al., 2020). Subsequently, the terms will be defined for this work.

Evapotranspiration is defined as "combined evaporation from the soil surface and transpiration from plants" (Thorntwaite, 1948). The author also introduced potential evapotranspiration as the "transfer [of water to the atmosphere] that would be possible under ideal conditions of soil moisture and vegetation" (Thorntwaite, 1948). In Penman (1956) potential evapotranspiration is defined as "the amount of water transpired in unit time by a short green crop, completely shading the ground, of uniform height and never short of water" (Penman, 1956).

$$\lambda ET = \frac{\Delta(R_n - G) + \rho_a c_p \frac{(e_s - e_a)}{r_a}}{\Delta + \gamma(1 + \frac{r_s}{r_a})} \tag{2.1}$$

ET: Reference evapotranspiration [mm/day], Δ: Slope vapor pressure curve [kPa/°C], R_n: Net radiation at the crop surface [MJ/m^2/day], G: Soil heat flux density [MJ/m^2/day], γ: Psychrometric constant [kPa/°C], T: Mean daily air temperature [°C], e_s: Saturation vapor pressure [kPa], e_a: Actual vapor pressure [kPa], ρ: , c_p: specific heat of air [unit], r_a: bulk surface resistance, r_s: bulk aerodynamic resistance, (Allen et al., 1998)

The United Nations Food and Agricultural Organization of the United Nations (FAO) defined reference (crop) evapotranspiration as "[t]he evapotranspiration rate from a reference surface, not short of water" (Allen et al., 1998). The reference surface is defined as "a hypothetical crop with an assumed height of 0.12 m having a surface resistance of 70 s m-1 and an albedo of 0.23, closely resembling the evaporation of an extension surface of green grass of uniform height, actively growing and adequately watered" (Allen et al., 1998). The FAO recommends the FAO Penman-Monteith equation to estimate reference (crop) evapotranspiration. This is described in detail in chapter 7.1.1 However, there are more approaches to approximating refecrence (crop) evapotranspiration Xiang et al. (2020). This means that reference crop evapotranspiration (ET_0) is a special type of potential evapotranspiration with a more detailed definition of the reference surface.

$$ET_0 = \frac{0.408 \cdot \Delta \cdot (R_n - G) + \gamma \cdot \frac{37}{T+273} \cdot U_2 \cdot (e_s - e_a)}{\Delta + \gamma \cdot (1 + 0.34 \cdot U_2)} \tag{2.2}$$

ET_0: Reference evapotranspiration [mm/day], Δ: Slope vapor pressure curve [kPa/°C], R_n: Net radiation at the crop surface [MJ/m^2/day], G: Soil heat flux density [MJ/m^2/day], γ: Psychrometric constant [kPa/°C], T: Mean daily air temperature [°C], U_2: Wind speed at 2 meters above the ground [m/s], e_s: Saturation vapor pressure [kPa], e_a: Actual vapor pressure [kPa], (Allen et al., 1998)

Another concept that is relevant in this context is open water evaporation, which refers to the amount of water that could potentially evaporate from the surface of an open water body. One way to estimate potential open water evaporation is introduced in Penman (1948) (equation 2.3)

$$E_o = \frac{H\Delta + E_a\gamma}{\Delta + \gamma} \tag{2.3}$$

E_o: open water evaporation rate [kg / m^2 / s], $\Delta = de_a/dT_a$, e_a: actual vapour pressure [mm. HG. or kPa], T_a: surface air temperature [°F or °C], γ: psychrometric constant [kPa / °C or mm. HG. / °F], E_a: isothermal evaporation rate [mm/d] (Penman, 1948)

2.2.2 Budyko Framework

The Budyko framework is a simple method to estimate the water balance of hydrological catchments. It relates the long-term evaporative ratio to its aridity index. The evaporative ratio is the ratio between actual evapotranspiration and precipitation (ET_a/P) and the aridity index is the ration between potential evapotranspiration and precipitation (ET_p/P). Over long periods of time it can be reasonable to assume that catchment water storage is negligible ($\overline{\Delta S} \approx 0$). With the assumption $\overline{\Delta S} = 0$, the relationship between long-term mean precipitation, long-term mean discharge and long-term mean evapotranspiration is $\overline{P} = \overline{E} + \overline{Q}$ (Chen et al., 2023; Reaver et al., 2020).

Budyko (1974), among others, suggested that long-teram mean actual evapotranspiration is a function of long-term mean precipitation and long-term mean potential evapotranspiration $\overline{ET_a} = f_0(\overline{P}, \overline{ET_p})$ (Budyko, 1974) as cited in Reaver et al. (2022). Hence, evapotranspiration in the original Budyko equation (equation 2.4) is assumed to be controlled by precipitation and potential evapotranspiration and does not consider catchment characteristics except long-term climatic conditions (Chen et al., 2023). Several mathematical expressions have been developed based on the behaviour of many catchments with different aridity indexes (Reaver et al., 2020).

The Budyko curve is constrained by the water limit and the energy limit. The evaporative index (ET_a/P) cannot exceed the water limit ($ET_a/P = 1, PET/P < 1$) as over a long period of time actual evapotranspiration cannot exceed precipitation unless there is additional water input to the catchment. The aridity index (ET_p/P) cannot exceed the energy limit ($ET_a/P = ET_p/P, ET_p/P < 1$) as actual evapotranspiration cannot exceed potential evapotranspiration unless the catchment loses water aside from discharge (Koppa et al., 2022). Catchments that plot close to the water limit are generally more humid while catchments that plot close the the energy limit are generally more humid.

$$\frac{ET_a}{P} = \left[\frac{ET_p}{P} tanh(\frac{ET_p}{P})^{-1}(1 - exp(-\frac{ET_p}{P}))\right]^{0.5} \tag{2.4}$$

ET_p: potential evapotranspiration [mm / yr], ET_a: actual evapotranspiration [mm / yr], P: precipitation [mm / yr] (Budyko, 1974) as cited in (Chen et al., 2023)

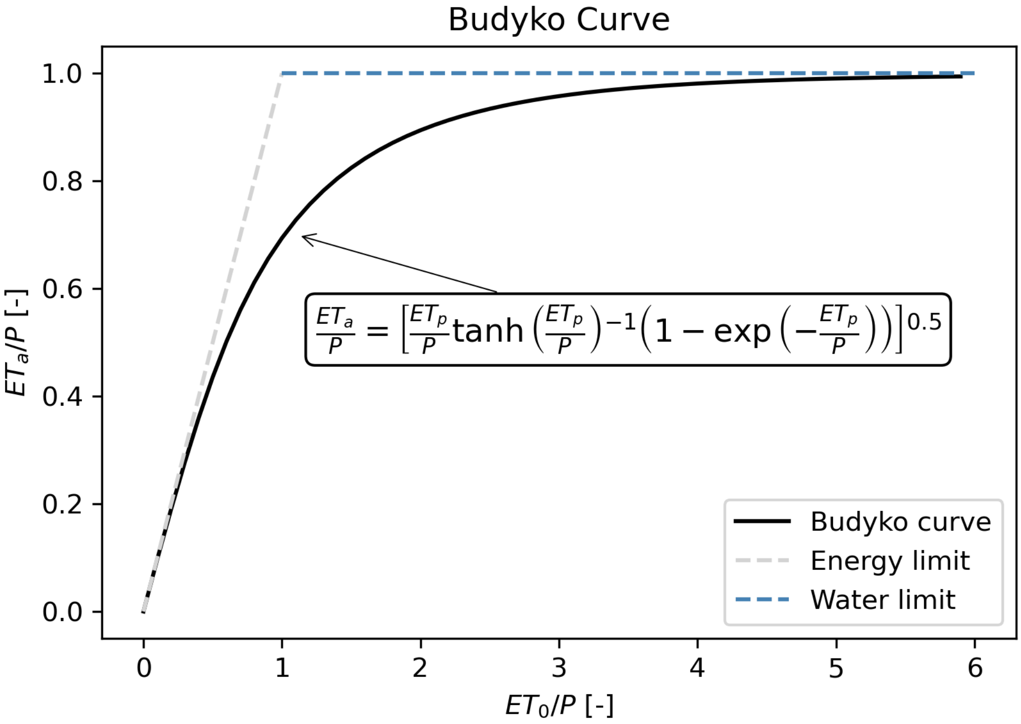

Figure 2.1: ET_p: potential evapotranspiration [mm / yr], ET_a: actual evapotranspiration [mm / yr], P: precipitation [mm / yr] (Budyko, 1974) as cited in (Chen et al., 2023)

Alternatives to the non-parametric original Budyko equation are often parametric equations that consider catchment characteristics have been developed. For example in Zhang et al. (2001) which includes a parameter that accounts for different types of vegetation (Chen et al., 2023).

Budyko (1974) considered the Budyko curve semi-empirical. The physical component is given by the conservation of mass and energy while the empirical component lies in the functional form of the curve (Budyko, 1974) as cited in Reaver et al. (2020). The original Budyko curve was developed based on a large number of catchments with different aridity indexes. Current interpretations of Budyko mostly acknowledge that the Budyko curves are semi-empirical and at the same time often associate physical meaning to the mathematical expressions. The trajectory of the curve is interpreted as the path that a catchment follows if its aridity index changes (Reaver et al., 2020).

2.3 Methods for Numerically Solving Ordinary Differential Equations

2.3.1 Euler's Method

Differential Equations are equations that relate a function to one or more of its derivates, describing the change over one or more independent variables. If the differential equation contains only one independent variable, it is referred to as *Ordinary Differential Equations.* The independent variable in this study is time. Simple differential equations can be solved analytically by computing the antiderivative and an arbitrary integrations constant C. If additionally to the ODE the value of the unknown function is known at a given point in time, it is called an *Initial Value Problem.* If the initial condition and the antiderivative is known, the Initial Value Problem (IVP) can be solved analytically (Atkinson et al., 2009).

If the antiderivative is not known, or if it is computationally or w.r.t complexity not feasible to compute it, or for another reason not desirable to analytically solve an ordinary differential equation, it can be solved numerically. The simplest approach to numerically solving ODEs is *Euler's method.* The *forward difference approximation* to the derivative numerically approximates the solution of an IVP. This is briefly illustrated using an example from Atkinson et al. (2009).

$$\frac{dy}{dt} \approx \frac{y_{n+1} - y_n}{h} \tag{2.5}$$

h: step size, (Atkinson et al., 2009)

$$y_{n+1} = y_n + h \cdot f(t_n, y(t_n)) \tag{2.6}$$

h: step size, (Atkinson et al., 2009)

2.3.2 Runge-Kutta Method

Multiple other methods for numerically solving IVPs exist. The *Runge-Kutta Method* is more accurate approximation of the analytical solution to an IVP compared to Euler's method. The concept of the Runge-Kutta method will be introduced briefly with focus on the Runge-Kutta 4th order method as this version is used in this work. A more detailed explanation can be found in Atkinson et al. (2009)

$$z_1 = hy_n, \tag{2.7}$$

$$z_2 = y_n + \frac{1}{2}h \cdot f(t_n, z1) \tag{2.8}$$

$$z_3 = y_n + \frac{1}{2}h \cdot f(t_n + \frac{1}{2}h, z2) \tag{2.9}$$

$$z_4 = y_n + h \cdot f(t_n + \frac{1}{2}h, z3) \tag{2.10}$$

$$y_{n+1} = y_n + \frac{1}{6}h[f(t_n, z_1) + 2f(t_n + \frac{1}{2}h, z_2) + 2f(t_n + \frac{1}{2}h, z_3) + f(t_n + h, z_4)] \tag{2.11}$$

z_1, z_2, z_3, z_4: Intermediate stages (increments) in the RK4 method,
t_n: current time, y_n: estimate of the solution at time t_n, t_{n+1}: following time step, y_{n+1}: estimate of the solution at time t_{n+1}, h: step size, (Atkinson et al., 2009)

2.4 Scientific Machine Learning

Scientific Machine Learning is an emerging discipline combine mechanistic modelling and machine learning. In this section important concepts and terms relevant for this work are introduced.

2.4.1 Universal Differential Equations & Neural Ordinary Differential Equations

Universal Differential Equations are a concept in scientific machine learning and defined as the combination of a mechanistic modelling part and a universal approximator. If the differential equation is an ODE then it is called an *Universal Ordinary Differential Equation* and if the universal approximator is a neural network, it is also referred to as *Neural Differential Equation* or, if the differential equation is an ODE it is referred to as *Neural Ordinary Differential Equation* (Rackauckas et al., 2020). Or as Bolibar et al. (2023) puts it: "[Universal or Neural Differential Equations] combine the physical simulation of a differential equation using a numerical solver with machine learning". Accordingly, Neural Ordinary Differential Equation models are models using an ordinary differential equation in which terms are partly or completely replaced by neural networks (Höge et al., 2022).

2.4.2 Universal Approximation Theorem

The *Universal Approximation Theorem* for neural networks states that any continuous function can be approximated arbitrarily well by a neural network. Different proofs with different assumptions exist. The approximation of a feed-forward neural network with one hidden layer is addressed in Cybenkot (1989) and Hornik (1991).

2.4.3 Automatic Differentiation

Optimization for Universal Differential Equations, and hence also Ordinary Differential Equations algorithms require a differentiable framework (Bolibar et al., 2023). For this work PyTorch is used to enable differentiability (Paszke et al., 2019).

3 Material & Methods

3.1 Data

3.1.1 LamaH-CE: LArge-SaMple Data for Hydrology and Environmental Sciences for Central Europe

The LamaH dataset includes catchments in nine European countries (Austria, the Czech Republic, Germany, Hungary, Italy, Liechtenstein, Slovakia, Slovenia and Switzerland) and covers an area of approximately 170,000 km^2. The gauge furthest downstream of the Austrian Danube represents the lowest point at 130 meters above sea level (m.a.s.l.) and the highest point is the peak of Piz Bernina at 4049 m.a.s.l. in Switzerland. The area is divided into 18 river regions based on the main tributaries of the Danube (Klingler et al., 2021).

Large-sample Hydrology (LSH) refers to datasets that include many catchments which often is paired with a wide variation of catchment types. They may also include data from different data sources and aim to improve the robustness of conclusions on hydrological processes and models. Examples for LSH datasets include the CAMELS datasets, such as the CAMELS-US dataset, which contains data from 671 in the contiguous United States (Addor et al., 2017), the CAMELS-CL dataset which covers 516 catchments in Chile Alvarez-Garreton et al. (2018), the CAMELS-GB dataset which covers 671 catchments in Great Britain (Coxon et al., 2020) or the CAMELS-BR dataset which contains 897 catchments in Brazil (Chagas et al., 2020). The CAMELS datasets contain hydrometeorological as well as static catchment attributes, both aggregated to the catchment polygons. They follow a consistent structure and data preparation and the LamaH-CE dataset is generally based on the same structure (Klingler et al., 2021).

Basin Delineation

The meteorological time series in the LamaH-CE dataset are aggregated means across the respective catchments. The catchment polygons were created based on the Hydrological Atlas of Austria and the HydroATLAS. While sub-bsins outlets of the Hydrological Atlas of Austria match the location of gauges, sub-basins from the HydroATLAS were in some cases manually adjusted to ensure "that the basin outlets of the polygons agree with the gauging station locations" (Klingler et al., 2021; Federal Ministry of Agriculture, 2007; Linke et al., 2019). The division into sub-catchments is based on data from the Hydrological Atlas of Austria and from the HydroATLAS. Three different delineation techniques (A,B, and C) and accordingly three sets of sub-catchments are included in the LamaH-CE dataset. In catchment delineation B, which contains 859 catchments, computes sub-catchments by subtracting the upstream area of a gauge from the upstream area above the downstream gauge . The median basin size in basin delineation B is 114 km^2 with a range from 1 km^2 to 2500 km^2. Basin delineation C, which is used in this study, is based on basin delineation B but includes only the 454 catchments with little or no anthropogenic influence (Klingler

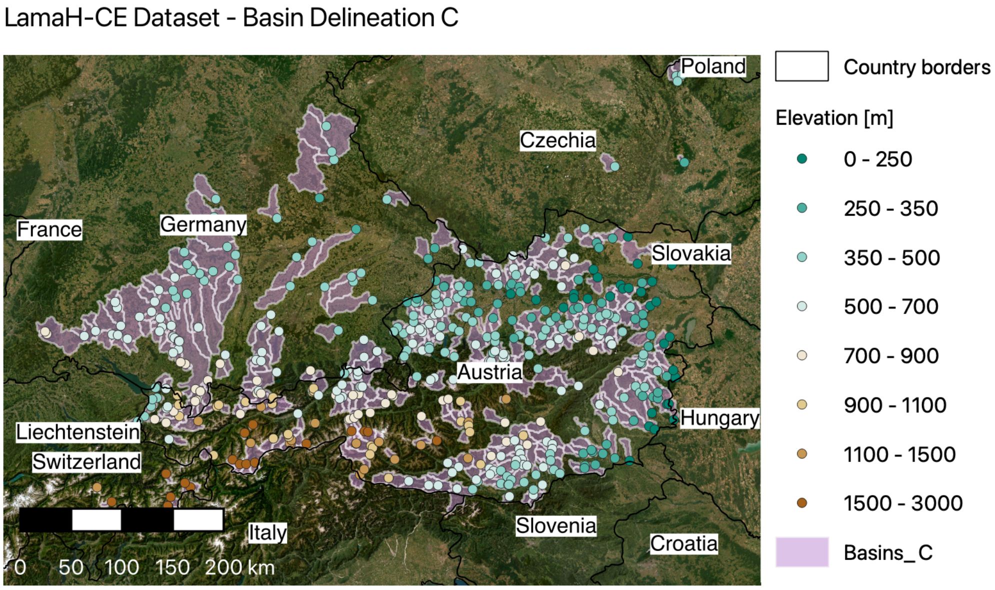

Figure 3.1: Catchments and gauges from catchment delineation C in the LamaH-CE dataset (Klingler et al., 2021), the color-coding represents the mean elevation of the catchments, basemap "ESRI Satellite" (Copyright 1995–2023 Esri)

et al., 2021). The catchments including their mean elevation are shown in figure 3.1.

The aggregation of coarsely gridded or vector data to the catchment shapefiles is performed by calculating the area-weighted arithmetic mean (upscaling approach 1). A second approach was used, mostly for data sources with a high spatial resolution (¡ 1km grid size) the arithmetic mean was calculated for all cells whose centroid lies inside the polygon (upscaling approach 2). However, if no centroid was within the polygon, upscaling 1 was used (Klingler et al., 2021).

Meteorological Data

The meteorological timeseries in the LamaH-CE dataset are based on the ERA5-Land dataset. Gap-free timeseries were obtained for 15 meteorological variables from 1981-01-01 to 2019-12-31 with daily and hourly resolution. The resolution of the underlying ERA5-Land data is 0.1°x 0.1°spatially and the temporal resolution is hourly and daily. Meteorological timeseries were created for all three basin delineations and the meteorological variables used in this work are described in table 3.1 (Klingler et al., 2021).

potential evapotranspiration (PET) values from the ERA5-Land dataset showed significantly too high values for large parts of the study area, therefore no PET were included in the LamaH-CE dataset.

Runoff Data

Runoff in LamaH-CE were obtained from obtained from 882 gauges in Austria, Germany, the Czech Republic and Switzerland. The dataset contains only 859 catchments as for 23 of the gauges it was not possible to clearly define the catchment area. Runoff are also available

Table 3.1: Meteorological variables used in this work from LamaH-CE dataset which are based on the ERA5-Land dataset. Temperatures at 2 m above the surface (Klingler et al., 2021; Muñoz-Sabater et al., 2021)

Variable	Unit
Minimum temperature	°C
Mean temperature	°C
Maximum temperature	°C
Precipitation	mm/d
Forecast albedo	-
Total evapotranspiration	mm/d

at daily and hourly resolution and most of the are derived via the relationship between water level and discharge (rating curve). The runoff were obtained from the Hydrographic Central Bureau of Austria (609 gauges), from the hydrological services of Bavaria and Baden-Württemberg (125 and 61 gauges), from the hydrological office of Switzerland (25 gauges) and the Czech Hydrometeorological Institute (61 gauges). Runoff are available in hourly and daily resolution, however, the at hourly resolution often cover shorter time periods. If hourly resolution runoff were available for the same period as daily runoff or longer, daily runoff data was computed based on hourly runoff data. The maximum time period for runoff data is from 1981 to 2017 as 1981 marks the start of the meteorological from the ERA5-Land dataset and 2017 marks "last year for quality-controlled runoff data from Austria at the point of request" (Klingler et al., 2021). Approximately 80 % of the runoff contain no gaps after gaps of less than 6 h were filled using linear interpolation (Klingler et al., 2021).

Catchment attributes

Catchment attributes describe the physical and georgaphical properties of a catchment. This includes landcover, hydrology, climate and vegetation characteristics. The catchment attributes in the LamaH-CE dataset were obtained using datasets with European to global coverage and include 10 topographic attributes, twelve attributes related to climate characteristics, fourteen attributes characterizing runoff , seven land cover attributes, six attributes describing vegetation indices, ten attributes characterizing soil properties, 16 geological attributes as well as thirteen types of (anthropogenic) impact (Klingler et al., 2021).

Climatic indices include mean daily total precipitation from Oct. 1989 to Sept. 2009 and mean daily total evapotranspiration from Oct. 1989 to Sept. 2009, both based on ERA5-Land data. Furthermore, mean daily reference evapotranspiration from 1970 to 2000 is included. However, due to shortcomings of PET from ERA5-Land, which are explained in chapter 3.1.1, based on the Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2 ?Klingler et al. (2021).

Data Preparation

The meteorological timeseries in the LamaH-CE dataset, which are derived from the ERA5-Land dataset, are gap free and were not processed any further Klingler et al. (2021).

Table 3.2: Variables in the E-OBS dataset. Temperature and relative humidity usually measured at 2 m above the surface. Wind speed measured at 10 m above surface. Precipitation data based on rain gauge data (Cornes et al., 2018).

Variable	**Unit**
Minimum temperature	°C
Mean temperature	°C
Maximum temperature	°C
Precipitation	mm/d
Relative humidity	%
Sea level pressure	hPa
Surface shortwave downwelling radiation	W/m^2
Wind speed	m/s
Land surface elevation	m

Catchment delineation C contains 454 catchments and accordingly also 454 runoff gauges and runoff timeseries. 371 out of these runoff timeseries are free of gaps at daily resolution. Furthermore, 246 runoff timeseries in catchment delineation C cover the whole time series from 1981-01-01 to 2017-12-31 without gaps. Only these 246 catchments were used in this study.

ERA5-Land

The ERA5-Land dataset is produced within the Copernicus Climate Change Service (C3S) by the European Centre for Medium-Range Weather Forecasts (ECMWF) and, as described in chapter 3.1.1 forms the basis of the meteorological forcings in the LamaH-CE dataset. It covers the global land surface and is based on ERA5, the firth generation of European ReAnalysis. The ERA5-Land dataset spans a time period fromm 1950 to present and is continuously updated. It has a spatial resolution of 0.1°x 0.1°and a temporal resolution of one hour Klingler et al. (2021); Muñoz-Sabater et al. (2021); Hersbach et al. (2020).

3.1.2 E-OBS

E-OBS is a gridded meteorological dataset covering the land area in Europe. The dataset is based on station data from the European Climate Assessment & Dataset (ECA&D) project and compromises more than 23000 meteorological stations (Klein Tank et al., 2002; Klok and Klein Tank, 2009; van den Besselaar and Copernicus Climate Change Service, 2023). It has a daily temporal resolution and a spatial resolution of 0.1°x 0.1°as well as 0.25°x 0.25°and covers a timespan from 1950 to 2023. The variables of the E-OBS dataset are listed and described in table 3.2.

The latest version at the time of writing is v27.0e, however, in this work version v26.0e was used as it was the most recent version available at the time of processing the data. The E-OBS dataset can be downloaded as an ensemble with 20 members, accounting for the uncertainty introduced by interpolating the station data into the grid, or as ensemble mean. In this work only the ensemble mean was used. The different variables in the E-OBS dataset are derived mostly from station data. The elevation data in E-OBS is from the Global 30 Arc-Second Elevation Data Set by the United States Geological Survey (USGS). It is a raster Digital Elevation Model (DEM) with global coverage.

The variables relative humidity and shortwave (global) radiation were post-processed in the E-OBS dataset. While the gridding approach already ensured that no relative humidity values above 100 % occurred, further post-processing was performed setting all values below 5 % to missing values. The shortwave (global) radiation data was also post-processed in cases of unrealistic values. This occurred especially in areas of low station density. Daily shortwave (global) radiation sums below 3 % compared to the radiation at the top of the atmosphere and daily sums that exceeded the expected shortwave (global) radiation on a clear sky day were set to missing values (van den Besselaar and Copernicus Climate Change Service, 2023).

Data Preparation

The E-OBS data was clipped spatially to cover the area from 8°E to 18°E longitude and 46°N to 50°N latitude, matching the approximate extent of the LamaH-CE dataset. Furthermore, the data was clipped to cover the time period from 1981-01-01 to 2022-12-31. The reduced dataset was inspected for missing values, and it was found that only relative humidity and wind speed had missing values. The missing values are a result of the post-processing described in chapter 3.1.2.

For relative humidity 556148 values in total were missing over the area from 8°E to 18°E longitude and 46°N to 50°N latitude and over the time period from 1981-01-01 to 2022-12-31, meaning that 0.928 % of values were missing. For wind speed a total of 14664 values were missing over the same area and time period, meaning that 0.024 % of values were missing.

For each time step with missing values a spatial interpolation was performed using the *rioxarray.interpolate_na* module. The method is based on the *scipy.interpolate.griddata* function and a nearest neighbour interpolation was performed (Virtanen et al., 2020). The interpolated relative humidity and wind speed data in combination with the other gap-free E-OBS variables could then be used to estimate potential evapotranspiration (reference evapotranspiration) without missing values.

Calculation of Potential Evapotranspiration

To obtain a dataset that is comparable to the LamaH-CE dataset and can be used with a similar version of the EXP-HYDRO-bases conceptual model, potential evaporation was estimated as an additional variable for the E-OBS dataset. This was done using the FAO recommended Penman-Monteith equation as described in Annex A 7.1.1. The FAO guideline on computing reference evapotranspiration with the FAO Penman-Monteith equation also includes guidance on estimating missing climatic data. The process of computing potential evapotranspiration (reference evapotranspiration) is described in Annex 7.1.1. One variable, albedo, was obtained from the ERA5-Land dataset to compute net shortwave radiation in equation 7.8 from incoming shortwave radiation (Muñoz Sabater, 2019; Muñoz-Sabater et al., 2021).

The final dataset which throughout this study is referred to as E-OBS dataset hence includes E-OBS variables without missing values, E-OBS variables with interpolated missing values (relative humidity and wind speed), and potential evapotranspiration (reference evapotranspiration) estimated using the FAO Penman-Monteith approach.

The gridded data was then aggregated to the catchments of catchment delineation C from the LamaH-CE dataset using *xagg*[1] and *geopandas* (Jordahl et al., 2020). The area weighted aggregation to the catchments in the shapefile was performed using *xagg.pixel_overlaps* and *xagg.aggregate*.

3.2 Data Analysis

3.2.1 Clustering based on Catchment Characteristics

Clustering is a technique to group datapoints from a dataset in a way that datapoints within a cluster have the greatest similarity within the cluster and the clusters have the greatest dissimilarity between each other. There are multiple approaches to clustering, one of them is unsupervised machine learning for which K-means is a popular clustering algorithm (Sinaga and Yang, 2020). The goal in this work is to form groups of catchments that with similar catchment characteristics which we expect to show similarity in their hydrological response to meteorological inputs. These groups are referred to as *hydrological response units* and are mentioned in a similar way for example in Flügel (1995).

K-means clustering, minimises the within-cluster sum of squares criterion:

$$\sum_{i=0}^{n} \min_{\mu_j \in C_i} ||x - \mu_i||^2 \tag{3.1}$$

n: number of clusters, C_j: j-th cluster, x: data point within j-th cluster, μ_j: centroid of j-th cluster, $||x - \mu_i||^2$: squared Euclidean distance between the data point x and the centroid μ_j, (Pedregosa et al., 2011)

To implement the K-means clustering in Python the *sklearn.cluster.KMeans* class is used. To implement the algorithm, the number of clusters to be formed (n) has to be chosen (Pedregosa et al., 2011).

To evaluate which number of clusters results in the best clustering of the data points, the sum of squared distances of the datapoints to the cluster center, also called *inertia*, was computed. As a second evaluation metric the *silhouette score* was computed, which calculates the mean nearest-cluster distance (b) and the mean intra-cluster distance (a) and then computes the silhouette score as $(b - a)/max(a, b)$. The optimal value for the silhouette score is 1 while the worst value is -1 and values close to 0 suggest overlapping clusters (Pedregosa et al., 2011).

The LamaH-CE dataset contains 62 catchment attributes. To cluster the catchments using a K-means method, two approaches are used. In the first approach, all catchment with numerical values (59 out of 62) were used as features in the K-means clustering. To reduce the number of catchment characteristics used as features in the K-mean clustering, a subset has been selected for the second approach. In the second approach ten catchment characteristics have been selected based on a sensitivity analysis of catchment characteristics by Kratzert et al. (2019b) using the CAMELS-US dataset and the Morris method. The model used by Kratzert et al. (2019b) for this was an entity-aware LSTM model. Subsequently, the ten catchment characteristics with the highest sensitivity as well

[1] https://xagg.readthedocs.io/en/latest/

as their sensitivity score (in paranthesis) in decreasing order are listed: precipitation (68), aridity (56), area (50), mean elevation (46), high precipitation duration (41), fraction of snow (0.41), high precipitation frequency (0.38), mean slope (0.37), geological permeability (0.35) and fraction of carbonate sedimentary rock (0.34) (Kratzert et al., 2019b).

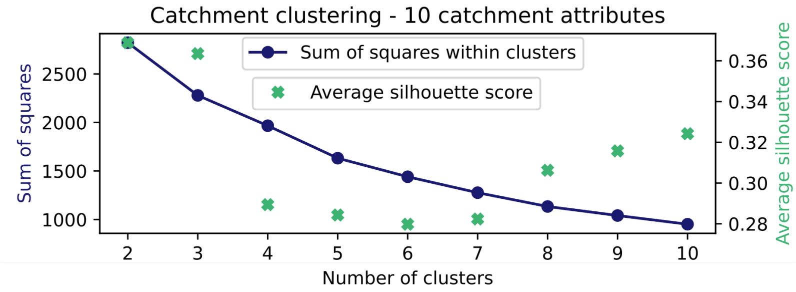

Figure 3.2: Illustration of K-means clustering

The corresponding, or most similar catchment attributes in the LamaH-CE dataset are: mean daily precipitation (p_mean), aridity (arid_1), area (area_calc), mean elevation (elev_mean), high precipitation duration (high_prec_du), fraction of snow (frac_snow), high precipitation frequency (high_prec_fr), mean slope (slope_mean), subsurface permeability (geol_perme) and fraction of carbonate sedimentary rock (gc_sc_fra) (Klingler et al., 2021). Mean daily precipitation, high precipitation duration, fraction of snow and high precipitation frequency are calculated based on precipitation data from ERA5-Land and in the case of fraction of snow, additionally mean temperature from ERA5-Land is used to calculate it. Aridity is computed as the ratio of mean reference evapotranspiration from Evapotranspiration (ET0) Climate Database v2 and mean precipitation from ERA5-Land (Trabucco and Zomer, 2019; Muñoz Sabater, 2019; Klingler et al., 2021).

Figure 3.2 illustrates the K-means clustering results for two to ten clusters as well as the values for the sum of square values and the silhouette scores using the second approach with ten catchment attributes as features. The result of the clustering is shown in figure 3.3 and will be discussed in the results and discussion chapters.

Based on the sums of squares as well as the silhouette score, K-Means clustering results for three, eight and then clusters were taken into further consideration and were paired with an analysis of the number of catchments per cluster as well as a manual analysis using satellite imagery to confirm geographical similarity. While showing overall inferior scores in terms of sum of squares and average silhouette score, the pattern across the number of clusters shows similarity. Despite a high silhouette score, catchment clustering with only two clusters was not considered as a minimum amount of three clusters was deemed necessary to represent the diversity of catchments in the LamaH-CE dataset catchment delineation C in the clustering process.

3.2.2 Selection of Representative Catchments

For the conceptual models BaseEOBS (E-OBS data) and BaseERA5L (LamaH-CE data) a representative catchment was selected for each cluster based on the KGE values closest to the median KGE value of that cluster. For the BaseEOBS model, the following catchments where selected: cluster 0: ID 241, cluster 1: ID 215, cluster 2: ID 581, cluster 3: ID 21, cluster 4: ID 277, cluster 5: 797, cluster 7: ID 432. For the BaseERA5L model the following catchments were selected: cluster 0: ID 334, cluster 1: 743, cluster 2: ID 439,

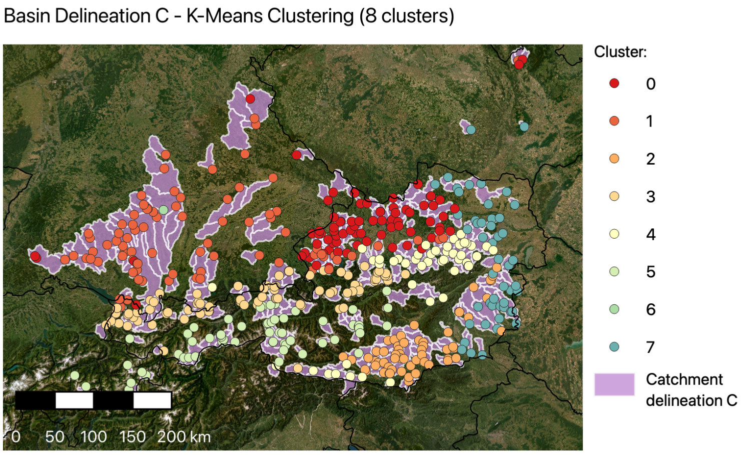

Figure 3.3: Catchments from basin delineation C (LamaH-CE dataset), the points represent the gauges of the respective catchments and the color-coding indicates the cluster as a result of the K-Means clustering with eight clusters based on catchments characteristics from the LamaH-CE dataset (Klingler et al., 2021)

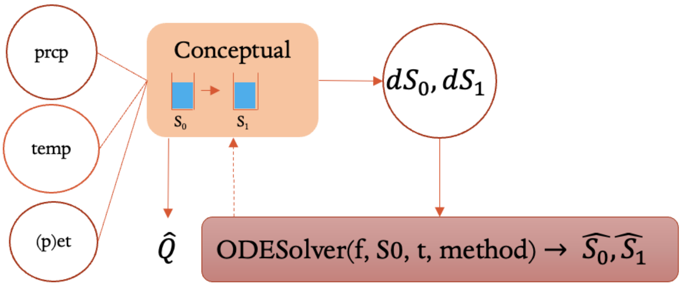
prcp
temp
(p)et
Conceptual
S_0
S_1
dS_0, dS_1
$\hat{Q}$
ODESolver(f, S0, t, method) → $\widehat{S_0}, \widehat{S_1}$

Figure 3.4: Conceptual model based on EXP-Hydro

cluster 3: ID 24, cluster 4: ID 330, cluster 5: ID 75, cluster 7: ID 383.

Additionally, one catchment for each dataset was chosen based on the KGE closest to the median KGE of all catchments from the conceptual model runs. For the E-OBS dataset this is catchment ID 572, for the LamaH-CE dataset this is catchment ID 79.

3.3 Models

3.3.1 Conceptual Model based on EXP-Hydro

EXP-Hydro

The EXP-Hydro model developed by Patil and Stieglitz (2014) is a rainfall-runoff bucket model that is spatially lumped and operates at a daily time-step. The model operates with two buckets, a catchment bucket and a snow accumulation bucket. Equation 3.2 describes the water balance of the catchment (or water) bucket and equation 3.3 describes the water balance of the snow bucket. The original model has three input variables: precipitation (P), temperature (T) and length of day (L_{day}) and two state variables: snow storage

(S_0) and catchment (or water) storage (S_1). The model contains six free calibration parameters which are listed in table 3.3 Patil and Stieglitz (2014). The conceptual model used in this work is based on EXP-Hydro but includes modifications. The current version of EXP-Hydro differes from the version described in the paper (Patil and Stieglitz, 2014) as the current version takes as input precipitation (P), temperature (T) and potential evapotranspiration (PET) instead of precipitation (P), temperature (T) and length of day (L_{day}) which means that equation 3.5 is not used in the EXP-Hydro version that is used in this work.

$$\frac{dS_1}{dt} = P_r + M - ET - Q_{bucket} - Q_{spill} \tag{3.2}$$

S_1: water storage catchment bucket [mm], P_r: precipitation as liquid rainfall [mm/d], M: snowmelt [mm/d], ET: evapotranspiration [mm/d], Q_{bucket}: generated runoff based on S [mm/d], Q_{spill}: capacity excess runoff [mm/d], (Patil and Stieglitz, 2014; Jiang et al., 2020)

$$\frac{dS_0}{dt} = P_s - M \tag{3.3}$$

S_0: storage snow bucket [mm], P_s: precipitation as snow [mm/d], M: snowmelt [mm/d] (Patil and Stieglitz, 2014; Jiang et al., 2020)

Table 3.3: Calibration parameters in the EXP-Hydro model including the calibration ranges as used by Patil and Stieglitz (2014)

Parameter	Description	Units	Lower lim.	Upper lim.
$S_{\max}$	Maximum storage catchment bucket	mm	100.0	1500.0
$Q_{\max}$	Maximum subsurface runoff	mm/d	10.0	50.0
D_f	Thermal degree-day factor	mm/d/°C	0.0	5.0
f	Rate of decline in runoff from catchment bucket	mm^{-1}	0.0	0.1
$T_{\max}$	Temperature above which snowmelt starts	°C	0.0	3.0
$T_{\min}$	Temperature threshold rain snow	°C	-3.0	0.0

Daily streamflow Q is the sum of Q_{bucket} and Q_{spill}. Q_{spill}, the capacity excess runoff, occurs when snowmelt and/or excess precipitation is available but the catchment storage S_1 is at maxmimum capacity S_{max}.

Potential evapotranspiration (ET_p, equation 3.5) in the original EXP-Hydro model is calculated based on Hamon's formula (Hamon, 1963). Actual evapotranspiration (ET_a, equaiton 3.4) is calculated based on potential evapotranspiration, catchment (or water) storage (S) and maximum water storage (S_{max}). Actual vapor pressure (e_a, equaiton 3.6) which is needed to calculate potential evapotranspiration is calculated based on temperature (T_a) (Patil and Stieglitz, 2014).

$$ET_a = ET_p(S/S_{max}) \tag{3.4}$$

ET_a: actual evapotranspiration [mm/d], ET_p: potential evapotranspiration [mm/d], S: water storage catchment bucket [mm], S_{max}: maximum storage capacity catchment bucket [mm], (Patil and Stieglitz, 2014)

$$ET_p = 29.8 L_{day} \frac{e_{sat}(T_a)}{T_s + 273.2} \tag{3.5}$$

ET_p: potential evapotranspiration [mm/d], L_{day}: length of day [h], e_{sat}: saturation vapor pressure [kPa] (Patil and Stieglitz, 2014)

$$e_a(T_a) = 0.611 exp(\frac{17.3 T_a}{T_a + 273.3}) \tag{3.6}$$

e_{sat}: saturation vapor pressure [kPa], T_a: temperature [°C] (Patil and Stieglitz, 2014)

Snowmelt in EXP-Hydro is calculated using a thermal degree-day model. The snow bucket accumulates only precipitaiton that falls as snow based on the temperature threshold (T_{min}). Id $S_0 > 0$ and $T_a > T_{max}$ snowmelt component (M) is calculated as follows:

$$M = min\{S_0, D_f(T_a - T_{max})\} \tag{3.7}$$

Else:

$$M = 0 \tag{3.8}$$

M: snowmelt [mm/d], S_0: snow storage [mm], D_f: thermal degree-day factor [mm/d/°C], T_a: temperature [°C], T_{max}: temperature above which snowmelt starts [°C] (Patil and Stieglitz, 2014)

BaseEOBS & BaseERA5L

The models *BaseEOBS* and *BaseERA5L* were developed based on the the current EXP-Hydro model which takes as input precipitation (P), temperature (T) and potential evapotranspiration (PET). To enable Automatic Differentiation, both models were created to by compatible with PyTorch. This was done by replacing the custom ODE solver by (Patil and Stieglitz, 2014), which is based on *scipy* (Virtanen et al., 2020) with ODE solvers from the *torchdiffeq* library (Chen, 2018), which is Python library providing a PyTorch implementation of differentiable ODE solvers. Furthermore, the loss functions or evaluations metrics were replaced by differentiable loss functions. The evaluation metrics used are described in chapter 3.5.

The model BaseEOBS was developed to run with E-OBS data it takes as input precipitation (P), temperature (T) and potential evapotranspiration (PET). The model BaseERA5L was developed to run with LamaH-CE data. Since potential evapotranspiration (PET) is not included in the LamaH-CE dataset, instead actual evapotranspiraiton (ET) is used as third input to the model. Accordingly, equation 3.4 is not used in this model as ET is a direct input to the model. Apart from this, the two models (BaseEOBS & BaseERA5L) are the same. The ODE solver solves an IVP and hence requires an initial condition. The equivalent of the differential equation in the model is the change in soil bucket and snow bucket storage over time depending on the current level of the storages and the current values of the meteorological forcings. Accordingly, the ODE solver requires an inital condition for the storages which are assumed to be empty in this work.

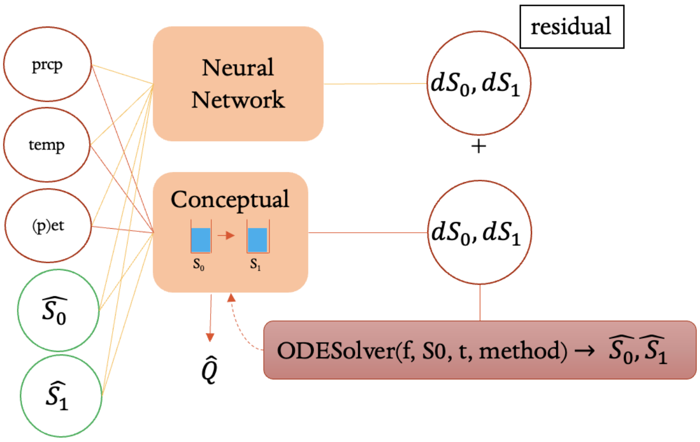

Figure 3.5: Hybrid model based on EXP-Hydro

3.3.2 Hybrid Models

Subsequently, a demonstration of the theoretical approach to creating the hybrid models in this work is introduced.

An IVP consists of an ODE and an initial condition:

$$\frac{dy}{dt} = f(t, y(t)) \tag{3.9}$$

$$y(0) = y_0 \tag{3.10}$$

The function f can be replaced by a neural network according to the Universal Approximation Theorem for Neural Networks (chapter 2.4.2):

$$\frac{dy}{dt} = NN(t, y(t), \theta) \tag{3.11}$$

In the context of this work the ODE is dS/dt, the change in system state over time. Furthermore, the system state and the change in system state also depends on the meteorological forcing variables $x(t)$:

$$\frac{dS}{dt} = NN(t, S(t), \theta, \overrightarrow{x}(t)) \tag{3.12}$$

Based on the *Universal Differential Equation* approach introduced in Rackauckas et al. (2020) a combination of a theoretical model and a neural network is created:

$$\frac{dy}{dt} = g(t, y(t)) + NN(t, y(t), \theta) \tag{3.13}$$

Applied to the system state and including the meteorological variables this becomes:

$$\frac{dS}{dt} = f(t, S(t), \overrightarrow{x}(t) + NN(S(t), \theta, x(t)) \tag{3.14}$$

t: time, $\overrightarrow{x}(t)$: array of meteorological forcings, θ: neural network parameters, $S(t)$: storage state at time t

Two hybrid models with two versions each, one with the Euler method in the ODE solver and on with the RK4 method in the ODE solver, were evaluated in this work. HybridEOBS for the modelling with E-OBS data and HybridERA5L for the modelling with the ERA5-Land-based LamaH-CE data. The HybridEOBS and HybridERA5L models were created by combining the conceptual base model (BaseEOBS and BaseERA5L) with a neural network. The neural network consists of an input layer with five inputs, a *Tanh* activation function, a hidden layer with sixteen nodes, a *LeakyReLU* activation function and an output layer with 2 output nodes. The neural network takes as input the current value of the three meteorological forcings precipitation (P), mean temperature (T_{mean}) and for the Hybrid V1 model potential evapotranspiration (PET) and for the Hybrid V2 model actual evapotranspiration (ET). Additionally, the neural network, for each time step, takes as input the storage values for the soil bucket and snow bucket which is computed by the ODE solver as it it solves the IVP which is initialized with the initial storages and which is here, as it is done for the conceptual models, assumed to be initially empty.

To account for the fact that many catchments receive very little snow fall, each of the hybrid models was also developed so that the neural network only connects to the water storage. Accordingly, the neural network has four input nodes and one output node. Due to minor differences in performance and the computational limitations which are discussed in the discussion, the models e excluded from the model comparison in this study.

The conceptual part of the hybrid models uses the calibrated parameters from the calibration of the conceptual models. Hence, only the neural network's weights and biases are optimized in training. The weights of the neural network are initialized with a $mean = 0$ and $std = 0.1$ to ensure that the random initialization of the neural network does not result in too large deviations of the prediction from the hybrid model from the streamflow prediction that would be produced by the conceptual part of the model.

To create a differentiable modelling framework, the conceptual part was re-written to replace aspects of the model that are not compatible with gradient tracking which is needed for the backpropagation. However, the conceptual part of the hybrid model performs the same processes as the current version of the original EXP-Hydro model.

3.3.3 Long Short-term Memory (LSTM) Model

The LSTM uses a moving window of 365 days and predicts one time step ahead. Accordingly, the input data has to be prepared by create sequences of 365 days and then shifting them by one day for the whole training, validation and testing period.

The LSTM models LstmEOBS and LstmERA5L are using E-OBS and LamaH-CE data respectively. LstmEOBS takes three meteorological forcings as input: precipitation (P), mean temperature (T_{mean}) and potential evapotranspiration (PET). Instead of potential evapotranspiration, LstmERA5L takes as third input actual evapotranspiration (ET). In different experiments, additional input forcings were added to the LSTM models.

Apart from that, the LSTM models have the same structure. They consist of an input layer with three or six input nodes, two LSTM layers with ten nodes, a dropout layer to avoid overfitting, a linear layer and a *LeakyReLU* activation function. The dropout rate during training is set to 30

Table 3.4: Training/calibration, validation and testing time periods

	Training / Calibraiton	Validation	Testing
Start	1981-10-01	2001-10-01	2005-10-01
Stop	2001-9-30	2005-9-30	2017-9-30

3.4 Calibration and Training

The conceptual models were calibrated from 1981-10-01 to 2001-10-01 and evaluated from 2005-10-01 to 2017-10-01. The hybrid models and LSTM models were trained on the same time period and also evaluated on the same time period. KGE is used as loss function for the calibration of the conceptual models as this is also used as loss function in the current version of EXP-HYDRO Patil and Stieglitz (2014). The hybrid models and LSTM models were trained and evaluated on the same time period. The validation period is used to observe the training process and to save the best set of parameters for both the conceptual models as well as the best weights and biases for the neural network for the hybrid and LSTM models. For the hybrid and LSTM models the MSE is used as loss function.

The meteorological forcings and discharge observations used for the conceptual and hybrid models were not scaled as the static parameters for the conceptual models and hence also for the conceptual component of the hybrid models were calibrated within the calibration ranges provided in (Patil and Stieglitz, 2014).

Scaling
The meteorological input forcings and discharge observations for the LSTM models were scaled using the *MinMaxScaler* from *scikit-learn* to ensure that the features are in the same value range to optimise the training process. The data is normalised by using the value ranges of the training sequences of the input forcings and discharge observations. Subsequently, the validation and testing timeseries are scaled using the scaler that was fitted to the training data. Equations 3.15 and 3.16 are used for this (Pedregosa et al., 2011). After the training process, the data is re-scaled using the inverse of the described process.

$$X_{\text{std}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{3.15}$$

$$X_{\text{scaled}} = X_{\text{std}} \times (\text{max} - \text{min}) + \text{min} \tag{3.16}$$

$x_{i,t}$: *MinMaxScaler* equations from *scikit-learn* (Pedregosa et al., 2011)

3.4.1 Particle Swarm Optimization

Particle Swarm Optimization is a paradigm of Swarm Intelligence (SI) and it is a stochastic method for optimization of continuous non-linear functions. It is used in different scientific fiels and is conceptually simple and computationally inexpensive (Kennedy and Eberhart, 1995; Gad, 2022).

A swarm is made of of many agents that locally interact with their environment and the whole concept is based on the behaviour of animals such as a fish school. An underlying concept in SI systems is that local interactions of agents has global impacts. In the context of Particle Swarm Optimization (PSO) the agents are referred to as particles which move through the space of the problem with a certain speed and direction. As a combination of a particles best location, current location as well as the location of the other particles, the direction for movement in the next iteration is determined. The particle swarm moves through the space and approaches the objective function minimum or maximum (Gad, 2022).

If the objective function is to be minimized, the PSO algorithm can be illustrated as follows:

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{3.17}$$

$$v_i(t+1) = w \cdot v_i(t) + c_1 \cdot r_1 \cdot (p_{best_i} - x_i(t)) + c_2 \cdot r_2 \cdot (g_{best} - x_i(t)) \tag{3.18}$$

$$\text{If } f(x_i(t+1)) > f(p_{best_i}), \rightarrow p_{best_i} = x_i(t+1) \tag{3.19}$$

$$\text{If } f(p_{best_i}) > f(g_{best}), \rightarrow g_{best} = p_{best_i} \tag{3.20}$$

$x_{i,t}$: particle i and time t, $v_{i,t}$: velocity of particle i at time t, p_{best}: best position of particle i, g_{best}: best position of all particles, w, c_1, c_2: controlling paramters, (Patil and Stieglitz, 2014)

3.5 Evaluation Metrics

Mean Squared Error

The MSE computes the average squared difference between prediction and observation. The lower the MSE the smaller the error between prediction and observation.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Q_{obs,i} - Q_{sim,i})^2 \tag{3.21}$$

$Q_{sim}(i)$: simulated discharge, $Q_{obs}(i)$: observed discharge

Nash-Sutcliffe Efficiency

The NSE (equation 3.22) is a commonly used performance metric with an upper limit of $NSE = 1$ indicating a perfect fit with Q_{sim} being identical to Q_{obs} whereas a value of $NSE = 0$ indicates "that the model simulations have the same explanatory power as the mean of the observations" (Knoben et al., 2019). Values $NSE < 0$ indicate a model performance that is inferior to the mean of the observations.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Q_{sim,i} - Q_{obs,i})^2}{\sum_{i=1}^{n}(Q_{obs,i} - \bar{Q_{obs}})^2} \tag{3.22}$$

$Q_{sim}(i)$: simulated discharge, $Q_{obs}(i)$: observed discharge, $\bar{Q}_{obs}$: mean observed discharge (Knoben et al., 2019; Nash and Sutcliffe, 1970)

Kling-Gupta Efficiency
Values of $KGE = 1$ indicate a perfect fit, $KGE = 0$ indicates that the explanatory power of the model is equivalent to the mean of the observations while values of $KGE < 0$ indicate inferior explanatory power of the model compared to the mean of the observations (Knoben et al., 2019).

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{3.23}$$

$$r = \frac{\sum_{i=1}^{n}(Q_{obs,1} - \bar{Q}_{obs})(Q_{sim,i} - \bar{Q}_{sim})}{\sqrt{\sum_{i=1}^{n}(Q_{obs,i} - \bar{Q}_{obs})^2} \cdot \sqrt{\sum_{i=1}^{n}(Q_{sim,i} - \bar{Q}_{sim})^2}} \tag{3.24}$$

$$\alpha = \frac{\sigma_{sim}}{\sigma_{obs}} = \frac{\sqrt{\sum_{i=1}^{n}(Q_{sim,i} - \bar{Q}_{sim})^2}}{\sqrt{\sum_{i=1}^{n}(Q_{obs,i} - \bar{Q}_{obs})^2}} \tag{3.25}$$

$$\beta = \frac{\mu_{sim}}{\mu_{obs}} = \frac{\bar{Q}_{sim}}{\bar{Q}_{obs}} \tag{3.26}$$

r: linear correlation between observations and simulations, α: flow variability error, β: bias term, σ: standard deviation, μ: mean, Q_{sim}: discharge prediction, Q_{obs}: discharge observation (Knoben et al., 2019; Gupta et al., 2009)

4 Results

4.1 Data Analysis

In this chapter the the catchments are analyzed using the Budyko framework to analyze the hydrological properties of the catchments. Furthermore, the deviation of the catchments from the Budyko curve is analyzed to draw conclusions on the in the ERA5-Land-based meteorological forcings LamaH-CE dataset.

4.1.1 Budyko Framework

Figure 4.1 visualises the catchments from basin delineation C (LamaH-CE dataset) in the Budyko framework based on mean actual evapotranspiration and mean precipitation (from Oct 1989 to Sept 2009) from the LamaH-CE dataset, based on data from ERA5-Land as well as mean potential evaporation also from the LamaH-CE dataset based on the ERA5-Land dataset (Muñoz Sabater, 2019; Klingler et al., 2021). In figure 4.2 the catchments of basin delineation C are shown using the same data for mean actual evapotranspiration and mean precipitation but now using mean reference evapotranspiration from the LamaH-CE dataset based on the Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2 (Muñoz Sabater, 2019; Klingler et al., 2021; Trabucco and Zomer, 2019). It can be seen that using the evapotranspiration from the ERA5-Land dataset results in significant deviation of the catchments from the Budyko curve which is not the case when using the reference evapotranspiration based on the Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2.

4.1.2 Catchment Clustering

In this chapter the results of the catchment clustering which is described in chapter 3.2.1 are examined. At first the location of the catchment clusters in the Budyko framework is analysed, this will later be used to discuss the model performance on the different types of clusters. Furthermore, the catchment characteristics and hydrological signatures for the respective clusters are analyzed. The analysis of the catchment characteristics in figure 4.4 and the hydrological signatures in figure 4.5 is complemented with the statistical metrics (median, mean, 95th and 5th percentile) in Appendix A. The clustering used in this work is based on ten catchment attributes and divides the catchments into 8 clusters. Each cluster contains between 20 and 59 catchments and the median catchment size of the clusters ranges from 67,2 km^2 to 126,1 km^2.

Figure 4.3 displays the the catchments from basin delineation C in the Budyko framework based on the same data as figure 4.2, which is mean actual evapotranspiration and mean precipitation (from Oct 1989 to Sept 2009) from the LamaH-CE dataset, based on data from ERA5-Land as well as mean reference evapotranspiration (from 1970 to 2000) from the LamaH-CE dataset based on the Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2 (Muñoz Sabater, 2019; Klingler et al., 2021; Trabucco

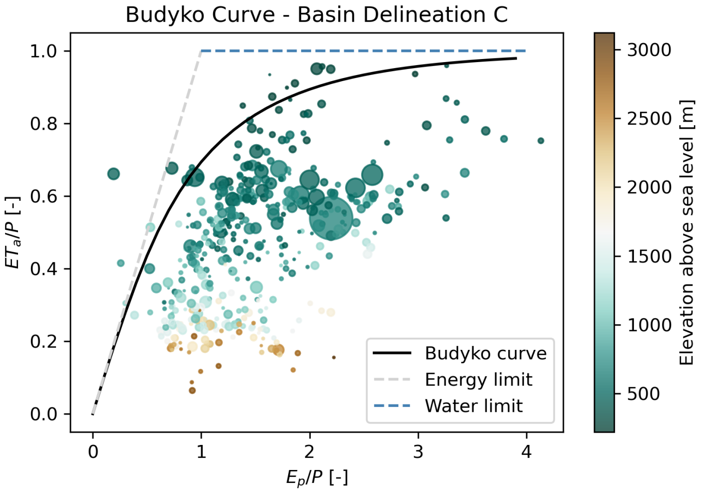

Figure 4.1: ET_a: mean actual evapotranspiration [mm/d] (from ERA5-Land)[1], Ep: potential evaporation (from ERA5-Land) [1], P: preciptation [mm/d] (from ERA5-Land)[1], the color indicates the mean elevation of the catchments above sea level [m] and the size of the circles indicates the relative catchment area

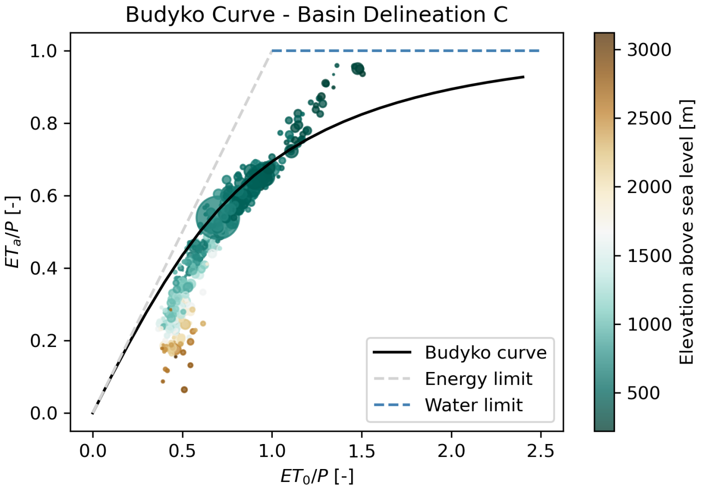

Figure 4.2: ET_a: mean actual evapotranspiration [mm/d] (from ERA5-Land)[1], ET_0: reference evapotranspiration (from Global Aridity Index and Potential Evapotranspiration [mm/d] (ET0) Climate Database v2)[2], P: preciptation [mm/d] (from ERA5-Land)[1], the color indicates the mean elevation of the catchments above sea level [m] and the size of the circles indicates the relative catchment area

and Zomer, 2019). The color-coding indicates the result of the K-Means clustering with eight clusters based on the catchment characteristics. It should be noted that the K-Means clustering included mean daily precipitation (from 1 October 1989 to 30 September 2009) as feature, hence the same variable, averaged over the same period of time, was an input feature to the clustering as well as part of the plot of the Budyko curve, explain part of the relatively clear distinction of catchment clusters in the Budyko curve. It can be seen in figure 4.3 that catchments from cluster 7 and cluster 5 form relatively distinct conglomerates at the top and bottom of the spread. Catchments from cluster 4 span a relatively wide range from top to bottom parallel to the energy limit line. Catchments from cluster

0, 1 and 2 seem to lie in a similar region within the Budyko framework and catchments from cluster 3 lie distinctly clustered below the middle. Cluster 6 only contains one large catchment.

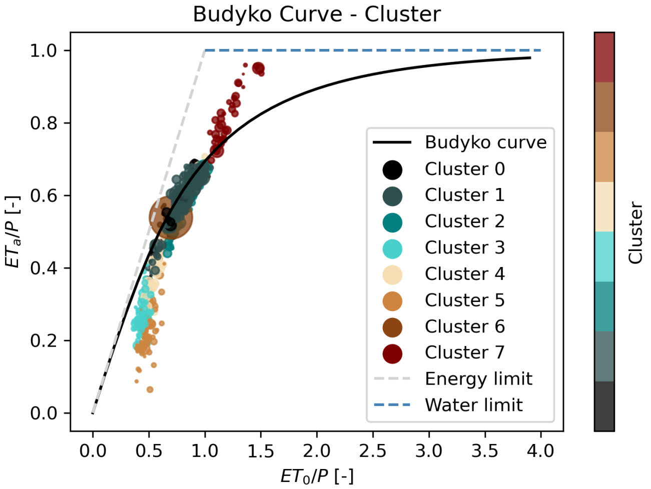

Figure 4.3: ET_a: mean actual evapotranspiration [mm/d] (from ERA5-Land)[1], ET_0: mean daily reference evapotranspiration (from Global Aridity Index and Potential Evapotranspiration [mm/d] (ET0) Climate Database v2)[2], P: mean daily preciptation [mm/d] (from ERA5-Land)[1], the colors indicate the respective cluster based on the KMeans clustering with 10 catchment characteristics and 8 clusters, the size of the circles indicates the relative catchment area

Cluster 0

The catchments from cluster 0 are mostly located in Northern Autria near the border to Germany and the Czech Republic. The catchments lie at a low elevation and the mean slope is small. The catchment receives very little snow.

Cluster 1

The catchments from cluster 1 are located primarily in Southern Germany and display in many ways similar catchment characteristics and hydrological signatures as catchments from cluster 0. Differences between the two catchment clusters are particularly clear in the subsurface permeability and in the fraction of carbonate sedimentary rock.

Cluster 2

Also the catchments from cluster 2 display in many aspects similarities to catchments from cluster 0 and 1 but does display a higher fraction of precipitation falling as snow and a higher mean elevation. All three cluster also plot in a similar region of the Budyko framework (figure 4.3).

Cluster 3

Parallel to the highest section of the mountain range lie catchments from cluster 3 with a mean elevation of 1339.1 m.a.s.l. The cluster display the second highest mean slop and is characterized by high mean precipitation, however with a low frequency of high precipitation events (defined as precipitation events five times larger than the mean precipitation). This indicates that these catchments likely experience relatively frequent and moderate precipitation events. The catchments experience a lot of snow fall with 28 % of precipitation falling as snow on average. The catchments have a high runoff ratio and plot low in the Budyko framework, indicating that a relatively small fraction of precipitation

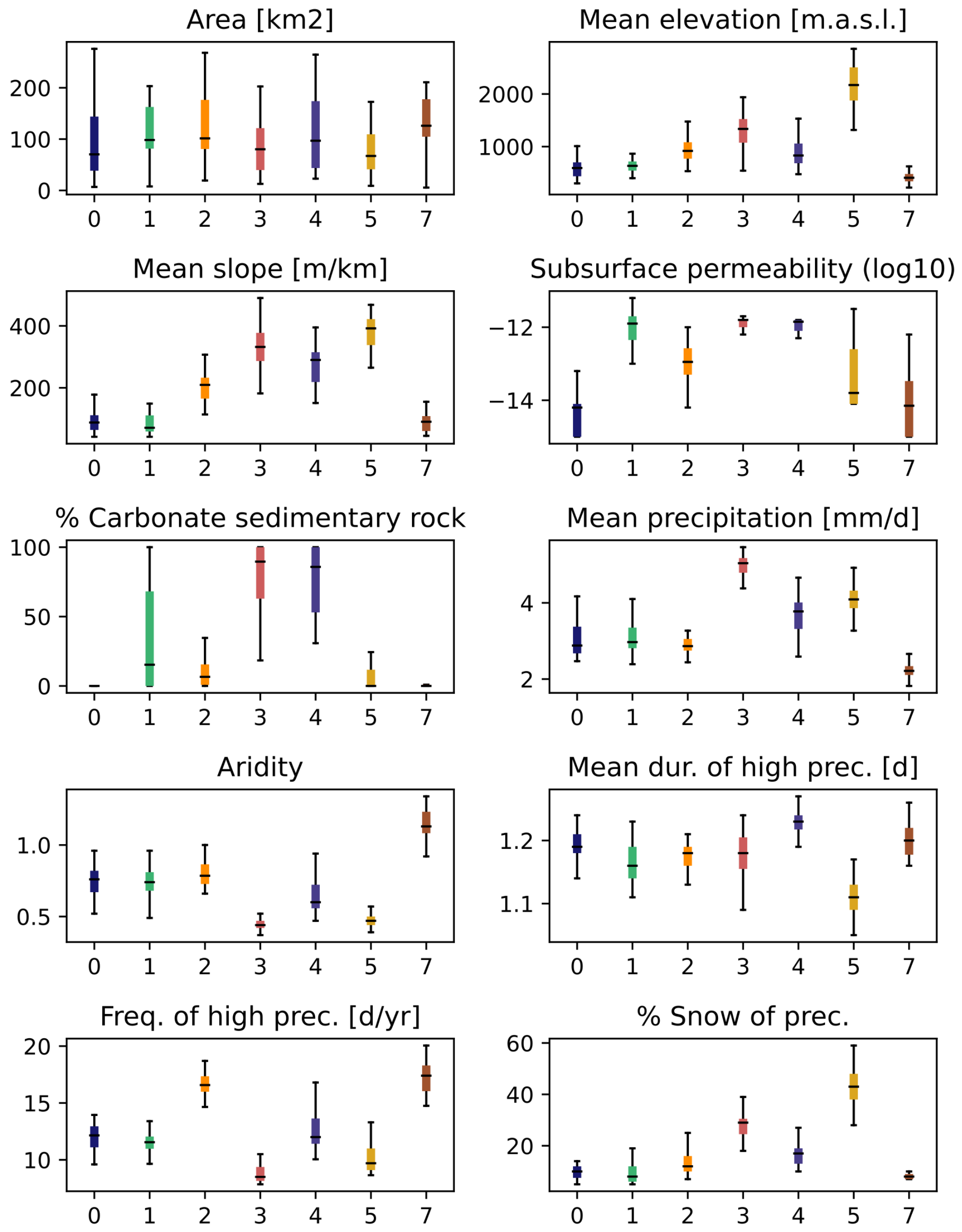

Figure 4.4: Boxplots of catchment characteristics per cluster
Variables: area_calc: Area [km^2], elev_mean: Mean elevation [m.a.s.l.], slope_mean: Mean slope [m/km], p_mean: Mean precipitation [mm/d], arid_1: Aridity as the ratio of mean daily reference evapotranspiration to mean precipitation, hi_prec_cu: Mean duration of high-precipitation events(number of consecutive days with 5 times mean daily precipitation) [d], hi_prec_fr: Frequency of dry days (<1 [mm/d] precipitation) [d/yr], geol_perme: Subsurface permeability [-], gc_sc_fra: Fraction of carbonate sedimentary rocks [-], frac_snow: Fraction of precipitation falling as snow [-]

evaporates and a high fraction is translated into runoff.

Cluster 4

The catchments which are mostly located in the lower mountain regions are characterized by a relatively high mean slope and relatively high mean precipitation as well as relatively high mean runoff and accordingly a comparatively high runoff ration. The discharge is characterized by comparatively high low-flow values and moderate high-flow values indicating a relatively moderate flow regime.

Cluster 5

The catchments from cluster 5 have by far the highest mean elevation (2165.9 m.a.s.l.) and are located in the high mountain regions of the alps. The catchments have the highest mean slope and experience relatively much precipitation on average paired with high runoff and accordingly high runoff ratios. It can be seen that at least part of the catchments experience both high-flow and low-flow events frequently and with high duration, indicating peaky runoff regimes.

Cluster 7

Cluster 7 contains the catchments with the lowest mean elevation (411.7 m.a.s.l.) and the second-lowest mean slope. The catchments receive the least amount of precipitation, on average 2.22 [mm/d]. However, in comparison with the other catchment clusters the mean daily runoff is even lower (0.44 [mm/d]) which results in a very low runoff ratio with an average value of 0.19, which means that on average 19 % of precipitation is translated into runoff. The catchments plot close to the water limit in the Budyko curve and show the highest aridity of all clusters, indicating that a significant portion of the precipitation evaporates. Catchments in cluster 7 at the same time also receive high precipitation most frequently on average and also experience high-flow relatively often, indicating a dynamic runoff regime.

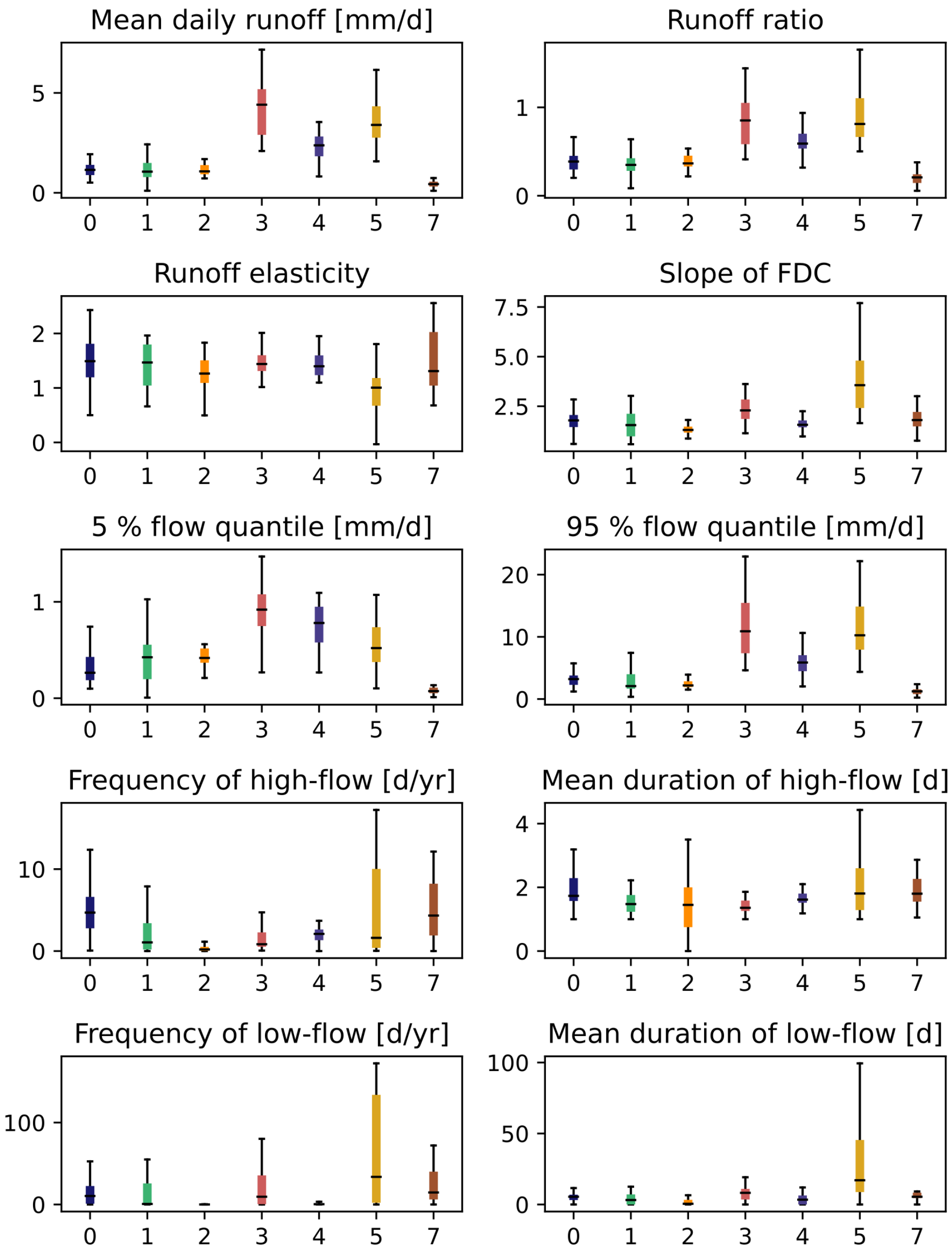

Figure 4.5: Boxplots of hydrological signatures per cluster
Variables: q_mean: Mean daily runoff [mm/d], runoff_ratio: ratio of mean daily runoff and mean daily precipitation [-], baseflow_index_ladson: ratio of mean daily – baseflow and mean daily discharge; hydrograph separation is performed using the digital filter [-], Q5: 5% flow quantile (low flow) [mm/d], Q95: 95% flow quantile (high flow) [mm/d], high_q_freq: Frequency of high flow days (>9 times median daily flow) [d/yr], high_q_dur: Mean duration of high-flow events [d], low_q_freq: Frequency of low-flow days (<0.2 times mean daily flow) [d/yr], low_q_dur: Mean duration of low-flow events [d]

4.2 Model Analysis

In the following section, the six models developed in this study are evaluated and compared. Table 4.1 gives an overview of the models, including the dataset used and the meteorological forcings. The conceptual models and the hybrid models were implemented using two different ODE solvers, (1) the Euler method and (2) the fourth-order Runge-Kutta method. Chapter 4.2.1 examines the impact of the method of the ODE solver as well as the selected data source on model performance. Chapter 4.2.2 compares the performance of the conceptual models, hybrid models and LSTM models and assess the performance on different types of catchments.

Table 4.1: Overview of the models including the data source, method for the ODE solver, input variables and the loss function used during calibration (conceptual models) and trainin (hybrid and LSTM models).

Model Name	Dataset		Solver		Input Forcings	Loss Function
	ERA5L	E-OBS	Euler	RK4		
BaseERA5L	🟦		🟧	🟧	p, et, t	KGE
BaseEOBS		🟦	🟧	🟧	p, pet, t	KGE
HybridERA5L	🟦		🟧	🟧	p, et, t	MSE
HybridEOBS		🟦	🟧	🟧	p, pet, t	MSE
LstmERA5L	🟦				p, et, t	MSE
LstmEOBS		🟦			p, pet, t	MSE

4.2.1 Impact of ODE Solver and Meteorological Forcings on Model Performance

ODE Solver

To examine the effect of the method of the ODE solver, the models BaseERA5L and BaseEOBS were compared in their performance across all catchments that are selected in this study. To assess the impact of the data source of the meteorological input forcings, the models were compared using the LamaH-CE (figure 4.7) and E-OBS (figure 4.6) dataset.

The median NSE and KGE values for the BaseERA5L model using the Euler method for the ODE solver are 0.05 and 0.46 on the test period while they are 0.30 and 0.54 when using the RK4 method for the ODE solver. This shows that the use of the RK4 method leads improvements in performance.

For the BaseEOBS model, the median NSE and KGE values on the test period using the Euler method are 0.25 and 0.56 while for the RK4 method they are 0.40 and 0.60. This confirms the performance improvement when using the RK4 method over the Euler method for the ODE solver. For a more detailed overview see table 7.9 (Appendix B).

Comparison of the HybridEOBS model with the Euler and RK4 method on the seven representative catchments based on the median KGE from the BaseEOBS model runs reveal that also for the hybrid model the performance is generally superior in most cases when using the RK4 method. These results are shown in Appendix B in table 7.10. The results and discussion of the hybrid models therefore focus on the models using the RK4 solver.

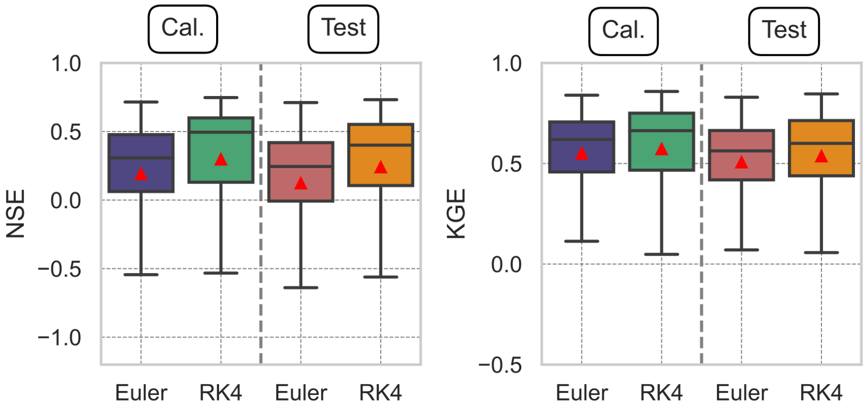

Figure 4.6: E-OBS Dataset: Boxplots of the Nash-Sutcliffe Efficiency and Kling-Gupta Efficiency values for all catchments from the Base model using E-OBS data. Boxplots for model runs using an ODE solver with Euler and RK4 method, outliers are excluded in the plot. The red triangle represents the mean.

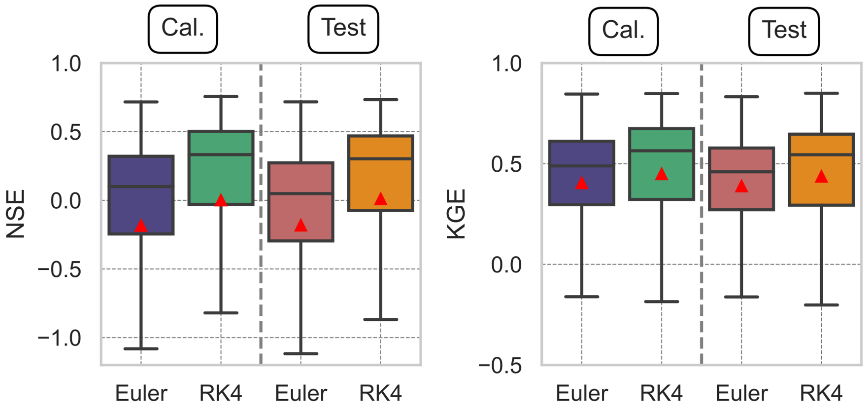

Figure 4.7: ERA5-Land Dataset: Boxplots of the Nash-Sutcliffe Efficiency and Kling-Gupta Efficiency values for all catchments from the Base model using ERA5-Land data. Boxplots for model runs using an ODE solver with Euler and RK4 method, outliers are excluded in the plot. The red triangle represents the mean.

Meteorological Forcings

Furthermore, higher median and mean NSE and KGE values as well as a lower MSE values when comparing the E-OBS dataset to the LamaH-CE (ERA5-Land), paired with smaller interquartile ranges show that the base models perform better using the E-OBS dataset as opposed to the LamaH-CE (ERA5-Land) dataset.

Similar patterns of performance improvement when using the E-OBS dataset over the LamaH-CE (ERA5-Land) dataset have been identified when comparing the hybrid models (HybridERA5L, HybridEOBS) on catchments 241, 215, 581, 21, 277, 797, 432, 572. For reasons of conciseness these results are not included in this study and further experiments with the hybrid models were restricted to the E-OBS dataset.

4.2.2 Model Comparison and Performance on Different Type of Catchments

The conceptual models (BaseEOBS, BaseERA5L) and LSTM models (LstmEOBS, LstmERA5L) were calibrated/trained and evaluated on all 246 catchments selected for this study. Due to computational limitations which are discussed in 5.2 the hybrid model was trained and evaluated on 14 catchments. In chapter 4.2.2 the models are compared on the individual catchments and in chapter 4.2.2 the model performance is analysed considering the different catchment characteristics of the catchment clusters.

Model Comparison on Individual Catchments

At first the models are compared on one representative catchment per cluster based on the catchment with the KGE value closest to the median KGE value of the respective cluster from the BaseEOBS model runs. Based on the findings in chapter 4.2.1 the RK4 solver is used for the ODE solver of the base model and hybrid model, Furthermore, the E-OBS dataset is chosen for the meteorological forcings. The results as displayed in table 4.2. The best NSE value for a certain catchment in most cases corresponds also to the best KGE value for that catchment. The base model scores best on catchment 277, the hybrid model scores best on catchments 581, 797, and for catchment 215 the KGE value is highest too. The best NSE value for 215 is achieved by the LSTM model and the LSTM model furthermore scores best on catchments 241, 215 and 21 for botch NSE and KGE. Overall the hybrid model outperforms the conceptual model on both NSE and KGE in four cases while the opposite is true in two cases. The largest performance (NSE) improvements of HybridEOBS over BaseEOBS are achieved on catchments 581 and 432 were the conceptual model achieves low NSE scores of 0.16 and 0.13. The LSTM model achieves both the highest NSE and KGE scores (catchment 215 and 21) as well as the lowest NSE and KGE scores (catchment 432).

The three models are compared on another seven catchments, in this case each catchment represents the catchment with the KGE value closest to the 75th percentile of the BaseEOBS runs. The results are displayed in tabel 4.3.

The BaseEOBS and HybridEOBS mdoels perform almost identical on catchments 220 and 219 with respect to NSE and KGE values. A particularly large performance improvement of HybridEOBS over BaseEOBS is again observed where the conceptual model scores lowest (catchment 586). The hybrid model improves upon the conceptual model also on catchment 214 where the conceptual model scores relatively high. Overall the hybrid model performs better in four out of seven cases in NSE score, equal in two cases and performs worse in one case. For the KGE scores the hybrid model performs better in three out of

Table 4.2: Comparison of the BaseEOBS, HybridEOBS and LstmEOBS models using the NSE and KGE values for the test period (Oct. 2005 - Oct. 2017). One representative catchment per cluster based on the median KGE of the BaseEOBS model for the respective cluster. Coloring indicates performance relative to the other models

	BaseEOBS		HybridEOBS		LstmEOBS	
ID	NSE	KGE	NSE	KGE	NSE	KGE
241	0.54	0.64	0.52	0.58	0.52	0.50
215	0.45	0.62	0.53	0.68	0.69	0.69
581	0.16	0.56	0.58	0.74	0.40	0.49
21	0.31	0.59	0.33	0.49	0.68	0.82
277	0.52	0.56	0.50	0.54	0.49	0.48
797	0.46	0.61	0.54	0.68	0.47	0.63
432	0.13	0.58	0.24	0.63	-0.68	-0.77

seven cases, worse in three out of seven cases and once it performs identical. The LSTM achieves the overall highest scores (catchment 534) and lowest scores (catchment 348).

Table 4.3: Comparison of the BaseEOBS, HybridEOBS, and LstmEOBS models using the NSE and KGE values for the test period (Oct. 2005 - Oct. 2017). One representative catchment per cluster based on the 75th percentile KGE value of the BaseEOBS model for the respective cluster. Coloring indicates performance relative to the other models.

	BaseEOBS		HybridEOBS		LstmEOBS	
ID	NSE	KGE	NSE	KGE	NSE	KGE
220	0.56	0.74	0.56	0.73	0.58	0.62
219	0.55	0.66	0.55	0.67	0.62	0.58
586	0.23	0.62	0.54	0.77	0.48	0.71
214	0.69	0.73	0.72	0.80	0.68	0.69
330	0.61	0.70	0.65	0.66	0.64	0.58
534	0.57	0.75	0.65	0.75	0.82	0.90
348	0.32	0.65	0.25	0.61	-0.05	-0.32

Model Performance and Catchment Characteristics

As discussed in chapter 4.1.2, the study area encompasses a wide range of catchment characteristics. It includes catchments at mean elevations as low as 221 m.a.s.l. and as high as 2858 m.a.s.l. There are large differences in mean slope and mean precipitation. Some catchments have a high snow coverage and others have relatively high aridity. As described in chapter 4.1.2, the catchments have been clustered based on ten catchment characteristics into groups that are expected to show similarities in their hydrological responses. In this chapter an analysis of performance of the conceptual and LSTM models on the different clusters as well as the performance of the hybrid model on the representative catchments of these clusters are analysed. Furthermore the hydrological signatures of the discharge observations are considered in the analysis. See figure 4.4 for an overview of the catchment characteristics analysis and figure 4.5 for an overview of the hydrological signature analysis.

The performance of the BaseEOBS and the LstmEOBS model on the catchment clusters is

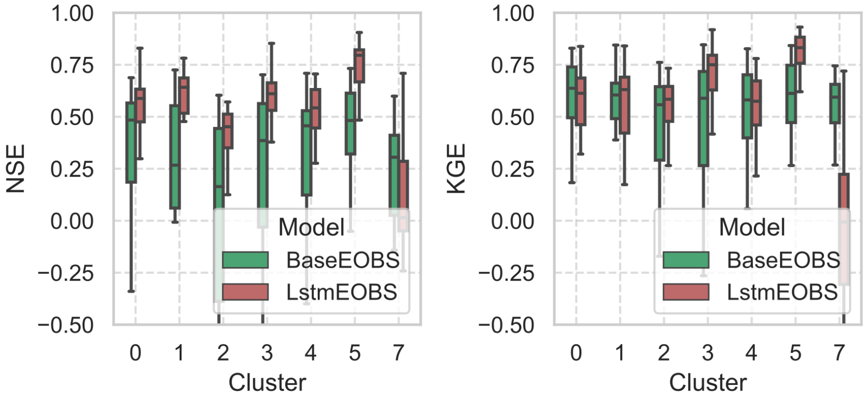

Figure 4.8: Comparison of the performance of the BaseEOBS and LstmEOBS accross the catchment clusters. Outliers are excluded in the plot and the lower limit of the y-axis is set to -0.5.

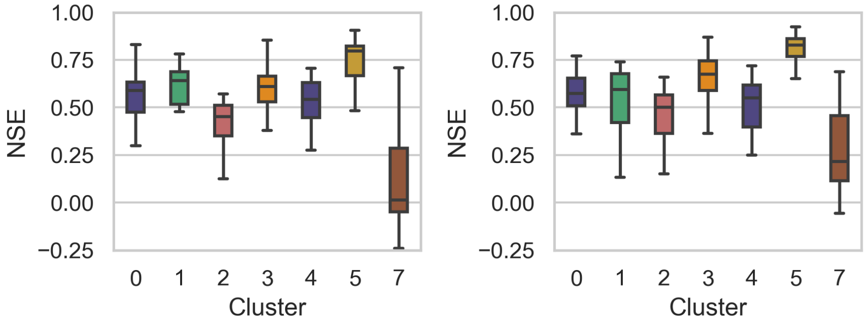

Figure 4.9: Comparisong of the original LSTM model (left) to a LSTM model with additional input forcings (right)
Data source: E-OBS dataset.

visualised in figure 4.8. It can be seen in both the boxplots for both performance metrics that their performance follows in a similar pattern across the cluster. The LstmEOBS model significantly outperforms the BaseEOBS model with respect to NSE on all clusters except for cluster seven, showing higher median values as well as smaller interquartile ranges.

The model performance for the catchments in cluster 0 and 1 is relatively similar which matches the findings in chapter 4.1.2 that these clusters display similarities across several catchment characteristics and hydrological signatures.

4.2.3 Impact of Adding additional Meteorological Input Forcings

Conceptual models can only take the meteorological variables as input that are considered in the model structure. LSTM models are very flexible with respect to their inputs. The hybrid models add this flexibility to the conceptual model without changing the inputs to the conceptual component of the model. Figure 4.9 compares the original LstmEOBS model which takes precipitation, mean temperature and potential evapotranspiration as input to an expanded version that takes as additional inputs minimum temperature, maximum temperature, humidity, pressure, wind speed, solar radiation, albedo, length of day.

The addition of additional input forcings improce the median NSE of the LSTM model on five of the seven clusters, it remains approximately the same on cluster zero while it deteriorates on cluster one.

Adding the same additional meteorological input forcings to the neural network component of the hybrid model while feeding the same meteorological input forcings to the conceptual part of the model did not lead to any performance improvements in experiments on the representative catchments. Observations of the training progress indicate that the additional inputs to the neural network lead to a larger initial deviation from the conceptual prediction. Additionally the training process converged significantly slower and was more prone to overfitting on the training data. Accordingly, these results are not included here but are instead discussed in chapter 5.1.3.

5 Discussion

5.1 Model Results

The aim of this study was to develop a framework for hybrid modelling in hydrology using a Neural Ordinary Differential Equations Approach. Subsequently, the results presented in chapter 4.2 will be discussed and the research questions will be answered. The section does not strictly follow the order of the research questions as different different parts of the results contribute to the same research questions. Answering SQ3 is divided into improvement techniques that were implemented leading up to the HybridEOBS model that is used for comparison with the conceptual and LSTM model and into an exploration of further approaches to improve the performance.

5.1.1 Model Performance and Comparison (SQ1 & SQ3)

ODE Solver
The conceptual models as well as the conceptual components contain as central element an ODE solver which then numerically approximates the trajectory of the system states (snow storage and water storage) using an initialization for each storage as initial condition. Accordingly, the method chosen for the ODE solver was explored first in the assessment of pathways to improving the conceptual and hybrid models. The Euler solver as a simple method and the RK4 as a more advanced method, which is also used by Patil and Stieglitz (2014) for the EXP-Hydro model and is recommended as fixed time step solver for the *torchdiffeq* solver (Chen, 2018).

The results in chapter 4.2.1 demonstrate that the chosen method for the ODE solver has a significant impact on the performance of the conceptual model. For the BaseERA5L model the median performance (NSE) on the testing data increases from 0.05 to 0.30 (Euler vs. RK4). For the BaseEOBS model, the performance (NSE) on the same time period increases from 0.25 to 0.50 (Euler vs. RK4). The results in table 7.10 in Appendix B comparing the BaseEOBS and HybridEOBS with the Euler method vs. the RK4 method for the ODE solver show that also for the hybrid model the ODE solver is of crucial importance and with the training setup used in this study, the effect of the ODE solver cannot be compensated for by the neural network in the hybrid model.

Meteorological Forcings
The analysis of the long-term water balance of the catchments using the meteorologic forcings from the LamaH-CE dataset in comparison with data from the Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2 (Trabucco and Zomer, 2019) revealed that there might be an overestimation of (potential) evapotranspiration in the ERA5-Land dataset. This is also mentioned by Klingler et al. (2021). Based on these findings, the E-OBS dataset was used additionally to assess the performance of the models using another dataset. The results show that the performance of all models significantly improved when using the E-OBS dataset over the ERA5-Land-based LamaH-CE dataset.

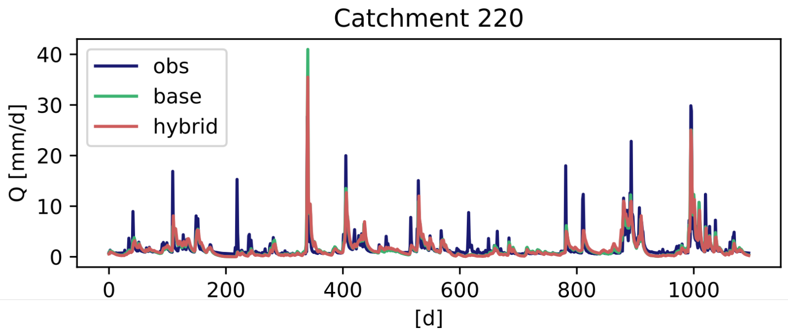

Figure 5.1: Hydrograph analysis of the conceptual model, hybrid model and the observations over a period of three years on catchment 220

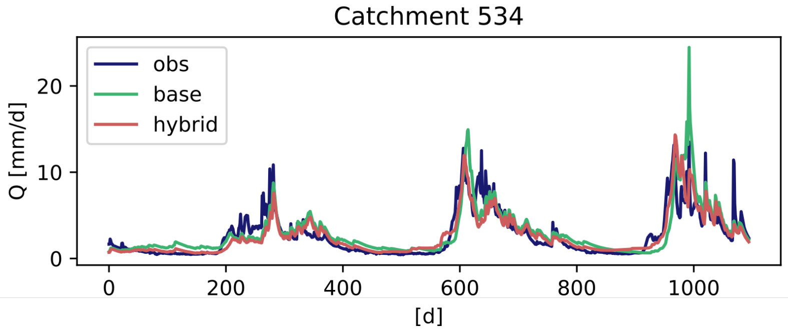

Figure 5.2: Hydrograph analysis of the conceptual model, hybrid model and the observations over a period of three years on catchment 534

Model Comparison

In figure 5.1 and 5.2 the hydrographs of the BaseEOBS and HybridEOBS as well as the observations are visualized of a period of three years. Catchment 220 represents a catchment were little to no improvement was achieved by the hybrid model while catchment 534 represents a catchment with a large performance improvement through the hybrid model. The hybrid model improves the prediction of baseflow and in the time period around day 1000 it can be seen that the hybrid model avoids the spike which the conceptual model predicts.

Figures 5.3 and 5.4 show the contrbution to the neural network to the changes in snow storage (S0) and water storage (S1) as well as the storage respective storage levels. This demonstrates that the hybrid modelling framework introduced in this study can be used to analyze both the model states of the model as well as the contributions of the neural network, which bears the potential of identifying limitations of the conceptual model which is underlying.

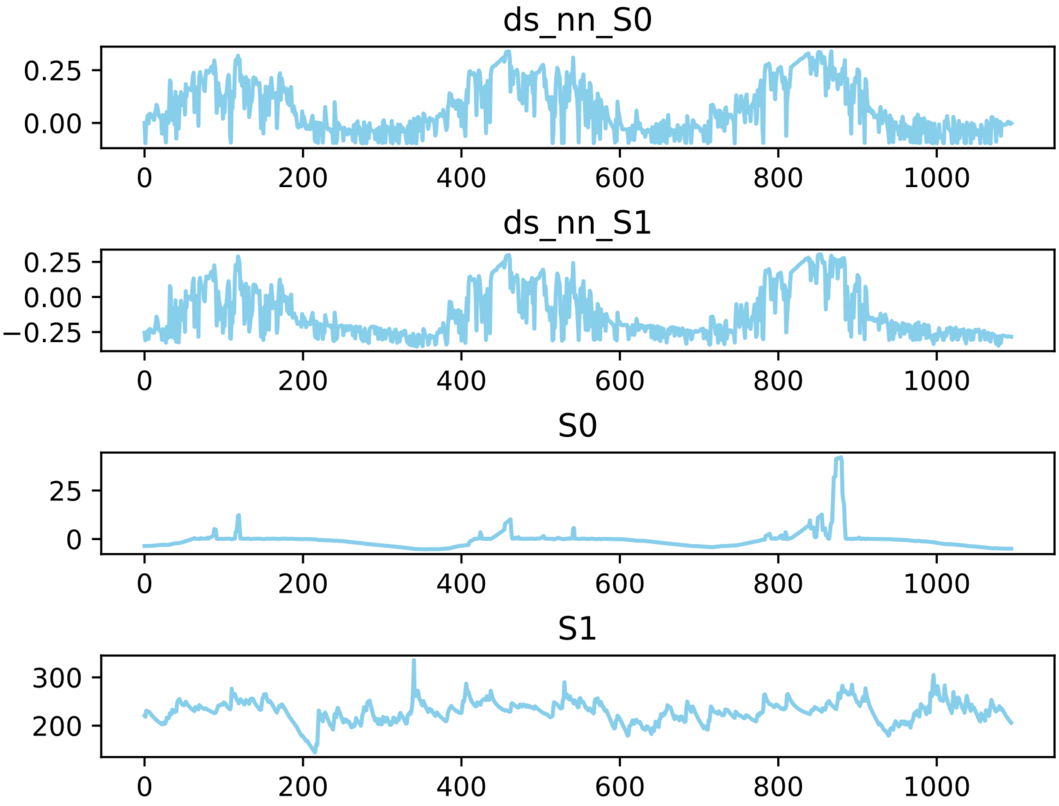

Figure 5.3: NN contribution to changes in storages and storage levels catchment 220

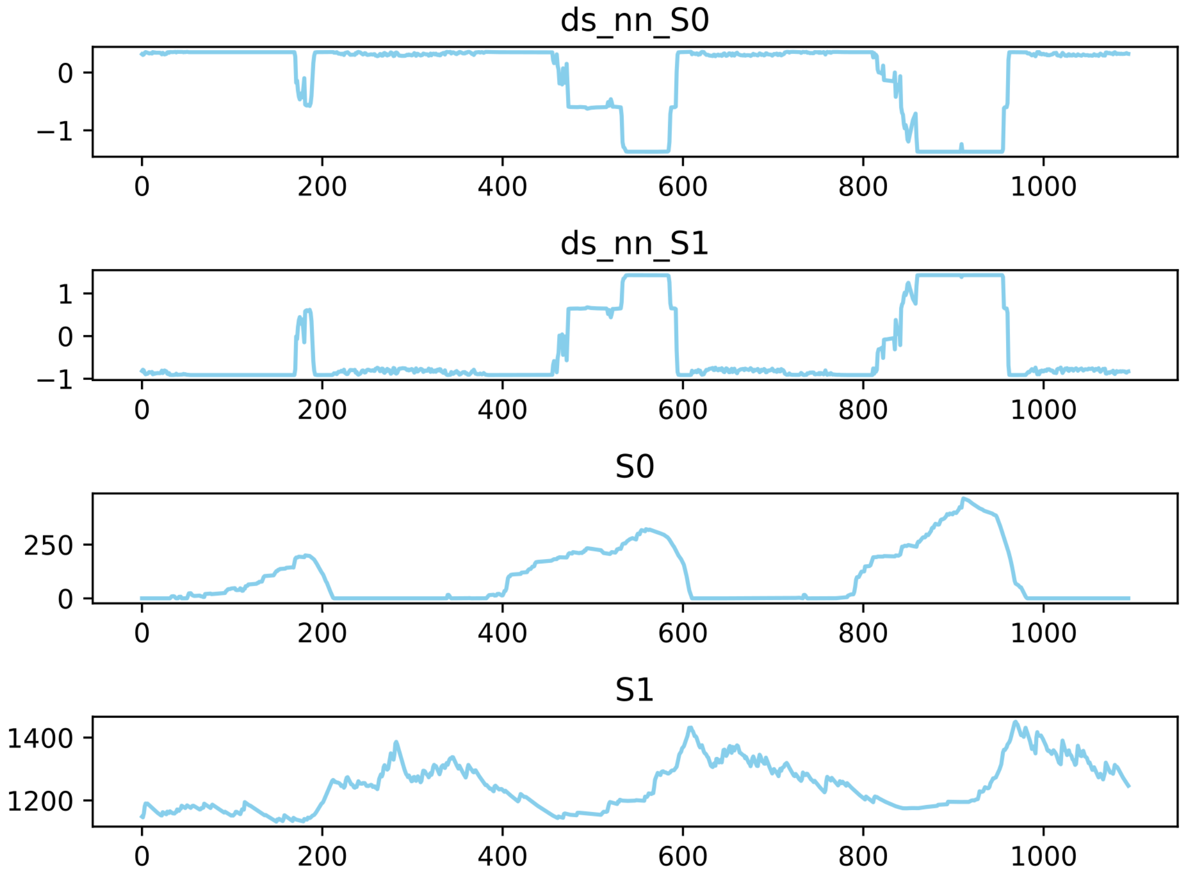

Figure 5.4: NN contribution to changes in storages and storage levels catchment 534

5.1.2 Model Performance on Catchments with Different Catchment Characteristics (SQ2)

Several findings in this study indicate that the catchment clustering performed using ten catchment attributes results in a meaningful creation of so-called hydrological response units, groups of catchments that respond to similar meteorological inputs in a similar way. Figure 3.3 shows that clusters form distinct groups of catchments in the same region even though no spatial information is included in the clustering. Even though the large differences in elevation throughout the study area are likely to contribute to the spatial agglomeration. Figure 4.4 shows however, that other catchment attributes must be decisive in this as well as clusters zero, one and seven show significant similarities in elevation.

The similarity of the hydrological response, at least with respect the NSE and KGE values becomes obvious in figure 4.8 which compares the BaseEOBS and the LstmEOBS models across clusters. A distinct pattern that is similar for both models can be seen. Similar performance of the respective models on clusters zero and one paired with the findings in the cluster analysis strengthens the hypothesis, that these catchment behave similarly, despite the fact that clusters zero and one are relatively far apart when looking at the map of the study area.

Two catchment clusters display particularly bad performance metrics, cluster two and cluster seven. Here performance on cluster seven is significantly worse compared to cluster two. When comparing the catchment attributes and hydrological signatures it stands out that both catchments receive high precipitation events much more frequently than the other catchment clusters, in both cases however, this is paired with a relatively low mean daily runoff and accordingly a low runoff ratio. At the same time both clusters also show relatively high aridity values. This indicates that is likely challenging for both the BaseEOBS and the LstmEOBS model to predict streamflow in arid catchments with a peaky discharge regime and a low runoff ratio. This is also in line with findings in Guo et al. (2020) with respect to the challenges for conceptual rainfall-runoff models.

Catchment cluster five represents the second best catchment with respect to the performance of the BaseEOBS model when look at the median of NSE and KGE values and it represents the best performance of the LstmEOBS model. The catchments in cluster five are located in the mountains at very high elevation and with steep slopes. The mean snow fraction is 42 % and mean precipitation is relatively high. Furthermore, mean daily runoff is also relatively high and accordingly, the runoff ratio as well. Paired with a low aridity this indicates the inverse of the findings above. Both models seem to perform well when the runoff ratio is high and aridity is low. Additionally both models seem to perform well in a setting where a large fraction of precipitation falls as snow. For the LSTM this is in line with findings in Sabzipour et al. (2023) where the performance of LSTM models is compared to distributed physically-based models. For the BaseEOBS performance on the three catchment clusters with a snow coverage higher than 15 % (cluster three, four and five) the performance is especially good for clusters four and five as cluster three display a larger interquartile range. Nevertheless, performance on the catchments with significant snow coverage is better than on the ones with little snow coverage. While reasons for this may be complex, on explanation for this is the model structure. The model only serves as a two-bucket model if there is precipitation falling as snow, else it basically acts as a single-bucket model. This could be a reason for the improved performance in catchments with higher snow coverage.

The performance of the hybrid model (HybridEOBS) is compared to the conceptual model

(BaseEOBS) on fourteen catchments, on representative for the median performance of the conceptual model and one representative for the performance near the 75th percentile of the conceptual model. Altogether the hybrid model outperforms the conceptual model with respect to NSE in nine out of fourteen cases while it performs the same in two cases and worse in three cases. For the KGE values the hybrid model performs better in seven cases, worse in 6 cases and the same in one case. No clear pattern is distinguishable that could relate the ability of the hybrid model to certain catchment attributes as there is a case for each cluster, except cluster zero, in which the hybrid model outperforms the conceptual model when looking at the fourteen comparisons and both evaluation metrics.

The results show that there is a clear difference in performance of the models on catchments with different catchment attributes. However, the ability of the hybrid model to improve the performance of the underlying conceptual model could not be identified as particularly strong or weak on certain types of catchments.

5.1.3 Exploration of Improvement Pathways (SQ3)

In an experiment of the effects of adding additional meteorological input forcings to the LSTM model it was shown that it increases the overall median NSE performance as well as increases performance on five out of seven of the catchment clusters. The goal of this experiment was to explore the option of adding additional meteorological input forcings to the hybrid model. However, the adding the same meteorological forcings to the neural network component of the hybrid did not lead to any improvements in performance. The observed tendency for overfitting the training data when adding the additional forcings is likely explained by the increasing number of neurons in the neural network while applying no regularization. Additionally, the added meteorogical forcings display a wider range of values, which in a modelling setup without scaling for the hybrid model is likely to be an additional factor explaining the slow and inconsistent convergence of the model during training. However, it has been shown in previous experiments that the hybrid models are able to capture additional information from the input forcings and it has been shown that additional forcings can improve the performance of the LSTM model. Accordingly an optimized setup of the hybrid model training including regularization and scaling of the additional input forcings as well as a more thorough selection of the number and type of input forcings is necessary to inverstigate whether this has potential to further improve the performance of the hybrid model. The same is true for experiments providing additional meteorolical forcings from another dataset which lead to the same described issues in the training process.

In a similar but nevertheless different approach Höge et al. (2022) have shown that an alternative implementation of a Neural Ordinary Differential Equations approach based on the same conceptual model (EXP-Hydro) can predict discharge on individual basins in the US using the CAMELS dataset as well as state-of-the-art deep learning models. A possible advantage of the approach is that in the *M50* model a part of the processes in the conceptual model and in the *M100* all of the processes in the conceptual model are replaced by a neural network. This allows the neural network to more flexibly to enhance the model structure of the base model. The modelling framework introduced in this study allows for a flexible setup in which parts or the entire neural network can be replaced by a neural network, hence to reproduce the approach by Höge et al. (2022). Exploring a hybrid model in which parts of the neural network are replaced by the neural network accordingly represents another pathway to improving the performance of the hybrid modelling approach introduced in this study.

5.2 Limitations and Recommendations

Data

One limitation of this study is related to the data preparation of the E-OBS dataset, more specifically regarding the approximation of reference evapotranspiration using the FAO Penman-Monteith equation. While variables from E-OBS that have been used as direct inputs to any of the models were gap-free. Some variables from the E-OBS dataset that were used to compute the reference evapotranspiration included gaps. The E-OBS dataset is based on interpolated station data so gaps in the time series are most likely left consciously as their temporal or spatial extent was too large or interpolation. In this study the variables with gaps were interpolated anyways to produce a gap-free dataset which introduces uncertainty to the reference evapotranspiration values.

The experiments with the conceptual models and LSTM models on the clustered catchments as well as their regional clusters which has been shown on a map, indicates that the clustering does group the catchments to a certain extent by hydrological similarity into so-called Hydrological Reponse Units. However, the K-means clustering which was the basis for the clutering, was based on ten catchment attributes that have been determined to be hydrologically relevant by (Kratzert et al., 2019a). However, this was performed on the CAMELS-US dataset for the contiguous united states and does not necessarily reflect upon their hydrological importance in central Europe, especially since the study area in the LamaH-CE dataset, despite significant variability in catchment characteristics, does not include some types of catchments/landscapes that are present in the United States.

Models and Training

While meteorological forcings were scaled for the use in the LSTM model, no scaling was applied to the meteorological forcings for the conceptual model and hybrid model. To optimize the training process of the models, future experiments should explore the added value of data scaling.

The EXP-Hydro model which is used as base model for the hybrid models in this study is a very simple conceptual model, especially in catchments with little to now snow fall it acts as a single-bucket model. Exploration of the effect of adding more complex conceptual hydrological models as base model would grant further insights into the potential of the approach. To further improve model performance, static parameters could be replaced partially or entirely by dynamic parameterization with time-dependent parameters that are also learnt by the neural network. This approach is implemented in Feng et al. (2023).

The differentiable modelling framework created in this study comes creates significant computational costs when training the models. To create a differentiable version of the conceptual model the model structure was changed in order to allow tracking the gradients. The *torchdiffeq* package is used to create a differentiable modelling framework. Depending on the modelling setup, tracking the gradients increases the time for the forward pass of the ODE solver by 10x or more, severely limiting the experimental setup executed in this study. Future studies to explore Neural Ordinary Differential Equations should be performed in a setting that is better designed for automatic differentiation, for example using the programming langueage *Julia*, which was done in Höge et al. (2022).

Because of the described computational limitations only a very limited hyper parameter optimization was performed. More experiments on the effect of different loss functions on the training of the hybrid models as well as the impact of different learning rates and training for more epochs is recommended to fully explore the potential of this approach.

Reflection on the Modelling Approach

The hybrid model introduced in this study aims to improve discharge predictions by improving the representation of the storages in the conceptual model. For each time step, the conceptual component of the model which is composed of the mechanistic processes and calibrated parameters can be seen as the right hand side of a differential equation which is integrated by numerical approximation using an ODE solver and an initial condition, in this case the initialization of the storages. The neural network component of the hybrid model receives for each time step the meteorological input forcings as well as the prediction for the two storage from the conceptual component. The conceptual component has been calibrated before, therefore the neural network learns to improve the computation of the changes in storages (dS0/dt and dS1/dt). The neural network is then trained by tracking the gradients and using a loss function to compute the difference between the simulated and the observed discharge. The process of improving results by adjusting storages to better capture the observed discharge is comparable to data assimilation.

Multiple examples exist in hydrological modelling for the use of data assimilation to improve discharge predictions. For example, terrestrial water storage from the Gravity Recovery and Climate Experiment (GRACE) has been used to adjust model storages at each time step to arrive at better discharge predictions (Wu et al., 2022). Another application of data assimilation for streamflow prediction is introduced in Boucher et al. (2020). The approach shows large similarities to the approach implemented in this study. GR4J, another simple conceptual model is chosen as base model. Similarly to EXP-Hydro it also has two model states that represent storages. These are also connected to a neural network and training is performed on streamflow observations as, similar to the case in this study, no observations for the storages exist (Boucher et al., 2020).

The approach in this study displays similarity to data assimilation but it is still to be assessed whether it has the same potential of improving hydrological models as data assimilation does. Therefore, it is suggested to conduct further experiments which compare the performance of the present neural ODE approach to data assimilation on conceptual models.

6 Conclusion

As part of this work it has been demonstrated that hybrid models using a Neural Ordinary Differential Equations approach represent a promising way to enhance the performance of conceptual models while preserving the ease of interpretability. At the same time it has been shown that machine learning-based approaches such as LSTM models are in many cases superior when it comes to accuracy of predictions. This shows that hybrid hydrological models using a Neural Ordinary Differential Equations approach can contribute to the field of hydrological modelling especially since the implementation in this work is only one of the many ways to combine physical/conceptual models and nural networks.

Key elements in developing hybrid models using a NODE approach are the choice of base model as well as the choice of ODE solver, the decisions in how to combine the conceptual part with the neural network and how to train the parameters of the conceptual part of the model as well as the neural network part. Another key point is the computational efficiency of the models which represented a bottle-neck in this work.

The field of Universal Differential Equations is gaining traction and has recently shown promising approaches across fields and will certainly benefit the field of hydrological modelling as well.

Code availablility: https://github.com/JonathanSchieren/MasterThesis

Bibliography

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

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M. (1998). Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. Technical report.

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

Atkinson, K., Han, W., and Stewart, D. (2009). *NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS.*

Bolibar, J., Sapienza, F., Maussion, F., Lguensat, R., Wouters, B., and Pérez, F. (2023). Universal Differential Equations for glacier ice flow modelling.

Bouaziz, L. J., Fenicia, F., Thirel, G., De Boer-Euser, T., Buitink, J., Brauer, C. C., De Niel, J., Dewals, B. J., Drogue, G., Grelier, B., Melsen, L. A., Moustakas, S., Nossent, J., Pereira, F., Sprokkereef, E., Stam, J., Weerts, A. H., Willems, P., Savenije, H. H., and Hrachowitz, M. (2021). Behind the scenes of streamflow model performance. *Hydrology and Earth System Sciences*, 25(2):1069–1095.

Boucher, M., Quilty, J., and Adamowski, J. (2020). Data Assimilation for Streamflow Forecasting Using Extreme Learning Machines and Multilayer Perceptrons. *Water Resources Research*, 56(6).

Budyko, M. I. (1974). Climate and Life. *Academic Press.*

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

Chen, R. T. Q. (2018). torchdiffeq.

Chen, Y., Chen, X., Xue, M., Yang, C., Zheng, W., Cao, J., Yan, W., and Yuan, W. (2023). Revisiting the hydrological basis of the Budyko framework with the principle of hydrologically similar groups. *Hydrology and Earth System Sciences*, 27(9):1929–1943.

Cornes, R. C., van der Schrier, G., van den Besselaar, E. J., and Jones, P. D. (2018). An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets. *Journal of Geophysical Research: Atmospheres*, 123(17):9391–9409.

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., and Woods, R. (2020). CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. *Earth System Science Data*, 12(4):2459–2483.

Cybenkot, G. (1989). Mathematics of Control, Signals, and Systems Approximation by Superpositions of a Sigmoidal Function*. *Math. Control Signals Systems*, 2:303–314.

Devia, G. K., Ganasri, B., and Dwarakish, G. (2015). A Review on Hydrological Models. *Aquatic Procedia*, 4:1001–1007.

Federal Ministry of Agriculture (2007). *Hydrological Atlas of Austria*.

Feng, D., Beck, H., Lawson, K., and Shen, C. (2023). The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment. *Hydrology and Earth System Sciences*, 27(12):2357–2373.

Feng, D., Fang, K., and Shen, C. (2020). Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales. *Water Resources Research*, 56(9).

Feng, D., Liu, J., Lawson, K., and Shen, C. (2022). Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy. *Water Resources Research*, 58(10).

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

Flügel, W.-A. (1995). Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the River Bröl, Germany. *Hydrological Processes*, 9:423–436.

Frame, J. M., Kratzert, F., Raney, A., Rahman, M., Salas, F. R., and Nearing, G. S. (2021). Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics. *Journal of the American Water Resources Association*, 57(6):885–905.

Gad, A. G. (2022). Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review. *Archives of Computational Methods in Engineering*, 29(5):2531–2561.

Guo, D., Zheng, F., Gupta, H., and Maier, H. R. (2020). On the Robustness of Conceptual Rainfall-Runoff Models to Calibration and Evaluation Data Set Splits Selection: A Large Sample Investigation. *Water Resources Research*, 56(3).

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

Hamon, W. R. (1963). Computation of direct runoff amounts from storm rainfall. *International Association of Hydrological Sciences Publication*, 63:52–62.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Höge, M., Scheidegger, A., Baity-Jesi, M., Albert, C., and Fenicia, F. (2022). Improving hydrologic models for predictions and process understanding using neural ODEs. *Hydrol. Earth Syst. Sci*, 26:5085–5102.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.

Jiang, S., Zheng, Y., and Solomatine, D. (2020). Improving AI System Awareness of Geoscience Knowledge: Symbiotic Integration of Physical Approaches and Deep Learning. *Geophysical Research Letters*, 47(13).

Jordahl, K., Van den Bossche, J., Fleischmann, M., Wasserman, J., McBridge, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., and Eubank, N. (2020). geopandas/geopandas: v0.8.1.

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, pages 1942–1948. IEEE.

Klein Tank, A. M. G., Wijngaard, J. B., Können, G. P., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Szalai, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van Engelen, A. F. V., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., Antonio López, J., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L. V., and Petrovic, P. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, 22(12):1441–1453.

Klingler, C., Schulz, K., and Herrnegger, M. (2021). LamaH-CE: LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe. *Earth System Science Data*, 13(9):4529–4565.

Klok, E. J. and Klein Tank, A. M. G. (2009). Updated and extended European dataset of daily climate observations. *International Journal of Climatology*, 29(8):1182–1191.

Knoben, W. J. M., Freer, J. E., and Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, 23(10):4323–4331.

Koppa, A., Rains, D., Hulsman, P., Poyatos, R., and Miralles, D. G. (2022). A deep learning-based hybrid model of global terrestrial evaporation. *Nature Communications*, 13(1).

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M. (2018). Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G. (2019a). Benchmarking a Catchment-Aware Long Short-Term Memory Network (LSTM) for Large-Scale Hydrological Modeling. *Hydrology and Earth System Sciences*.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G. (2019b). 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.

Lindstrom, G., Johansson, B., Persson, M., Gardelin, M., and Bergstr6m, S. (1997). Development and test of the distributed HBV-96 hydrological model. Technical report.

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., and Thieme, M. (2019). Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific Data*, 6(1):283.

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

Moretti, G. and Montanari, A. (2008). Inferring the flood frequency distribution for an ungauged basin using a spatially distributed rainfall-runoff model. *Hydrology and Earth System Sciences*, 12(4):1141–1152.

Muñoz Sabater, J. (2019). ERA5-Land hourly data from 1950 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS).*

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J. N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383.

Nash, J. E. and Sutcliffe, J. V. (1970). RIVER FLOW FORECASTING THROUGH CONCEPTUAL MODELS PART 1 - A DISCUSSION OF PRINCIPLES*. Technical report.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelsheim, N., Antiga, L., Desmaison, A., Koepf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library.

Patil, S. and Stieglitz, M. (2014). Modelling daily streamflow at ungauged catchments: What information is necessary? *Hydrological Processes*, 28(3):1159–1169.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Penman, H. (1956). Evaporation: an introductory survey. *Netherlands Journal of Agricultural Science*, 4(1):9–29.

Penman, H. L. (1948). Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 193(1032):120–145.

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

Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A., and Edelman, A. (2020). Universal Differential Equations for Scientific Machine Learning.

Reaver, N. G. F., Kaplan, D. A., Klammler, H., and Jawitz, J. W. (2020). Reinterpreting the Budyko Framework. *Hydrology and Earth System Sciences.*

Reaver, N. G. F., Kaplan, D. A., Klammler, H., and Jawitz, J. W. (2022). Theoretical and empirical evidence against the Budyko catchment trajectory conjecture. *Hydrology and Earth System Sciences*, 26(5):1507–1525.

Sabzipour, B., Arsenault, R., Troin, M., Martel, J. L., Brissette, F., Brunet, F., and Mai, J. (2023). Comparing a long short-term memory (LSTM) neural network with a physically-based hydrological model for streamflow forecasting over a Canadian catchment. *Journal of Hydrology*, 627.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer International Publishing, Cham.

Seibert, J. and Bergström, S. (2022). A retrospective on hydrological catchment modelling based on half a century with the HBV model. *Hydrology and Earth System Sciences*, 26(5):1371–1388.

Sinaga, K. P. and Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, 8:80716–80727.

Singh, A. (2018). A Concise Review on Introduction to Hydrological Models. *GRD Journal for Engineering*, 3(10).

Staudinger, M., Stahl, K., Seibert, J., Clark, M. P., and Tallaksen, L. M. (2011). Comparison of hydrological model structures based on recession and low flow simulations. *Hydrology and Earth System Sciences*, 15(11):3447–3459.

Thorntwaite, C. W. (1948). An Approach toward a Rational Classification of Climate. *The Geological Review*, 38:55–94.

Trabucco, A. and Zomer, R. (2019). Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2. *CGIAR Consortium for Spatial Information (CGIAR-CSI) [dataset].*

Tsai, W. P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., and Shen, C. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, 12(1).

van den Besselaar, E. and Copernicus Climate Change Service (2023). E-OBS daily gridded observations for Europe from 1950 to present: Product user guide. *ECMWF Confluence Wiki.*

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, , Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., and Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272.

Wu, W.-Y., Yang, Z.-L., Zhao, L., and Lin, P. (2022). The impact of multi-sensor land data assimilation on river discharge estimation. *Remote Sensing of Environment*, 279:113138.

Xiang, K., Li, Y., Horton, R., and Feng, H. (2020). Similarity and difference of potential evapotranspiration and reference crop evapotranspiration – a review. *Agricultural Water Management*, 232:106043.

Zealand, C. M., Burn, D. H., and Simonovic, S. P. (1999). Short term streamflow forecasting using artificial neural networks. *Journal of Hydrology*, 214(1-4):32–48.

Zhang, L., Dawes, W. R., and Walker, G. R. (2001). Response of mean annual evapotranspiration to vegetation changes at catchment scale. *Water Resources Research*, 37(3):701–708.

7 Appendix

7.1 Appendix A - Data & Data Analysis

7.1.1 Estimating Potential Evapotranspiration

This chapter shows the equations used to estimate potential evapotranspiration using the FAO Penman-Monteith Equation including the equations to estimate missing climatic data.

FAO Penman-Monteith Equation

From the Penman-Monteith equation , the equation for aerodynamic resistance and the equaiton for surface resistance the FAO Penman-Monteith equation (equation 7.1) was developed.

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma\frac{900}{T+273}u_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \tag{7.1}$$

ET_0: reference evapotranspiration [mm day^{-1}], R_n: net radiation at the crop surface [MJ m^{-2} day $^{-1}$], G: soil heat flux density [MJ m^{-2} day^{-1}], T: mean daily air temperature at 2m height [°C], u_2: wind speed at 2m height [m s^{-1}], e_s: saturation vapour pressure [kPa], e_a: actual vapour pressure [kPa], $e_s - e_a$: saturation pressure deficit [kPa], Δ: slope vapour pressure curve [kPa / °C], γ: psychrometric constant [kPa / °C] Allen et al. (1998)

Missing Climatic Data

Saturation vapour pressure and mean saturation vapour pressure

Saturation vapour pressure (e_s) can be calculated using equation 7.2.

$$e^\circ(T) = 0.60108exp(\frac{17.27T}{T + 273}) \tag{7.2}$$

$e(T)$: saturation vapour pressure at air temperature T [kPa], T: air temperature [°C] Allen et al. (1998)

The daily mean of the saturation vapour pressure can then be computed with equation 7.3.

$$e_s = \frac{e^\circ(T_{max}) + e^\circ(T_{min})}{2} \tag{7.3}$$

$e^\circ(T)$: saturation vapour pressure at air temperature T [kPa], T: air temperature [°C] Allen et al. (1998)

Slope of saturation vapour pressure curve

The slope of the relationship between temperature T and saturation vapour pressure is calculated as shown in equation 7.4.

$$\Delta = \frac{4098(0.6018exp(\frac{17.27T}{T+237.3}))}{(T+237.3)^2} \tag{7.4}$$

Δ: slope of saturation vapour pressure curve at air temperature T [kPa / °C], T: air temperature [°C] Allen et al. (1998)

Determine actual vapour pressure from relative humidity Actual vapour pressure can be determined from the daily mean relative humidity as well as from saturation vapour pressure at daily minimum and maximum temperature. This is shown in equation 7.5.

$$e_a = \frac{RH_{mean}}{100}(\frac{e°(T_{max}) + e°(T_{min})}{2}) = \frac{RH_{mean}}{100} * e_s \tag{7.5}$$

e_a: actual vapour pressure [kPa], RH_{mean}: mean relative humidity [%], $e°(T_{min})$: saturation vapour pressure at daily minimum temperature [kPa], $e°(T_{max})$: saturation vapour pressure at daily maximum temperature [kPa]

Vapour pressure deficit

The vapour pressure deficit is calculated as the difference between mean saturation vapour pressure (equation 7.3) and actual vapour pressure (equation 7.5).

Albedo

Albedo values are required to calculate net shortwave radiation (R_ns) and were ontained from ERA5-Land, making it the only observed variable that was added from a dataset other than E-OBS for the computation of PET.

Soil heat flux

The soil heat flux is small compared to other components of the energy balance and will be ignored in this work Allen et al. (1998).

Net radiation

Net radiation (R_n) is the difference between incoming net shortwave radiation (R_{ns}) and outgoing net longwave radiation (R_{nl}).

$$R_n = R_{ns} - R_{nl} \tag{7.6}$$

R_n: net radiation [MJ / m^2 / day], R_{nl}: outgoing net longwave radiation [MJ / m^2 / day], R_{ns}: incoming net shortwave radiation [MJ / m^2 / day] Allen et al. (1998)

To compute net radiation, first incoming net shortwave radiation (R_{ns}) and outgoing net longwave radiation (R_{nl}) need to be determined.

Net longwave radiation

$$R_{nl} = \sigma(\frac{T^4_{max,K} + T^4_{min,K}}{2})(0.34 - 0.14\sqrt{e_a})(1.35\frac{R_s}{R_{so}} - 0.35) \tag{7.7}$$

R_{nl}: net outgoing longwave radiation [MJ / m^2 / day], σ: Stefan-Boltzman constant [4.903 * 10$^{-}9MJK^4/m^2/day]$,T$_{max,K}$: maximum daily air temperature [K], $T_{min,K}$: minimum daily air temperature [K], e_a: actual vapour pressure [kPa], R_S/R_{so}: relative shortwave radiation, R_s: shortwave radiation [MJ / m^2 / day], R_{so}: clear-sky radiation [MJ / m^2 / day] Allen et al. (1998)

Incoming shortwave radiation

Incoming shortwave radiation (R_s) is part of the E-OBS dataset.

Net shortwave radiation

Net shortwave radiation is calculated based on incoming shortwave radiation (R_s) and albedo (α).

$$R_{ns} = (1 - \alpha)R_s \tag{7.8}$$

R_{ns}: net shortwave radiation [MJ / m^2 / day], α: albedo [-], R_s: incoming shortwave radiation [MJ / m^2 / day] Allen et al. (1998)

Clear-sky shortwave radiation

The calculation of clear-sky shortwave radiation is required to compute the net longwave radiation.

$$R_{so} = (0.75 + 2 * 10^{-5}z)R_a \tag{7.9}$$

R_{so}: clear-sky shortwave radiation [MJ / m^2 / day], z: station elevation above sea level [m], R_a: extraterrestrial radiation [MJ / m^2 / day] Allen et al. (1998)

Extraterrestrial radiation

$$R_a = \frac{24(60)}{\pi} G_{sc} d_r (\omega_s sin(\phi) sin(\delta) + cos(\phi) cos(\delta) sin(\omega_s)) \tag{7.10}$$

R_a: extraterrestrial radiation [MJ / m^2 / day], G_{sc}: solar constant = 0.0820 [MJ / m^2 / day], d_r: inverse relative distance Earth-Sun, ω_s: sunset hour angle [rad], ϕ: latitude [rad], δ solar declination [rad] Allen et al. (1998)

To calculate extraterrestrial radiation (R_a) also the inverse relative distance Earth-sun and solar declination are required (equation 7.11 & 7.12).

$$d_r = 1 + 0.033cos(\frac{2\pi}{365}J) \tag{7.11}$$

$$\delta = 0.409sin(\frac{2\pi}{365}J - 1.39) \tag{7.12}$$

J: day of the year Allen et al. (1998)

$$\omega_s = arccos(-tan(\phi) - tan(\delta)) \tag{7.13}$$

J: day of the year Allen et al. (1998)

Wind speed Wind speed is part of the E-OBS dataset at 10 m above the surface. It is corrected to 2 m above the surface with equation 7.14.

$$u_2 = u_z \frac{4.87}{ln(67.8z - 5.42)} \tag{7.14}$$

u_2: wind speed at 2m above surface [m / s], u_z: measured wind speed z m above surface [m / s], z: height of wind speed measurement above surface[m] Allen et al. (1998)

Atmospheric parameters

The FAO Penman-Monteith equation also requires the psychometric constant (γ) which again requires the atmospheric pressure (P).

$$P = 101.3(\frac{293 - 0.0065z}{293})^{5.26} \tag{7.15}$$

P: atmospheric pressure [kPa], z: elevation above sea level [m] Allen et al. (1998)

$$\gamma = \frac{c_p P}{\epsilon \lambda} = 0.665 * 10^{-3} P \tag{7.16}$$

γ: psychrometric constant [kPa / °C], P: atmospheric pressure [kPa], λ: latent heat of vaporization: 2.45 [MJ / kg], c_p: specific heat at constant pressure: 1.013 * 10^{-3} [MJ / kg / °C], ϵ: ratio molecular weight of water vapour / dry air: 0.622 Allen et al. (1998)

7.1.2 Catchment Clustering & Analysis

Table 7.1: Median catchment characteristics across clusters based on data from (Klingler et al., 2021). Variables: area_calc: Area [km^2], elev_mean: Mean elevation [m.a.s.l.], slope_mean: Mean slope [m/km], p_mean: Mean precipitation [mm/d], arid_1: Aridity as the ratio of mean daily reference evapotranspiration to mean precipitation, hi_prec_cu: Mean duration of high-precipitation events(number of consecutive days with 5 times mean daily precipitation) [d], hi_prec_fr: Frequency of dry days (<1 [mm/d] precipitation) [d/yr], geol_perme: Subsurface permeability [-], gc_sc_fra: Fraction of carbonate sedimentary rocks [-], frac_snow: Fraction of precipitation falling as snow [-]

Cluster	Count	area_calc	elev_mean	slope_mean	p_mean	arid_1	hi_prec_du	hi_prec_fr	geol_perme	gc_sc_fra	frac_snow
0	59	70,298	593,0	88,0	2,88	0,76	1,19	12,15	-14,2	0,0	0,1
1	31	98,303	634,0	71,0	2,97	0,74	1,16	11,55	-11,9	0,153	0,08
2	24	101,433	918,0	209,5	2,865	0,785	1,18	16,575	-12,95	0,066	0,12
3	43	80,218	1336,0	332,0	5,04	0,44	1,18	8,5	-11,8	0,896	0,29
4	32	96,852	831,0	290,0	3,775	0,6	1,23	12,0	-11,85	0,858	0,17
5	37	67,166	2170,0	392,0	4,09	0,47	1,11	9,7	-13,8	0,0	0,43
7	20	126,046	407,5	90,5	2,215	1,13	1,2	17,4	-14,15	0,0	0,08

Table 7.2: Mean catchment characteristics across clusters based on data from (Klingler et al., 2021).

Cluster	Count	area_calc	elev_mean	slope_mean	p_mean	arid_1	hi_prec_du	hi_prec_fr	geol_perme	gc_sc_fra	frac_snow
0	59	95,505	584,068	94,576	3,053	0,746	1,19	11,988	-14,39	0,003	0,1
1	31	169,534	622,968	86,645	3,179	0,727	1,164	11,392	-11,981	0,318	0,088
2	24	130,626	949,625	205,667	2,887	0,8	1,175	16,731	-13,058	0,121	0,135
3	43	90,771	1339,14	328,698	4,936	0,442	1,177	8,908	-11,912	0,8	0,28
4	32	123,376	870,281	280,312	3,661	0,646	1,229	12,972	-11,938	0,771	0,169
5	37	92,688	2165,892	383,459	4,058	0,476	1,123	10,15	-13,37	0,099	0,422
7	20	156,172	411,7	91,15	2,222	1,162	1,201	17,252	-14,02	0,056	0,082

Table 7.3: 5th percentile of catchment characteristics across clusters based on data from (Klingler et al., 2021).

Cluster	Count	area_calc	elev_mean	slope_mean	p_mean	arid_1	hi_prec_du	hi_prec_fr	geol_perme	gc_sc_fra	frac_snow
0	59	13,016	366,0	49,8	2,54	0,538	1,149	9,92	-15,0	0,0	0,06
1	31	45,933	449,5	49,5	2,495	0,525	1,115	9,675	-12,85	0,0	0,05
2	24	30,873	580,05	123,7	2,47	0,672	1,132	15,062	-14,2	0,0	0,082
3	43	15,312	847,8	213,7	4,155	0,38	1,111	7,955	-12,29	0,35	0,182
4	32	28,941	556,75	183,75	2,696	0,496	1,19	10,315	-12,245	0,366	0,115
5	37	19,269	1508,6	307,0	3,27	0,398	1,068	8,77	-14,1	0,0	0,288
7	20	56,458	258,05	46,9	1,877	0,939	1,16	14,75	-15,0		

Table 7.4: 95th percentile of catchment characteristics across clusters based on data from (Klingler et al., 2021).

Cluster	Count	area_calc	elev_mean	slope_mean	p_mean	arid_1	hi_prec_du	hi_prec_fr	geol_perme	gc_sc_fra	frac_snow
0	59	268,988	835,1	175,1	3,96	0,901	1,221	13,55	-13,39	0,027	0,14
1	31	363,888	785,5	145,5	4,35	0,905	1,225	12,4	-11,2	0,924	0,145
2	24	258,251	1451,4	301,45	3,258	0,958	1,2	18,632	-12,3	0,345	0,231
3	43	200,524	1799,6	441,5	5,307	0,499	1,23	10,455	-11,8	1,0	0,36
4	32	257,361	1228,95	375,4	4,43	0,907	1,26	18,122	-11,8	1,0	0,229
5	37	283,884	2718,4	462,2	4,8	0,578	1,24	12,98	-12,0	0,519	0,52
7	20	386,182	583,15	148,35	2,632	1,482	1,241	19,718	-12,2	0,374	0,101

Table 7.5: Median hydrological signatures across clusters based on data from (Klingler et al., 2021). Variables: q_mean: Mean daily runoff [mm/d], runoff_ratio: ratio of mean daily runoff and mean daily precipitation [-], baseflow_index_ladson: ratio of mean daily – baseflow and mean daily discharge; hydrograph separation is performed using the digital filter [-], Q5: 5% flow quantile (low flow) [mm/d], Q95: 95% flow quantile (high flow) [mm/d], high_q_freq: Frequency of high flow days (>9 times median daily flow) [d/yr], high_q_dur: Mean duration of high-flow events [d], low_q_freq: Frequency of low-flow days (<0.2 times mean daily flow) [d/yr], low_q_dur: Mean duration of low-flow events [d]

Cluster	Count	q_mean	runoff_ratio	baseflow_index_ladson	Q5	Q95	high_q_freq	high_q_dur	low_q_freq	low_q_dur
0	59	1,147	0,387	0,617	0,264	3,213	4,694	1,732	10,361	5,46
1	31	1,058	0,349	0,715	0,425	2,084	1,056	1,475	0,639	3,143
2	24	1,072	0,366	0,776	0,418	2,188	0,194	1,449	0,014	0,5
3	43	4,409	0,852	0,665	0,919	10,896	0,833	1,357	9,444	8,2
4	32	2,376	0,59	0,67	0,78	5,872	2,097	1,615	0,361	3,375
5	27	3,397	0,812	0,68	0,52	10,248	1,611	1,806	33,694	17,061
7	20	0,436	0,208	0,638	0,074	1,24	4,334	1,8	14,653	5,436

Table 7.6: Mean hydrological signatures across clusters based on data from (Klingler et al., 2021).

Cluster	Count	q_mean	runoff_ratio	baseflow_index_ladson	Q5	Q95	high_q_freq	high_q_dur	low_q_freq	low_q_dur
0	59	1,257	0,4	0,607	0,327	3,49	5,243	1,952	15,472	5,034
1	31	1,262	0,372	0,692	0,411	3,327	2,828	1,964	13,519	3,907
2	24	1,158	0,392	0,773	0,472	2,572	0,441	1,326	0,7	2,14
3	43	4,285	0,872	0,652	0,931	12,279	2,214	1,5	23,185	7,484
4	32	2,324	0,622	0,665	0,758	6,0	2,577	1,603	3,251	3,86
5	27	3,627	0,89	0,682	0,558	11,617	11,149	2,524	64,742	28,072
7	20	0,446	0,196	0,642	0,093	1,257	5,453	1,835	26,032	6,18

Table 7.7: 5th percentile of hydrological signatures across clusters based on data from (Klingler et al., 2021).

Cluster	Count	q_mean	runoff_ratio	baseflow_index_ladson	Q5	Q95	high_q_freq	high_q_dur	low_q_freq	low_q_dur
0	59	0,654	0,246	0,458	0,118	1,783	0,497	1,35	0,0	0,0
1	31	0,388	0,132	0,452	0,075	1,007	0,0	0,0	0,0	0,0
2	24	0,754	0,275	0,667	0,295	1,7	0,0	0,0	0,0	0,0
3	43	2,209	0,473	0,528	0,307	5,795	0,091	1,0	0,0	0,0
4	32	1,19	0,423	0,528	0,314	3,075	0,392	1,196	0,0	0,0
5	27	2,169	0,523	0,599	0,205	5,372	0,028	1,0	0,0	0,0
7	20	0,133	0,064	0,485	0,031	0,244	0,528	1,0	0,712	2,332

Table 7.8: 95th percentile of hydrological signatures across clusters based on data from (Klingler et al., 2021).

Cluster	Count	q_mean	runoff_ratio	baseflow_index_ladson	Q5	Q95	high_q_freq	high_q_dur	low_q_freq	low_q_dur
0	59	2,854	0,668	0,731	0,663	7,305	12,589	2,828	53,355	9,162
1	31	2,69	0,66	0,859	0,801	7,646	11,736	6,454	62,056	10,79
2	24	1,671	0,532	0,849	0,904	3,897	1,904	2,656	4,257	9,418
3	43	6,788	1,381	0,788	1,619	21,452	9,497	2,64	95,263	16,648
4	32	3,39	0,89	0,77	1,063	9,423	6,642	2,05	15,506	11,817
5	27	6,055	1,309	0,773	0,95	21,876	54,289	7,189	160,634	72,907
7	20	0,754	0,323	0,802	0,214	2,408	12,429	2,856	73,218	9,517

7.2 Appendix B - Models & Model Results

7.2.1 Impact of ODE Solver and Forcing Data on Model Performance

Table 7.9: Median NSE, KGE, and MSE values for the BaseERA5L and BaseEOBS model using ERA5-Land data and E-OBS data

	LamaH-CE				E-OBS			
	Calibration		Test		Calibration		Test	
	Euler	RK4	Euler	RK4	Euler	RK4	Euler	RK4
NSE	0.10	0.33	0.05	0.30	0.31	0.50	0.25	0.40
KGE	0.49	0.56	0.46	0.54	0.62	0.66	0.56	0.60
MSE	4.71	3.25	4.86	3.39	3.54	2.33	3.82	2.70

Table 7.10: Impact of the method for the ODE solver
Training with starting learning rate of 0.01, 25 epochs for the hybrid model with the RK4 solver and 50 epochs for the hybrid model with the Euler solver.

Model	BaseEOBS				HybridEOBS			
Solver	Euler		RK4		Euler		RK4	
ID	NSE	KGE	NSE	KGE	NSE	KGE	NSE	KGE
241	0.47	0.62	**0.54**	**0.64**	0.44	0.53	**0.52**	**0.58**
215	0.39	**0.65**	**0.45**	0.62	0.40	0.61	**0.53**	**0.68**
581	0.12	0.55	**0.16**	**0.56**	0.48	0.71	**0.58**	**0.74**
21	0.27	**0.59**	**0.31**	**0.59**	0.06	0.36	**0.33**	**0.49**
277	0.36	**0.59**	**0.52**	0.56	0.31	**0.54**	**0.50**	**0.54**
797	0.37	0.59	**0.46**	**0.61**	0.37	0.54	**0.54**	**0.68**
432	-0.12	0.49	**0.13**	**0.58**	-0.13	0.50	**0.24**	**0.63**