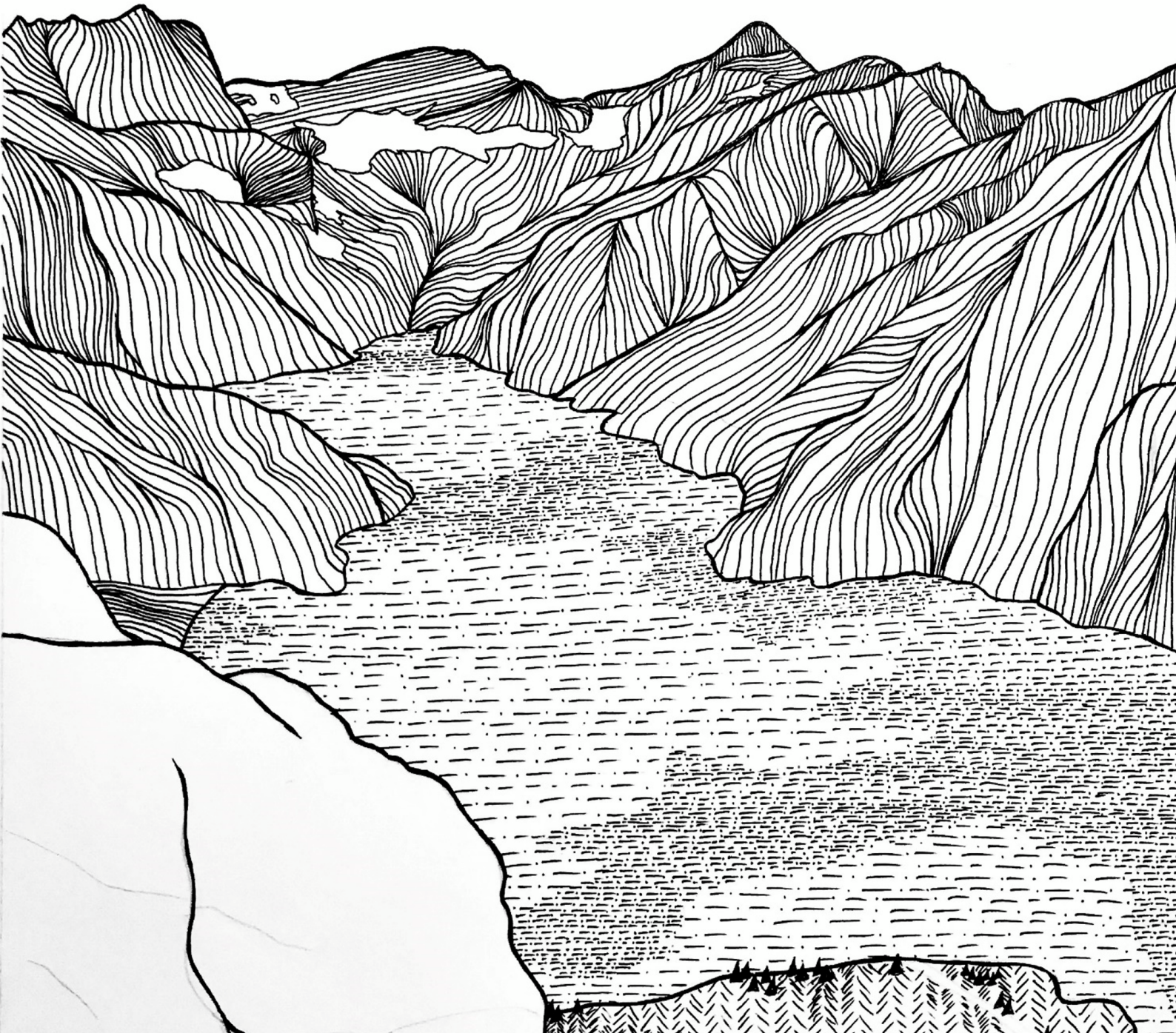
Assessing Global Applicability of a Long Short-Term Memory (LSTM) Neural Network for Rainfall-Runoff Modelling

Katharina Wilbrand



GRADUATION REPORT

Master of Science

ASSESSING GLOBAL APPLICABILITY OF A LONG SHORT-TERM MEMORY (LSTM) NEURAL NETWORK FOR RAINFALL-RUNOFF MODELLING

by Katharina Wilbrand, 5085713

Committee TU Delft: Dr. Riccardo Taormina (Chair)
Dr. Ir. Marie-Claire ten Veldhuis
Dr. Markus Hrachowitz

Supervisors Deltares: Martijn Visser
Ruben Dahm
Jonathan Nuttall

Department of Water Management
Faculty of Civil Engineering and Geosciences,
November 2021

TU Delft Deltares

ACKNOWLEDGEMENTS

The past nine months have been an intense, challenging and rewarding time for me. A year ago I had no idea about machine learning and the many possibilities in hydrological science. Now I am able to build such models, work with them and understand their relevance to hydrology. I have learnt so much on this exciting journey.

This thesis would never have happened without Riccardo's enthusiasm, confidence and trust in my abilities. Thank you for taking so much time, sharing your fascination and always encouraging me to get to the root of every problem. I enjoyed a lot having you as my supervisor. Also, I would like to thank Marie-Claire and Markus. Your critical attitude towards machine learning models and your constructive criticism helped me to always keep in mind the importance of my work for the field of hydrology. Thank you for your time, interest and attentive listening.

Equally huge thanks go to Martijn, Ruben and Jonathan. Despite a 99% online internship, you were able to give me a great insight into the work at Deltares. Each of you, with your own way, contributed a lot to my thesis and my personal growth. It is an honour for me that my work is valued so much, that I could be part of your team and could learn from you and your colleagues. Also thanks to Laurène and Jerom for the valuable conversations and for providing your data and results.

The intense home office time would have been much more difficult if I hadn't made such wonderful friends in Delft during my first Master's year. Thank you all for the many experiences, the fun, and the support in every situation. A big part of this is the time I spent in Board 66 & 67 of the Dispuut Water & Environment. Many thanks to you boardies and to Wim, for the activities and practical experiences as a balance to the computer work, and for the countless conversations over offline and online coffees. All of you made me feel at home in Delft.

I would especially like to thank Basti. No one has experienced all the ups and downs of this intense time with me like you. You always believed in me, even when I had the biggest doubts, you celebrated every success with me and inspired me with your critical and creative way of thinking. I am deeply grateful that you are part of my life and I look very much forward to our time together in Munich.

Finally, I want to thank my family and friends in Germany, without whom I would not have been able to study in Delft and who have been and are always there for me. A special thanks goes to Pia and Isi, above all for the last few weeks.

Katharina Wilbrand
Delft, November 2021

ABSTRACT

Rainfall-runoff modelling is essential for short- and long-term decision-making in the water management sector. The accuracy of streamflow predictions of hydrologic models increases with the availability of and the access to streamflow observations. Therefore, one of the key challenges in the field of hydrology is to produce Predictions in Ungauged Basins (PUB), where observations are lacking. Recent research has shown the potential of deep learning neural networks as an alternative approach to conceptual and process-based hydrologic models for this purpose.

In this study, the existing Multi-Timescale LSTM (MTS-LSTM) architecture is used to investigate if such a deep learning network is able to learn universal hydrologic behaviour. Therefore, a MTS-LSTM is trained on a large variety of >500 US catchments and subsequently tested outside the US, in the European Meuse river basin. The model is not re-trained or finetuned to simulate an ungauged situation and to asses whether the streamflow predictions can compete with those from the uncalibrated distributed model *wflow_sbm*.

Results indicate that the MTS-LSTM trained on US data cannot compete with *wflow_sbm* in the Meuse catchments. The simulated streamflow time series can be unrealistically shifted and scaled compared to the time series of observed streamflow due to sensitivity regarding static model input. This means, the values for catchment characteristics cannot be extremer in the testing data than in the training data. Therefore, it is recommended to select catchments for the model training such that the most extreme conditions are covered. In the case that the MTS-LSTM is trained for a specific region, results clearly compete with or outperform the distributed model. For the Meuse test catchments, the neural network achieves Nash-Sutcliffe Efficiency (NSE) values >0.46 where the application of *wflow_sbm* is problematic and yields negative NSE values. To exploit the potential of an LSTM, the model should be trained on all available data of the entire Meuse basin instead of on the subset of catchments used here. For some water management applications it is important to accurately predict high flow events. Training the MTS-LSTM with the Mean Quadrupled Error (M4SE) loss function showed that the peak flow representation can improve for catchments where the use of a NSE loss already leads to good predicting performance. Thus – for gauged and ungauged catchments – an implementation of a combined loss function appears a valuable follow-up research.

For ungauged catchments, these results imply that the global neural network model as tested in this study should be supplemented with finetuning. Thereby, the global model could not yet be applied everywhere, however, in regions where only few years of streamflow records are available. Alternatively, a regional neural network model trained on nearby catchments could be applied for PUB, if streamflow observations are accessible in the surrounding area. Finally, it is of high importance to maintain and extend the network of streamflow gauging stations globally, and to ensure easy access to the data.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AET	Actual Evaporation and Transpiration	18
ANN	Artificial Neural Network	2
CAPE	Convective Available Potential Energy	
CDF	Cumulative Density Function	36
CONUS	Conterminous United States	12
DEM	Digital Elevation Model	6
DL	Deep Learning	2
EA-LSTM	Entity-Aware - LSTM	14
ECMWF	European Center for Medium-Range Weather Forecasts	15
FDC	Flow Duration Curve	20
FHV	FDC High Segment Volume	
FNN	Feed-forward Neural Network	8
GRDC	Global Runoff Data Centre	14
GVF	Green Vegetation Fraction	35
HFD	Half-Flow Date	
HRU	Hydrologic Response Unit	12
HUC	Hydrologic Unit Code	12
KGE	Kling-Gupta Efficiency	19
LAI	Leaf Area Index	25
LSTM	Long Short-Term Memory	2
ML	Machine Learning	5
MSE	Mean Squared Error	11
MTS-LSTM	Multi-Timescale LSTM	3
M4SE	Mean Quadrupled Error	26
NLDAS-2	North American Land Data Assimilation System Phase 2 Dataset	15
NSE	Nash-Sutcliffe Efficiency	ii
PET	Potential Evaporation and Transpiration	6
PUB	Predictions in Ungauged Basins	3
RNN	Recurrent Neural Network	5
SAC-SMA	Sacramento Soil Moisture Accounting Model	38
SWE	Snow Water Equivalent	6
USGS	United States Geological Survey	14

1 INTRODUCTION

1.1 RESEARCH MOTIVATION

Hydrologic modelling of streamflow plays a key role in answering a variety of water resource related questions. The extreme flooding this summer (2021) in Western Europe and extremely high water levels in rivers like the Meuse river and tributaries of the Rhine river have demonstrated the importance of timely and precise streamflow forecasting to react and evacuate accordingly. Furthermore, decision making for flood-prevention and the development of emergency plans depends on good knowledge of the upstream catchment and how the discharge will change in the near and far future.

The current streamflow can be monitored with different gauging methods and these momentary observations support real-time decision making. These collected observations over many years serve to calibrate hydrologic models which then aim to predict the river discharge. Despite being widely applied, modelling and forecasting methods are exposed to various uncertainties rooted in the complex structure of the water cycle. Gaining in-depth understanding of the underlying, strongly interconnected processes like e.g. evaporation, infiltration in the root zone or surface and sub-surface flow is still a major challenge in the field of hydrology and hydrologic modelling (Blöschl et al., 2019).

This challenge is even greater when no field observations of present river discharge are available and when historical streamflow records are scarce, because then the calibration and evaluation of a hydrologic model becomes difficult or impossible and streamflow predictions are exposed to high uncertainties (Bouaziz et al., 2021; Hrachowitz et al., 2013). While Europe and North America show a great coverage of gauging stations for meteorological data and streamflow, data collection in other parts of the world is comparably low (see Figure 1.1) (Kratzert et al., 2019b).
Missing data complicates the development and calibration of physically based conceptual hydrologic models as hydrologic processes are hypothesised without possi-

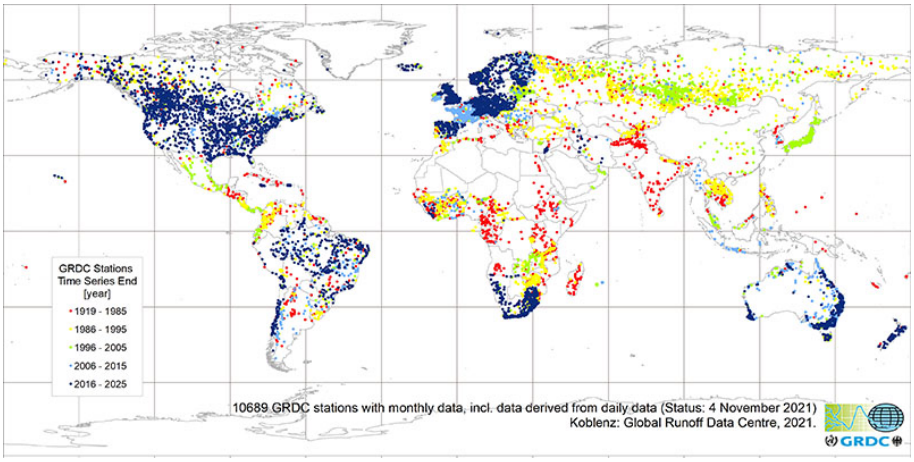


Figure 1.1: Global distribution of streamflow gauging stations (GRDC, 04.11.2021)

bilities for adequate validation (Bouaziz et al., 2021). Further, the transfer of existing models calibrated in a sufficiently gauged catchment into an ungauged region requires available data for re-calibration of the model as the characteristics of runoff behavior differ with each basin (Kratzert et al., 2019b).

However, the problem of accurate rainfall-runoff predictions has recently been successfully approached from a different perspective. The observed data – streamflow and meteorological parameters – incorporates much more than the measured value but tells about the combination and interaction of processes related to the characteristics of a catchment. Thus, the challenge is to extract this additional information from the observed data (Luce, 2014; Kratzert et al., 2019b).
Artificial Neural Network (ANN) models have shown in various fields their ability to "learn" from data and mimic non-linear relations - without implementation of physics-based processes or hydrologic knowledge of the programmer. While process-driven models perform best when calibrated with catchment specific observational data, a single neural network can learn rainfall-runoff relations with data from a variety of catchments at the same time, thus obtaining knowledge on patterns in a large range of different climatic and geological conditions. Kratzert et al. (2019a) and Ayzel et al. (2020) have shown that a so called Long Short-Term Memory (LSTM) neural network can outperform process-based hydrologic models. Long Short-Term Memory (LSTM)s are a special form of ANNs, allocated to the field of Deep Learning (DL).

The recent experiences with LSTM models offer opportunities to counteract the disadvantages for ungauged regions. The results of Kratzert et al. (2019b) and Ayzel et al. (2020) have proven that LSTM networks are able to yield on average good performances in ungauged basins. By creating reasonable streamflow forecasts for data-scarce basins with models trained on data from a variety of gauged catchments, the need for calibration data from the basins of interest decreases significantly. Regionalization or re-training of a model becomes obsolete if a global model can successfully be trained with sufficient data from other basins.

Investigating the potential of LSTM models to enhance streamflow predictions in regions with reduced access to observational data is very important. Those regions are to a significant share found in African, Asian and South-American countries with poor economies, political instabilities, social equality and equity problems, where prospects to extension of gauging station network are considerably low (Hrachowitz et al., 2013). Even when the network of stations can be extended it is not possible to gather historical meteorological and runoff data over the past decades which would be required to calibrate a (process-based) hydrologic model. Therefore, a global trained model ready to be applied in any place on earth with only requiring recent meteorological data can be a tool to counteract these societal injustices by enabling access to enhanced streamflow predictions.

1.2 PROBLEM STATEMENT

Commonly applied methods to predict streamflow in ungauged basins are hydrologic model parameters regionalization or nearest-neighbour regionalization (Ayzel et al., 2020). However, results of different methods are very catchment dependent (Razavi and Coulibaly, 2013; Ayzel et al., 2020) and the question of how to estimate runoff in ungauged basins as in how spatial heterogeneity in hydrologic processes is created remains one of the unsolved problems in hydrology (Blöschl et al., 2019).

The work of Kratzert et al. (2019b) and Ayzel et al. (2020) has shown the potential of LSTM networks to enhanced predictions in ungauged basins. However, these researchers worked with regionally trained networks, i.e. a model trained on US catchments tested in other US catchments or a model trained on Russian catchments tested in other Russian catchments. Testing a model in a completely different geographical region has not been done to our knowledge. Investigating the possibility of a readily trained model to be applied independently from the training region and identifying regional characteristics of catchments that affect model performance can contribute to enhanced access to Predictions in Ungauged Basins (PUB).

A prerequisite for neural network models to learn the relation between meteorological conditions and streamflow magnitude is the coverage in the training data of diverse hydrologic behaviours and climatic and environmental conditions that influence the streamflow generation. To allow for a globally applicable model, input data from a global dataset is required, such that the model input comes from the same data source regardless of the catchment to model. So far, LSTMs for hydrologic application have been trained with local datasets.

As the calibration and application of a process-based hydrologic model requires large amount of time, another reason to explore the suitability of neural networks in the field of hydrology is an extreme reduction in training and computing time while being able to work with a larger amount of input data. A short computing time is very important in cases of real-time streamflow predictions for e.g. the assessment of impact-based flood forecasting.

1.3 RESEARCH OBJECTIVE

The overall goal of this research is to investigate the generalization of a LSTM model for streamflow predictions across continents. As for operational cases like e.g. flood forecasting modelled streamflow time series on a sub-daily time scale are required, this research is done with the Multi-Timescale LSTM (MTS-LSTM) developed by Gauch et al. (2021). A single MTS-LSTM model is trained on data from a large variety of catchments to then generate streamflow predictions without further training for an arbitrary catchment anywhere around the globe. It is hypothesised that the performance is at least comparable to the distributed model *wflow_sbm*.

Since a spatially and temporal consistent data source is required for the model input, the first research objective is to assess the applicability of the global meteorological dataset ERA5. By comparing the performance of an MTS-LSTM model trained with ERA5 data in US catchments to previously achieved performance of the same model with the regional dataset NLDAS-2 through Kratzert et al. (2018), the suitability of the ERA5 dataset is assessed. Due to the coarser spatial resolution of ERA5 a deterioration in model performance is expected, however comparable performance to the process-based distributed model *wflow_sbm* is hypothesised. Wflow_sbm is considered as benchmark model, as it is regularly operated within Deltares and generated streamflow predictions are input for further hydrologic and hydraulic models.

The second research objective is the potential of the model trained on US ERA5 data to function as a globally applicable hydrologic model for streamflow predictions. The research is based on the assumption that a large variety of US catchments covers enough climate zones to represent the diversity in globally occurring hydrologic behaviour. Therefore, the MTS-LSTM model trained on US catchments functions as global model and is then tested in catchments outside of the US, more precisely in individual catchments of the Meuse basin in Europe. Here, hourly streamflow observations and predictions from the process-based distributed model *wflow_sbm*

are available to differentiate the performance of the neural network model. Further, a regional MTS-LSTM will be trained exclusively on Meuse data to (1) quantify the difference in performance between gauged and ungauged conditions and (2) work out the difference in performance and computing efficiency compared to *wflow_sbm*.

As the MTS-LSTM of the reference studies is trained on a NSE loss function, the third research objective of this study is to investigate model performance when training on a different loss function that is more sensitive to the magnitude of high flows.

With this research it is aimed to take a step towards improved access to streamflow predictions for ungauged basins with the approach of a globally applicable model. Moreover, the potential of neural networks in general and LSTM models in particular to supplement process-based hydrologic model approaches is investigated.

1.4 RESEARCH QUESTIONS

Does a MTS-LSTM trained on data from US catchments prove to be globally applicable as a hydrologic model?

The research question will be addressed by answering the following sub-questions:

- SQ1: How is model performance affected when using a globally available but lower resolution dataset (ERA5) as model forcing compared to a regional high resolution dataset (NLDAS-2)?
- SQ2: How does the trained MTS-LSTM model perform when applied in catchments in the Meuse river basin, simulating ungauged catchments?
- SQ3: Can model performance regarding high flow representation be improved by training the MTS-LSTM with a different loss function?

The model performance of the MTS-LSTM is benchmarked against the performance of the process-based distributed model *wflow_sbm* from Deltares.

Explanation of used terms:
River basin or *basin* refers to the the whole area from which rainfall contributes to the streamflow of a main river such as the Meuse. The term *catchment* is used to refer to a smaller unit within a river basin, describing the contributing area upstream of a streamflow gauge. *Ungauged basins* refers to basins where access to streamflow observations is not given. *Globally applicable* means that a single LSTM network is trained with data of a number of catchments with diverse characteristics and then ready to be applied in ideally any arbitrary basin of choice around the globe. Here, this is tested exemplary in catchments where actually observations are available to validate the predictions against.

1.5 READING GUIDE

The structure of the report is as follows: Chapter 2 gives theoretical information on hydrologic modelling approaches and background on machine learning methods in general and the chosen model in particular. Chapter 3 presents the study domain and used datasets including pre-processing methods, and describes the methodology for each research sub-question. Chapter 4 serves to present and discuss the results of the data analysis and the experiments for each sub-question to answer the overall research aim. The last Chapter 5 concludes the findings and states recommendations for ongoing research in the field.

2 THEORETICAL BACKGROUND

In this chapter, background information is given on the principles of hydrologic modelling and on the differences between conceptual or physics-based to data-driven approaches in Section 2.1. The spatially distributed hydrologic model *wflow_sbm* model is described as it serves as benchmark model for the experiments of this research. The concept of Machine Learning (ML) is introduced in Section 2.2 and related to the field of hydrology. The underlying principle of a Recurrent Neural Network (RNN) leads to the functioning of LSTMs and the special form of MTS-LSTMs.

2.1 HYDROLOGIC MODELLING

2.1.1 Conceptual and Physics-based Hydrologic Models

Hydrologic models are a simplified mathematical representation of the water cycle in the way the modeler perceives the hydrologic processes. Therefore, a model is a conceptualization of reality, which requires comprehension of interactions between rainfall-runoff relations and storage of water. Physically based approaches attempt to comply with the laws of conservation of mass, energy and momentum which demand knowledge of boundary conditions, system stages and system parameters (Hrachowitz and Clark, 2017). The accurate representation of how the water moves and distributes on local, basin, regional or global scale requires knowledge on geographical and geological parameters like climate zone, surface cover, vegetation and soil type, pores, slope and elevation.
In conceptual modelling, the before mentioned system parameters are regionalised or generalised, and processes or storages are spatially lumped and can be represented by empirically found relations. Thereby, what is perceived as reality is simplified and described with a higher degree of abstraction (Hrachowitz and Clark, 2017).

Hydrologic models in practice combine both physical-based (or process-driven) and conceptual methods to a certain degree in their representation of the water cycle. Thus, there exist various approaches diverging in how to deal with non-linearity of the rainfall-runoff relation and heterogeneity of system parameters (e.g. land cover, hydraulic conductivity, areal and temporal rainfall distribution) (Hrachowitz and Clark, 2017; Bouaziz et al., 2021). The chosen model structure and equations determine a number of parameters with unknown value which are calibrated with observational data of mostly rainfall, evaporation and streamflow. This introduces uncertainty due to the problem of equifinality, as one model can result in the same output with different constellations of parameter values, without knowledge on which parameter set is the representation closest to reality. As well, many models can yield the same or similar outputs despite distinctive internal structures. Bouaziz et al., 2021 compared a variety of process-driven conceptual models for the Meuse catchment and found "substantial dissimilarities [...] for annual and seasonal evaporation and interception rates, days of year with water stored as snow, mean annual max snow storage, size of root zone storage capacity".

2.1.2 Deltares *wflow_sbm* Process-based Hydrologic Model

The *wflow_sbm* model is a spatially distributed hydrologic model, based on topog_sbm, a process-based storm flow generation model by Vertessy and Elsenbeer (1999), schematised in Figure 2.1 (Verseveld et al., 2020). Distributed models work with a grid structure and raise the lumped representation to a higher resolution, e.g. from catchment-scale to one kilometer grid cells. Distributed models are an approach that is a combination of lumped conceptual and physics-based models.

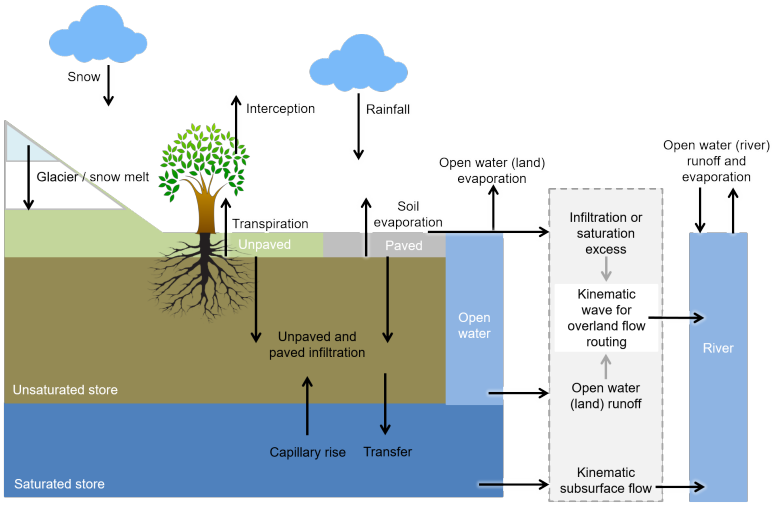


Figure 2.1: Overview of lumped processes and storages per grid cell and lateral flow representation in the *wflow_sbm* model (Verseveld et al., 2020)

Based on spatial time series of precipitation, Potential Evaporation and Transpiration (PET) and temperature the processes in each grid cell are computed and the discharge at the catchment outlet is modelled. The area of the catchment is determined from a Digital Elevation Model (DEM) and the derived flow direction of each cell in the raster grid which covers the catchment area. In *wflow_sbm* the hydrologic processes are represented by storage units which are connected through equations representing water fluxes. By modelling how the water moves through a grid cell, the discharge at the catchment outlet can be determined as a combined runoff from all grid cells with the help of kinematic wave routing for lateral subsurface, overland and river flow (Imhoff et al., 2020). Processes modelled by *wflow_sbm* are mentioned briefly in Table 2.1.

Process	Implementation
reference evaporation	determined based on input of potential evaporation and land use
snow	air temperature threshold (e.g. 0°C) to determine rate of snow melt, rate of refreezing and Snow Water Equivalent (SWE)
glaciers	based on snow modelling snowpack and thickness threshold ($\leq$ 10mm) to determine ice fraction of snowpack and rate of glacier melt
interception	analytical Gash model (daily time steps) or numerical Rutter model (sub-daily): NetInterception = Precipitation - Throughfall - Stemflow
soil	soil water accounting scheme, as in *Topog_sbm* (Vertessy and Elsenbeer, 1999)
kinematic wave	overland and river flow kinematic wave routing to determine flow across grid cells

Table 2.1: Model processes of *wflow_sbm* (Verseveld et al., 2020)

2.1.3 Data-driven Hydrologic Models

A distinct method of hydrologic modelling uses data-driven approaches which do not consider natural processes and physical laws however statistically determine a relation between input and output data i.e. between a set of meteorological forcing parameters and streamflow observations. However, those methods are limited in their ability of recognizing and reproducing non-linearity between rainfall and runoff which naturally occurs in the hydrologic system when rainfall events cause flushing of different storage types associated with their activation thresholds. For example, effective precipitation describes the through-fall of precipitation to the ground initiated when the interception storage of the vegetation cover is filled. Here, concurrent evaporative processes have to be considered and introduce further non-linearity. Therefore, it is not realistic to infer a runoff value from one unique value of rainfall amount or intensity. Observations show that similar rainfall events can cause significantly different response in river water levels depending on previous conditions within the catchment and the spatial rainfall distribution.

Where the statistical analysis of linear rainfall-runoff relations has its boundaries, ANNs allow to represent complex non-linearity between input and output through ML methods (Kratzert et al., 2018). With ML a system can identify patterns between independent input variables (meteorological parameters) and a dependent output variable (streamflow) and based on these learned patterns predict values for the output.

Since in nature the runoff can depend on processes in the water cycle of the preceding days, weeks or months, a special type of RNN finds application in the field of hydrologic modelling, the so called LSTM. While LSTM networks have been widely and successfully used for speech recognition or language modelling, the past years have proven that the memory function in the network is applicable for rainfall-runoff modelling, as well. Research has shown that hydrologic modelling with LSTM networks yield as good or better runoff predictions compared to process-driven hydrologic models (Kratzert et al., 2018).

2.2 HYDROLOGIC MODELLING WITH LSTM NEURAL NETWORKS

2.2.1 Structure of Neural Networks

The internal structure of a neural network is not related to the hydrologic cycle but consists of nodes ordered into layers. Each node has a hidden state h that determines the processing of the input data. The hidden state of a neuron is composed of its weight w_i and a bias term b_i. Figure 2.2 depicts how input information is processed in a single neuron, described by the following equation:

$$y = \phi(b + \sum_{i=1}^{m}(\omega_i x_i)) \tag{2.1}$$

where m different input variables x_i, also called features, reach the neuron simultaneously, are multiplied by a weight w_i, summed up and added to the bias b. An activation function ϕ subsequently transforms the weighted field into a continuous output signal y (Nielsen, 2015). Different functions can be chosen as activation function, among them *sigmoid, ReLU, tanh* or *Softmax* (Sharma, 2017).

Each network has one input layer with a neuron for each input feature, one output layer with a neuron for each variable to predict and at least one hidden layer

with an arbitrary number of neurons as shown in Figure 2.3 for a Feed-forward Neural Network (FNN). The sum of all weight and bias terms in a network is the number of trainable parameters which are tuned during the training of the network. In a FNN each neuron of a layer is connected to all neurons of the previous layer. Feed-forward denotes that information is only passed one-directional and no back-propagation or feedback loops are involved like in LSTM networks (LeCun, Bengio, and Hinton, 2015). A network with two or more hidden layers is called deep neural network, as the level of abstraction increases with the number of neurons and layers and thereby enabling to represent more complex patterns between input and output. However, a deeper network or a higher number of neurons per layer does not necessarily result in improved pattern recognition and depends on the amount of available training data. The optimal architecture in terms of number of layers and neurons therefore has to be determined for each application individually through hyperparameter-tuning.

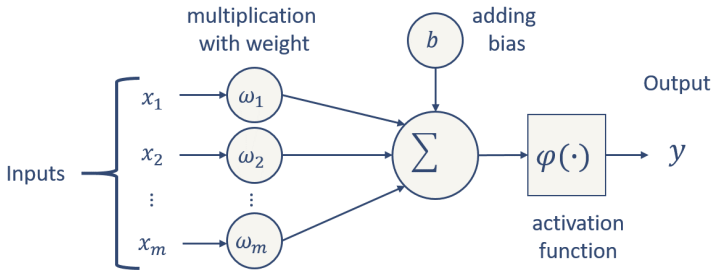


Figure 2.2: Artificial Neuron

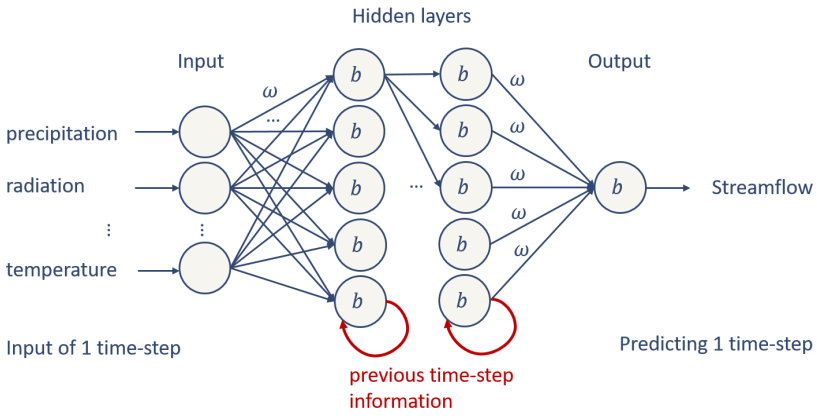


Figure 2.3: Basic FNN with 2 layers (blue). Adding information flow from previous time step changes the network into an RNN (red). Own figure.

2.2.2 Recurrent Neural Networks

Recurrent Neural Networks are suitable to process sequential data. Additional to optimization through backpropagation, previous hidden states are considered when updating the weights and bias values. Thereby, a memory function is added to the network which enables analysis of time series, as the output for the current input depends also on the hidden state resulting from input of the previous time step which is shown in Figure 2.3 in red and in Figure 2.4 on the left.

Vanishing gradient problem
Backpropagation as a method for updating model parameters of an RNN introduces the problem of vanishing gradients (Shen, 2018). The further away from the output layer a layer is located in the network, the more terms are multiplied when updating the parameters of its neurons. This results in exponentially small differences between current and new hidden states which slows down the training process significantly (Shen, 2018).

2.2.3 Long Short-Term Memory Neural Networks

Long Short-Term Memory neural networks are a specific form of RNNs with the ability to solve the vanishing gradient problem through an adjusted structure of the neurons. A neuron does not only have a hidden state but additionally has three *gates* which determine the information flow through the neuron and define an additional *cell state*. This cell state turns the short-term memory function of a simple RNN into a memory over longer time-periods. Thus providing an essential feature for modelling of rainfall-runoff relations where runoff depends on meteorological conditions over the past months or years (Kratzert et al., 2018).

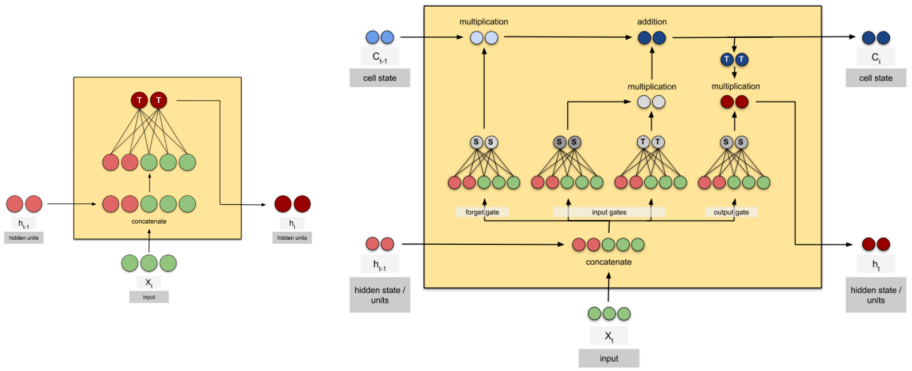


Figure 2.4: Simple RNN (left) and LSTM cell (right) (Karim, 2018)

The three gates determine if information is deleted (forget gate), if information is added (input gate) and which information is outputted (output gate) (Kratzert et al., 2018). Figure 2.4 shows the information flow through a LSTM cell at time t and how hidden h_t and cell c_t state are updated based on the current input x_t as well as on h_{t-1} and c_{t-1}. The information flow through the three gates is influenced by the *sigmoid* or *tanh* activation function shown by S and T respectively. The current input x_t can thereby be composed of various features, in the figure exemplary shown by three green nodes.

The following equations describe how each element $x[t]$ of the input sequence $x = [x[1], ..., x[T]]$ at a time step $1 \leq t \leq T$ is processed through a LSTM network, while $x[t]$ is composed of a number of time-evolving input features (Kratzert et al., 2019c):

$$i[t] = \sigma(W_i x[t] + U_i h[t-1] + b_i) \tag{2.2}$$
$$f[t] = \sigma(W_f x[t] + U_f h[t-1] + b_f) \tag{2.3}$$
$$g[t] = tanh(W_g x[t] + U_g h[t-1] + b_g) \tag{2.4}$$
$$o[t] = \sigma(W_o x[t] + U_o h[t-1] + b_o) \tag{2.5}$$
$$c[t] = f[t] \odot c[t-1] + i[t] \odot g[t] h[t] = o[t] \odot tanh(c[t]) \tag{2.6}$$

with:

$i[t], f[t], o[t]$ = input, forget and output gate
$g[t]$ = cell input
$h[t-1]$ = recurrent input (hidden state of previous time step)
$c[t-1]$ = cell state of previous time step
W, U, b = learnable parameters per gate
$\sigma(\cdot)$ = sigmoid activation function
$tanh(\cdot)$ = hyperbolic tangent activation function
$\odot(\cdot)$ = element-wise multiplication

Dynamic and static input
The LSTM is further able to work with both input evolving over time x_d and static input x_s simultaneously. In that case, the equation for the input gate i changes as follows:

$$i[t] = \sigma(W_i x_s + b_i) \tag{2.7}$$

All other equations remain as mentioned before while $x[t]$ is to be replaced with $x_d[t]$. Thus, the input gate does not change over time anymore but is initially determined by the static input x_s. When using an LSTM for hydrologic applications, static inputs can be characteristics describing the conditions within a catchment e.g. the average vegetation cover, soil composition or slope. The additional input results in the mapping of the meteorological time-series into streamflow being conditioned by the values of the static input and begin different for each catchment (Kratzert et al., 2019c).

2.2.4 Multi-Timescale LSTM

Multi-Timescale LSTM is a LSTM network with an individual branch for each time scale introduced by Gauch et al. (2021) and is used for all experiments in the here presented research study. The model is available in the GitHub repository of NeuralHydrology (2020). In the case of predicting streamflow on a daily and hourly scale, the MTS-LSTM network has two branches. First, the coarser branch processes daily input time series with a look-back window of T_D (e.g. 365 days) to produce a streamflow prediction for one day. Then, to get hourly streamflow predictions for the same day, the hidden and cell states of the daily branch from point $T_D - T_H/24$ are transferred to the hourly LSTM branch. T_H is the look-back window based on which the hourly predictions for one day are generated, e.g. 2 weeks = 336 hours. See an exemplary visualization with $T_H = 72$ hours in Figure 2.5.

Advantage over regular LSTM
This architecture brings the advantage of producing predictions on temporal higher resolutions without an extreme increase in the size of the input time series, assuming an unchanged look-back window: A regular LSTM would have to process 8760 hourly sequences of meteorological forcing time series to produce 24 hourly streamflow predictions of a single day. The MTS-LSTM generates those predictions with 365 daily + 336 hourly = 701 sequences and has proven to be more than twice as efficient regarding computation time (Gauch et al., 2021).

Multiple time scale input datasets
Moreover, as depicted in Figure 2.5, each of the branches can work with an individual input forcing dataset. Meaning, that the input to the daily branch can be an entirely different forcing dataset with other meteorological parameters than the hourly branch. Additionally, all branches can process meteorological forcing from multiple datasets simultaneously.

Loss function and consistency across time scales

Neural Networks are optimized with a loss function (Nielsen, 2015). The loss function is chosen by the user and can e.g. be the Mean Squared Error (MSE). For conceptual hydrologic models the NSE is often used to calibrate and optimise the model parameters. Also for neural networks the NSE is a possible loss function to assess if the simulated streamflow matches the real, observed streamflow. The MTS-LSTM with each one model branch for daily and hourly streamflow predictions is evaluated on both time scales with the NSE loss function (Gauch et al., 2021). To ensure that the daily predictions match with the daily mean of the hourly predictions, Gauch et al. (2021) implemented a regularization in the loss function of the MTS-LSTM model:

$$NSE_{reg}^{D,H} = \underbrace{\frac{1}{2} \sum_{\tau \in \{D,H\}} \left(\frac{1}{B} \sum_{b=1}^{B} \sum_{t=1}^{N_b^T} \frac{(\hat{y}_t^T - y_t^T)^2}{(\sigma_b + \epsilon)^2} \right)}_{\text{NSE per time scale}} + \underbrace{\frac{1}{B} \sum_{b=1}^{B} \frac{1}{N_b^D} \sum_{t=1}^{N_b^D} \left(\hat{y}_t^D - \frac{1}{24} \sum_{h=1}^{24} \hat{y}_{t,h}^H \right)^2}_{\text{mean squared difference regularization}} \tag{2.8}$$

with:

B = number of basins
N_b^T = number of samples for basin b at time scale τ
y_t^T = observed streamflow values
$\hat{y}_t^T$ = predicted streamflow values
σ_b = observed streamflow variance of basin b over whole training period
ϵ = small value to guarantee stability

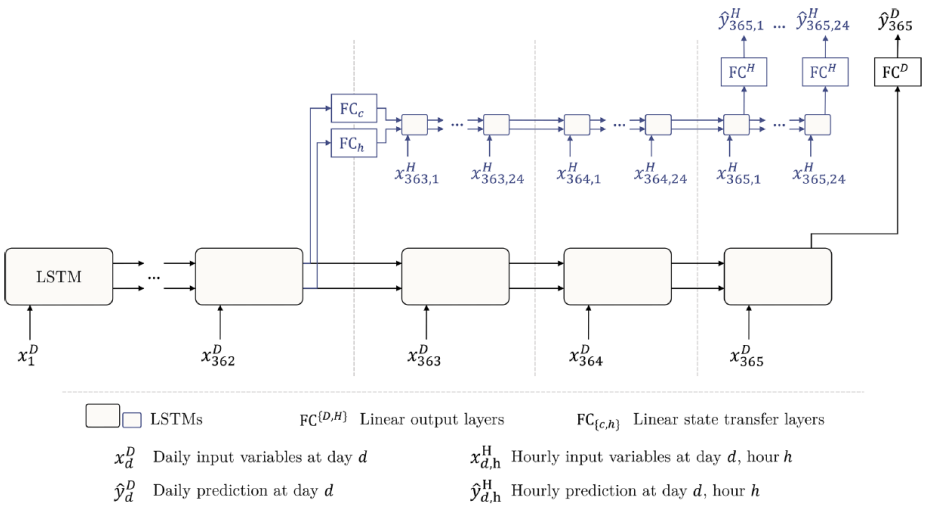


Figure 2.5: MTS-LSTM architecture with one branch for daily and one branch for hourly predictions. Here with T_D = 365 days and T_H = 72 hours. The model weights from the daily branch are shared with the hourly branch through a linear state transfer layer $FC_{\{c,h\}}$ (Gauch et al., 2021).

3 MATERIALS AND METHODS

In this chapter, the underlying data for the research is described. In section 3.1 the study domain, the sources of the input and output datasets and the pre-processing methods applied to this data are specified. The model performance is assessed by catchment clustering, evaluation metrics and hydrologic signatures which are presented in Section 3.2. Then, Sections 3.3, 3.4 and 3.5 explain the methodology used to answer each of the three research sub-questions.

3.1 DATA

3.1.1 Study Domain

For the training of the MTS-LSTM 516 catchments included in the Camels US dataset are chosen. Those catchments are headwater catchments with an area smaller 2000 km^2. Due to little to no reservoir presence and their upstream location the catchments are considered as *near natural*. This is important as the neural network is supposed to learn and mimic natural water cycle processes. Anthropogenic intervention in the rainfall-runoff process would ingest irregularities in the input-output relation and thereby deteriorate model performance. The Conterminous United States (CONUS) are structured into 18 Hydrologic Response Unit (HRU) which can be identified by a 2-digit Hydrologic Unit Code (HUC). Figure 3.1 depicts the distribution of size and number of catchments per HRU used for this study. Only 2 (5) catchments of HRU 9 (16) are included in this study and all catchments of this unit have the overall largest (smallest) area. The catchment area and shape are derived based on the global digital elevation model *Merit DEM* with a resolution of 3 arc seconds (Yamazaki et al., 2017) and the resulting flow direction is computed with the package *pyflwdir*[1].

To test the trained model regarding its applicability in out-of-sample catchments, five near-natural headwater catchments in a European river basin, the Meuse basin, are chosen (see Figure 3.2). The Meuse is a rain-dominated river with seasonal runoff varying largely throughout the year, mainly due to large evaporation differences between summer (high) and winter (low) (Bouaziz et al., 2020). The Belgian catchments can have snow every year, however, the contribution to runoff is low with maximum 15 mm/yr SWE (Bouaziz et al., 2020). The catchments are described more detailed in Table 3.1. Catchment 6, Treignes has the highest forest fraction, largest area and no fissured aquifers. Opposed to that, catchment 702, Yvoir, shows fissured aquifers and agricultural land use. The two catchments 701 and 703, Hastiere and Warnant, have been part of the severe floods in July, 2021. In Hastiere, fissured aquifers occur, the forest fraction is high and less land is used by agriculture. In Warnant, no fissured aquifers occur while agriculture is more present and forest area smaller. The flashiness index is higher for catchment 701. Catchment 13, Huccorgne, has the smallest forest fraction and highest agricultural use and is located in the flattest area. Mean precipitation and streamflow are here lowest from those five selected catchments while the silt fraction in the soil is highest.

1 https://github.com/Deltares/pyflwdir

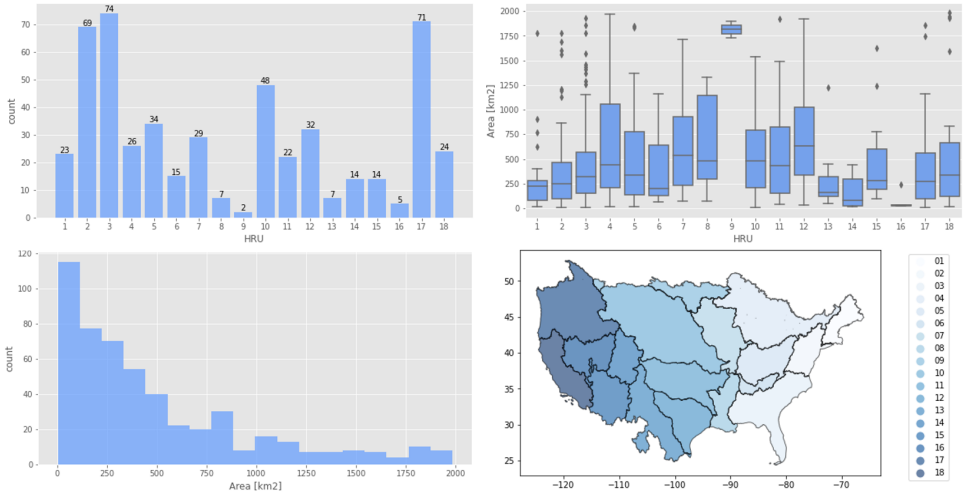


Figure 3.1: Number and size of Camels US catchments of each HRU, data source: Addor et al. (2018)

Name	ID	P	PET	Q	$\frac{PET}{P}$	$\frac{Q}{P}$	flash.	forest	agriculture
	-	$\frac{mm}{a}$	$\frac{mm}{a}$	$\frac{mm}{a}$	%	%	%	%	%
Treignes	6	985	579	398	60	41	28	54	27
Huccorgne	13	737	593	181	82	25	19	3	80
Hastiere	701	802	582	285	73	36	32	41	40
Yvoir	702	865	577	264	68	31	13	16	60
Warnant	703	819	586	275	72	34	12	20	64

Name	ID	fissures	clay	sand	silt	area	slope
	-	%	%	%	%	km^2	%
Treignes	6	0	22	21	57	551	6.6
Huccorgne	13	16	21	9	70	307	2.6
Hastiere	701	0	24	21	55	169	5.4
Yvoir	702	71	24	13	63	226	6.4
Warnant	703	56	24	15	61	127	6.2

Table 3.1: Attributes for selected catchments of Meuse basin, from (Bouaziz et al., 2020). flash.: Flashiness-index, fissures: fissural aquifers.

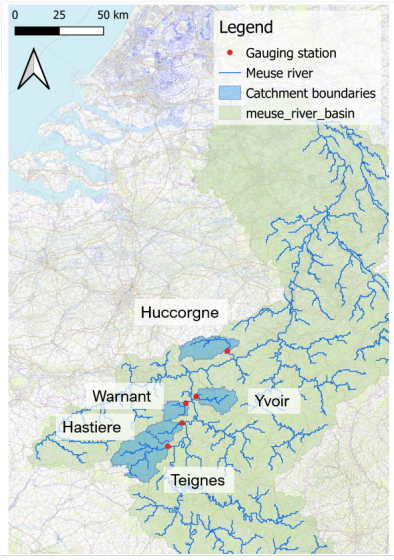


Figure 3.2: Five test catchments from the Meuse river basin in the Belgian Ardennes.

3.1.2 Datasets

Streamflow
Neural Networks and process-based hydrologic models require streamflow observations over many years for calibration. The United States Geological Survey (USGS) Water Information System provides sub-daily streamflow observations for the 516 selected US catchments. Gauch et al. (2020) provide the time series averaged to hourly and daily time steps over the period from 01.10.1979 until 30.09.2018. The time series from 1990 until 2003 serves as training data. For ten catchments the time series starts not before 2008 (see catchment IDs in Appendix B.1), which therefore are treated as 'ungauged catchments' within the US.
For the Meuse test catchments, streamflow observations are required to assess model performance when using the US trained MTS-LSTM to simulate ungauged catchments, as well as for the training of the regional Meuse MTS-LSTM. Observations are available from the Global Runoff Data Centre (GRDC) [2].

Catchment attributes
The Camels US dataset provides a range of catchment characteristics describing climate, topography, soil, land cover and hydrology of US catchments. Kratzert et al. (2019c) performed an attribute ranking based on the sensitivity of an Entity-Aware - LSTM (EA-LSTM) regarding 27 Camels US attributes shown in Table 3.2. The climatic indices are calculated from the Daymet dataset over the period from 01.10.1989 until 30.09.2009. The hydrologic indices are based on USGS streamflow data over the same period, the land-cover data is derived from *MODIS* over the period from 2002 until 2014 and PET is determined using Priestley-Taylor method calibrated per catchment (Addor et al., 2017).
To work on a global scale, global datasets are used to derive equal or similar attributes. For this purpose, the HydroMT package[3] from Deltares is used. Therefore, the created static attributes dataset is referred to as *HydroMT dataset*. All underlying data sources are shown in Table 3.2. Attributes describing the soil composition are excluded since, in addition to their lower sensitivity ranks, the soil composition is respected when determining the *Saturated hydraulic conductivity* with the pedotransfer functions that are implemented in HydroMT (Imhoff et al., 2020). As the *Precipitation seasonality* is an attribute with lower sensitivity (0.27) and is derived from temperature and precipitation, which are included in the forcing, it was decided to exclude this attribute. Although some attributes are derived from the forcing time series (e.g. mean precipitation or mean PET) or dependent on other attributes (e.g. aridity as the ratio of mean precipitation to mean PET) they are not excluded due to higher sensitivity rankings. As this study focuses on the generalization of an MTS-LSTM and applicability in the case of ungauged basins, the search for independent attributes with the execution of a new sensitivity analysis is refrained from.

Meteorological parameters
To predict streamflow, the input parameters for the MTS-LSTM need to describe meteorological conditions over time. Process-based hydrologic models require specific input parameters depending on the implemented conceptualisation of the water cycle. The forcing for the *wflow_sbm* model includes precipitation, PET and temperature time series. The forcing parameters are derived from the ERA5 dataset, as it is globally available. As the performance of the LSTM model trained with ERA5 data will be compared to the same model trained with NLDAS-2 data, the forcing parameters available from NLDAS-2 are chosen.

2 https://www.bafg.de/GRDC/EN/Home/homepage_node.html
3 https://github.com/Deltares/hydromt

Rank	Catchment characteristic	Sensitivity	dependent	excluded	Camels US data source	HydroMT data source
1	Mean precipitation	0.68	yes		Daymet	ERA5
2	Aridity	0.56	yes		Daymet	ERA5
3	Area	0.50			USGS	Merit Hydro
4	Mean elevation	0.46			USGS	Merit Hydro
5	High precipitation duration	0.41	yes		Daymet	ERA5
6	Fraction of snow	0.41	yes		Daymet	ERA5
7	High precipitation frequency	0.38	yes		Daymet	ERA5
8	Mean slope	0.37			USGS	Merit Hydro
9	Geological permeability	0.35			GLHYMPS	GLHYMPS
10	Fraction of carbonate sedimentary rock	0.34			GLiM	GLiM
11	Clay fraction	0.33		yes	STATSGO	
12	Mean PET	0.31	yes		Daymet	ERA5
13	Low precipitation frequency	0.30	yes		Daymet	ERA5
14	Soil depth to bedrock	0.27		yes	Pelletier	
15	Precip. seasonality	0.27	yes		Daymet yes	
16	Fraction of forest	0.27		yes	USGS	
17	Sand fraction	0.26		yes	STATSGO	
18	Saturated hydraulic conductivity	0.24			STATSGO p.transfer	soilgrids,
19	Low precip. duration	0.22	yes		Daymet	ERA5
20	Max. GVF	0.21			MODIS	vito
21	Annual GVF diff.	0.21			MODIS	vito
22	Annual LAI diff.	0.21			MODIS	MODIS
23	Volumetric porosity	0.19			STATSGO	GLHYMPS
24	Soil depth	0.19			STATSGO	soilgrids
25	Max. LAI	0.19			MODIS	MODIS
26	Silt fraction	0.18		yes	STATSGO	
27	Max. water content	0.16			STATSGO	soilgrids

Table 3.2: Camels US catchment characteristics used as static attributes by Kratzert et al. (2019c) who determined model sensitivity. Attributes that depend on forcing time series or other attributes are indicated as *dependent*. Attributes which are not used in the range of this study are indicated as *excluded*. The right column shows the data sources from which the attributes are derived for the HydroMT attribute dataset.

NLDAS-2: The North American Land Data Assimilation System Phase 2 Dataset (NLDAS-2) provides hourly data for eleven different forcing parameters from 1979 to present with a spatial resolution of 0.125 degree ($\sim 12km$). This reanalysis dataset is based on different infiltration, soil moisture and land surface models (Xia et al., 2012). Precipitation is based on temporal disaggregation of daily observations supported with disaggregation-weights derived from hourly radar precipitation estimates.

ERA5: The ERA5 dataset is the climate reanalysis (fifth generation) of the European Center for Medium-Range Weather Forecasts (ECMWF) providing atmospheric parameters with global coverage. All parameters are available from 1979 to present, but a preliminary back-extension to January 1950 is already available (Bell et al., 2020). Moreover, climate prediction datasets up to the end of the 21st century[4] based on the ERA5 dataset are in development. This facilitates research on climate scenarios and developments with a model trained on ERA5 data . ERA5 is of 0.25

4 https://climate.copernicus.eu/high-resolution-climate-projections

degree ($\sim 31km$) resolution and provides forcing time series on hourly time steps. The data origins from the reanalysis of observations and the model output from the ECMWF Integrated Forecast System.

Table 3.3 shows all eleven forcing parameters included in the NLDAS-2 dataset and the equivalent parameters derived from the ERA5 dataset. Seven of the parameters are the same in both datasets, given in identical units. Precipitation, and evaporation as well as convective precipitation from ERA5 are measured in m instead of mm and are converted to the same unit as given in the NLDAS-2 data. The convective fraction of precipitation is the ratio of convective precipitation to the total precipitation. Specific humidity is not provided in ERA5, however, the possible amount of water taken up by the air depends on the dew point temperature which is available in the ERA5 data. The correlation between both parameters can be seen exemplary in Figure 3.3 for one catchment with a correlation coefficient of 0.92. Not exactly matching absolute values but a strong correlation between the two parameters is required, since the output of the neural network depends on the relative change of the input parameters over time, not on the absolute values.

Nr.	NLDAS-2 name	Unit	ERA5 name	Unit
1	Total precipitation	$\frac{kg}{m^2h}$	Total precipitation	$\frac{m}{h}$
2	2m air temperature	K	2m air temperature	K
3	Surface pressure	Pa	Surface pressure	Pa
4	Surface downward longwave radiation	$\frac{W}{m^2}$	Mean surface downward longwave radiation flux	$\frac{W}{m^2}$
5	Surface downward shortwave radiation	$\frac{W}{m^2}$	Mean surface downward shortwave radiation flux	$\frac{W}{m^2}$
6	2m specific humidity	$\frac{kg}{kg}$	2m dew point temperature	K
7	Convective Available Potential Energy (CAPE)	$\frac{J}{kg}$	CAPE	$\frac{J}{kg}$
8	Potential evaporation	$\frac{kg}{m^2h}$	Potential evaporation	$\frac{m}{h}$
9	Convective fraction	–	Convective precipitation	$\frac{m}{h}$
10	10m u wind component	$\frac{m}{s}$	10m u-component of wind	$\frac{m}{s}$
11	10m v wind component	$\frac{m}{s}$	10m v-component of wind	$\frac{m}{s}$

Table 3.3: Forcing parameters available from NLDAS-2 and equivalent parameters from ERA5 dataset. Parameters with gray background differ between both datasets in their unit or definition.

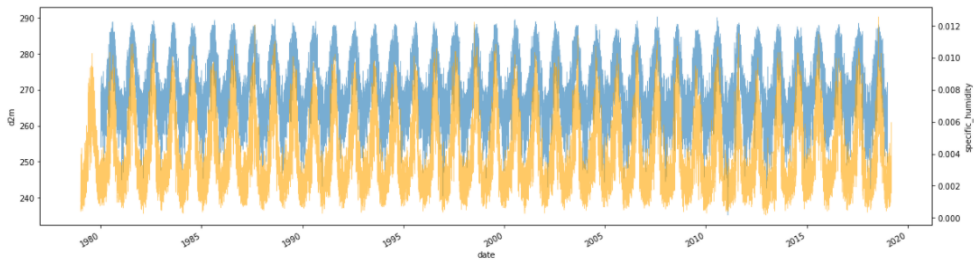


Figure 3.3: Time series plot of 2m dew point temperature in [K] (blue, ERA5) and specific humidity in [kg/kg] (orange, NLDAS-2) for catchment *07208500* with a correlation coefficient of 0.92.

3.1.3 Data Pre-processing

The data preparation for training the MTS-LSTM requires a check for missing values and outliers in observed streamflow, catchment attributes and dynamic forcing. As the NLDAS-2, Camels US and USGS data has been used by the NeuralHydrology (2020) group, a quality check showed that the provided data does not require further pre-processing. A comparison of monthly and yearly climatologies based on NLDAS-2 and ERA5 forcing revealed deviation in monthly means and yearly sums of e.g. precipitation, however, the differences were in the expected range.
For the Meuse, the forcing of two catchments (701 and 703) required handling of outliers. Values above or below a certain threshold or within a range around the mean are replaced with the mean of the whole time series. The forcing raw and cleaned time series for catchment 703 are exemplary shown in Appendix D.3.

Scaling
The units of the different input parameters often differ from each other (e.g. mm/h, $°C$, Pa) which can result in different scales for each input parameter. If one parameter varies in units, e.g. from m to km, and contains a large range of values across all catchments, the model sensitivity to this parameter would suffer and subsequently the model performance would decrease. A crucial step of data pre-processing is therefore the scaling of numerical input parameters. Here, the scaling is done by standardization. The time series of the input parameters are taken over the training period only, excluding all data from validation and testing period, and the mean $\bar{x_d}$ and standard deviation σ_{x_d} are calculated for each parameter. Then, for each value in the time series the standardised value $x_{d,scaled}$ is calculated with Equation 3.1.
The scaler determined with the training data for a parameter thus consists of its mean and standard deviation and is stored to the local drive to later scale the input data for validation and testing purposes with this exact scaler. The same scaling method is applied to the target parameter y (streamflow). $\bar{y}$ and σ_y of observed streamflow are determined over the training period of all catchments included in the model training. The scaler is first applied to back-scale the modeled streamflow predictions in the training process and compare the output with the observed streamflow. Later, the back-scaling is applied in the same way during validation and testing with Equation 3.2. The same standardization is applied on the static attributes.

$$x_{d,scaled} = \frac{x_d - \bar{x}_d}{\sigma_{x_d}} \tag{3.1}$$

$$y = y_{scaled} * \sigma_y + \bar{y} \tag{3.2}$$

3.2 EVALUATION METHODS

To enable the analysis of the large group of 516 US catchments, a clustering of these catchments based on their characteristics is done before the model performance assessment. The modeled streamflow time series is compared to the observed time series and analysed by the means of metrics and flow signatures, with a focus on the representation of high flow events. Grouping the US catchments based on the metrics and signatures will show if and where the model performance can be related to the catchment characteristics. In this section, the clustering method, applied metrics and signatures are described.

3.2.1 Clustering based on Catchment Characteristics

The selected 516 US catchments are clustered based on the 21 attributes described under 3.1.2. The chosen clustering method is k-means clustering where k describes the number of clusters to divide the given data into. The goal is to find a number of catchments with similarities across certain attributes that build a cluster. Furthermore, these clusters are plotted together in the Budyko Framework.

K-means clustering
To determine the optimal number of clusters, the silhouette coefficient[5] is calculated. The higher the coefficient the better the individual clusters are distinguishable. Figure 3.4 shows the silhouette coefficient for clustering the 516 catchments based on Camels US attributes and HydroMT attributes in up to ten clusters. With the Camels US attributes, clustering into six or seven clusters has the highest silhouette coefficient (disregarding the option of only two clusters). As grouping into seven clusters still leads to groups with more than 30 catchments, k=7 is chosen. For HydroMT the optimal choice is k=8 clusters. The clustering is done with the *sklearn.cluster* and *sklearn.metrics* module from *scikit-learn* [6].

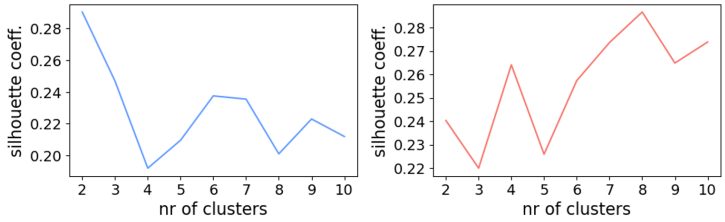


Figure 3.4: Silhouette coefficients for clustering 516 US catchments on 21 Camels US attributes (left) and 21 HydroMT attributes (right)

Budyko Framework
The Budyko Framework, developed by Budyko (1974), is based on the hypothesis that the long term water balance of a catchment is the partitioning of mean precipitation $\bar{P}$ into mean evaporation and transpiration $\bar{ET}$ and discharge $\bar{Q}$. The underlying assumption is that storage changes impact the water balance on a short term time scale but balance out over long time periods $\Delta(\bar{S}) = 0$. Thus,the water balance equation results in:

$$\bar{P} = \bar{ET} + \bar{Q} \tag{3.3}$$

The Budyko curve then sets the evaporative index $\frac{AET}{P}$, as the fraction of Actual Evaporation and Transpiration (AET) to mean precipitation P, in relation to the aridity index $\frac{PET}{P}$, signifying the PET as fraction of mean precipitation:

$$\frac{AET}{P} = \sqrt{\frac{PET}{P} * \tanh \frac{PET}{P}^{-1} * (1 - e^{-\frac{PET}{P}})} \tag{3.4}$$

The Budyko Framework is limited by the water limit and the energy limit. The former means that not more water can evaporate or be transpired than is available through precipitation ($AET \leq P \Leftrightarrow \frac{AET}{P} \leq 1$). The energy constraint means that the available thermal energy through radiation can maximally result in the potentially possible evaporation and transpiration PET, therefore $AET \leq PET \Leftrightarrow \frac{AET}{P} \leq \frac{PET}{P}$. When a catchment plots higher in the Budyko frame on the vertical axis, it is an indication for less mean discharge and larger mean evaporation and transpiration.

5 https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
6 https://scikit-learn.org/stable/modules/classes.html

Catchments plotting between 0 and 1 on the horizontal axis are energy limited, catchments plotting > 1 are water limited while a shift to the right signifies a change in climate towards less precipitation and warmer temperatures.
Annual values of $\frac{AET}{P}$ as a function of $\frac{PET}{P}$ can deviate from the long term value towards more arid or humid conditions. Furthermore, the partitioning of P in AET and Q can fluctuate depending on the meteorological conditions in each year (Jones et al., 2012). Reasons for catchments not plotting on or close to the Budyko curve are missing and low quality measurements of precipitation, temperature, evaporation and streamflow, anthropogenic interventions like land use change, or effects of and response to climate change. Existing methods to determine evaporation consider different meteorological parameters (temperature, radiation, wind speed and direction) and vary in results for PET. Changing climatological conditions result in variations of rainfall and temperature magnitudes, patterns, frequencies and duration upon which the natural system responds. Subsequently, the main vegetation type can change and affect evaporation, transpiration, storage and runoff processes.

3.2.2 Evaluation Metrics

Evaluation metrics help to assess model performance by describing how well the simulated streamflow represents the observed streamflow. In hydrology, a multi-criteria assessment of model performance is required as similarity of both hydrographs (simulated and observed) can be interpreted regarding various indicators, depending on the overall modelling goal and intended application for the streamflow time series. Therefore, not only the commonly used NSE and its further development the Kling-Gupta Efficiency (KGE), but also metrics regarding peak flows are considered and described subsequently. The following notation is chosen for all equations:

Q_s = simulated flow
Q_o = observed flow
t = time step
σ = standard deviation
h = index for high flow

To assess the impact on model performance when using a lower resolution dataset to derive forcing parameters, the results are systematically analysed. First, all US catchments are divided into groups regarding the peak timing, high flow bias and peak magnitude as shown in Table 3.4. In addition to the metrics used for grouping, NSE and KGE are considered for the analysis. Then, the differences between model results from NLDAS-2 forcing and from ERA5 forcing are determined and quantified per group. The metrics are described in more detail in the following sections.

Metric or Signature	Performance groups	
	Best	Lowest
Peak timing error, daily [days]	$\Delta t < 1$	$2 \leq \Delta t$
Peak timing error, hourly [hours]	$\Delta t < 3$	$6 \leq \Delta t$
High flow bias (FDC) [%]	$FHV < \lvert\pm 15\rvert$	$\lvert\pm 30\rvert < FHV$
Peak magnitude, rel. difference [-]	$\epsilon_{rel} < 0.1$	$0.5 < \epsilon_{rel}$

Table 3.4: Grouping catchments according to model performance based on different metrics.

NSE
The Nash-Sutcliffe Efficiency (NSE) is a broadly used measure of performance for hydrologic models. The dimensionless score describes how good the simulated values fit the observed flow values. The MSE between observation and simulation is

normalised to the sum of the observations' variance σ_o^2 (Eq. 3.5) (Gupta et al., 2009). Subtracting this ratio from 1 results in NSE values ranging from $-\infty$ to 1 while a value closer to 1 indicates a better fit. An NSE > 0.75 is regarded as "very good" while NSE values below 0 indicate that the observed average is a better fit than the simulated values (Nash and Sutcliffe, 1970).
The assessment of a model on a large number of catchments is done by determination of the mean NSE and visualization in a cumulative density function of all NSE values.

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Q_s^t - Q_o^t)^2}{\sum_{t=1}^{T}(Q_o^t - \overline{Q}_o)^2} = 1 - \frac{MSE}{\sigma_o^2} \tag{3.5}$$

KGE
The Kling-Gupta Efficiency (KGE) is a performance measure developed from a decomposition of the NSE (Eq. 3.6). The NSE has been criticised to overestimate model performance in highly seasonal regions and therefore being less suitable for comparison of model performance across basins with different seasonality (Gupta et al., 2009). The three components of the NSE are the linear correlation coefficient r (Eq. 3.7), the bias α as the ratio of standard deviations of observed and simulated flow (Eq. 3.8) and the coefficient of variances β as the ratio of mean observed and simulated flow (Eq. 3.9). For r, values can reach from -1 to 1 while values closer to 1 indicate a stronger correlation, 0 means no correlation and negative values an inverse correlation. The α_{NSE} can range from 0 to ∞ and values closer to 1 are desirable. The β_{NSE} can range from $-\infty$ to ∞ and values closer to 0 indicate better performance.

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{3.6}$$

$$r = \frac{\sum_{t=1}^{T}(Q_o^t - \overline{Q_o})(Q_s^t - \overline{Q_s})}{\sqrt{\sum_{t=1}^{T}(Q_o^t - \overline{Q_o})^2 (Q_s^t - \overline{Q_s})^2}} \tag{3.7}$$

$$\alpha_{NSE} = \sigma_s / \sigma_o \tag{3.8}$$

$$\beta_{NSE} = (\mu_s - \mu_o) / \sigma_o \tag{3.9}$$

Flow duration curve
The Flow Duration Curve (FDC) is a representation of all flows of a hydrograph over a certain period of time in descending order, disregarding their time of occurrence (Searcy, 1959). An FDC is more expressive when based on longer time series. Thereby, full hydrologic years should be considered to not capture e.g. a wet winter without the corresponding dry summer season. FDCs can be computed for different time scales like individual months, years, decades or the entire length of an available time series. The time scale should be considered when analysing and interpreting the FDC. The curve shows which percentage of the time the flow rate is equal to or higher than a certain value. Q95 refers to the flow rate equalled or exceeded in 95% of the time i.e. the low flows, Q5 respectively refers to the high flow rates equalled or exceeded in 5% of the time. Additional model performance metrics are deduced from the FDC and explained in the following.

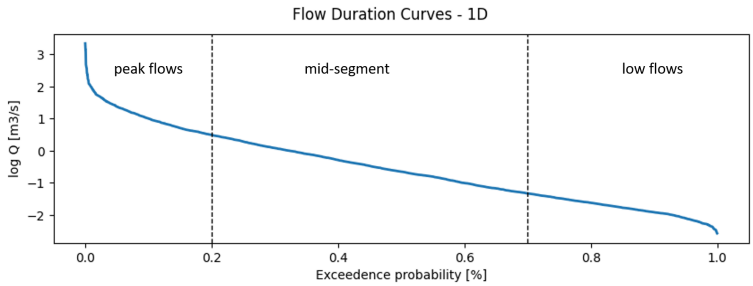


Figure 3.5: Typical FDC shapes with logarithmic y-axis.

High flow bias
To assess how well the magnitude and occurrence of peak flows are represented in the simulated flow, the peak flow bias of the FDC is determined. The sum of all deviations between simulated and observed high flows with a maximum exceedance percentage of 2% is related to the sum of the observed 2% exceedance flow rate (Eq. 3.10). The closer the bias is to 0, the better is the peak flow fit. As simulations can be higher or lower than the observations, the peak flow bias can range from $-\infty$ to ∞ (Yilmaz, Gupta, and Wagener, 2008). FHV stands for FDC High Segment Volume.

$$FHV = \frac{\sum_{t_h=1}^{T_h}(Q_{S,h} - Q_{O,h})}{\sum_{t_h=1}^{T_h} Q_{O,h}} * 100 \tag{3.10}$$

Peak timing error
The time difference between observed and modeled peak flows is determined with the same method as implemented by NeuralHydrology (2020). All flows higher than the mean flow plus one standard deviation in the observed streamflow time series are selected. Only peaks with a distance of 100 days (hours) are kept for the analysis. Therefore, the lowest peaks are rejected until the distance criterion is fulfilled. Then, the simulated time series is searched for a peak around the same point in time $\pm$ three days (twelve hours), for all peaks found in the observed daily (hourly) time series. The peak timing error is the mean of all absolute time differences between observed and simulated peaks for the analysed time series of one catchment.

Peak magnitude error
The peaks are found with the same method as for the peak timing error metric. Then the absolute and relative differences between the observed and the simulated peak are determined, with Equation 3.11 and 3.12. The final metric is the average of the absolute ϵ_{abs} and ϵ_{rel} values per catchment. Additionally, in the simulation the recognised peaks are counted and the sum of peaks that are higher (lower) than the observed time series is determined, referred to as *count*, *sim<obs* (*sim>obs*).

$$\epsilon_{abs} = Q_o - Q_s \tag{3.11}$$

$$\epsilon_{rel} = \frac{\epsilon_{abs}}{Q_o} \tag{3.12}$$

3.2.3 Hydrologic Signatures

The following five flow signatures regarding high flows are calculated based on the observed and the modeled streamflow time series. A comparison of both values

per signature (observed and modeled signature) serves as model performance indicator. For the US model a correlation coefficient between observed and modeled signatures can be calculated based on the results of a large number of catchments to see if and which parts of the high flow are better represented by the model. For the model evaluation with signatures, methods from NeuralHydrology (2020) are used.

High flow frequency describes how often the $threshold \cdot \bar{Q}$ is exceeded continuously. The *threshold* is set to 0.9. High flow frequency is the average frequency measured in d/a.

High flow duration describes how many time steps on average the $threshold \cdot \bar{Q}$ is exceeded continuously. The *threshold* is set to 0.9. High flow duration is measured in d or h according to the modeled time scale.

95 % flow quantile ($Q95$) describes the flow magnitude which occurs in 95 % of the time. This is the streamflow value in m^3/s (or mm/h or mm/d) of the FDC at 5 % on the x-axis, as it is exceeded in 5 % of the time.

Mean half flow date (HFD) describes the day of the year at which half of the cumulative yearly discharge is reached. The resulting day is the average of all considered hydrologic years starting on October, 1st.

Runoff ratio ($\bar{Q}/\bar{P}$) describes the unit-less ratio of mean discharge to mean precipitation over the considered time period.

3.3 SQ1: US MTS-LSTM MODEL

The first research sub-question is dedicated to the comparison of the model performance when training on data of the global dataset ERA5 compared to a model trained on the local dataset NLDAS-2. Figure 3.6 shows a scheme of the applied methodology. The number (1) shows the location of the 516 training catchments within the US which are clustered into groups showing similar characteristics (2), see Section 3.2.1. The characteristics serve as static model input (3), once derived from the Camels US dataset, once from HydroMT. The meteorological data from NLDAS-2 and ERA5 serve as dynamic model input (4). All models are trained with (calibrated against) observed streamflow from USGS stations (5). To assess the effect on model performance, all four combinations of dynamic and static inputs are used to train a MTS-LSTM, resulting in the four models A1, A2, B1, and B2 (6). These models receive spatially averaged data as input. The simulated streamflow time series from the benchmark model *wflow_sbm* (7) are obtained with gridded ERA5 and HydroMT input data. The performance of all models is analysed using metrics and hydrologic signatures defined in Section 3.2 (8) and the 516 catchments are then clustered again based on these metrics (9). Thus, the overlap between previously obtained clusters based on catchment characteristics and the clusters based on model performance can be assessed (10).

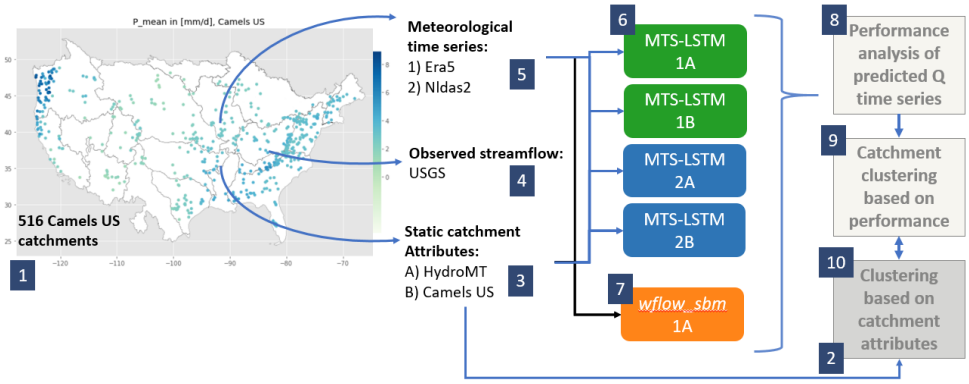


Figure 3.6: Overview of methodology for sub-question 1.

3.3.1 Training - Validation - Testing Ratio

The available time range covered by the forcing parameter datasets NLDAS-2 and ERA5 and by the observed streamflow for the US catchments reaches from 1979 until 2018. This data is split into three sets, one for training, one for validation during the training process and one for testing of the model, see Figure 3.7. These ratios are chosen to make results more comparable with those from (Gauch et al., 2021) who have used the same ratios with NLDAS-2 forcing data and Camels US attributes.

Figure 3.7: Training - Validation - Testing ratio for data from US catchments

3.3.2 Hyperparameter Tuning

The tuning of hyperparameters is a method to improve model performance through adjustments of parameters that define the learning process of the model. Those hyperparameters can be selected manually by the user or optimised through methods. Opposed to internal model states (the weight and bias values, cell states of LSTM networks) the hyperparameters do not change throughout the model training. As the hyperparameters of the MTS-LSTM have been optimised by Gauch et al. (2021), a less extensive tuning is performed for the hyperparameters and values shown in Table 3.5. MTS-LSTM networks with all possible combinations of hyperparameter values are trained and evaluated and the configuration with the best results regarding the mean NSE of all US catchments is chosen as the final hyperparameter configuration. Due to time limitations, not all hyperparameters are tuned again for this study and thus some parameters are taken over from Gauch et al. (2021), shown in Table 3.6.

The entire configuration set-up for model training and evaluation is defined in a *.yml* file. An exemplary configuration file for one of the models of this research can be found in Appendix (A).

Hyperparameter	Values			MTS-LSTM		
				NLDAS-2	ERA5	Gauch
hidden size	32	64	128	128	**64**	64
dropout	0.2	0.4	0.6	0.2	**0.2**	0.4
epochs	30	50		30	**30**	30
batch size	256	2048	6000	2048	**2048**	256

Table 3.5: Values for hyperparameter tuning of MTS-LSTM with NLDAS-2 forcing and Camels US statics, ERA5 forcing and HydroMT statics and values resulting from tuning by Gauch et al. (2021). Values in bold are used for all model training experiments with US data.

Hyperparameter	Values
regularization	yes
sequence length	365 days, 336 hours
learning rate*	1: $5e^{-4}$, 10: $1e^{-4}$, 25: $5e^{-5}$
loss	NSE
optimiser	Adam

Table 3.6: Values for hyperparameters taken over from tuning by Gauch et al. (2021). *learning rate for epochs 1 to 9, 10 to 24 and 25 to final epoch

3.3.3 Training Experiments

To allow for an assessment of the effect on model performance when using a lower resolution dataset, the MTS-LSTM has to be trained two times, ones with each forcing dataset. For this purpose the combined data of all US catchments is used, no data from the European test catchments is included. To also assess the quality of the HydroMT dataset for catchment attributes, two experiments with NLDAS-2 and ERA5 as forcing are performed.

Experiment	Dynamic forcing dataset	Static attributes
1A	ERA5	HydroMT
1B	ERA5	Camels US
2A	NLDAS-2	HydroMT
2B	NLDAS-2	Camels US

Table 3.7: MTS-LSTM training for dataset comparison.

3.3.4 Benchmark Predictions of *wflow_sbm*

To benchmark the MTS-LSTM results against a process-based model, already existing results from the distributed model *wflow_sbm* are considered for the US catchments and the Meuse catchments. The model is uncalibrated but parameters are estimated through pedotransfer functions for each grid-cell (see Section 2.1.2) (Imhoff et al., 2020).
For the US, the model has been applied over 484 of the 516 catchments that are basis for the MTS-LSTM experiments. The grid cell size is 1 *km* and the *wflow_sbm* simulations reach from 2009 to 2019. The model input is derived from HydroMT datasets like *soilgrids*, *vito* and *MODIS* and the catchment area is determined with Merit Hydro, similar to the procedure for the static attributes for the MTS-LSTM model (see Table 3.2). The PET is determined with the method of Bruin et al. (2016). The precipitation forcing data comes from the *MSWEP* dataset (Beck et al., 2017), PET and temperature from the ERA5 dataset. The results are on daily time scale.
For the Meuse *wflow_sbm*, the forcing data comes from the ERA5a dataset and PET is determined with the Makkink formula. The grid cell size is 600 *m* x 925*m* and

used datasets are among others CORINE land cover[7], soilgrids and MODIS for Leaf Area Index (LAI). Streamflow predictions and observations for the years from 2013 to 2018 are available.

3.4 SQ2: TESTING US MTS-LSTM FOR PUB

The second research sub-question targets the applicability of the trained MTS-LSTM outside the US with the example of the Meuse basin. Figure 3.8 illustrates the methodology to answer this sub-question. The number of catchments in the Meuse basin (1) is limited to five to allow for a more in-depth analysis of the tested models. Each of these catchments is assigned to one of the clusters from SQ1 according to the present characteristics, such that a statement regarding the expected model performance can be made (2). The trained MTS-LSTM model 1A from SQ1 is tested on these catchments with input from ERA5 and HydroMT (3). Additionally, a new MTS-LSTM is trained explicitly regional on the data from the Meuse catchments (4). The resulting streamflow time series are compared against observed streamflow and against streamflow predictions from *wflow_sbm* simulations obtained with ERA5 and HydroMT data (5). The performance analysis includes assessment of metrics and hydrologic signatures (6) to answer the question of whether the US trained MTS-LSTM is suitable as model for PUB and if a regional MTS-LSTM can compete with *wflow_sbm*.

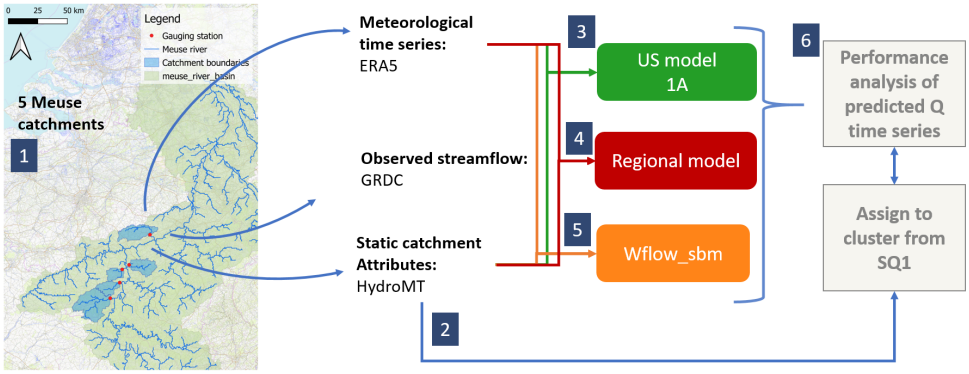


Figure 3.8: Overview of methodology for sub-question 2.

The US model trained with ERA5 forcing and HydroMT statics is tested on five test catchments in the Meuse basin to simulate an ungauged situation. Therefore, nothing is changed on the model, it receives the same eleven meteorological parameters derived from ERA5 that have been used for training and the same static attributes derived from HydroMT as input. Each test catchment is assigned to a cluster from the previous catchment clustering of the US catchments to find out to which US region the Meuse catchments share most similarities in their characteristics.

The model results for the Meuse catchments are assessed regarding high flow metrics and signatures and compared to those achieved with the $wflow_{sbm}$ model. Additionally, the results are compared to a regional MTS-LSTM trained on ERA5 forcing data and HydroMT statics. This regional model is once trained on data from all five test catchments and then tested on a test period of each catchment, see Table 3.8. An individual hyperparameter tuning based on the validation results resulted in the model configuration shown in Table 3.9. Subsequently, five models of the same configuration are trained on four catchments, each time excluding one of the five Meuse catchments which then serves as independent test catchment to simulate an

7 https://land.copernicus.eu/pan-european/corine-land-cover

ungauged situation, see Table 3.10. The training and validation period for these four models is the same as for the regional model. However, the testing is performed on the whole available time series from 2005 to 2017 of the excluded catchments. With this approach, the model is tested for streamflow predictions in the same period as the training data but also over five years lying ahead of the training period.

Period	Start	End
Training	10.10.2005	30.09.2012
Validation	01.10.2012	30.09.2013
Testing	01.10.2013	30.09.2017

Table 3.8: MTS-LSTM training, validation, testing periods for the Regional Meuse model.

Hyperparameter	Values		
hidden size	32	64	**128**
dropout	**0.2**	0.4	
epochs	30	**50**	
batch size	256	**512**	
learning rate*	1: $5e^{-4}$, 2: $1e^{-4}$, 5: $5e^{-5}$	1: $5e^{-4}$, 10: $1e^{-4}$, 20: $5e^{-5}$	

Table 3.9: Values for hyperparameter tuning of MTS-LSTM with ERA5 forcing and HydroMT statics for the Regional Meuse model. Bold values are used for subsequent model training.

Meuse	1		2		3		4		5	
PUB	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
6	x		x		x		x			x
13	x		x		x			x	x	
701	x		x			x	x		x	
702	x			x	x		x		x	
703		x	x		x		x		x	

Table 3.10: Method to test the Regional Meuse MTS-LSTM for PUB. Each of the five models is trained with data from four catchments and then tested on the fifth catchment.

3.5 SQ3: DIFFERENT LOSS FUNCTION

The third research sub-question tries to answer if replacing the NSE loss function leads to an improved high-flow representation. A loss function is chosen that is more sensitive to magnitude differences between observed and predicted peak height. Streamflow values are not normally distributed but can be better approximated through e.g. a Gumbel distribution due to extremely high values occurring less often than the baseflow. This has the effect that a NSE loss function leads to a better fit for the baseflow and neglection of peak height in the case that peaks are stronger related to the baseflow. Only, in catchments with a naturally high baseflow and less extreme peaks, the NSE loss function can result in a better fit regarding both baseflow and peaks. Therefore, a different loss function is tested to yield better performance regarding peak height, also in catchments with low baseflow and high peaks.

Figure 3.9 shows a scheme of the applied methodology. The chosen loss function is the Mean Quadrupled Error (M4SE) and is implemented in the NeuralHydrology (2020) code base. The experiments from SQ1 with ERA5 and HydroMT (1A) and with NLDAS-2 and Camels US (2B) are repeated. Each time a MTS-LSTM is trained on 516 catchments with the same training, validation and testing periods as for SQ1 and same hyperparameter settings. Equally, a regional MTS-LSTM for the Meuse

basin is trained as in SQ2 but with the new loss function. The model performance is analysed by means of the metrics and hydrologic signatures described in Section 3.2.2 and 3.2.3 to see if improvements regarding high flow can be observed. Further, differences in the clusters built for SQ1 are investigated.

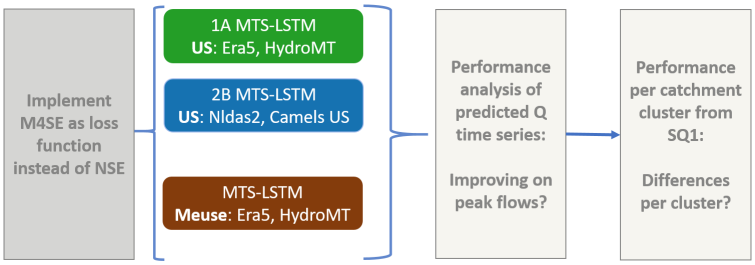


Figure 3.9: Overview of methodology for sub-question 3.

To answer the first two research questions, MTS-LSTM neural networks trained with the NSE as loss function are used. However, the NSE is not particularly sensitive regarding the magnitude of high flows. Therefore, a different loss function is chosen to see if model performance regarding metrics and signatures of high flow improve. The M4SE replaces the NSE as loss function, see Equation 3.13. As the difference between Q_s and Q_o is taken to the power of four instead of two, large differences in magnitude are translated into larger loss values with the M4SE compared to the MSE or NSE. Thus, a larger loss results in more internal weight and bias adjustments and a training with respect to the magnitude of higher flows instead of the overall overlap between simulated and observed streamflow.

$$M4SE = \frac{1}{N} * \sum_{i=1}^{N} (Q_s - Q_o)^4 \tag{3.13}$$

The new loss is implemented by extending the *BaseLoss* class in the neuralhydrology repository. The new loss function is tested by repeating experiments 1A and 2B from Table 3.7, ERA5 forcing with HydroMT statics and NLDAS-2 forcing with Camels US statics. Further, the new loss function is tested on a regional model trained on Meuse data with ERA5 forcing and HydroMT statics.

3.6 OVERVIEW OF MODELS

SQ	Model name	Model type	Training	Testing	Comments
1	1A	MTS-LSTM	US	US	ERA5, HydroMT
	1B	MTS-LSTM	US	US	ERA5, Camels US
	2A	MTS-LSTM	US	US	NLDAS-2, HydroMT
	2B	MTS-LSTM	US	US	NLDAS-2, Camels US
	wflow US	*wflow_sbm*	US	US	ERA5, HydroMT
2	1A PUB	MTS-LSTM	-	Meuse	Model 1A from SQ1
	US_no	MTS-LSTM	-	Meuse	no static input
	US_less	MTS-LSTM	-	Meuse	less static input
	Regional Meuse	MTS-LSTM	Meuse	Meuse	
	PUB Meuse	MTS-LSTM	Meuse, 4	Meuse, 1	4 different models
	wflow Meuse	*wflow_sbm*	Meuse	Meuse	
3	M4SE US	MTS-LSTM	US	US	Like 1A and 2B (SQ1)
	M4SE Meuse	MTS-LSTM	Meuse	Meuse	Like Regional Meuse (SQ2)

Table 3.11: Overview of models used for each research sub-question (SQ). The same colors are used in plots and tables.

4 RESULTS AND DISCUSSIONS

In this chapter, the results are presented and discussed in five sections: 4.1 Analysis of the datasets, 4.2 Clustering based on catchment characteristics, 4.3 Performance analysis of US MTS-LSTM (SQ1), 4.4 Performance analysis in Meuse basin (SQ2) and 4.5 Analysis of using a different loss function (SQ3). The first two sections each end with a summary of the relevance for the overall research focus. The three last sections each deal with one of the research sub-questions, first presenting the findings and then the discussion in the context of the research aim. The final section of this chapter points out relevant limitations of the research method.

4.1 DATASET ANALYSIS

4.1.1 Dynamic Input: Comparison NLDAS-2 and ERA5 Forcing

To compare the two forcing datasets NLDAS-2 and ERA5, the year long time series are re-sampled to create a monthly and yearly climatology per HRU. For this purpose the whole available time series from 01.10.1981 until 30.09.2018 are used.
The most significant differences regarding **monthly precipitation** occur in the HRUs 17 and 18, see Figure 4.1. Here, ERA5 underestimates precipitation compared to NLDAS-2 mean monthly values during wet winter months. However, the convective precipitation of ERA5 is up to two times as high as the NLDAS-2 convective precipitation. Reversely, in a region with dry and small catchments (e.g. HRU 13) ERA5 precipitation is overestimated compared to NLDAS-2 precipitation during wet months in summer. These two regions are located in the Northwest US in mountainous areas west of the Rocky Mountains.
Regarding **temperature**, the maximum difference between ERA5 and NLDAS-2 mean monthly values occurs in HRU 9 and 16 during warm summer months, see Figure 4.1. In HRU 9 ERA5 the mean monthly temperature is lower, in HRU 16 higher (up to 3°C). However, of HRU 9 only two large catchments are considered and of HRU 16 only five. For all other HRUs, the mean monthly temperature of ERA5 is comparable to the one of NLDAS-2.
Yearly precipitation climatologies (see Appendix D.2) show that the ERA5 long term mean precipitation deviates in a range from 0 to 100 mm/yr from the NLDAS-2 long term mean. Only for HRU 17 the ERA5 mean is around 400 mm/yr lower, resulting from larger discrepancies especially from 1981 until 2010. Later, until 2018 the yearly precipitation mean approaches the NLDAS-2 mean. HRU 17 is the region with the highest precipitation compared to all other HRUs with yearly means of >2500 mm/a (with NLDAS-2, >2100 mm/yr with ERA5).
Additionally, several iterations of a visual inspection showed satisfying overlap between ERA5 and NLDAS-2 time series. During each iteration, 20 catchments out of the 516 US catchments were selected and the ERA5 time series of each forcing parameter was plotted together with the NLDAS-2 time series. The correlation coefficient for each forcing parameter can be found in Table 3.3.

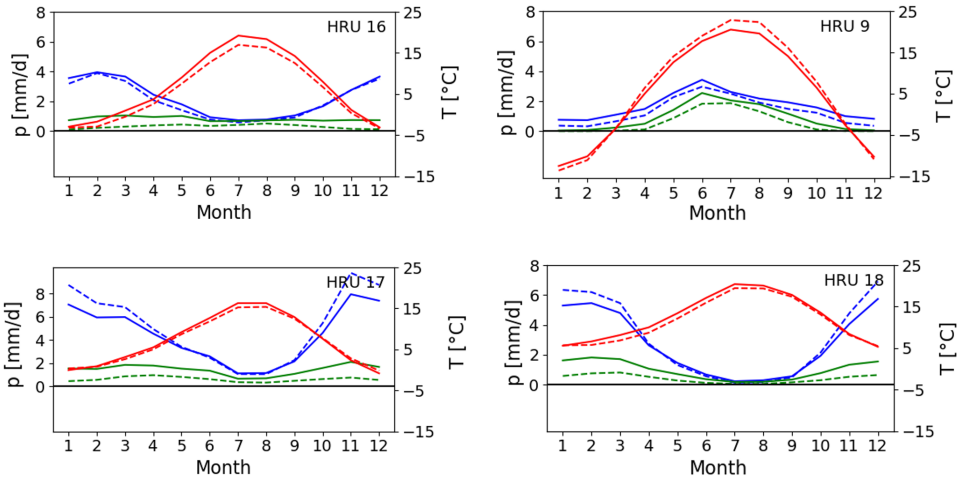


Figure 4.1: Monthly climatology for HRU 16, 9, 17 and 18. Continuous lines based on ERA5 forcing, dashed lines based on NLDAS-2 forcing. Blue: precipitation in mm/h, green: convective precipitation in mm/h, red: temperature in $°C$.

Summary and relevance of comparison:
The two meteorological datasets are compared regarding the eleven chosen forcing parameters. Significant differences were mainly found regarding the total precipitation and convective precipitation. In HRU 3 and 8, located in the South-East of the US precipitation of NLDAS-2 is slightly lower than ERA5 precipitation (monthly mean, and yearly sum). This accounts as well for HRU 13 located in the South at the border to Mexico. Here NLDAS-2 shows lower mean precipitation in the wet summer months. In mountainous and coastal regions in the Western US the monthly and yearly ERA5 precipitation is lower, especially in wet winter months. For those regions, the trend of the yearly precipitation sums approach each other for the more recent years. As well, the absolute difference seems larger than for other HRUs, however, those are the regions with the highest yearly precipitation and relatively seen the difference is not higher than for other HRUs.
The same pattern of ERA5 over- or underestimating monthly precipitation compared to the NLDAS-2 product is depicted in the convective part of the precipitation, while the relative difference between ERA5 and NLDAS-2 convective precipitation is larger than the difference in total precipitation. As convective precipitation is caused through the lifting of warm air, a difference in the monthly temperature between ERA5a and NLDAS-2 could be expected. However, differences of up to $5°C$ can only be observed for the two HRUs with two and five catchments. Thus, a deviation between the two datasets for one catchment is more relevant in these HRUs than in those with climatologies build from data of many catchments. For all other regions, ERA5 and NLDAS-2 do not differ significantly in temperature.
The comparison reveals that the main difference between the two datasets is not in the mean or cumulative precipitation but in capturing the absolute daily precipitation height. ERA5 smooths out precipitation height due to the coarser resolution. Due to the majority of catchments having an area smaller than a grid cell of the ERA5 raster ($31x31km^2$) and the NLDAS-2 precipitation being mainly based on observations, individual rainfall events are captured less precisely with ERA5. Therefore, it is expected that the lower precipitation precision affects the streamflow predictions negatively as this parameter is one of the most important forcing parameters.

4.1.2 Static Input: HydroMT Catchment Attributes

The 21 catchment attributes derived from global datasets with HydroMT are compared to the same attributes from the Camels US dataset to assess similarities and differences. For each attribute, a histogram shows its value distribution and the overlap between the two datasets (Figure 4.2). Those attributes which differ significantly in distribution and value range between the two datasets are shown again in separate histograms in Figure 4.3.
The high precipitation duration is lower (below one day) for the HydroMT than for Camels US (one to two days) and the high precipitation frequency has a narrower range (15 to 25 days) compared to Camels US (5 to 30 days). This results from ERA5 not capturing rainfall peaks as well as a dataset like NLDAS-2 which is based on ground observations. As ERA5 smooths out high and low extreme values, also the low precipitation duration is shorter for HydroMT (0 to 2.5 days) compared to Camels US (2 to 17 days). This results from the low precipitation being defined as values below 20% of the mean precipitation. With a lower mean, the baseflow threshold drops and thus the duration of continuous low precipitation events can shorten while the frequency increases.
The average value of each attribute per HRU and how those mean values differ across the HRUs is shown in Appendix B.3. An example is shown in Figure 4.4, revealing that the mean precipitation per HRU is up to $1 mm/d$ higher in the Camels US data compared to HydroMT mean precipitation for 13 of 18 HRUs. In the remaining five HRUs, the mean precipitation is maximal $0.3 mm/d$ higher than Camels US mean precipitation. Despite these differences, the differences between the HRUs are of the same range in both datasets,i.e. HRU has the highest mean precipitation and HRU 15 the lowest.
For the catchment slope, the mean value per HRU differs more significantly between the two datasets, as shown in the right plot of Figure 4.4. The slope derived from Merit DEM with HydroMT is up to 300 % higher than the Camels US slope. However, again the differences between the HRUs are represented similarly in both datasets. HRUs 6 and 13 to 18 have the steepest terrain while 4, 7 and 9 are flatter regions.

Summary and relevance of comparison:
The same pattern as observed in the comparison of the forcing precipitating time series can be seen in the mean precipitation attribute, as the ERA5 derived mean is lower for the mountainous regions in the Western US, but higher for e.g. HRU 9 with only two catchments. The lower recognition of extreme precipitation values from ERA5 becomes visible in the histograms of the attributes describing high and low precipitation frequency and duration.
Two other important attributes describing flow behavior and differing between the Camels US and the HydroMT dataset are the saturated hydraulic conductivity and the volumetric porosity of the soil. The HydroMT attributes show a broader distribution which could lead to a better differentiation of catchments and positively affect streamflow modelling with the MTS-LSTM. This could be counteracted by a narrower distribution of maximum water content values for the HydroMT dataset. However, the water content attribute is ranked lowest regarding the influence on LSTM streamflow predictions.
Further, differences in the mean slope per catchment between Camels US and HydroMT hint towards quality differences between different DEM datasets and approaches. The same accounts for the soil depth. However, as both slope and soil depth show distributions in the histograms of the same shape, little effect on the LSTM model performance is expected. This expectation is amplified through the same relative distribution of mean slope per HRU (see Figure B.3) for both attribute datasets and the low ranking of the soil depth attribute (see Table 3.2). The difference in mean PET originates from the chosen method to determine PET. Although

the height of this attribute differs between the two datasets, the distribution shown in the histograms is still comparable, as well as the mean value per HRU.

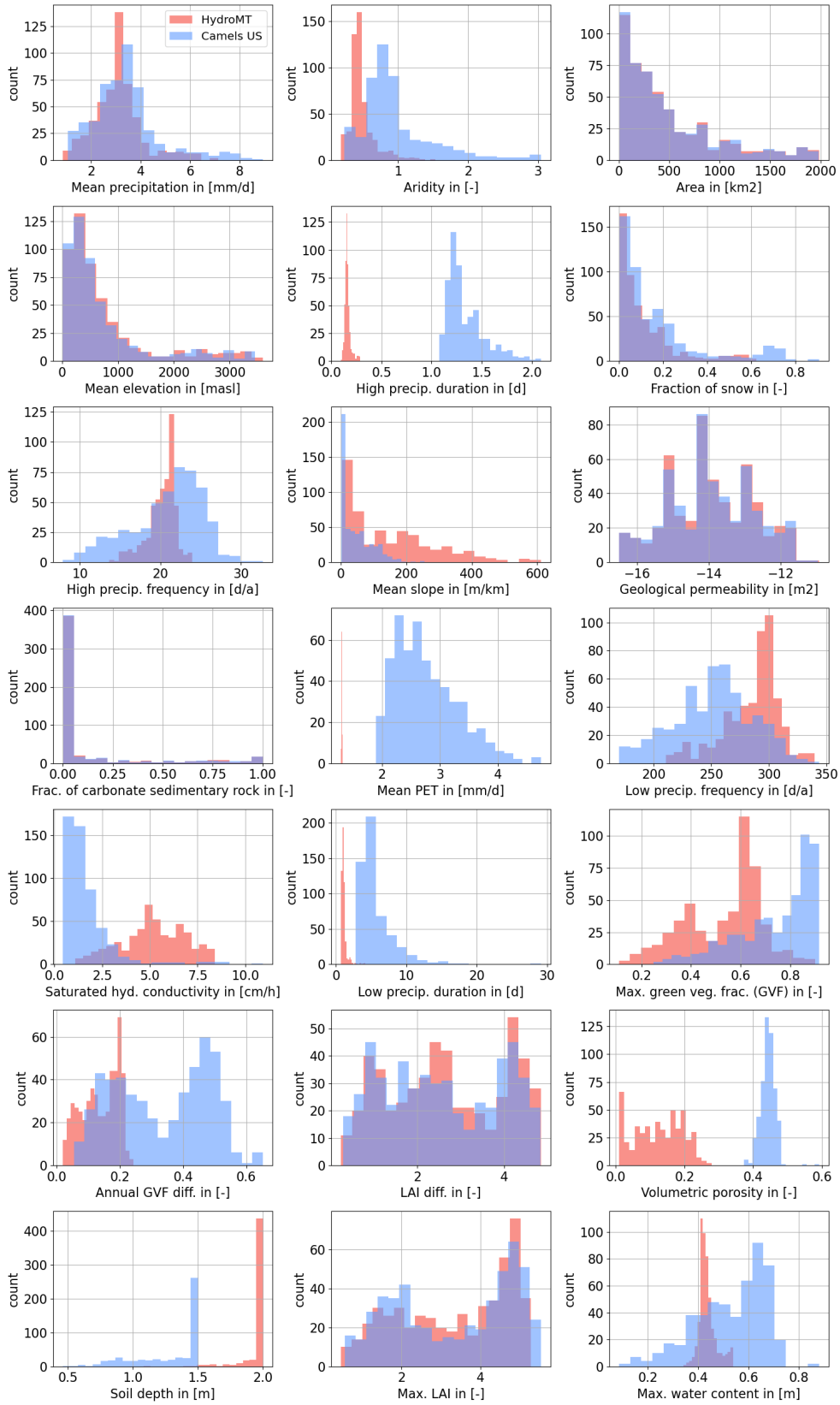


Figure 4.2: Histogram per catchment attribute based on data from US catchments.

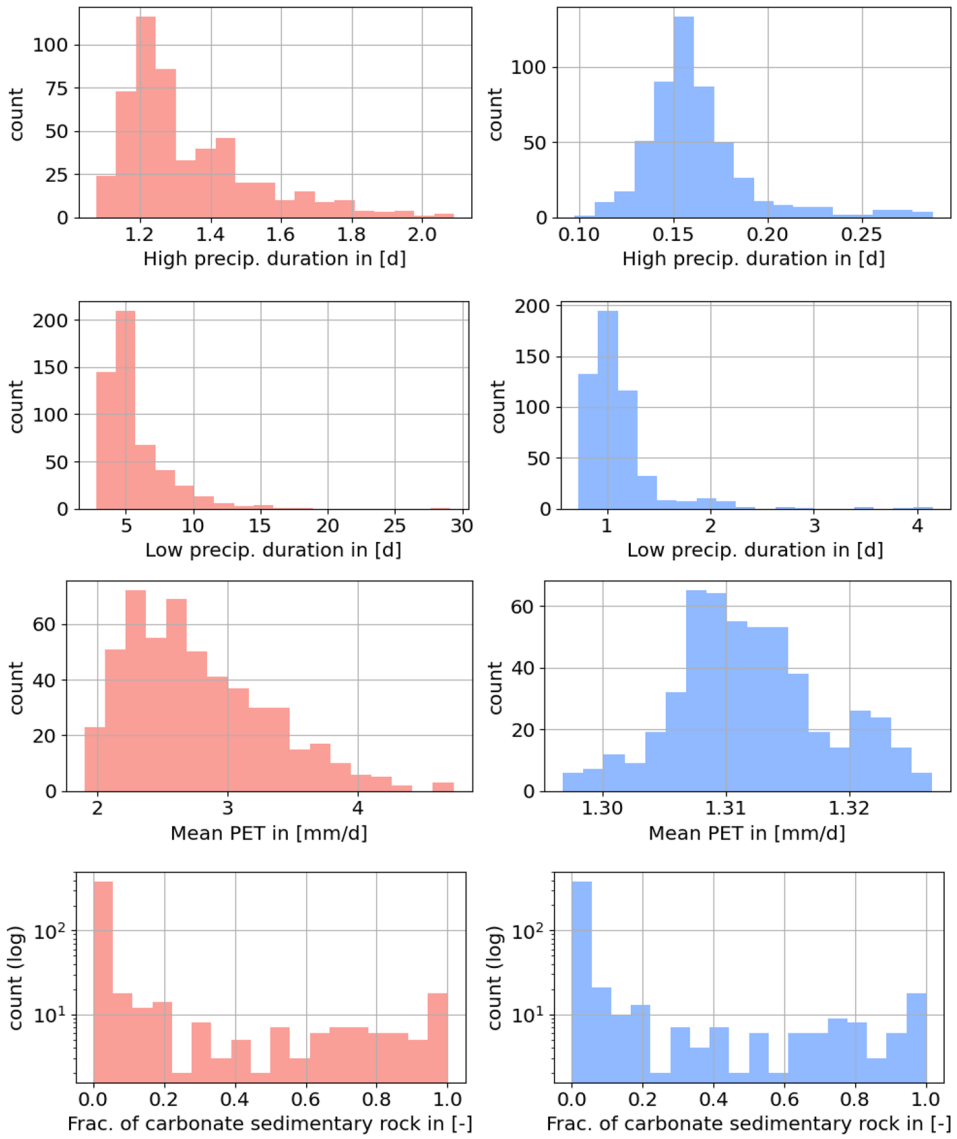


Figure 4.3: Histogram per catchment attribute based on data from 516 Camels US catchments for attributes that are significantly different between HydroMT (red) and Camels US (blue). *Fraction of carbonate sedimentary rock* with log-scale on y-axis.

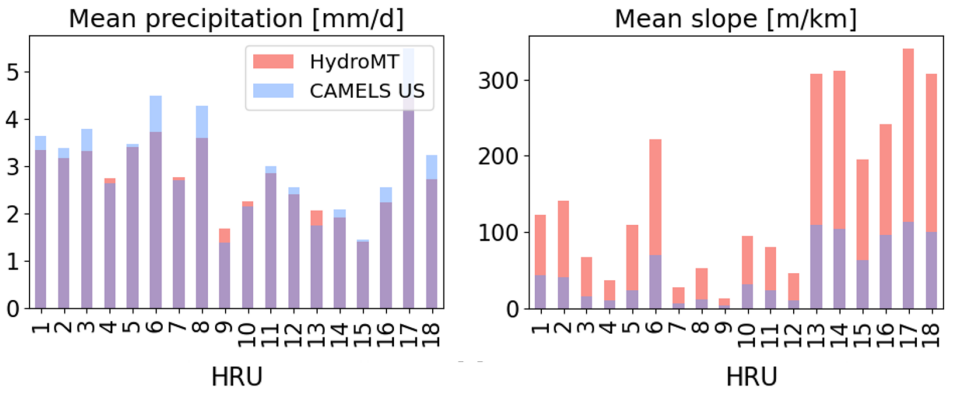


Figure 4.4: Mean attribute value per HRU, example plots for mean precipitation and slope. All other plots are in Appendix B.3

4.2 CLUSTERING BASED ON CATCHMENT CHARACTERISTICS

Clustering the US catchments results in seven (eight) clusters based on Camels US (HydroMT) attributes as shown in the maps in Figure 4.5. The additional cluster that results from the HydroMT attributes combines the most south catchments of cluster 6 and most north catchments of cluster 7 (pink and turquoise Camels US cluster, left map) into a new cluster located at the West Coast (black HydroMT cluster, right map). Catchments falling under cluster 5 with the Camels US attributes (left map), appear in cluster 2 or 3 with the HydroMT attributes (right map). Despite differently assigned groups of catchments and individual catchments, the general picture created through the HydroMT k-means clustering overlaps with the Camels US based clustering. Table 4.1 shows how many catchments belong to each cluster. Figure 4.6 shows where the catchments per cluster plot in the Budyko Framework, an individual Budyko Framework per cluster is shown in Appendix C.1.

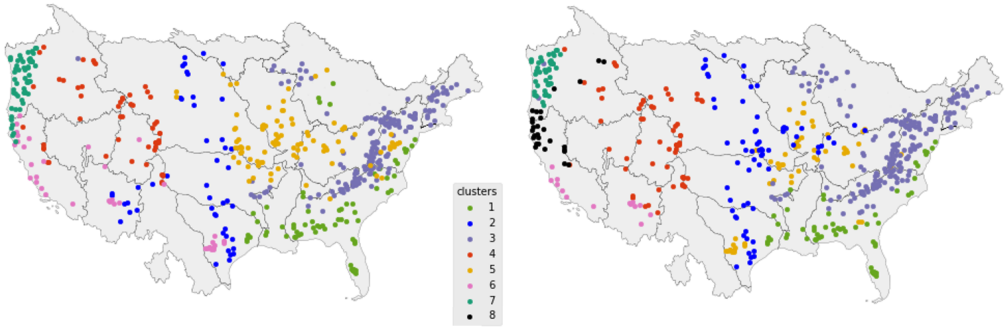


Figure 4.5: Clustering by Camels US (left) in k=7 clusters and HydroMT (right) catchment attributes in k=8 clusters.

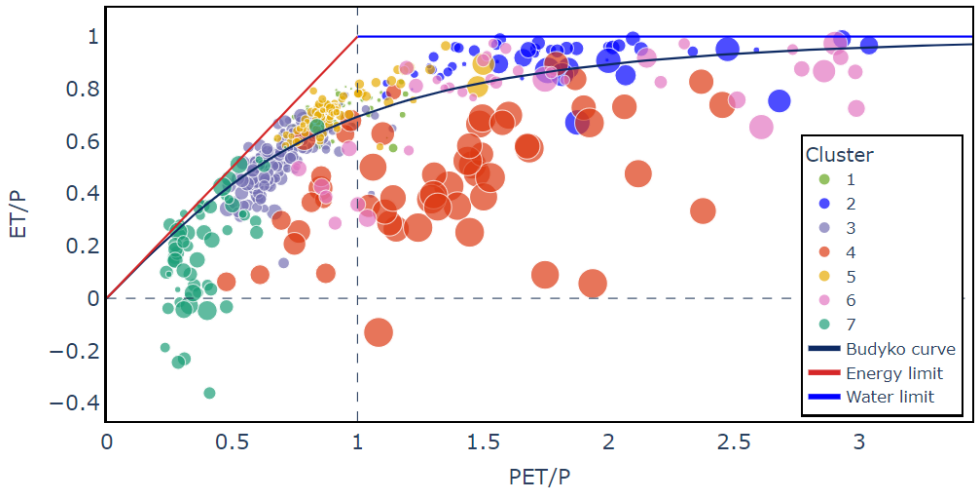


Figure 4.6: Clusters based on Camels US attributes plotted in Budyko Framework. Greater marker size relates to higher mean elevation.

Cluster	Camels US	HydroMT
1	67	55
2	46	65
3	164	192
4	54	61
5	89	54
6	38	14
7	58	47
8	-	28

Table 4.1: Number of catchments per cluster.

Cluster description

Cluster 1: The catchments of this cluster are located in the lowest elevation and have the flattest slope. Here, almost no snow occurs as the climate is subtropical and humid. The region has a forest cover of 50% and evaporation and transpiration are high and energy limited, as can be seen in the Budyko Framework in Figure 4.6. The soil conductivity is high compared to other clusters.

Cluster 2: The catchments of this cluster have the largest mean area and the driest climate conditions with high aridity. The cluster extends like a channel from north to south central US. In this region in the mid-latitude desert and semiarid steppe, the vegetation cover is low. Not only mean precipitation is low but also mean streamflow, the baseflow index, Q95 and Q5 are very low. The periods of low flow are longest as well as periods with no flow.

Cluster 3: These catchments of humid continental and subtropical climate are energy limited as they plot very to the left in the Budyko Framework ($\frac{PET}{P} < 1$. They are located in the Appalachian mountains (with 500 m mean elevation), have a high vegetation cover indicated by high forest fraction, high Green Vegetation Fraction (GVF) and high LAI. Vegetation cover changes with the seasons. A steep FDC indicates surface runoff dominated flow processes.

Cluster 4: As the catchments of this cluster are located in the Rocky Mountains in the North-West US, the mean elevation is highest and the mean slope steepest. This cluster has on average the smallest catchments with the highest daily snow fraction. The snow storage leads to a late HFD. The climate is dry with high aridity in more eastern catchments with semi-arid steppes. The western catchments have a more humid climate and count to the alpine highlands.

Cluster 5: These catchments lie in a region with a humid continental climate where summers are warm and rainfall peaks occur during summer. Vegetation cover is relatively high compared to other clusters. Catchments are rather large on average, with the flattest slope. The carbonate rock fraction, soil porosity and maximum water content of the soil are highest in this cluster, indicating hydrologic subsurface processes which can cause difficulties for streamflow modeling.

Cluster 6: This region is characterised by a low LAI and little changes in GVF throughout the year. The climate is Mediterranean with high aridity and evaporation processes are mainly water limited. Periods of low precipitation and low streamflow last longest of all clusters, no streamflow period as well. Snow occurrence is very little. A flat FDC slope indicates the presence of surface or groundwater storage which regulates runoff processes (see also map in Appendix E.1).

Cluster 7: Those catchments are located at the north marine West-coast, west of the Rocky Mountains. The elevation is relatively high with on average 760 m, daily snow fraction is high compared to other clusters while the mean precipitation is

highest in this cluster. These catchments show energy limited conditions ($\frac{PET}{P} < 1$ in the Budyko Framework), as the mean streamflow is very high as well as the runoff ratio. Rainfall peaks occur in winter, thus the half flow date occurs early in the year (average on day 145) for these catchments. A steep FDC indicates towards highly variable runoff response to rainfall and domination of surface runoff process over subsurface flow. These catchments show a high vegetation cover which does not differ extremely throughout the year. In the Budyko Framework it can be observed that some of the catchments plot below $0\frac{ET}{P}$. This means that $Q(+ET) > P$, thus the catchment runoff (in sum with actual evaporation) exceeds the incoming precipitation. A reason explaining this can be storage release through melting glaciers and long-term snow storage.

Cluster 8: The above cluster descriptions are based mostly on the clustering with Camels US attributes but relate as well with little deviations to the HydroMT clusters. This additional HydroMT cluster combines catchments from the cluster 6 and seven of the Camels US clusters. Elevation and slope are similar to cluster 6, while the vegetation cover is higher. Snow occurrence and mean precipitation are higher and aridity less intense than in cluster 6, but not as extreme as in cluster 7.

Summary and relevance of clustering results:
The resulting clusters and their spatial distribution overlap well with spatial grouping, although no information on latitude and longitude is included in the clustering approach. This underlines that geographical regions can be identified and separated by their climate, soil, vegetation and hydrology characteristics. The clustering is not strongly influenced through individual attributes but by combinations of attributes. Elevation and slope differ significantly between the clusters and attributes related to aridity and humidity play an important role as well, which also accounts for vegetation attributes. The attribute histograms already show an overall similar distribution with presence of some differences between individual attributes (see 4.1.2). This is also pictured in the overall congruence between Camels US and HydroMT clusters with little deviations for individual catchments.

4.3 SQ1: US MTS-LSTM MODEL

First, Section 4.3.1 presents the results from the US model experiments. Then, in Section 4.3.2 the findings are discussed to answer the first research sub-question:

How is model performance affected when using a globally available but lower resolution dataset (ERA5) as model forcing compared to a regional high resolution dataset (NLDAS-2)?

4.3.1 US Model Experiments

The individually achieved NSE values per catchment are shown in the US maps in Figure 4.7 and in the Budyko plot in 4.8. The Cumulative Density Function (CDF) of NSE values is plotted for all four models 1A, 2A, 1B and 2B, trained on ERA5 and NLDAS-2 with Camels US or HydroMT statics, see Figure 4.9. In the plots, the grey horizontal line marks the accumulation of 50% of the catchments, the vertical red line marks an NSE of 0.7 as a lower boundary for general good model performance. The plots in Figure 4.9 show that the performance is not significantly affected when replacing the static attributes from Camels US with those from HydroMT. The light and dark blue lines overlap (2A and 2B, NLDAS-2 forcing) as well as the light and dark green lines (1A and 1B, ERA5 forcing). This accounts for daily (left plot) and hourly (right plot) results. Only for higher NSE values, the models with Hy-

droMT statics perform slightly better than those with Camels US statics, as the CDF is shifted minimally to the right (green compared to blue).

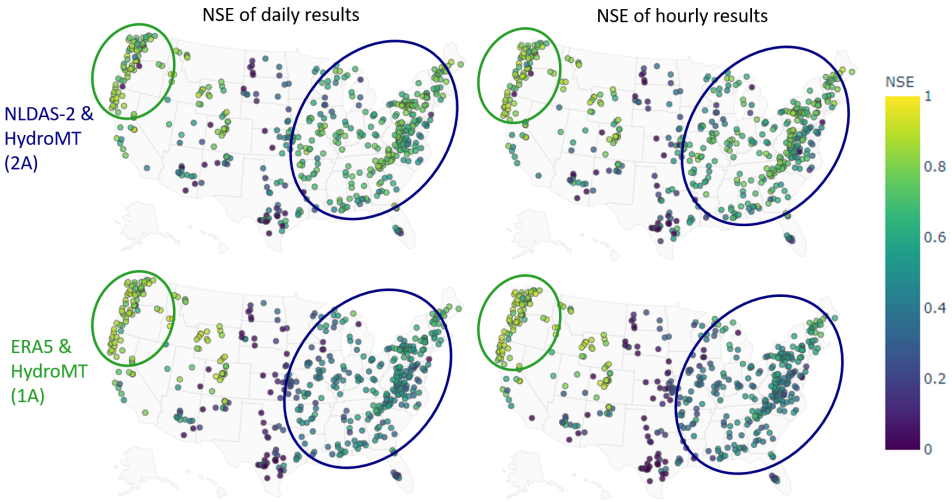


Figure 4.7: NSE per US catchment for model trained on NLDAS-2 and HydroMT (model 2A) and for model trained on ERA5 and HydroMT data (model 1A). NSE values below 0 are shown in same color as NSE=0. Green ellipses show regions where model 1A is better, blue ellipses show regions where model 2A is better.

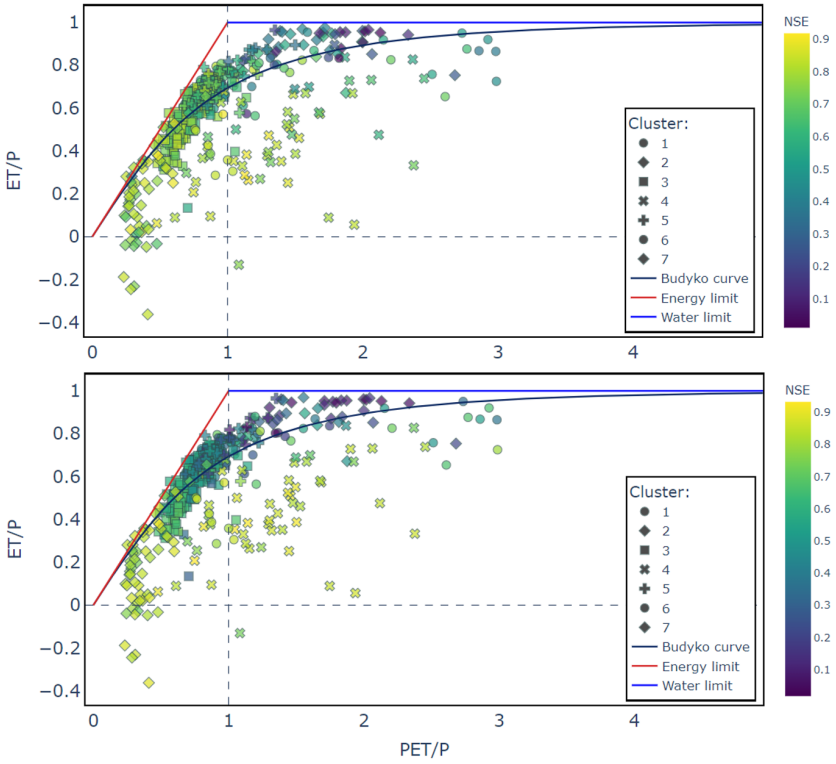


Figure 4.8: 516 US catchments in Budyko Framework, colored by NSE values. Upper plot: trained on NLDAS-2 data (model 2A). Lower plot: trained on ERA5 data (model 1A). Mean P, mean PET and mean Q from Camels US dataset (Addor et al., 2017). Budyko plots per cluster for model 1A can be found in Appendix C.1.

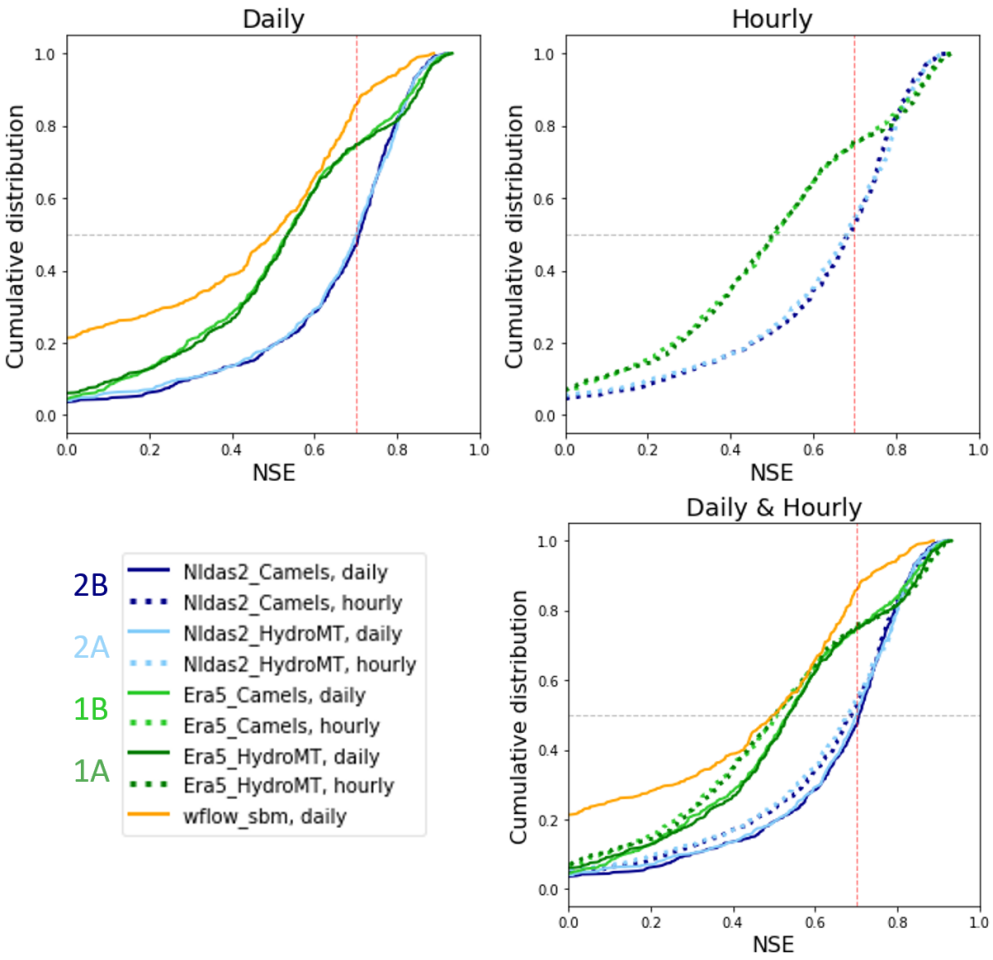


Figure 4.9: CDF of NSE of 516 US catchments based on test-period for all four combinations of forcing and statics as input for MTS-LSTM (models 1A, 1B, 2A, 2B). Continuous graphs for daily results, dashed graphs for hourly results. CDF for *wflow_sbm* results in orange.

However, the performance drops more significantly, when comparing ERA5 (models 1A, 1B) to NLDAS-2 forcing (models 2A, 2B): 50 % of the catchments have an NSE above 0.55 with ERA5 forcing compared to above 0.7 with NLDAS-2 forcing. Only for NSE values above 0.8 the model with ERA5 forcing is better (green lines shifted to the right of blue lines). The right plot shows that daily results (continuous lines) are better regarding the NSE compared to hourly results (dashed lines). This accounts especially for the 50 % of catchments with a lower NSE (below 0.7 with NLDAS-2 forcing and below 0.5 with ERA5 forcing).
The comparison with the *wflow_sbm* CDF reveals that all MTS-LSTM models outperforms the process-based model on daily and hourly results as the orange line appears to be most left in the plots. With 50 % the neural network models (2A, 2B) achieve a higher percentage of catchments with a good NSE compared to *wflow_sbm* with 15%. This confirms results of earlier research by Kratzert et al. (2019c) who showed that LSTM models are able to outperform several calibrated process-based hydrologic models in the US, among others the Sacramento Soil Moisture Accounting Model (SAC-SMA). The CDF plots give an indication to overall performance regarding the NSE, however do not exclude the possibility of the *wflow_sbm* or the ERA5 trained MTS-LSTM (1A) showing better results than the NLDAS-2 trained model (2A) for individual catchments.

The mean model performance is determined by averaging the metrics from all 516 US catchments over the test period or the models trained with HydroMT statics, as shown in Table 4.2. As already seen in the CDF plots, the performance of the ERA5 model is lower compared to the NLDAS-2 model regarding NSE which is also reflected in a lower KGE. Further, more catchments have a NSE < 0 with the ERA5 model while the number of catchments with a KGE < -0.41 remains similar to the NLDAS-2 model. Regarding the peak flow metrics, the NLDAS-2 model performs better, as well: Peak timing is more accurate, FHV is lower and the absolute and relative peak magnitude errors are lower compared to the ERA5 model.

Model	Freq	NSE	KGE	NSE<0	KGE<-0.41	FHV	Peak-Timing	ϵ_{abs}	ϵ_{rel}
1A ERA5	1D	0.54	0.60	31	**11**	-21	0.45	7.75	0.53
	1H	0.50	0.58	36	**15**	-21	4.89	0.28	0.53
2A NLDAS-2	1D	**0.70**	**0.71**	**20**	**11**	-15	**0.39**	**5.77**	**0.41**
	1H	**0.68**	**0.71**	**28**	16	**-12**	**4.38**	**0.24**	**0.46**
wflow_sbm	1D	0.46	0.52	136	85	**-1**	0.67	27.08	0.54

Model	Freq	Q_{high} freq. [d/a] or [h/a]	Q_{high} duration [d] or [h]	Q95 [mm/h]	HFD_{mean} [d]	$\bar{Q}/\bar{P}$ [-]
1A ERA5	1D	6	2.31	3.70	170	**0.39**
2A NLDAS-2	1D	**8**	**2.60**	3.77	**171**	0.37
wflow_sbm	1D	17	3.56	**4.57**	158	-
Obs. Q	1D	12	2.51	4.45	176	0.40
1A ERA5	1H	160	28.80	0.15	171	**0.38**
2A NLDAS-2	1H	**231**	**35.53**	**0.16**	**175**	0.37
Obs. Q	1H	280	34.91	0.18	176	0.41

Table 4.2: Median metrics and high flow signatures for US MTS-LSTM models 1A and 2A. Static attributes from HydroMT. Signatures based on observed streamflow shown in gray rows. As the precipitation input for *wflow_sbm* is MSWEP, no runoff ratio is determined with ERA5 precipitation. Best daily (1D) and hourly (1H) values shown in blue.

To assess for which catchments the US model performs best and worst, results are sorted into the groups with (1) a low peak timing error, low FHV and a low relative peak magnitude error and (2) higher peak timing error, high FHV and high relative peak magnitude error. An overview is given in Table 4.3. For the ERA5 trained model, more catchments fulfill the criteria for best performance as well as for lowest performance compared to the NLDAS-2 trained model. All catchments are performing well regarding the peak flow metrics show high NSE and KGE values. Likewise, low performance regarding the peak flow metrics correlates with NSE values below 0.5 and negative median KGE values. The simulated peak magnitude is on average more often underestimated (around 70% of the recognised peaks) than overestimated (around 30%), for daily and hourly predictions, for high and low performance, for the model trained with NLDAS-2 and ERA5 forcing.

Perf.	Peak-Timing	FHV	ϵ_{rel}	Model	Nr. of catchm. 1D	1H	both	NSE	median KGE
Best	<0.5d <3h	$<\lvert\pm 15\rvert$	<30%	1A ERA5 2A NLDAS-2	28 26	21 9	6 4	>0.8 >0.7	0.88 0.84
Worst	>1d >6h	$>\lvert\pm 30\rvert$	>50%	1A ERA5 2A NLDAS-2	13 10	29 15	7 4	<0.4 <0.5	<0 <0

Table 4.3: Best and lowest performance of US MTS-LSTM models 1A and 2A. Peak-Timing, FHV and ϵ_{rel} are conditions for best (worst) performance. Column *both* gives number of overlap between catchments with best (worst) performance on daily and hourly time scale. See also Figure 4.7 for location of catchments with good and poor performance.

4.3.2 Discussion of Results for SQ1

To answer the first research sub question, the effect on model performance of the global but lower resolution dataset ERA5 is compared to the use of NLDAS-2. The experimental modelling setup with the four different models, i.e. A1, A2, B1 and B2, allows to estimate the effect of the applied dataset which is used as the dynamic model input. Additionally, the effect of the choice of dataset for the static model input can be assessed and the quality of the composed HydroMT dataset of catchments attributes can be quantified. Finally, the clustering on model performance enables the derivation of conditions under which the MTS-LSTM can be expected to perform best, sufficiently well or poor, measured by the general metrics NSE and KGE as well as high flow metrics and flow signatures.

The use of the ERA5 forcing time series leads to a significant drop in NSE, visualised in the CDFs in Figure 4.9. As the predictions of more catchments show a negative NSE compared to the NLDAS-2 model and the average NSE and KGE are significantly lower with 0.54 and 0.60 respectively (see Table 4.2, it is concluded that ERA5 is not suitable as a forcing dataset for a model with the aim to model any catchment in the US. The model trained on the NLDAS-2 dataset (2A) scores on average higher for all high flow metrics i.e. the peak timing error is smaller and the magnitude of peaks is closer to the observed height as for ERA5 results (1A). This is confirmed with the average high flow signatures from the NLDAS-2 model which are closer to the signatures calculated on observed flow than those from the ERA5 results.
However, the CDF plots (Figure 4.9) reveal that for catchments with a high NSE larger than 0.8 general performance increases when using ERA5 forcing instead of NLDAS-2, regardless the choice of dataset for the static input. The catchment clustering based on the performance metrics uncovers that this increase can be observed for catchments of clusters 4, 7 and 8 (see Figure 4.7). This means, in the regions in the north-western US (cluster 4, 7 and 8) the ERA5 model (1A) outperforms the NLDAS-2 model (2B) on average. Here, the climatology discussed in Section 4.1.1 has shown that ERA5 monthly and yearly precipitation is lower compared to NLDAS-2. Thus, the latter could overestimate precipitation values in those regions, compared to the true unknown precipitation height, subsequently leading to less good streamflow predictions.

The distribution of NSE values over the US maps (see Figure 4.9 and Budyko plots in Figure 4.8) reveals that for cluster 2 (central US) the model performance is the worst of all clusters for either of the four models. Regions with an arid climate and little to no streamflow through the course of a year are subsequently those areas where the MTS-LSTM is not suited for rainfall-runoff modelling. Comparing the results to the *wflow_sbm* model shows, that also a process-based hydrologic model has difficulties in achieving acceptable streamflow predictions close to the observed values. This phenomena results inevitably from the missing signal in the observed

streamflow time series in periods where rivers are dried out entirely. From the results of these experiments, it can be concluded that the MTS-LSTM cannot learn which rainfall event of the few occurring events causes peaks in the streamflow based on the provided data. It can be hypothesised that a MTS-LSTM trained only for all catchments of cluster 2 together or for each of the catchments individually can achieve better performance, since the data from other climate zones would not be influencing the learning process of the network. However, as the main cause for the low NSE values is the magnitude of the simulated streamflow peaks, it is questionable if a regional or individual MTS-LSTM would be able to predict the peak height more accurately.
Additionally, for the central US long-term groundwater depletion has been observed and is caused by water withdrawals exceeding the refilling through natural processes (Richey et al., 2015). For catchments located in the widespread area where groundwater depletion occurs (see US map by (Konikov, 2013) in Appendix E.1), the water mass balance is likely to be disturbed. Such human influenced processes are difficult to respect within a hydrologic model of either type – process-based or data-driven – since the dynamics do not follow the physical laws present under near natural conditions but include a certain level of randomness.

That the model performance is lower in catchments of cluster 6 and the most south-western catchments of cluster 2, can be related to the flat slope of the FDCs for those catchments. This hints towards the presence of surface or groundwater storages regulating the runoff behaviour. Additionally, groundwater depletion occurs in the regions of these clusters (see Appendix E.1).

In the eastern half of the US, the NLDAS-2 model (2A) performs better than the ERA5 model (1A). This is especially true for catchment cluster 3 which has a humid continental subtropical climate with high but seasonally changing vegetation cover. Similar to the north-western US, these catchments are surface runoff dominated which appears to results in acceptably good model performance of the MTS-LSTM, given that the input forcing data is of good quality.

High performance with both models (1A, 2A) is only achieved in the north-western US for catchments of cluster 4, 7 and 8. Thus, good performance can be related to mountainous regions or alpine highlands with humid climate and occurrence of large vegetation cover, where runoff processes are surface dominated.

In conclusion, the ERA5 dataset does not lead to sufficiently good performance, so that it would be suitable for a MTS-LSTM (trained on US data) that is generally applicable within and outside of the US. The results imply that the model, if applied in catchments outside the US, will only perform acceptably in catchments similar to those in the north-western US, unless the the model is retrained (finetuned) on local data. Ungauged basins with characteristics similar to a cluster located in the eastern US (cluster 1 or 3) can potentially result in acceptable streamflow predictions, however with a higher uncertainty.

For the static model input, the catchment attributes which are derived from the datasets included in HydroMT prove to be similarly suitable compared to the Camels US attributes. In conclusion, an in depth research on the choice of relevant catchment attributes would be beneficial to clarify the influence of attributes (1) derived from the forcing time series, like mean precipitation, (2) attributes already respected through the derivation of other attributes, like the soil composition, and (3) attributes largely differing among different datasets. The latter is seen e.g. in the histogram of the maximum GVF and the annual difference in GVF, which show a significant different shape (Figure 4.2). GVF is derived from *MODIS* for the Camels US dataset and from the *vito* dataset for the HydroMT dataset.

4.4 SQ2: TESTING US MODEL FOR PUB

Firstly, in Section 4.4.1 the Meuse catchments are categorised according to the US catchment clustering. Then, the results from the testing of the US model in the Meuse are presented in Section 4.4.2, followed by the results for the Meuse catchments with a regional MTS-LSTM in Section4.4.3 and the testing of this regional model for PUB in Section 4.4.4. Finally, in Section 4.4.5 these results are discussed to answer the second research sub-question:

How does the trained MTS-LSTM model perform when applied in catchments outside the US, simulating ungauged catchments?

4.4.1 Meuse Catchment Classification

To estimate what model performance of the US trained MTS-LSTM model can be expected in the Meuse catchments (serving for the simulation of ungauged basins), they are assigned to a cluster based on their catchment characteristics. For the Meuse catchments catchment attributes are only available from the HydroMT dataset, therefore the clusters based on the HydroMT dataset are considered. Using the *kmeans.predict()* function on the Meuse attributes after fitting to the US HydroMT attributes, the Meuse catchments are all assigned to cluster 1 (see US map in Figure 4.5). The similarity of the individual attributes from the Meuse catchments to those of the catchments in cluster 1 is visualised with box-plots in Appendix D.4. A number of attributes are outside of the range of the values from the US catchments, namely mean precipitation, soil porosity, mean elevation, GVF, mean PET, aridity, low precipitation duration and frequency and soil depth.

Summary and relevance of catchment classification
All five test catchments from the Meuse basin are allocated to the same cluster, meaning they are, despite their differences, more similar to each other than to any other region in the US. This confirms the observation that the catchment clustering based on the selected characteristics leads to geographically cohere groups. The fact, that the Meuse catchments fall in cluster 1, hints towards topographically flat catchments with low elevation and little snow occurrence while being covered with forest up to around 50%. Looking at the before mentioned boxplots shows that elevation and slope are steeper for the Meuse catchments than for the average of the cluster 1 US catchments. The classification of the Meuse catchments into this cluster originates mainly from the catchment size, a relatively high LAI and a low daily snow fraction.
These results indicate that the conditions in the Meuse catchment are not entirely represented with the 516 US catchment, since the similarity of the Meuse catchments to the US catchments in cluster 1 is rather low. It can be hypothesised that other combinations of catchment characteristics are not sufficiently represented with the selected US catchments. Subsequently, a higher number of catchments with even more variability in catchment attributes and climate conditions could be used to train a MTS-LSTM model, to improve the PUB.

4.4.2 Performance of US Model in Meuse Basin

Testing the US MTS-LSTM trained on ERA5 and HydroMT data (model 1A) in the Meuse catchments resulted in the time series shown exemplary in Figure 4.10. The simulated time series covers a longer time-span, as the forcing is available from 1981 onward while the observed streamflow is only available from 2005 onward. The simulated streamflow shows unexpectedly negative values in a very small range of streamflow values compared to the observed streamflow.

This shifting and squeezing of the time series is assumed to be caused by the scaling of the input data. Therefore, the Meuse input data is analysed more in depth and compared to the distribution of forcing parameters from the US ERA5 data. The forcing parameters do not show any striking difference to the US data, which can be seen in histograms in Appendix F.1. Results for catchment 703 in the upper plot of Figure 4.10 show how the streamflow prediction react on outliers in the forcing time series. The peaks in the orange graphs are also visible in the forcing time series. After such a peak, the predicted streamflow drops lower compared to the rest of the time series. Furthermore, the Meuse static attributes are compared to the distribution of values per attribute for the US catchments. The attributes that fall out of the range from the US catchments are shown in Figure 4.11. All other Meuse attributes lay within the range of the US catchment attributes, as shown in the box-plots in Appendix D.5.

Re-training the US model

Based on the findings, a new MTS-LSTM is trained once without all static input (referred to as US_no in the following) and once with less static attributes (US_less) than the initial model 1A from SQ1, excluding mean PET, max. GVF and GVF difference. Apart from this, the training procedure is the same as for the models 1A and 2B from SQ1 on the 516 US catchments. The resulting time series plots are not shifted and squeezed anymore and lie in a similar value range like the observations. All plots are shown in Appendix F.6 and the resulting streamflow hydrograph for catchment 703 is shown in the lower plot of Figure 4.10. All metrics are given in Table F.1. Table 4.4 shows the NSE per catchment. The model without static attributes results in good daily (0.48) and hourly (0.53) performance for catchment six, while *wflow_sbm* is still better (0.75 and 0.78). The model with less static attributes gives good results for catchment 13, where agriculture is high and fissured aquifers are present, while *wflow_sbm* shows negative NSE values on both time scales. For all other catchments both US-trained models show negative NSE values.

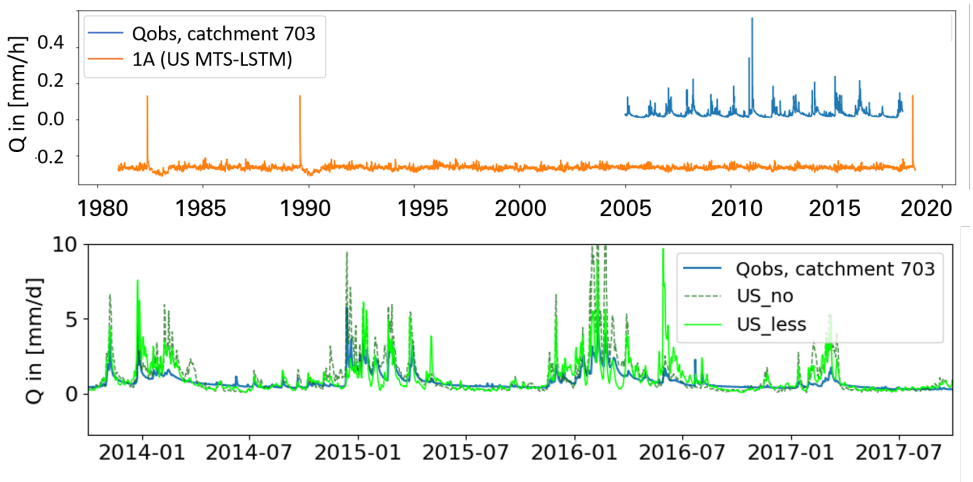


Figure 4.10: Upper plot: First result for test catchment 703 in Meuse basin with MTS-LSTM trained on ERA5 and HydroMT data of US catchments (model 1A from SQ1). Observed Q in blue, simulated Q in orange, in mm/h on the y-axis. Lower plot: Results with retrained US model without static input (US_no, dark green, dashed) and with less static attributes (US_less, lime).

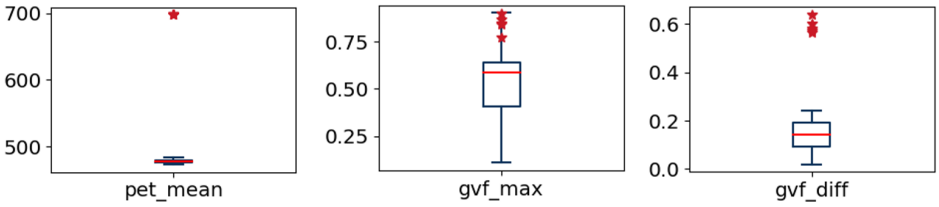


Figure 4.11: Attributes box-plots based on US catchments. Colored dots show values of Meuse catchments. These three attributes (mean PET [mm/yr], max. GVF [-] and GVF difference [-]) are subsequently excluded from the static attributes for model training and testing.

4.4.3 Regional Meuse MTS-LSTM

A regional MTS-LSTM neural network trained on all five test catchments is compared to the *wflow_sbm* model. Table 4.4 shows the NSE per catchment as an indicator for the overall performance. The resulting time series plots over the test period are shown in Appendix F.8 for both models compared to observed streamflow. All performance metrics are shown in Table F.1 and the peak flow signatures based on MTS-LSTM, *wflow_sbm* and observed streamflow are shown in Table F.3.

The MTS-LSTM out-performs the *wflow_sbm* model for the Meuse catchments 13, 702 and 703 (Huccorgne, Hastiere and Warnant) regarding all determined metrics. The NSE values are >0.45 for the MTS-LSTM while NSE values for *wflow_sbm* are around 0 or <0 for these catchments. The hydrographs for *wflow_sbm* show many peaks up to four times higher than the observed peak flow and are much spikier than the MTS-LSTM hydrographs. These three catchments (13, 702 and 703) are those with agriculture and fissured aquifers.
For catchment 701 (Hastiere) the MTS-LSTM shows a higher NSE for the daily results (0.53) than *wflow_sbm* (0.51), the KGE however is lower (0.60 opposed to 0.70). Both NSE and KGE are higher for *wflow_sbm* on the hourly time scale (NSE: 0.58 <0.74, KGE: 0.68 <0.86). For the hourly results, the *wflow_sbm* scores better on all metrics apart from the average peak timing error, which is 0.7 days larger compared to the peak timing error with the MTS-LSTM. This accounts as well for catchment 6 (highest forest fraction) on the daily time scale where *wflow_sbm* has a peak timing error of 1 day opposed to 0.71 days for the MTS-LSTM model, while the prior scores better on all other metrics. For the hourly time scale, the peak timing and magnitude are better for *wflow_sbm* while KGE and FHV are better for MTS-LSTM. NSE and KGE are high for both models (>0.7).

4.4.4 Regional Meuse MTS-LSTM for PUB

The regional Meuse MTS-LSTM is further tested in a simulated ungauged basin situation. Therefore, each of the five Meuse catchments is excluded once from the model training. The five resulting models (see Table 3.10) are each tested in the excluded catchment. Table 4.4 shows the NSE per catchment as an indicator for the overall performance. The performance metrics are shown in Appendix F.1 and the peak flow signatures based on MTS-LSTM, *wflow_sbm* and observed streamflow are shown in Appendix F.3 to allow for direct comparison with the metrics and signatures of the US models (1A, US_no, US_less), the regional Meuse model and *wflow_sbm*.

On the daily time scale, positive NSE values are achieved for all catchments. However, on the hourly time scale NSE values fall <0 for catchments 701, 702 and 703 (Hastiere, Yvoir, Warnant). Where *wflow_sbm* shows negative NSE values for the daily

Catchm.	*wflow_sbm*	US_no	US_less	Regional Meuse	PUB Meuse
6	0.75	0.48	-1.89	0.6	0.30
13	-4.21	-2.98	0.47	0.55	0.13
701	0.51	0.12	-2.53	0.53	0.31
702	-0.50	-2.93	-0.92	0.47	0.25
703	-0.34	-1.48	-1.03	0.61	0.33

Catchm.	*wflow_sbm*	US_no	US_less	Regional Meuse	PUB Meuse
6	0.78	0.53	-0.56	0.7	0.23
13	-3.65	-0.92	-71.02	0.46	0.07
701	0.74	0.39	-0.21	0.58	-1.80
702	0.01	-0.61	-17.44	0.33	-0.64
703	0.22	-0.13	-30.55	0.59	-0.61

Table 4.4: NSE of all models tested in Meuse catchments, based on daily results in upper table, based on hourly results in lower table. Dark blue indicates best result, light blue good result, red negative NSE values. The MTS-LSTM trained on US data without statics is US_no, the one with less statics is US_less.

results, the MTS-LSTM achieves positive NSE values (13: 0.13, 702: 0.25 and 703: 0.33). All peak metrics deteriorate in comparison to the regional MTS-LSTM. For catchment 6, Treignes, the hydrograph of the predicted streamflow shows peaks of too little amplitude while the baseflow is too high. Both, peaks and baseflow, are much better represented with the regional MTS-LSTM. For catchment 13, Huccorgne, this behavior is even more extreme and the hydrograph approaches a horizontal line, where the baseflow is too high. For catchments 701, 702 and 703 the simulated streamflow follows the curse of the observed streamflow better, however, peaks are simulated too low and the baseflow deviates from the observed one in many periods.
Due to the very flat hydrographs of predicted streamflow, the applied method and chosen thresholds to determine frequency and duration of high flow events does not enable to recognise peaks and thus the signatures cannot be calculated. Therefore, these values are missing in Table F.3.

4.4.5 Discussion of Results for SQ2

The application of the US MTS-LSTM in catchments of the Meuse basin in Europe enables to answer the second research sub-question *How does the trained MTS-LSTM model perform in catchments in the Meuse river basin, simulating ungauged catchments?* This experiment has the intention to simulate the situation of an ungauged basin where no access to historical records of streamflow is given to train or calibrate a hydrologic model. To assess the model performance against a reference, not only simulations from the model *wflow_sbm* are considered as benchmark but also those from a regional Meuse MTS-LSTM.

The classification of the Meuse catchments into a cluster from the US based on catchment characteristics reveals that the similarity is given for a minority of the 21 characteristics. Therefore, it is hypothesised that the selected US catchments do not cover sufficient combinations of the 21 characteristics to serve as a training base for a globally applicable model. The results indicate that other regions across the world could show as little similarity to US catchments as the Meuse. Since similarity to the training data and catchments is a pre-requisite for the global model to be transferable, this consequently reduces the (expected) model performance of PUB and the number of catchments outside of the US where the model is applicable.

Testing the MTS-LSTM in the five Meuse catchments results in scaled and shifted streamflow predictions far off from the observed streamflow when using the US model trained on ERA5 forcing and HydroMT static attributes (A1). The poor model performance is caused by a few Meuse attributes lying outside of the attribute range from the US catchments. This shows the restricted applicability of the US MTS-LSTM outside of the US and highlights again the under-representation of existing catchment types through the US catchments. Already one or a few static attributes being of different height influences the activation of different parts of the neural network in such a way that the resulting prediction is shifted and scaled significantly. Such errors can be caused through dataset specif biases (e.g. a different bias for ERA5 data in the US than in Europe). Subsequently, the high importance of model inputs originating from the same data source for training and testing purpose is proven.

Re-training the US MTS-LSTM with no static input and once with excluding the problematic static inputs yielded two models that give more reasonable streamflow predictions in the Meuse basin. However, regarding the NSE, the regional Meuse MTS-LSTM as well as the *wflow_sbm* outperform both US MTS-LSTM. Individual results are acceptable with a NSE of 0.48 (catchment 6) for the model without static input and 0.47 (catchment 13) with the model with less static attributes. This shows on one hand the potential of the global model approach, on the other hand indicates that the approach requires more research for improvement. Results for other catchments showing streamflow prediction too far off from the observed streamflow highlight the uncertainty inherent in the approach, especially for the operational PUB case where no observed streamflow are available to evaluate the prediction. As the Meuse test catchments are not very similar to the US catchment groups with best model performance (north-western US), it is suggested to search for catchments with similar characteristics and repeat the experiments for PUB. Thereby, it could be further assessed if the catchment attributes are a suitable quantifier for model performance.

The problematic results regarding PUB lead to the **suggestion of adjusted approaches** regarding the chosen datasets. A global dataset with static attributes should be used entirely to determine the scaler for data pre-processing. Then, the catchments for training should be chosen such that the entire value range of each attribute is covered and the global distribution is represented. This reduces the risk of attribute values of test catchments lying outside the range of the attribute distribution. The same method should be applied for the forcing parameters. However, decade-long time series on a sub-daily scale with global coverage would be an enormous amount of data to derive a scaler from. Therefore, the feasibility of the approach remains questionable regarding the forcing time series. The same problem holds for the streamflow observations. Theoretically, all available streamflow observation data should be used to determine the scaler of the worldwide distribution. However, as these observations are not originating from one dataset like static and dynamic forcing, the approach is unrealistic. Thus, expert knowledge is required to make a selection of catchments with extreme conditions that should serve as training catchments.

A regionally trained MTS-LSTM model for PUB seems to be a functional method to achieve reasonable well streamflow predictions for an ungauged catchment in the same basin. Following up on the idea of regionally trained models for PUB means to refrain from the goal of setting up a globally applicable model. Especially for the daily time scale and for catchments where the process-based model *wflow_sbm* shows negative NSE values, the regional approach performs well. However, the results from training a regional Meuse model on four of the selected test catchments and testing it on the fifth have shown that individual catchments can differ signifi-

cantly in their properties and runoff-behavior. Then, also the regional approach is not resulting in good performance, indicating that a thorough analysis of the catchment similarities within one basin is required to assess the expected performance for PUB of such a regional model.

The regional model trained on all five test catchments can compete with *wflow_sbm* or outperforms the process-based model on both time scales. This indicates that a regional model trained on the data of all catchments in the Meuse basin could perform even better. Alternatively, one MTS-LSTM model for each catchment can be trained individually with local datasets. A third option is a regional Meuse model, initially trained on all Meuse catchments together, then fine-tuned individually for each catchment. Which solution works best has to be assessed with appropriate experiments. Previous results from Kratzert et al. (2018) indicate that the best results are achieved with the third option.

Considering the more extreme conditions of meteorological parameters as well as catchment characteristics and streamflow height developing in the upcoming years due to changing climate conditions, the question evolves whether a MTS-LSTM could be suitable for streamflow forecasting and future scenario studies. Based on the conclusion that the distribution of the input parameters should ideally cover the full range of possible values, the training data for a forecasting scenario should be a composition of the meteorological forecast. This means the training data is a synthetic time series that represents more extreme events and does not come from historical data. The scalers for the forcing should be determined based on the whole forecasting time series. The static attributes are average values over several years, averaged for a whole catchment. To capture land-use change, vegetation cover change, climate change and other changing characteristics, the attribute value per catchment should be re-calculated regularly e.g. every year over the past 20 years. 20 years is the period over which the climate indices of the Camels US dataset have been computed. Otherwise, static attributes could be converted into dynamic input if time series are available. Then, a sensitivity analysis should be done to assess if the added value is larger compared to using the attributes statically.

4.5 SQ3: DIFFERENT LOSS FUNCTION

The first part of this Section 4.5.1 presents the results for the US model trained with the M4SE loss function, ERA5 forcing and HydroMT statics and tested in the US catchments. In the second part 4.5.2, the results for the regional Meuse model trained with the M4SE loss function, ERA5 forcing and HydroMT statics are presented. Here, the entire set of 21 static attributes is used. The last part 4.5.3 discussed the findings to answer the third research sub-question:

Can model performance regarding high flow representation be improved by training the MTS-LSTM with a different loss function?

4.5.1 US MTS-LSTM

All median performance metrics are shown in Table 4.5. Compared to the model of SQ1, the median NSE for the model trained on ERA5 drops by 0.3 (0.35) for daily (hourly) results, for the NLDAS-2 trained model the NSE drops by 0.2 (0.23) respectively. The KGE drops less by 0.3 for the ERA5 model and 0.17 for the NLDAS-2 model for both time scales. The number of catchments with a negative NSE increase up to over 200 for ERA5 and over 120 for NLDAS-2. Averaged over all catchments, peak timing and absolute error in peak magnitude increase slightly for the NLDAS-2 model while the absolute error improves slightly for the ERA5 model. For the

relative error in peak magnitude no significant change occurs. The median FHV improves by 12 (19) mm/h for daily (hourly) ERA5 results and by 11 (10) mm/h for NLDAS-2 results. The ratio of peaks that are underestimated in the modeled streamflow compared to the overestimated peaks is more balanced with the M4SE loss function, still more peaks are simulated lower compared to the observed peak height.

The results per US catchment cluster show improvements on high flow metrics for clusters where model performance has already been good with the NSE as a loss function. These were, in decreasing performance order, cluster 7, 3 and 5. All metrics per cluster are shown in Table E.1.
For **cluster 7** all peak flow metrics improve with the M4SE loss function, while the median NSE drops by 0.03 for all time scales and both models. The KGE drops as well by 0.03 for the ERA5 model and is not affected for the NLDAS-2 model. For **cluster 3** improvements occur for the FHV for all models and both time scales. The peak timing error decreases only for the NLDAS-2 model. The absolute peak magnitude error decreases, for the daily results more significantly (1.27 mm/h for ERA5 results) than for hourly results. The relative peak magnitude error decreases minimal (0 for hourly ERA5 up to -0.09 for daily ERA5 results). For **cluster 5** the relative peak magnitude error decreases for ERA5 results. The FHV improves for all models and both time scales. All other high flow metrics do not improve with the M4SE loss function for catchments of this cluster. The ratio of under- and overestimated peaks balances out with the new loss function for all clusters while still more peaks are underestimated compared to the true, observed peak height.

M4SE US		NSE	KGE	NSE<0	KGE<-0.41	FHV	Peak-Timing	ϵ_{abs}	ϵ_{rel}
Model	Freq								
1A ERA5	1D	0.54	0.60	31	11	-21	0.45	7.75	0.53
	1D	0.24	0.31	171	115	-8	0.46	6.87	0.48
	1H	0.50	0.58	15	-21	36	4.89	0.28	0.53
	1H	0.15	0.28	207	139	-2	5.05	0.27	0.54
2A NLDAS-2	1D	0.70	0.71	20	11	-15	0.39	5.77	0.41
	1D	0.50	0.55	110	66	-4	0.41	5.94	0.41
	1H	0.68	0.71	28	16	-12	4.38	0.24	0.46
	1H	0.45	0.54	129	65	-1	4.80	0.25	0.48

Table 4.5: Median performance metrics for MTS-LSTM models trained on data from 516 US catchments with M4SE as loss function. Static attributes from HydroMT. Compared to results from models 1A and 2A (see Table 4.2) here shown in gray rows.

M4SE US		Q_{high} freq.	Q_{high} duration	Q95	HFD_{mean}	$\bar{Q}/\bar{P}$
Model	Freq	[d/a], [h/a]	[d], [h]	[mm/h]	[d]	[-]
1A ERA5	1D	0.38	0.01	0.99	0.87	0.97
	1D	0.45	0.04	0.95	0.84	0.82
2A NLDAS-2	1D	0.45	0.03	0.97	0.91	0.94
	1D	0.39	0.09	0.95	0.84	0.87
1A ERA5	1H	0.34	-0.01	0.99	0.91	0.96
	1H	0.41	0.01	0.93	0.81	0.78
2A NLDAS-2	1H	0.47	-0.02	0.97	0.93	0.94
	1H	0.21	-0.05	0.95	0.80	0.86

Table 4.6: Resulting correlation between simulated signatures and signatures from observed streamflow for US catchments with MTS-LSTM models 1A and 2A. Results from model with M4SE loss function in white rows, results from model with NSE loss function in gray rows. When two units are given, the first one applies to daily (1D) results, the second one to hourly (1H).

4.5.2 Meuse MTS-LSTM

Hydrograph plots for the Meuse catchments with the regional Meuse MTS-LSTM model and the new loss function are shown for the test period in Appendix F.10. All resulting metrics and signatures can be found in Table F.1 and F.3 (model: M4SE).

Table 4.7 gives an overview of the general performance and improvements with the new loss function compared to the NSE loss function (for the Regional Meuse MTS-LSTM from SQ2). While the NSE is lower for all cases, apart from catchment 701 on the hourly time scale, the KGE improves for some catchments with the new loss function. Overall, the peak magnitude error improves for each catchment, however, only for catchment 13 on the hourly time scale all metrics improve. The FDCs in Appendix F.3 show that the baseflow representation deteriorates for all catchments and the high flow segment of the FDC is a similar or slightly better fit compared to the results of the model trained with the NSE loss.

Catchment	Loss (1D)		Loss (1H)		Improvements
	NSE	M4SE	NSE	M4SE	
6	0.60	0.50	0.70	0.63	KGE, FHV, peak magnitude (1D)
13	0.55	0.40	0.46	0.40	All peak metrics (1H)
701	0.53	0.32	0.58	0.59	Peak magnitude (1D, 1H)
702	0.47	0.25	0.33	0.32	KGE, peak magnitude (1D, 1H)
703	0.61	0.34	0.59	0.49	Peak magnitude (1D, 1H)

Table 4.7: NSE values per catchment for MTS-LSTM trained with M4SE loss function compared to the model trained with the NSE loss, for daily (1D) and hourly (1H) results. Dark blue indicates similar good NSE values, light blue indicates good but lower NSE values, beige indicates much lower NSE values with the new loss function.

4.5.3 Discussion of Results for SQ3

Replacing the NSE loss function with the M4SE loss function results in an overall deterioration in general model performance, measured by NSE and KGE. When however considering regions in the US where the initial model trained on the NSE loss function showed best performance, the new loss function results in little to now general deterioration but partly to improved high flow metrics and signatures. Thus, for an application of the MTS-LSTM with focus on peak flow representation, the new loss function is suitable. A combination of both loss functions could lead to a higher general performance, regarding high flows as well as low flows. Then, an assessment of how to condition the application of either one of the loss functions during the training of a LSTM is required. For example, the M4SE loss could be applied to values in the upper quartile of the observed streamflow distribution, while for all lower values the NSE loss determines the training procedure.
The metrics of the MTS-LSTM trained on and tested in the Meuse catchments do indicate towards an improvement regarding high flow representation on the hourly time scale when training with the M4SE loss. The hydrographs and FDCs however show that the improvement is not valid for all test catchments and the baseflow representation deteriorates. The results confirm the observation that improvement is mainly achieved when the model the is trained with the NSE loss function performs well already.
For the experiment with the new loss function, no additional hyperparameter tuning has been done. This could potentially lead to improved results compared to the here presented streamflow simulations and metrics.

4.6 LIMITATIONS

The findings and interpretations resulting from this research have to be seen in the scope of the selected methodology and the reader should be aware of the following limitations and remarks.

The US and European catchments chosen as training and testing catchments are to a high degree near natural, small headwater catchments. The findings are not proven to be scalable to significantly larger catchments or to be applicable to catchments which are under strong human influence.
The *wflow_sbm* model serves as a benchmark. A direct and 'fair' comparison is not feasible due to the internal structures of a process-based model and the training (calibration) concept behind a ML model being of very different nature. By choosing data sources for the MTS-LSTM that are regularly used in simulations with *wflow_sbm* and have mainly been used for the production of the benchmark results referred to in this research, it is intended to approximate a reasonable base for a delimitation between the results of the two model types.
As the *wflow_sbm* results for the US catchments have not been explicitly computed for this study but have been provided from a different project, the training, validation and testing periods do not match exactly with those applied in this research. However, the general picture of MTS-LSTM results outperforming *wflow_sbm* results is expected to be correct and not differ significantly when matching those periods exactly.
All streamflow predictions and performance metrics computed with the MTS-LSTM for US and Meuse catchments are results from a single model run. A method to reduce the generalization error of the predictions is *ensemble modeling*. Then the same model is trained on different training datasets and the results of each iteration are aggregated to get the final prediction.
For testing the M4SE loss function no tuning of the hyperparameters is done. An additional tuning could improve the achieved results. Especially the learning rate, number of epochs and batch size can have an effect on the model performance.
All results and conclusions related to the Meuse catchments are based on a selection of five test catchments. Thus, generalization and scalability to other river basins has to be shown individually, while the findings of this research serve as an indication which performance can be expected under the here applied research setting.

5 CONCLUSION & RECOMMENDATIONS

This research aims to investigate the transferability of LSTM neural networks trained to predict rainfall-runoff relations. Predicting streamflow in ungauged catchments is a major challenge in the field of hydrologic modelling. When no streamflow observations are available, process-based models can face uncertainties in their representation of hydrologic processes and chosen model parameters. Recent research with LSTM models establishes a new approach to PUB as such networks seem to learn hydrologic behavior exclusively from the training data. Supposing this detected behavior is equivalent to physical laws determining the processes of the water cycle, such a model trained on data from a large number of variant catchments should be able to apply this hydrologic knowledge to any other geographical region, thus being transferable and ideally globally applicable. State of the art research has tested spatial transferability of a US LSTM model to other catchments within the US or from a Russian LSTM to other Russian catchments (Kratzert et al., 2019b; Ayzel et al., 2020). The scope of this research goes beyond previous experiments and tests a US trained LSTM in European catchments to answer the research question:

Does a MTS-LSTM trained on data from US catchments prove to be globally applicable as hydrologic model?

5.1 CONCLUSION

For the purpose of global model applicability, datasets with worldwide coverage are required and therefore, in this research, their suitability is assessed by demarcating performance of a MTS-LSTM model trained on the global datasets (Era5 and a compilation of HydroMT datasets) against a MTS-LSTM model trained on local datasets for which good performance is documented in state-of-the-art literature.
The experiments to test the suitability of the ERA5 dataset in US catchments have shown that using global datasets as dynamic and static model input does not perform equally good like a model trained on a high resolution local dataset (NLDAS-2). Nevertheless, the neural network trained with ERA5 still outperforms the process-based distributed model *wflow_sbm* for US catchments. For humid and surface-runoff dominated catchments the MTS-LSTM model achieves average NSE values of > 0.8. Poor model performance occurs in arid regions, regions with groundwater depletion and high soil conductivity.

Subsequent application of the US trained model for streamflow predictions in catchments of the Meuse basin confirms these findings as catchments with fissured aquifers are more difficult to simulate than those with higher yearly precipitation and streamflow and with widespread forest cover. However, the US trained MTS-LSTM only leads to a higher NSE value (0.48) compared to the process-based model *wflow_sbm* (negative NSE) for one of the test-catchments. In the other catchments either none of the two models results in a positive NSE or the *wflow_sbm* is the better choice.
Initially, the US trained MTS-LSTM produced an unrealistic shifted and compressed streamflow hydrograph compared to the observed streamflow of the Meuse catchments, due to 3 out of 21 static input parameters not lying within the range of the

training data distribution. From the strong influence of single static parameters it is deduced that the use of a global dataset together with a training dataset derived from more than 500 catchments is not suitable to create a globally applicable model. First, a global dataset could show a different bias compared to reality per continent or climate zone due to incorporation of different local data products or varying model settings. Second, the training data does not include sufficient extreme attribute values and should therefore be composed of catchments covering the most extreme conditions to prevent offsets in the predicted streamflow time series.

Concluding, to improve the applicability of a MTS-LSTM model in ungauged basins, it follows that (1) the forcing dataset should be a global dataset providing sub-daily meteorological information of higher spatial resolution than Era5, e.g. the Era5 land dataset. (2) the most extreme catchments with accessible data worldwide – regarding the catchment characteristics represented in static attributes and regarding extreme events captured in the meteorological forcing – should function as training catchments while (3) the standardization of the model input data – as one of the most important prepossessing step for a machine learning model – should be done based on the mean and standard deviation of the entire global distribution for each static and dynamic parameter.

The promising performance of the MTS-LSTM within the US when trained on US data, as well as results from previous studies with locally trained LSTM models, were the reason for training a regional MTS-LSTM for the Meuse basins. The aim is to benchmark the regional Meuse MTS-LSTM against results from the process-based model *wflow_sbm*. Results confirm the hypothesis and show that the MTS-LSTM can compete with *wflow_sbm* and additionally achieve good NSE values of minimum 0.46 where the process-based model shows negative NSE values. In conclusion, this proves the ability of LSTM models to mimic local hydrologic processes at least as good and in some cases better than a process-based model. As these results are based on a model trained with merely a subset of the available data for the Meuse basin, accompanied by the use of the low resolution Era5 data as forcing, the possibly achievable performance with an LSTM model is far from being fully exploited.

For the case, that a hydrologic model is applied in a catchment with the intention to predict streamflow correctly with focus on the high flow peaks, a different loss function for the training of the MTS-LSTM has been tested. The M4SE yielded even more accurate peak height representation in the cases where already the model trained with a NSE loss function worked well. As the baseflow representation deteriorates compared to the model trained with the NSEloss function, a combination of both resulting streamflow time series could lead to an overall improvement. Thus – for gauged and ungauged catchments – an implementation of a combined loss function appears a valuable follow-up research.

Despite the model performance for PUB being lower than initially hypothesised, the here presented research findings have already further implications for the development of a model for PUB. According to the presented results, a regional or local trained LSTM model is the best choice for rainfall-runoff modelling and thus a model should always be tried to be fitted best to the catchment where it is applied. As for many catchments long historical time series of streamflow observations are not accessible the prerequisites to train an individual local model are not fulfilled. However, in the case that few years of streamflow observations are available, a global LSTM model pre-trained on a variety of catchments offers the advantage to be finetuned with these short time series of local data. Thereby fitting a global model to the local conditions comes closest to setting up a local model trained with long historical streamflow records.

5.2 RECOMMENDATIONS

Based on the conclusions of this research regarding the applicability of a global MTS-LSTM as hydrologic model for streamflow predictions in gauged and ungauged basins, the following summarises the recommended actions for improvement and further research in the field.

Regarding the **static model input**, a sensitivity analysis should be performed to assess which attributes add hydrologic relevant information and how redundant information given by attributes derived from the forcing time series influences the model output.

For a **Meuse MTS-LSTM**, all available data from each catchment should be used as training data for a Meuse basin model. Alternatively, one model per Meuse catchment should be set up, if possible with local high resolution forcing data.

In order to improve accurate modelling of **high flows**, the implementation of a loss function combining the NSE and a loss function more sensitive to high signals, like the M4SE, should be investigated.

For PUB with a global LSTM the training dataset should be of higher resolution than the Era5 dataset. The model should be trained on catchments representing the most extreme values of all forcing parameters and catchment characteristics. The resulting model should be finetuned locally if short time series of observed streamflow are accessible.

Last but not least, to improve streamflow predictions for ungauged basins, to counteract the development or increment of societal inequalities resulting from restricted access to such predictions, and to enhance our understanding of hydrologic processes, it is of high importance to continue **measuring streamflow in the field**. The global coverage with gauging stations should grow instead of shrink and the access to observational data should be easy and free.

A CONFIGURATION FILE EXAMPLE

```
additional_feature_files: None
allow_subsequent_nan_losses: 10
batch_size: 2048
clip_gradient_norm: 1
clip_targets_to_zero:
- qobs
data_dir:
      C:\Users\kwilbrand\MTS-LSTM\data\ hydromt
dataset: generic
device: cuda:0
dynamic_inputs:
- cp           # convective_fraction
- msdwlwrf # longwave_radiation
- cape        # potential_energy
- pev         # potential_evaporation
- press       # pressure
- kin         # shortwave_radiation
- d2m         # specific_humidity
- temp        # temperature
- precip      # total_precipitation
- u10         # wind_u
- v10         # wind_v
embedding_hiddens:
- 30
- 20
- 64
epochs: 30
experiment_name: 03_era5_hydromt
head: regression
hidden_size:
 1D: 64
 1H: 64
initial_forget_bias: 3
learning_rate:
 0: 0.0005
 10: 0.0001
 25: 5e-05
loss: NSE
metrics:
- NSE
- KGE
- RMSE
- FHV            # peak flow bias
- Peak-Timing
model: mtslstm
num_workers: 1
optimizer: Adam
output_activation: linear
output_dropout: 0.2
shared_mtslstm: false
predict_last_n:
 1D: 1
 1H: 24
regularization:
- tie_frequencies
seed: 110
seq_length:
 1D: 365
 1H: 336
static_attributes:
- elev_mean
- slope_mean
- area_km2
- lai_max
- lai_diff
- gvf_max
- gvf_diff
- soil_thickness
- soil_porosity
- ksatver_logmean
- theta_s        # max_water_content
- carb_rocks_frac
- geol_permeability
- p_mean
- pet_mean
- aridity
- frac_snow_daily
- high_prec_freq
- high_prec_dur
- low_prec_freq
- low_prec_dur
target_variables:
- qobs
test_basin_file: data/516_basins.txt
test_end_date: 30/09/2018
test_start_date: 01/10/2008
train_basin_file: data/516_basins.txt
train_data_file: None
train_end_date: 30/09/2003
train_start_date: 01/10/1990
transfer_mtslstm_states:
 h: linear
 c: linear
use_frequencies:
- 1D
- 1H
validate_every: 10
validate_n_random_basins: 516
validation_basin_file: data/516_basins.txt
validation_end_date: 30/09/2008
validation_start_date: 01/10/2003
```

Figure A.1: Configuration file for model 1A with ERA5 forcing and HydroMT static input.

B US CATCHMENTS

B.1 US UNGAUGED CATCHMENTS

Figure B.1: US catchments without streamflow observations in training period, therefore functioning as ungauged basins within the US in the experiments of SQ1: 01552000, 01552500, 01567500, 07301410, 07346045, 08050800, 08101000, 08104900, 08109700, 08158810

B.2 CATCHMENT ATTRIBUTE PER HRU

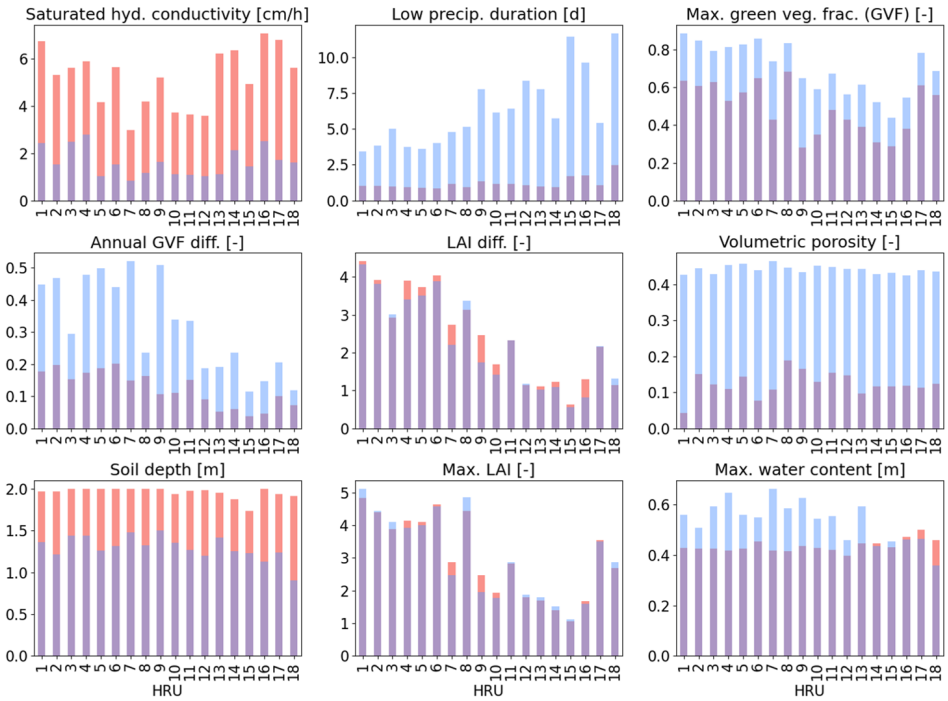


Figure B.2: Mean catchment attribute per HRU (part 1).

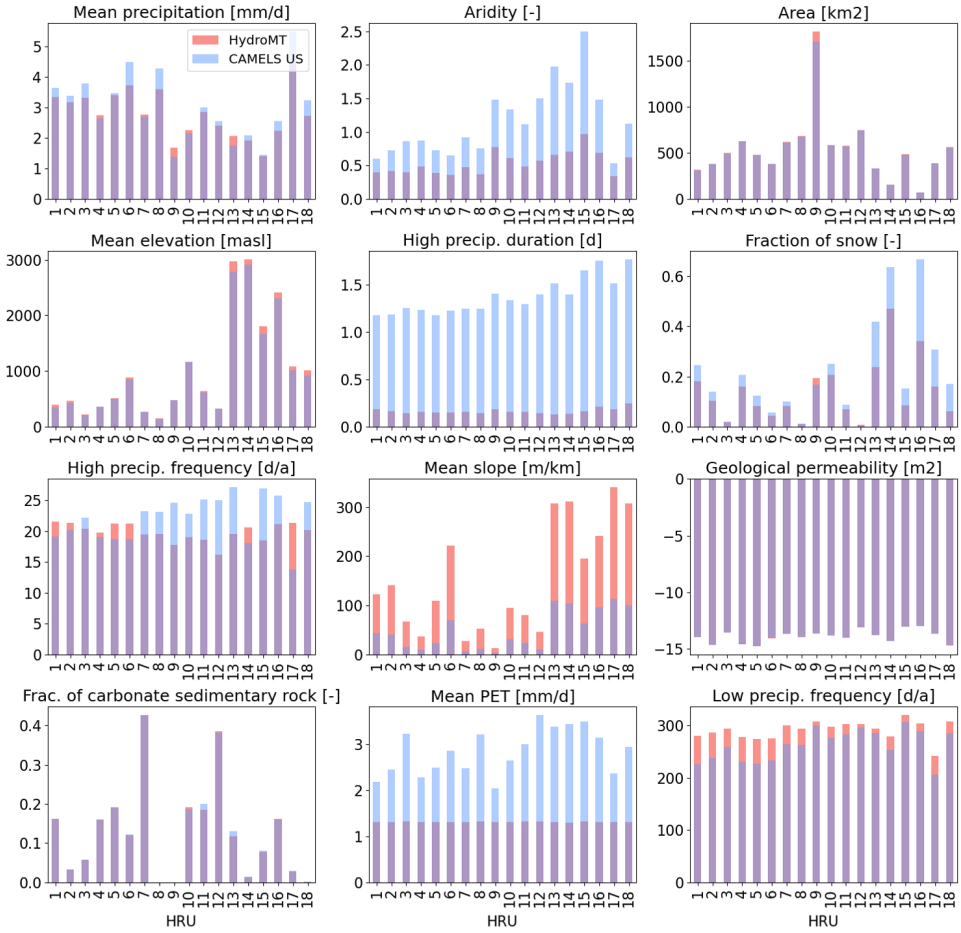


Figure B.3: Part 2 of Figure B.2. Red bars for HydroMT values, blue bars for Camels US, overlap in purple.

C CATCHMENT CLUSTERING

C.1 BUDYKO PLOT PER US CATCHMENT CLUSTER

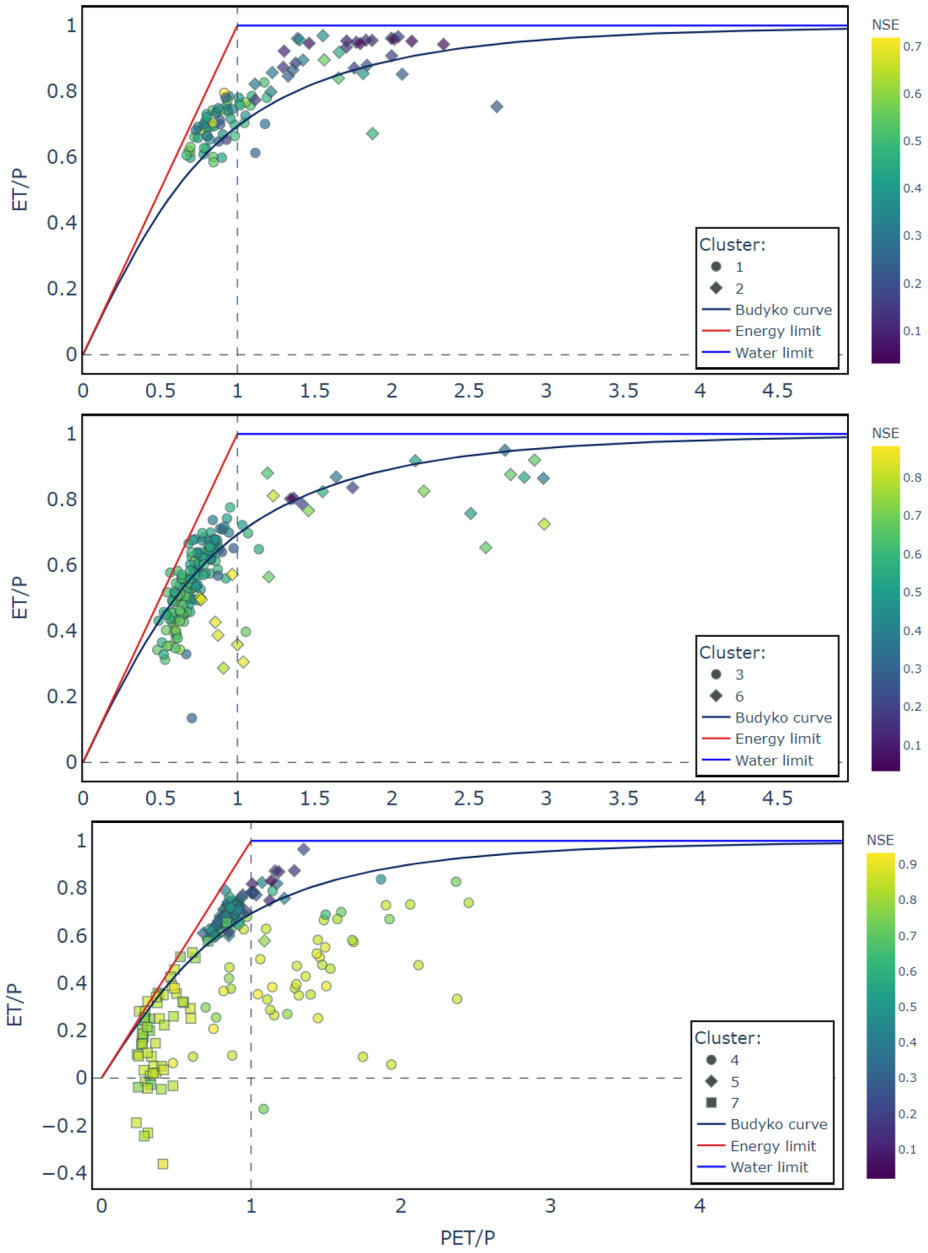


Figure C.1: Budyko plot per catchment cluster. Y-axis: $\frac{A\bar{E}T}{\bar{P}}$, x-axis: $\frac{P\bar{E}T}{\bar{P}}$. Colored by NSE achieved with model 1A on daily time scale.

Figure C.2: Catchments with negative $\frac{ET}{P}$ in Budyko plot: 06746095, 12040500, 12041200, 12054000, 12056500, 12167000, 12175500, 12178100, 12186000, 12147500, 14400000.

C.2 MEAN CATCHMENT ATTRIBUTES PER CLUSTER

Nr.	Area	Geological Permeability	Soil porosity	Carbonate rock fraction	Mean elevation	Mean slope	Max. GVF
1	597.18	-13.27	0.42	0.05	92.96	5.31	0.77
2	840.44	-13.74	0.44	0.03	853.56	19.82	0.52
3	405.03	-14.65	0.44	0.05	508.02	40.68	0.87
4	292.31	-13.78	0.43	0.07	2530.78	111.25	0.55
5	540.21	-13.73	0.47	0.41	323.52	9.58	0.73
6	533.40	-13.40	0.44	0.36	903.94	63.22	0.55
7	365.53	-13.71	0.44	0.00	719.09	111.64	0.84

	GVF difference	LAI difference	Max. LAI	Mean precipitation	Mean PET	Aridity	High precip. duration
1	0.25	2.61	3.80	3.74	3.32	0.89	1.27
2	0.24	0.93	1.35	1.84	3.11	1.78	1.46
3	0.46	4.09	4.77	3.57	2.49	0.71	1.20
4	0.22	1.11	1.67	2.42	3.01	1.36	1.45
5	0.47	2.25	2.61	2.89	2.59	0.92	1.25
6	0.13	0.96	1.83	2.23	3.37	1.73	1.62
7	0.18	2.34	4.03	6.27	2.30	0.39	1.55

	Low precip. duration	High precip. frequency	Low precip. frequency	Snow fraction	Soil depth to bedrock	Soil conductivity	Max. water content
1	5.14	22.48	261.11	0.02	1.50	3.20	0.61
2	8.43	25.72	297.81	0.11	1.41	1.14	0.57
3	3.82	19.50	233.90	0.15	1.25	1.52	0.53
4	6.29	18.95	254.53	0.64	1.25	2.01	0.47
5	4.96	23.13	264.15	0.09	1.41	0.93	0.64
6	11.27	25.35	301.52	0.09	0.84	1.30	0.29
7	5.58	13.81	200.56	0.19	1.24	1.65	0.46

Table C.1: Mean attribute values per cluster. Maximum values per attribute in blue, minimum values in red.

D DATASET COMPARISON

D.1 ERA5 VS. NLDAS-2 FORCING PER HRU IN US

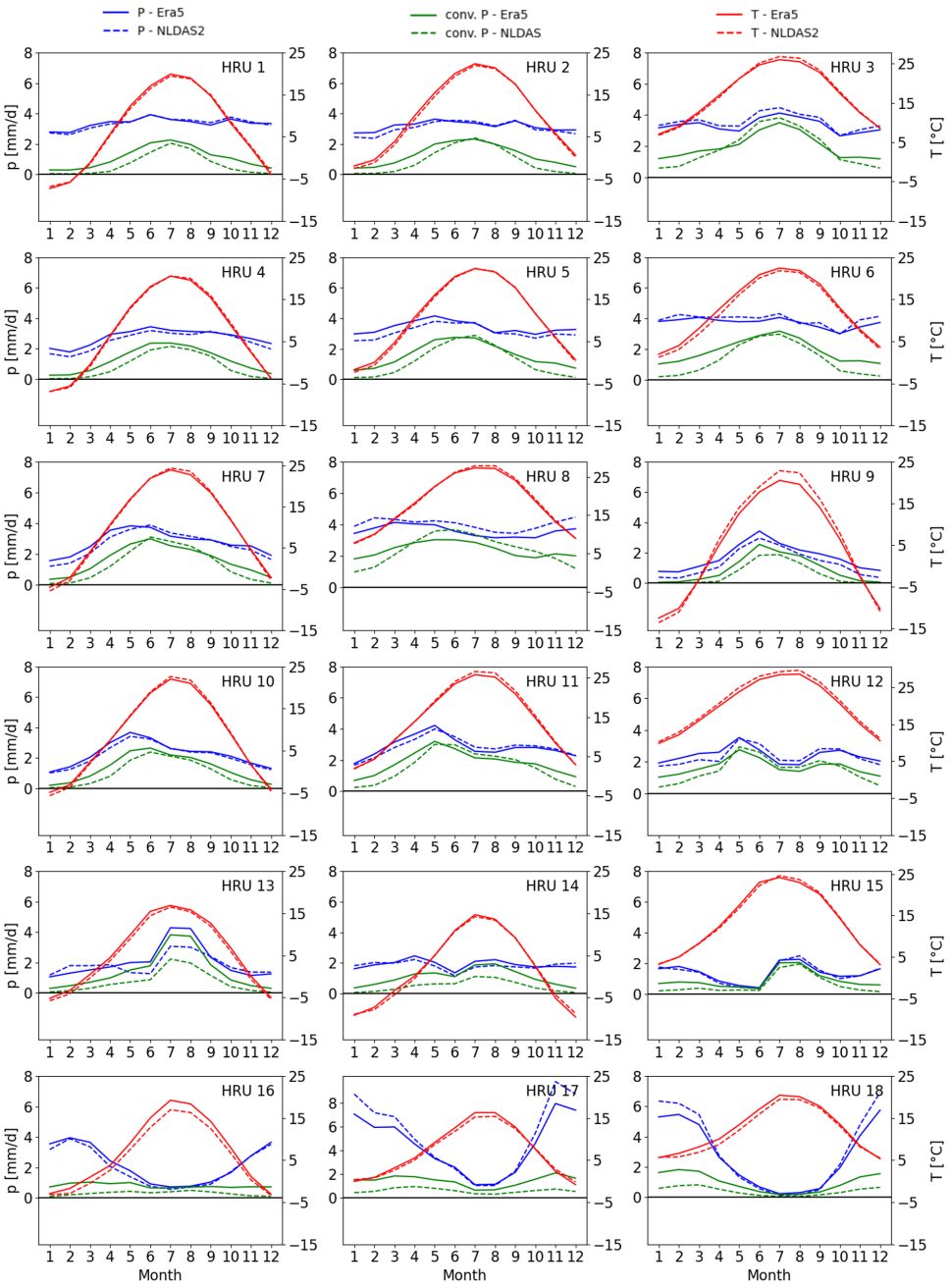


Figure D.1: Monthly climatology per HRU in US. Continuous line based on ERA5 forcing data, dashed line based on NLDAS-2 forcing data. Blue: precipitation in mm/d, green: convective precipitation in mm/d, red: temperature in $°C$.

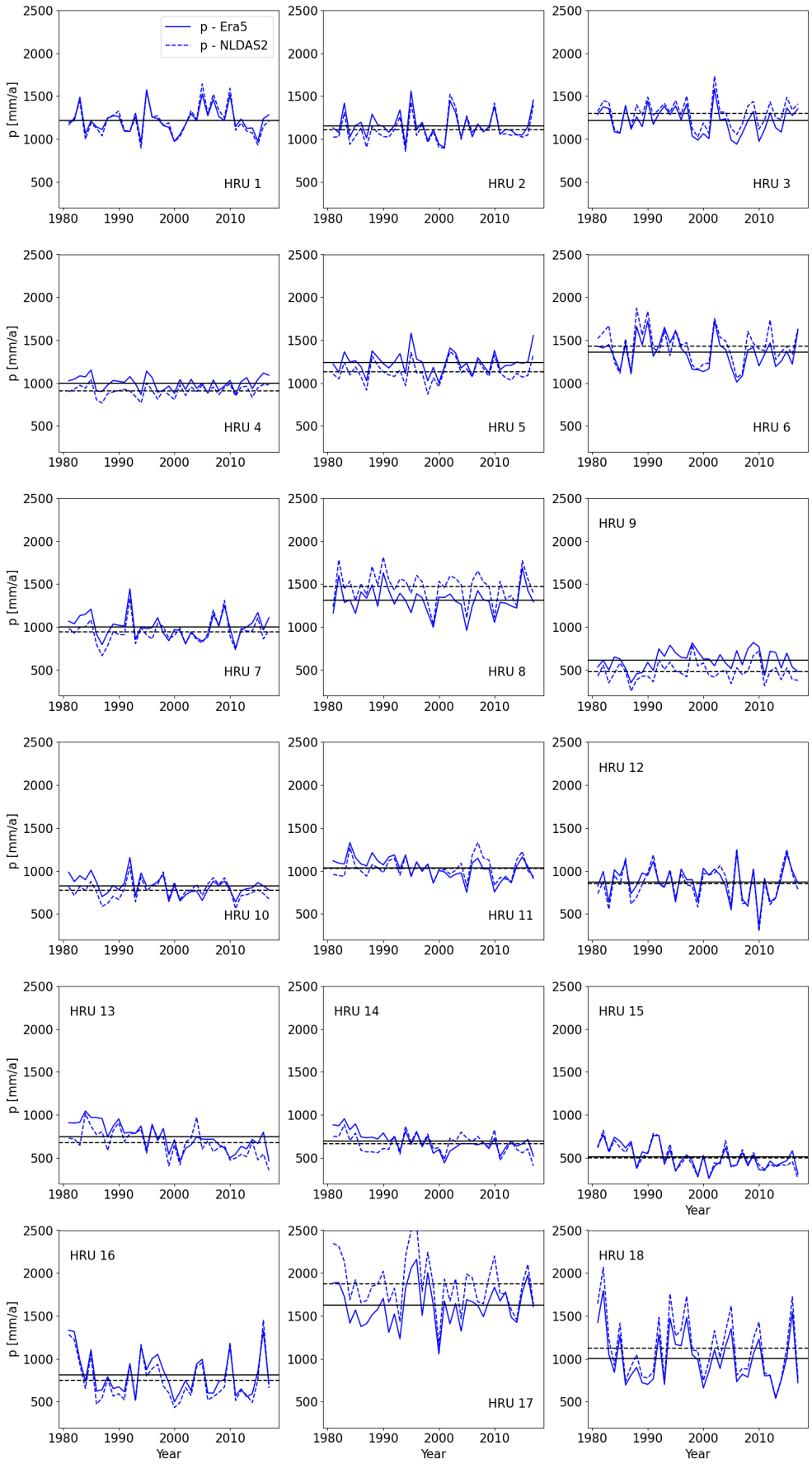


Figure D.2: Yearly climatology per HRU in US. Continuous line based on ERA5 forcing data, dashed line based on NLDAS-2 forcing data.

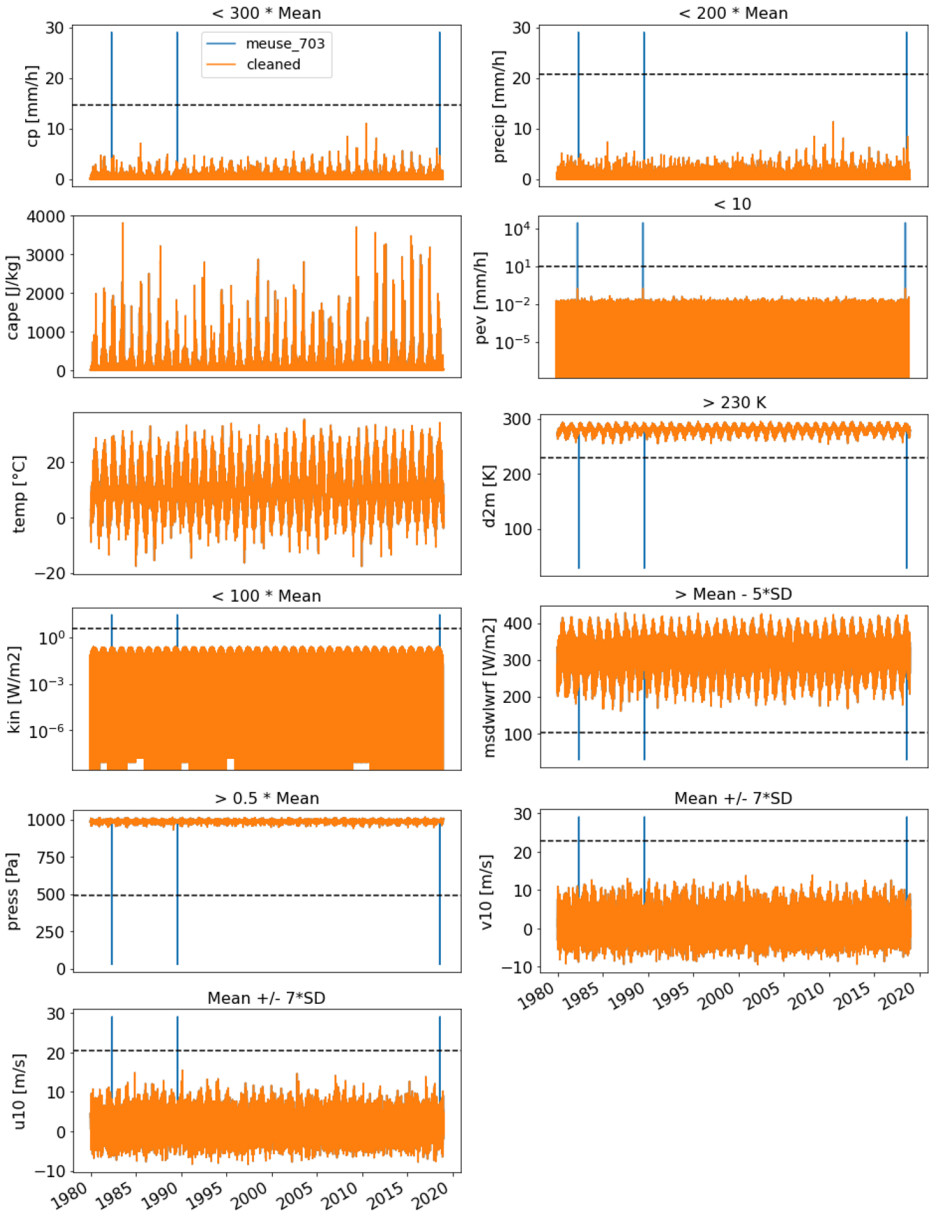


Figure D.3: Time-series of forcing parameters for Meuse catchment 703. Orange graphs are cleaned time series, blue vertical lines show outliers that were previously included in time series. Black dashed horizontal lines show thresholds applied to find outliers and replace with mean. Thresholds given in title of each subplot, if outliers were present. SD = standard deviation.

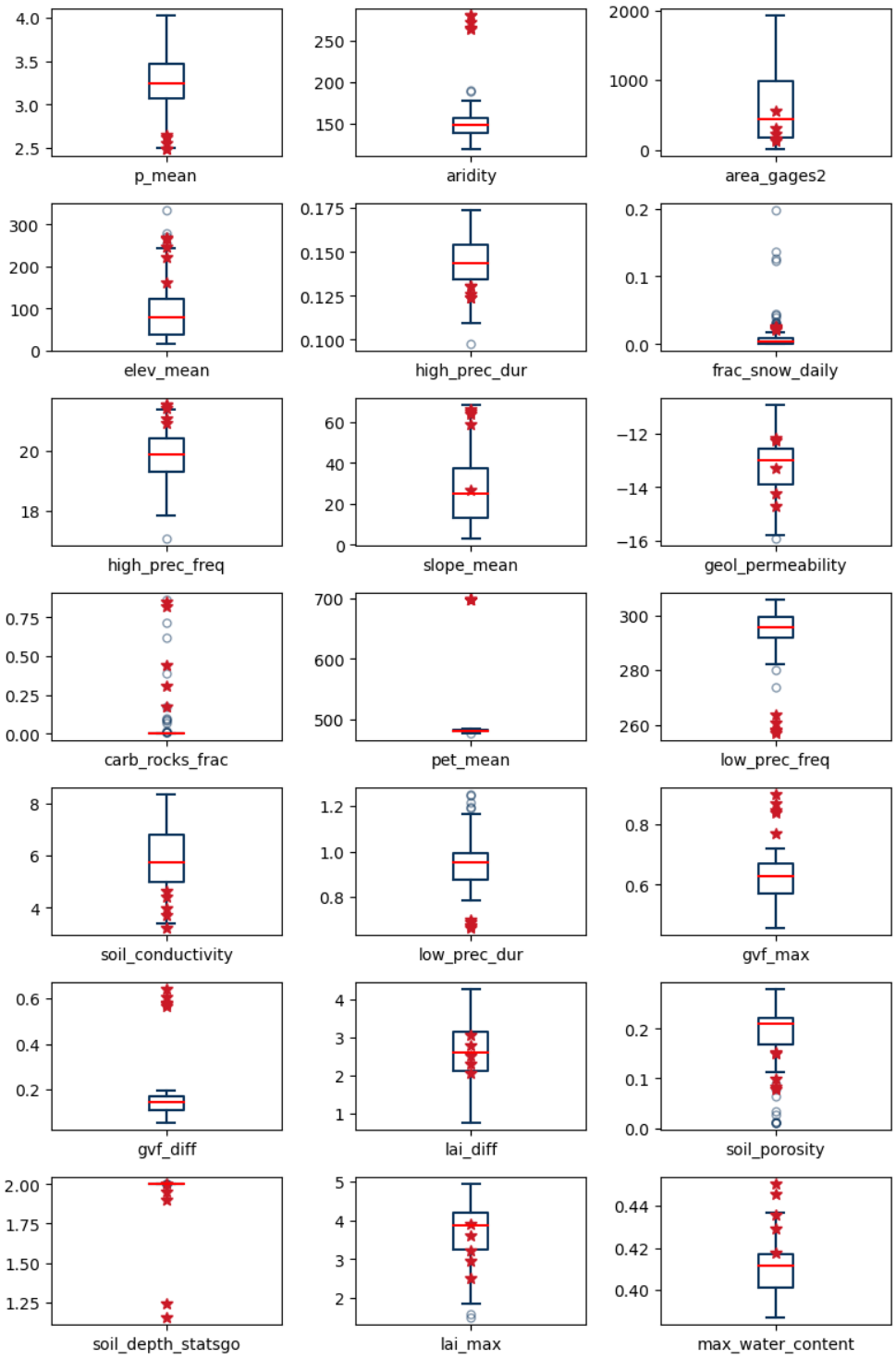


Figure D.4: Attributes box-plots based on US catchments of cluster 1. Red stars show values of Meuse catchments.

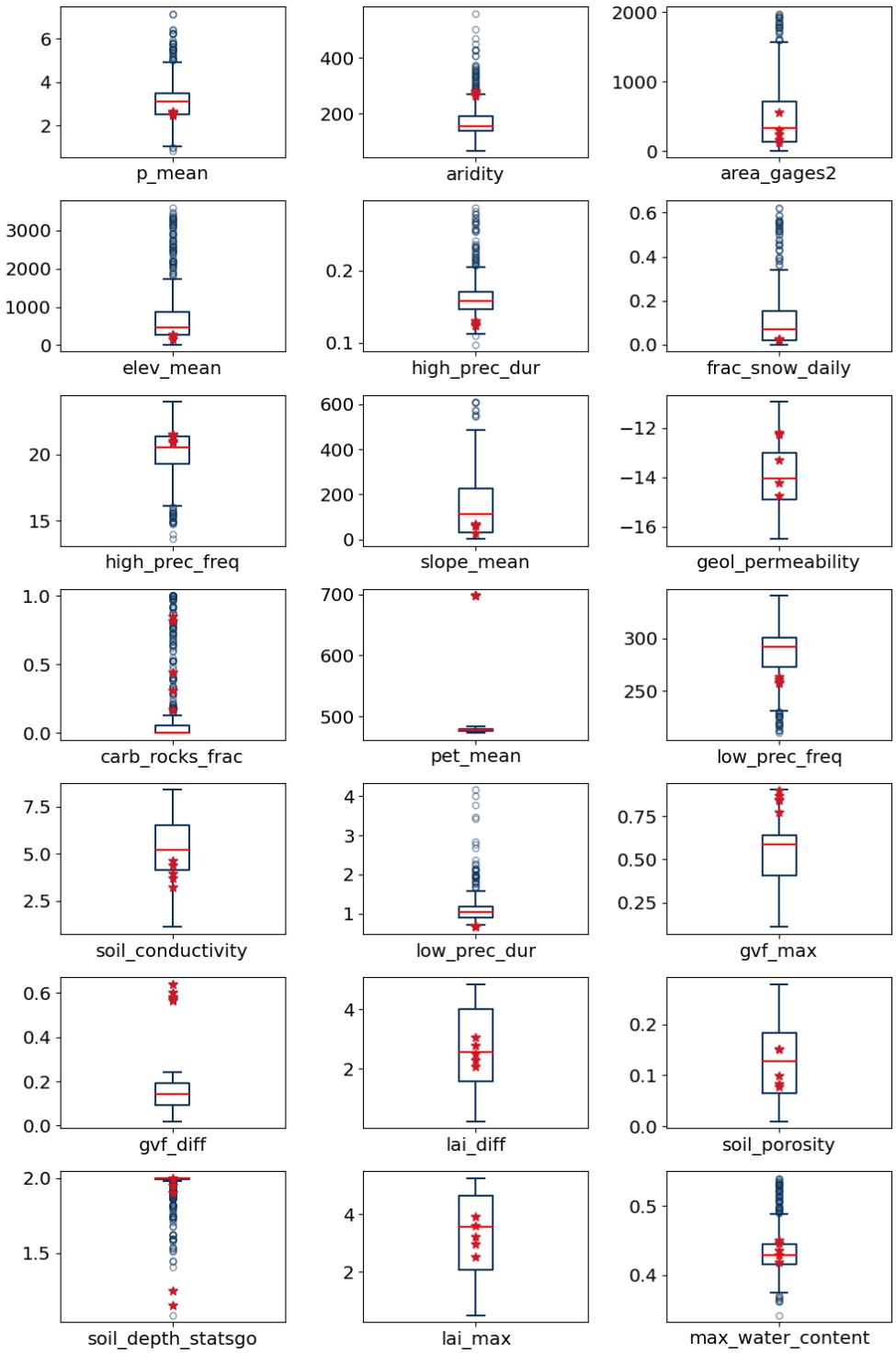


Figure D.5: Attributes box-plots based on all US catchments. Red stars show values of Meuse catchments.

E US MTS-LSTM PERFORMANCE

E.1 GROUNDWATER DEPLETION US

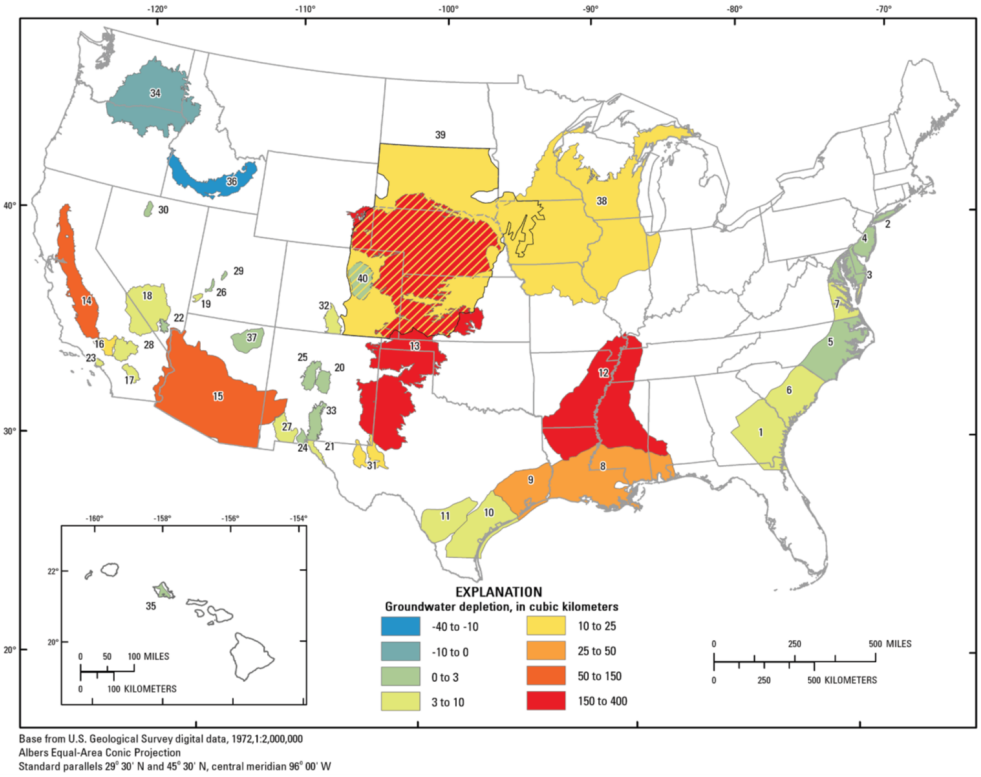


Figure E.1: US map showing cumulative groundwater depletion, 1990 - 2008 (Konikov, 2013). Strongest depletion of 150 - 400 km^3 in red.

Cluster	Forcing	Freq.	NSE	KGE	NSE <0	FHV	KGE <-0.41	Peak-Timing	ϵ_{abs}	ϵ_{rel}	Peaks	$Q_s < Q_o$	$Q_s > Q_o$
1 (67)	ERA5	1D	0.46	0.55	5	-21	0	0.64	7.36	0.56	17	14	3
		1H	0.40	0.52	4	-23	0	6.73	0.23	0.58	69	59	12
	NLDAS-2	1D	0.57	0.69	2	2	0	0.62	5.52	0.44	17	11	5
		1H	0.56	0.65	5	7	0	6.73	0.17	0.46	69	48	19
2 (46)	ERA5	1D	0.13	0.09	11	-49	7	0.86	3.82	0.79	10	8	1
		1H	0.06	-0.07	16	-53	10	6.61	0.17	0.80	30	28	2
	NLDAS-2	1D	0.25	0.31	8	-26	5	0.73	3.10	0.73	10	8	1
		1H	0.19	0.24	12	-31	10	5.51	0.14	0.73	30	25	4
3 (164)	ERA5	1D	0.57	0.65	0	-23	0	0.38	8.24	0.51	19	17	2
		1H	0.55	0.63	0	-22	0	4.70	0.30	0.52	111	92	18
	NLDAS-2	1D	0.74	0.77	0	-14	0	0.29	5.60	0.38	19	15	3
		1H	0.73	0.77	0	-9	0	3.71	0.23	0.42	111	84	26
4 (54)	**ERA5**	1D	0.86	0.83	1	-12	0	0.88	2.57	0.31	10	8	2
		1H	0.85	0.81	0	-15	0	2.11	0.12	0.33	39	36	5
	NLDAS-2	1D	0.80	0.75	1	-18	0	0.80	3.27	0.35	10	9	1
		1H	0.80	0.78	1	-15	0	3.52	0.15	0.37	39	36	4
5 (89)	**ERA5**	1D	0.42	0.47	4	-36	0	0.50	11.10	0.62	16	15	2
		1H	0.35	0.40	3	-38	0	5.70	0.43	0.64	81	67	10
	NLDAS-2	1D	0.66	0.68	1	-19	0	0.35	6.94	0.47	16	13	3
		1H	0.62	0.65	1	-15	0	4.73	0.29	0.53	81	65	14
6 (38)	ERA5	1D	0.52	0.33	10	-7	4	0.43	5.91	0.60	9	7	2
		1H	0.36	0.23	13	-7	5	3.91	0.33	0.63	26	22	5
	NLDAS-2	1D	0.53	0.48	8	-4	6	0.41	4.52	0.56	9	6	2
		1H	0.43	0.39	9	-4	6	3.85	0.31	0.72	26	21	7
7 (58)	**ERA5**	1D	0.84	0.86	0	-9	0	0.31	12.08	0.29	18	14	3
		1H	0.85	0.87	0	-5	0	3.69	0.41	0.28	108	76	33
	NLDAS-2	1D	0.81	0.77	0	-18	0	0.32	12.96	0.32	18	15	3
		1H	0.82	0.78	0	-17	0	4.20	0.44	0.32	109	91	21

Table E.1: Median performance metrics for MTS-LSTM models 1A and 2A trained on US data. Static attributes from HydroMT. Bold forcing dataset name marks model with better results per cluster. Number of catchments per cluster in parenthesis behind cluster number. Last to columns refer to identified peaks in testing period. Peak Timing in $[d]$ ($[h]$), FHV and absolute magnitude error ϵ_{abs} in $[mm/d]$ ($[mm/h]$), other metrics unit-less.

F MEUSE RESULTS

F.1 PERFORMANCE METRICS AND SIGNATURES

Nr.	Freq	Model	NSE	KGE	FHV	Peak Timing	ϵ_{abs}	ϵ_{rel}
6	1D	wflow	**0.75**	**0.80**	-16	1.0	**0.14**	**0.29**
		US_no	0.48	0.64	12	**0.54**	3.85	0.40
		US_less	-1.92	-0.45	63	0.65	4.13	0.59
		LSTM	0.60	0.67	-31	0.71	5.97	0.58
		PUB	0.30	0.21	-67	0.96	6.54	0.63
		M4SE	0.50	0.72	**-7**	1.71	5.15	0.52
	1H	wflow	**0.78**	0.72	-20	**4.23**	**0.07**	**0.30**
		US_no	0.53	0.68	**-12**	5.95	0.14	0.49
		US_less	-0.93	-0.06	34	7.70	0.24	0.98
		LSTM	0.70	**0.77**	**-12**	5.08	0.10	0.44
		PUB	0.23	0.14	-70	7.91	0.21	0.64
		M4SE	0.63	0.73	-17	5.67	0.12	0.45
13	1D	wflow	-4.21	-0.96	158	**0.57**	**0.19**	1.87
		US_no	-2.98	-0.66	129	0.92	2.46	1.03
		US_less	0.47	**0.64**	**-3**	0.92	0.97	**0.41**
		LSTM	**0.55**	0.55	-38	0.71	1.42	0.54
		PUB	0.13	0.04	-70	1.22	1.82	0.64
		M4SE	0.40	0.55	-40	1.14	1.59	0.59
	1H	wflow	-3.65	-0.79	159	9.90	0.13	1.72
		US_no	-0.92	0.12	63	7.91	0.05	0.54
		US_less	-71.18	-6.97	770	8.46	0.23	2.61
		LSTM	**0.46**	0.48	-35	7.10	0.05	0.54
		PUB	0.07	0.01	-71	**6.44**	0.05	0.54
		M4SE	0.40	**0.57**	**-28**	6.95	0.04	**0.40**
701	1D	wflow	0.51	**0.70**	**1**	1.17	**0.13**	**0.31**
		US_no	0.12	0.29	26	0.76	4.33	0.53
		US_less	-2.55	-0.93	46	1.14	3.39	0.38
		LSTM	**0.53**	0.60	-37	**0.50**	5.44	0.57
		PUB	0.31	0.24	-46	1.37	5.77	0.59
		M4SE	0.32	0.57	-20	1.33	5.03	0.52
	1H	wflow	**0.74**	**0.86**	-10	6.40	**0.09**	**0.39**
		US_no	0.39	0.58	**-4**	6.83	0.12	0.49
		US_less	-0.52	-0.10	-25	9.01	0.19	0.82
		LSTM	0.58	0.68	-18	**5.71**	0.11	0.49
		PUB	-1.80	-1.03	-9	6.25	0.11	0.58
		M4SE	0.59	0.71	-22	7.18	0.10	0.40

Table F.1: Metrics for Meuse catchments with *wflow_sbm*, US-trained model without statics (US_no) and with less statics (US_less), regional Meuse MTS-LSTM (LSTM), regional PUB simulations (PUB) and regional MTS-LSTM with M4SE loss function (M4SE). Best value per metric is marked in blue per catchment and time scale. (Part 1)

Catchm.	Freq	Model	NSE	KGE	FHV	Peak Timing	ϵ_{abs}	ϵ_{rel}
702	1D	wflow	-0.50	0.14	64	0.86	**0.07**	**0.50**
		US_no	-2.93	-0.55	130	0.96	2.69	0.78
		US_less	0.92	0.06	73	0.96	2.17	0.63
		LSTM	**0.47**	0.48	**-42**	**0.43**	1.94	**0.50**
		PUB	0.25	0.47	**-42**	1.29	2.08	0.54
		M4SE	0.25	**0.50**	**-42**	0.86	1.96	0.52
	1H	wflow	0.01	0.39	48	5.11	0.07	0.63
		US_no	-0.61	0.24	65	4.99	0.06	0.59
		US_less	-17.46	-2.77	342	6.00	0.22	2.19
		LSTM	**0.33**	0.43	-31	**3.91**	0.07	0.55
		PUB	-0.64	0.33	**24**	6.0	**0.05**	0.56
		M4SE	0.32	**0.50**	-36	5.47	0.06	**0.45**
703	1D	wflow	-0.34	0.16	67	1.00	**0.07**	0.61
		US_no	-1.48	-0.18	103	0.91	2.23	0.62
		US_less	-1.03	0.05	77	0.95	2.32	0.68
		LSTM	**0.61**	0.59	-31	**0.57**	2.06	**0.55**
		PUB	0.33	0.41	-50	1.60	2.61	0.62
		M4SE	0.34	**0.65**	**-18**	1.14	2.01	0.61
	1H	wflow	0.22	0.43	50	7.91	0.07	0.60
		US_no	-0.13	0.45	47	4.67	0.07	0.56
		US_less	-30.60	-3.98	453	6.71	0.31	2.50
		LSTM	**0.59**	**0.62**	**-22**	5.44	**0.05**	0.43
		PUB	-0.61	0.04	-55	6.22	0.08	0.58
		M4SE	0.49	0.59	-33	**4.50**	0.06	**0.39**

Table F.2: Part 2 of Table F.1

Catchm.	Freq	Model	Q_{high} frequency [d/yr]	Q_{high} duration [d], [h]	Q95 [mm/h]	HFD_{mean} [d]	$\bar{Q}/\bar{P}$ [-]
6	1D	wflow	7	2	3.84	124	0.42
		US_no	6	3	5.91	141	0.69
		US_less	2	2	7.63	137	0.96
		LSTM	4	4	3.72	128	0.39
		PUB	0	-	2.42	149	0.52
		M4SE	6	5	4.23	121	0.48
		obs	17	2	3.89	125	0.42
	1H	wflow	11	28	0.14	128	0.37
		US_no	6	1	0.20	139	0.57
		US_less	166	1	0.10	102	0.20
		LSTM	9	30	0.16	128	0.39
		PUB	0	-	0.07	153	0.30
		M4SE	3	33	0.14	123	0.41
		obs	17	36	0.16	125	0.42

Table F.3: Hydrologic signatures with *wflow_sbm*, US-trained model without statics (US_no) and with less statics (US_less), regional Meuse MTS-LSTM (LSTM), regional MTS-LSTM for PUB simulations (PUB), regional MTS-LSTM with M4SE loss function (M4SE) and observed streamflow (obs). In case two units are given, the first one applies to daily (1D) results and the second one to hourly (1H). Best value per metric is marked in blue per catchment and time scale. (Part 1)

Catchm.	Freq	Model	Q_{high} frequency [d/yr]	Q_{high} duration [d], [h]	Q95 [mm/h]	HFD_{mean} [d]	$\bar{Q}/\bar{P}$ [-]
13	1D	wflow	59	5	4.23	109	0.38
		US_no	5	3	4.00	138	0.50
		US_less	1	1	2.02	149	0.30
		LSTM	0	-	1.22	146	0.18
		PUB	0	-	0.81	174	0.24
		M4SE	0	-	1.44	141	0.26
		obs	2	2	1.70	134	0.23
	1H	wflow	58	96	0.16	108	0.33
		US_no	4	1	0.12	137	0.38
		US_less	58	5	0.50	138	0.38
		LSTM	0.3	23	0.05	149	0.18
		PUB	0	-	0.03	178	0.23
		M4SE	0	-	0.06	138	0.25
		obs	2	22	0.07	135	0.23
701	1D	wflow	17	2	3.53	120	0.37
		US_no	6	3	5.45	141	0.65
		US_less	0	1	6.54	168	0.92
		LSTM	7	4	2.46	131	0.29
		PUB	0	-	2.94	146	0.49
		M4SE	0	1	3.37	129	0.40
		obs	16	2	2.96	121	0.31
	1H	wflow	22	46	0.13	120	0.32
		US_no	6	1	0.18	139	0.52
		US_less	24	1	0.04	139	0.12
		LSTM	13	33	0.11	126	0.28
		PUB	0	-	0.25	140	0.94
		M4SE	4	27	0.12	118	0.33
		obs	17	30	0.12	122	0.31
702	1D	wflow	2	2	3.07	129	0.44
		US_no	6	2	4.76	139	0.56
		US_less	2	2	3.21	154	0.46
		LSTM	0	-	1.32	155	0.26
		PUB	0	-	1.48	150	0.25
		M4SE	0	-	1.56	129	0.25
		obs	0	1	1.96	151	0.32
	1H	wflow	2	16	0.11	134	0.38
		US_no	5	1	0.14	138	0.42
		US_less	31	2	0.30	149	0.66
		LSTM	0	-	0.05	156	0.25
		PUB	7	36	0.11	135	0.27
		M4SE	0	-	0.06	139	0.28
		obs	1	16	0.08	152	0.32
703	1D	wflow	5	1	3.24	126	0.41
		US_no	6	3	4.52	141	0.56
		US_less	2	2	3.42	152	0.48
		LSTM	0	-	1.48	150	0.29
		PUB	0	-	1.42	156	0.26
		M4SE	0	-	2.02	129	0.30
		obs	0	1	2.11	145	0.33
	1H	wflow	6	24	0.12	131	0.35
		US_no	5	1	0.14	140	0.43
		US_less	66	3	0.41	153	0.81
		LSTM	0	-	0.06	153	0.29
		PUB	14	26	0.05	147	0.12
		M4SE	0	-	0.07	136	0.29
		obs	0	8	0.09	145	0.34

Table F.4: Part 2 of Table F.3

F.2 HISTOGRAMS OF ERA5 FORCING PARAMETERS

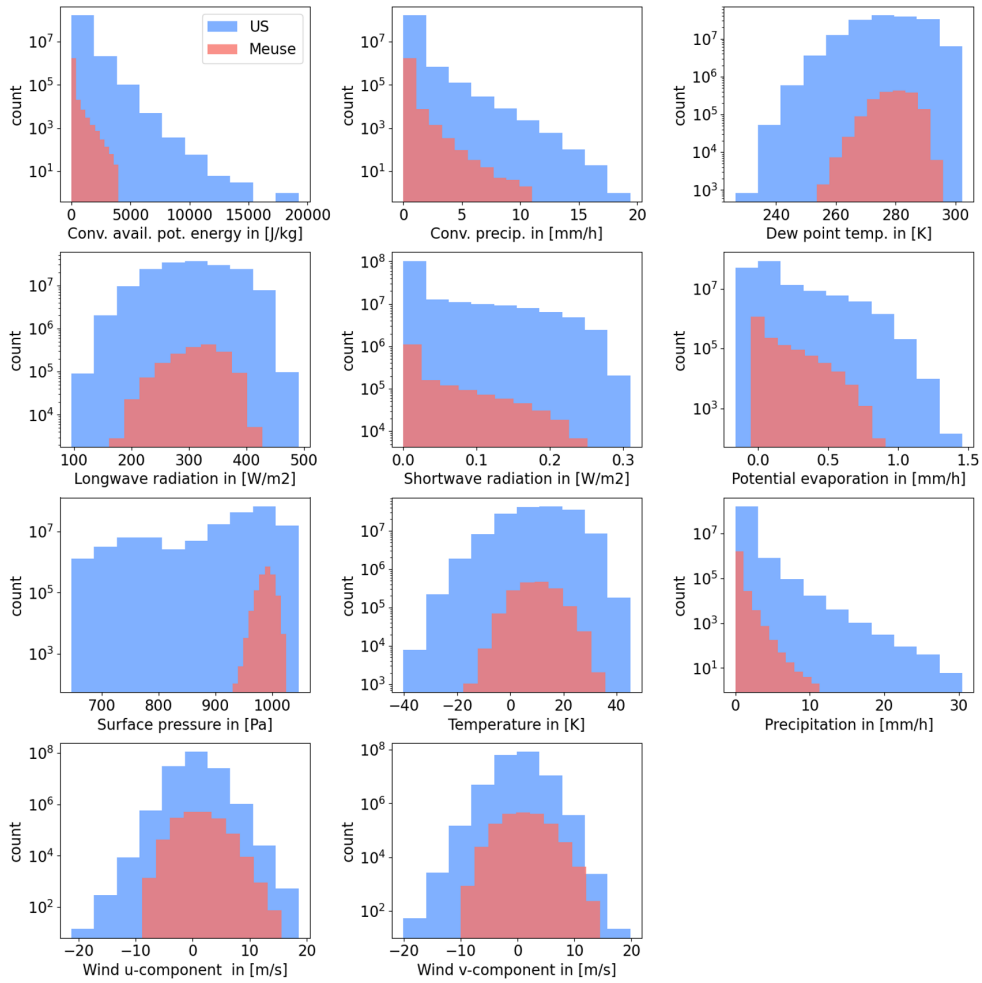


Figure F.1: Histograms of ERA5 forcing parameters based on data from 516 US catchments (blue) and 5 Meuse catchments (red).

F.3 FLOW DURATION CURVES

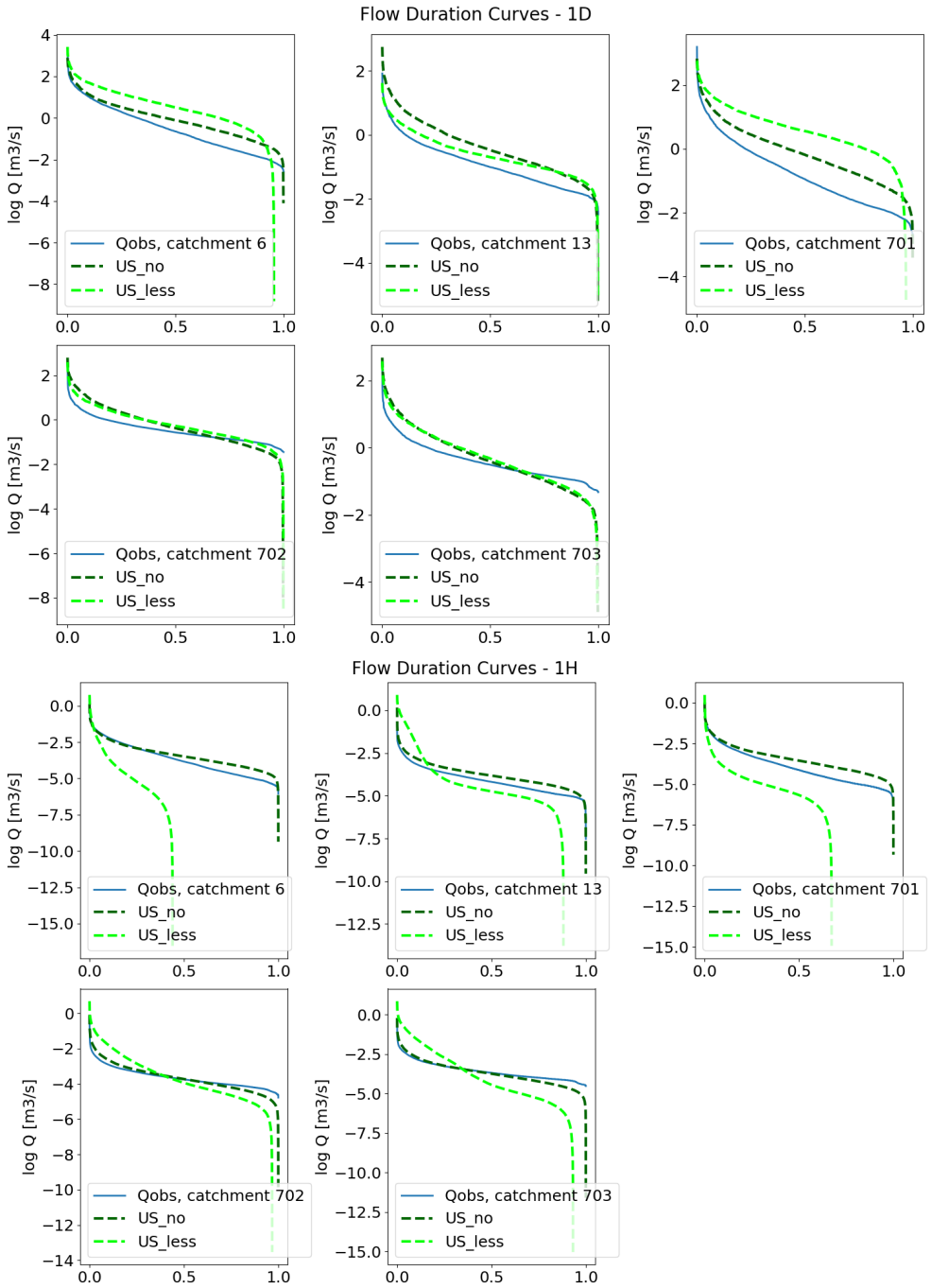


Figure F.2: Logarithmic FDCs for Meuse catchments with US-trained MTS-LSTM without static input (US_no, dark green) and with less static attributes (US_less, lime).

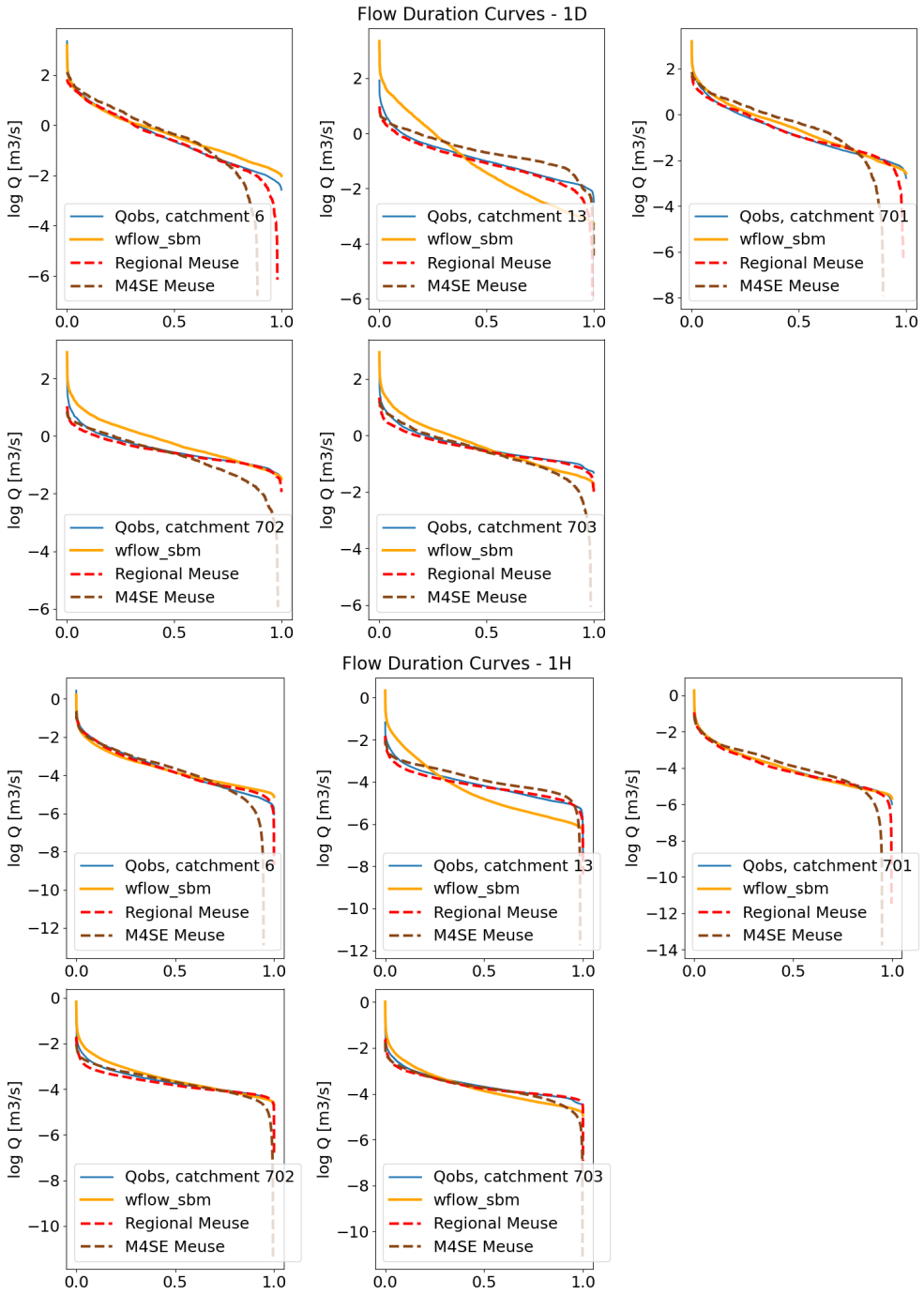


Figure F.3: Logarithmic FDCs for Meuse catchments, derived from observed streamflow (blue), predictions from *wflow_sbm* (orange), regional Meuse MTS-LSTM trained with NSE loss function (red) and regional Meuse MTS-LSTM trained with M4SE loss function (brown).

F.4 TIME SERIES PLOTS

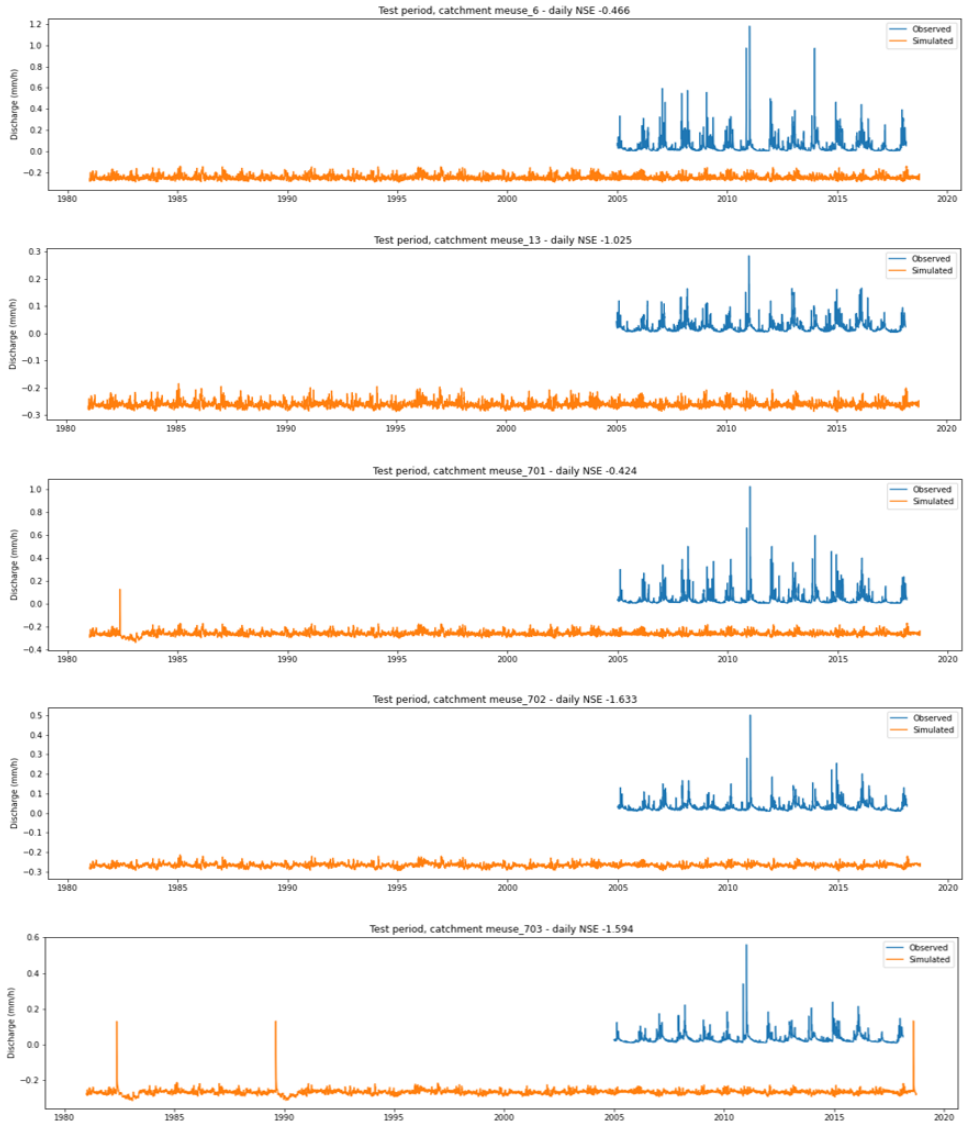


Figure F.4: First result of shifted and scaled streamflow simulations for test catchments in Meuse basin with MTS-LSTM trained on ERA5 and HydroMT data of US catchments

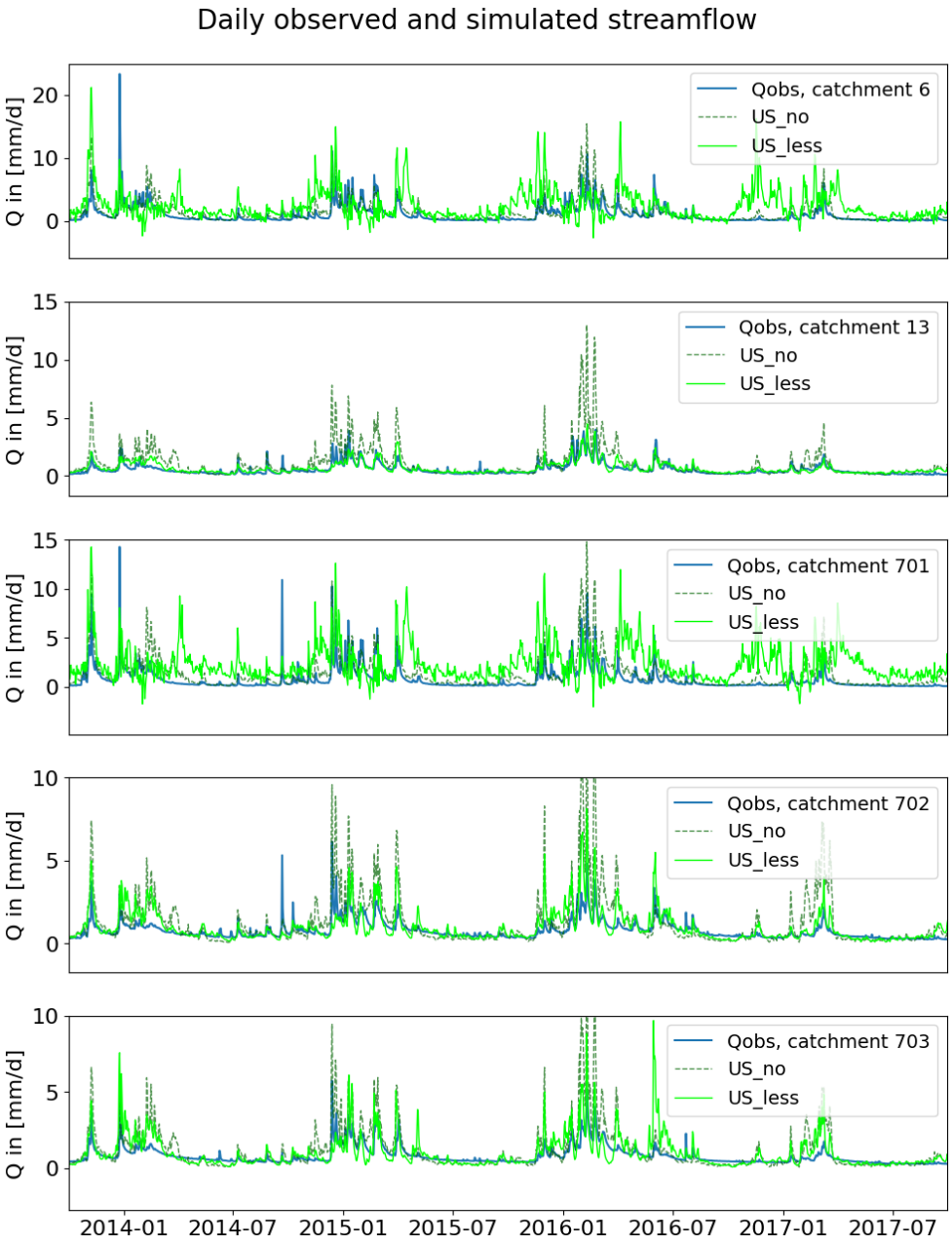


Figure F.5: Hydrographs for streamflow from US-trained MTS-LSTM with no static input (US_no, dark green, dashed plot) and less static input parameters (US_less, lime) compared to observations, daily results.

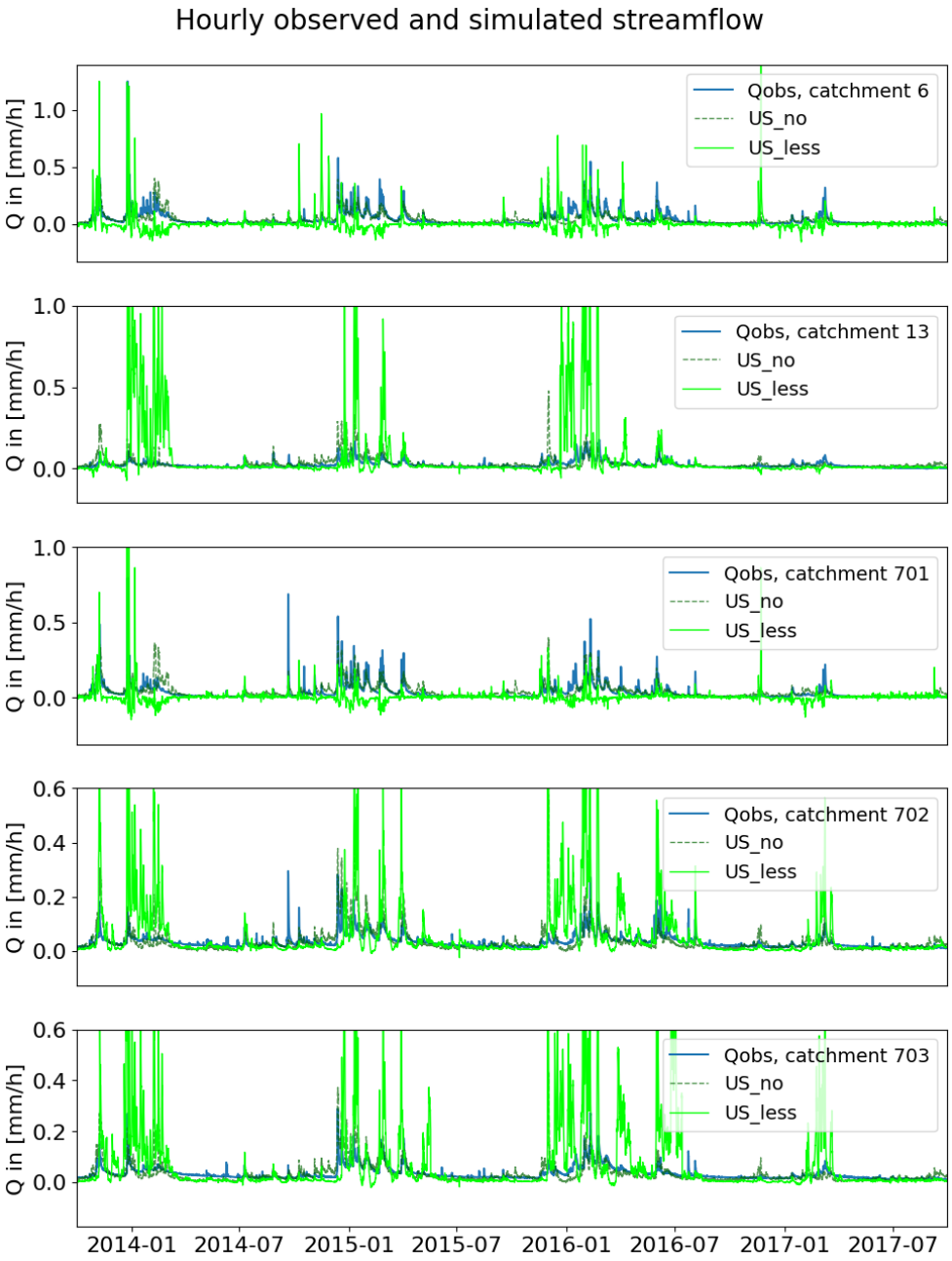


Figure F.6: Hydrographs for streamflow from US-trained MTS-LSTM with no static input (US_no, dark green, dashed plot) and less static input parameters (US_less, lime) compared to observations, hourly results.

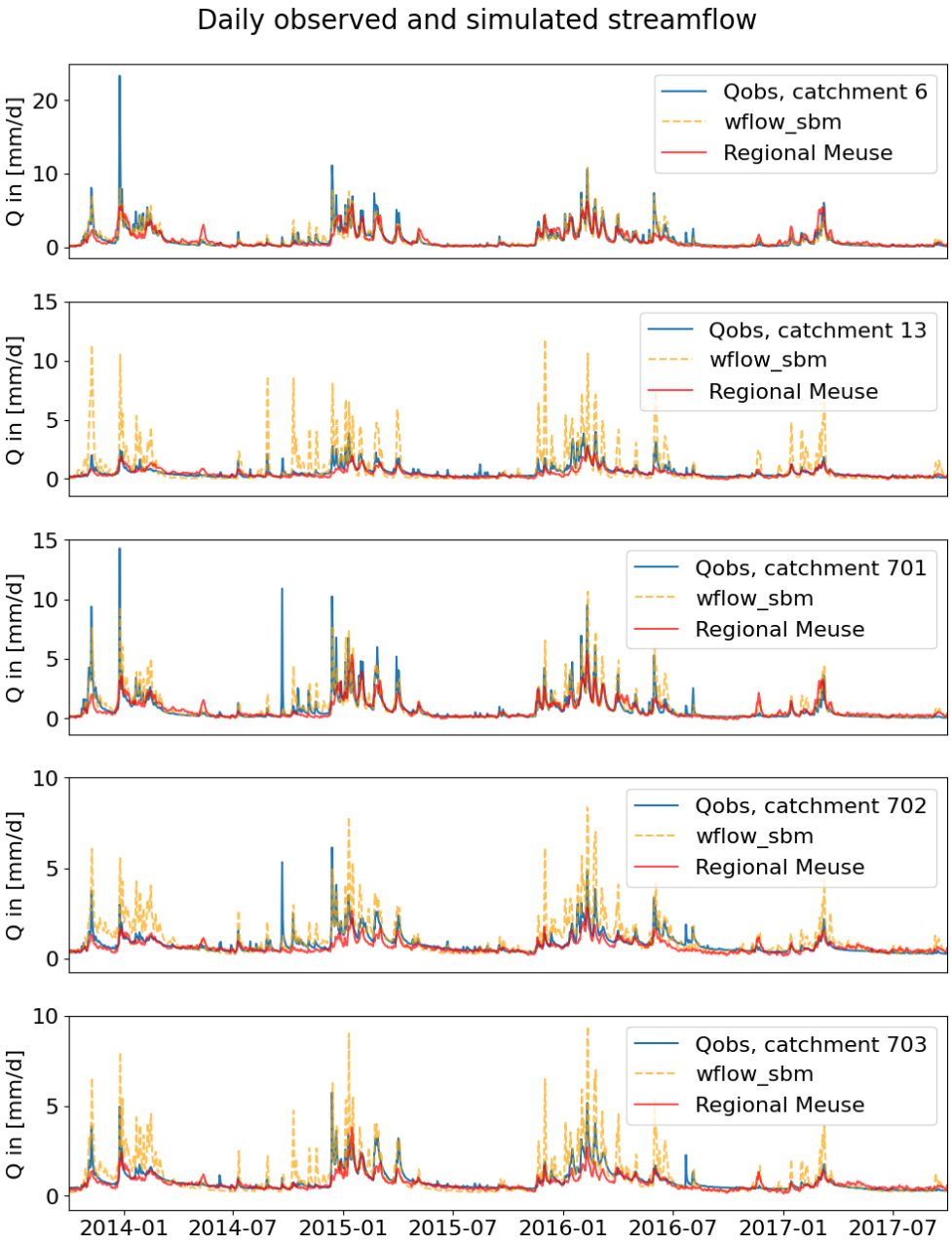


Figure F.7: Hydrographs for streamflow for Meuse catchments from observed streamflow (blue) and streamflow modeled with *wflow_sbm* (orange) and the regional Meuse MTS-LSTM (red), daily results.

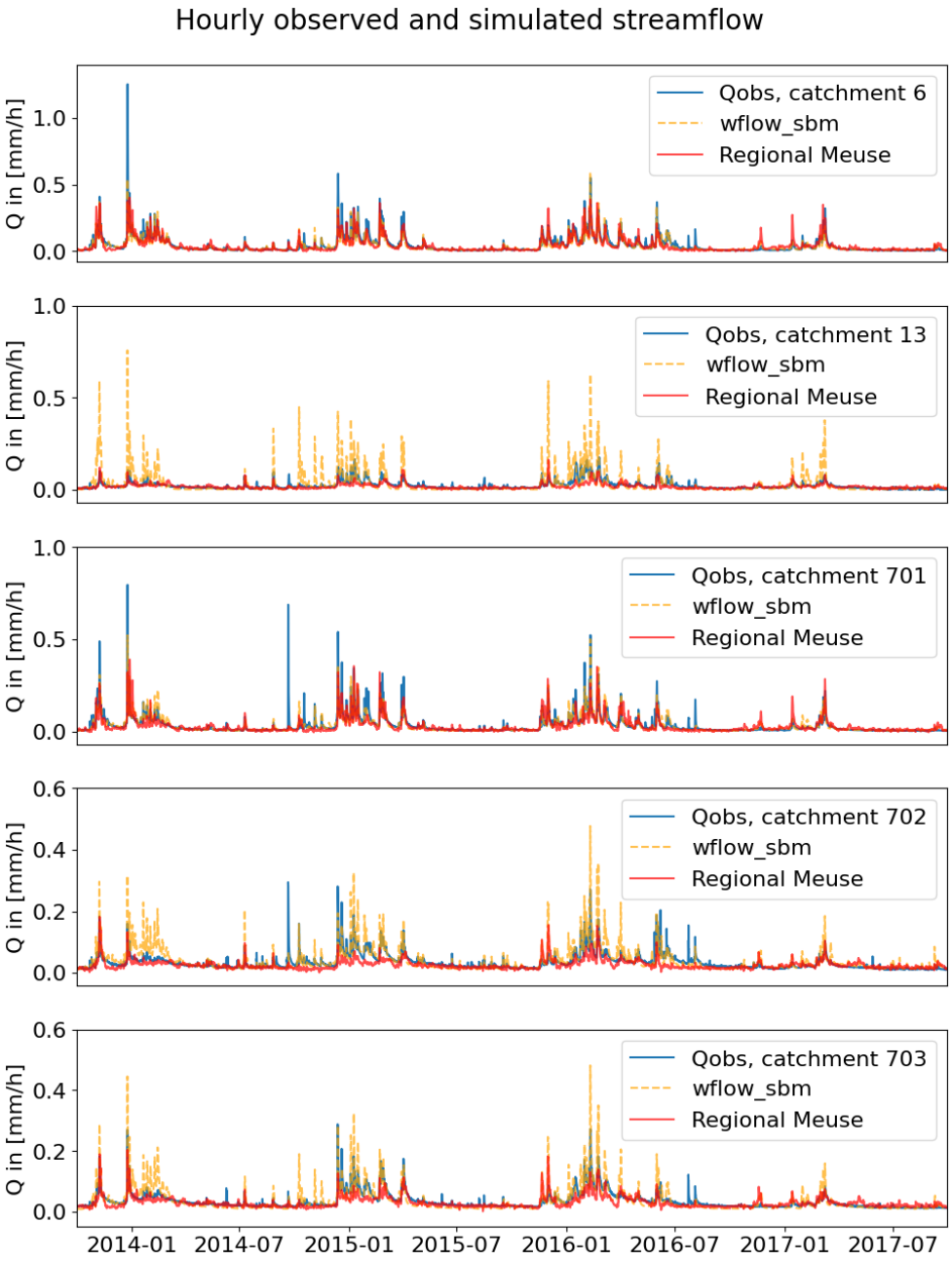


Figure F.8: Hydrographs for streamflow for Meuse catchments from observed streamflow (blue) and streamflow modeled with *wflow_sbm* (orange) and the regional Meuse MTS-LSTM (red), hourly results.

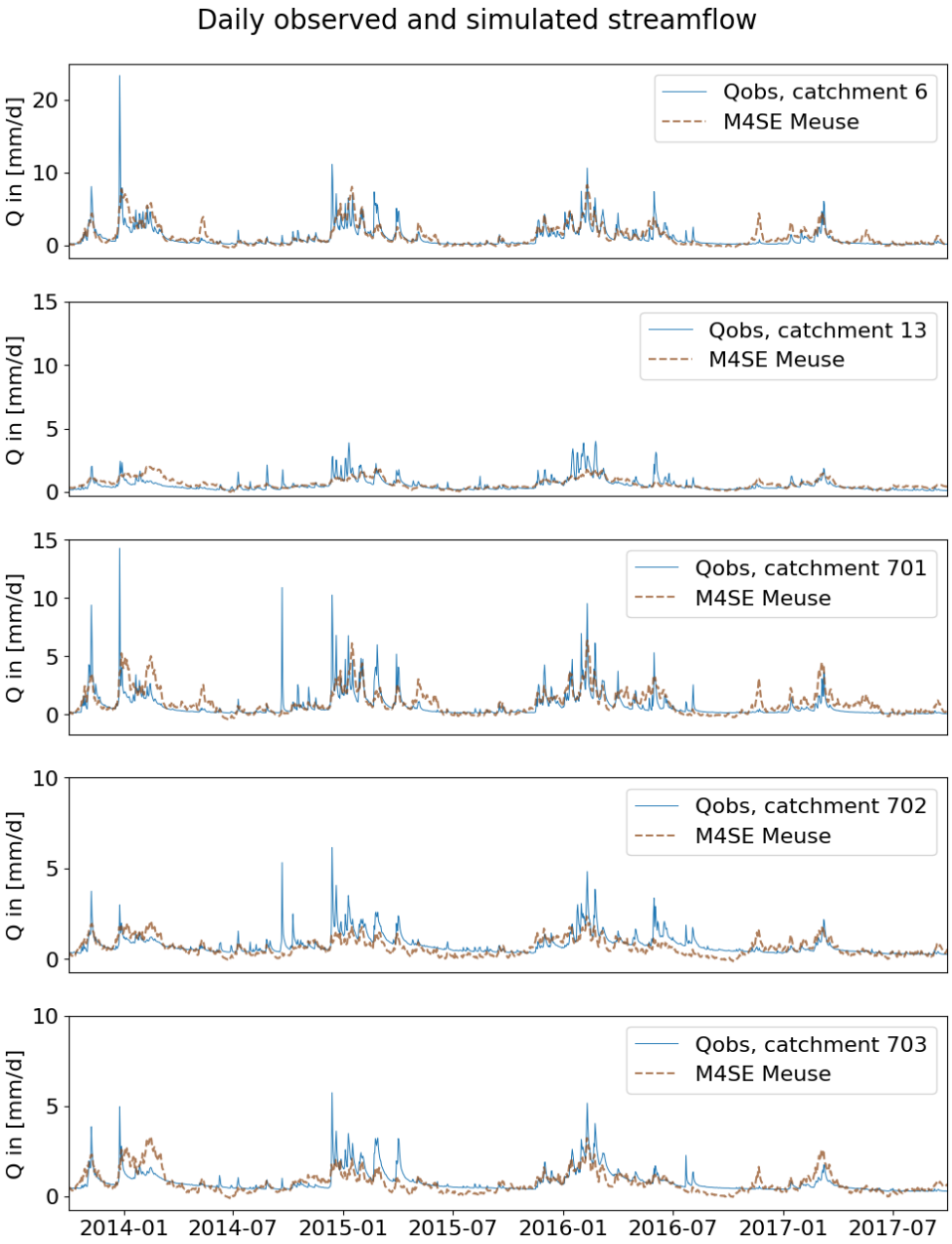


Figure F.9: Hydrographs for streamflow from MTS-LSTM with M4SE as loss function (brown) compared to observations (blue), daily results.

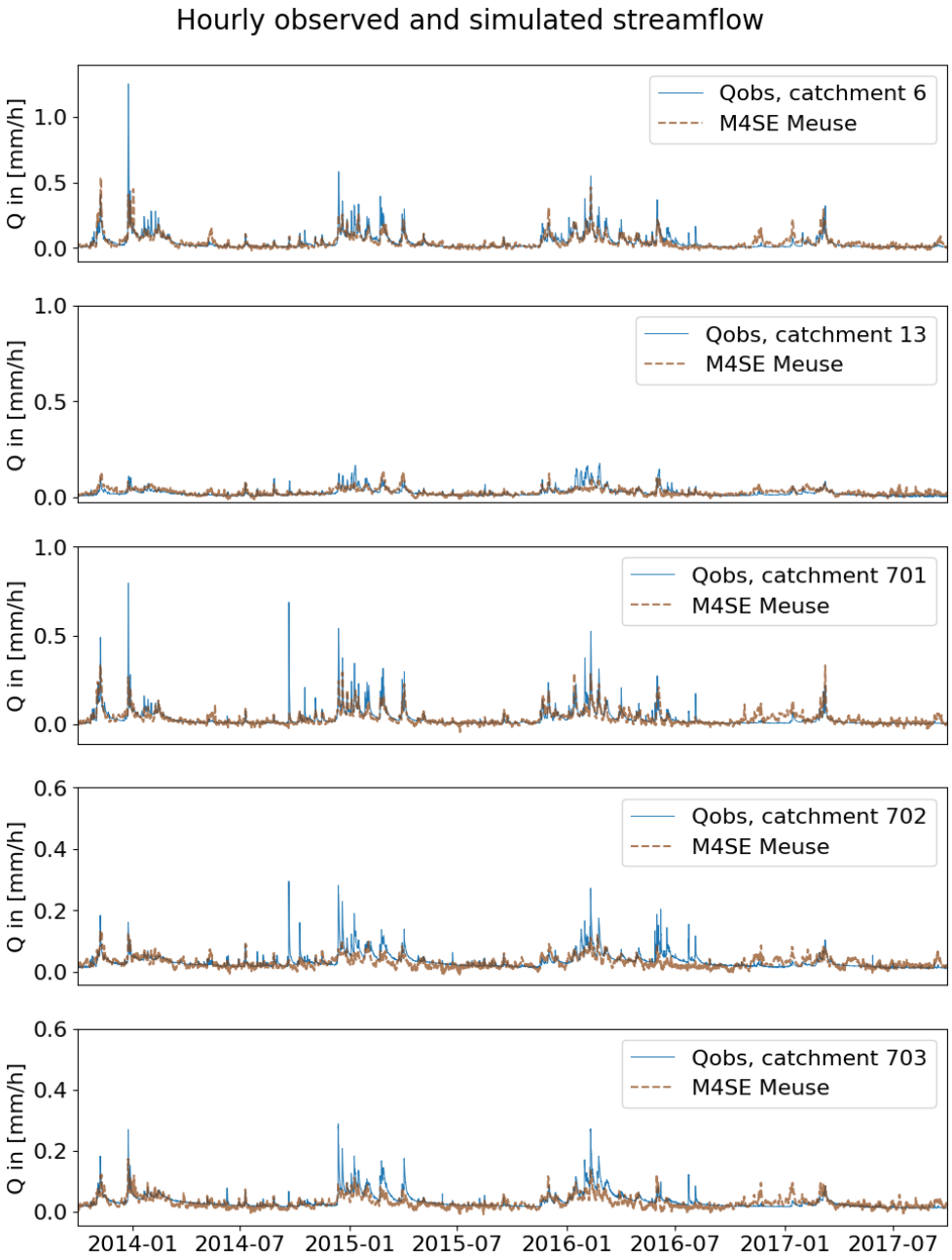


Figure F.10: Hydrographs for streamflow from MTS-LSTM with M4SE as loss function (brown) compared to observations (blue), hourly results.

G BIBLIOGRAPHY

Addor, N. et al. (2017). "The CAMELS data set: catchment attributes and meteorology for large-sample studies". In: *Hydrology and Earth System Sciences* 21.10, pp. 5293–5313. DOI: 10.5194/hess-21-5293-2017.

Addor, N. et al. (2018). "A Ranking of Hydrological Signatures Based on Their Predictability in Space". In: *Water Resources Research* 54.11, pp. 8792–8812. DOI: 10.1029/2018WR022606.

Ayzel, G. et al. (2020). "Streamflow prediction in ungauged basins: benchmarking the efficiency of deep learning". In: *E3S Web Conf.* 163, p. 01001. DOI: 10.1051/e3sconf/202016301001.

Beck, H. E. et al. (2017). "MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data". In: *Hydrology and Earth System Sciences* 21.1, pp. 589–615. DOI: 10.5194/hess-21-589-2017.

Bell, B. et al. (2020). "ERA5 hourly data on pressure levels from 1950 to 1978 (preliminary version)". In: Copernicus Climate Change Service (C3S) Climate Data Store (CDS). URL: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels-preliminary-back-extension?tab=overview.

Blöschl, G. et al. (2019). "Twenty-three unsolved problems in hydrology (UPH) – a community perspective". In: *Hydrological Sciences Journal* 64.10, pp. 1141–1158. DOI: 10.1080/02626667.2019.1620507.

Bouaziz, L. J. E. et al. (2020). "Improved Understanding of the Link Between Catchment-Scale Vegetation Accessible Storage and Satellite-Derived Soil Water Index". In: *Water Resources Research* 56.3. DOI: 10.1029/2019WR026365.

Bouaziz, L. J. E. et al. (2021). "Behind the scenes of streamflow model performance". In: *Hydrology and Earth System Sciences Discussions* 25, 1069–1095. DOI: 10.5194/hess-25-1069-2021.

Bruin, H. A. R. de et al. (2016). "A Thermodynamically Based Model for Actual Evapotranspiration of an Extensive Grass Field Close to FAO Reference, Suitable for Remote Sensing Application". In: *Journal of Hydrometeorology* 17.5, pp. 1373–1382. DOI: 10.1175/JHM-D-15-0006.1.

Budyko, M. I. (1974). *Climate and life*. Academic press.

Gauch, M. et al. (Oct. 2020). *Data for "Rainfall-Runoff Prediction at Multiple Timescales with a Single Long Short-Term Memory Network"*. DOI: 10.5281/zenodo.4072701.

Gauch, M. et al. (2021). "Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network". In: *Hydrology and Earth System Sciences* 25.4, pp. 2045–2062. DOI: 10.5194/hess-25-2045-2021.

GRDC (4.11.2021). *GRDC stations with monthly data*. Accessed: 2021-11-14. URL: https://www.bafg.de/GRDC/EN/Home/homepage_node.html.

Gupta, H. V. et al. (2009). "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling". In: *Journal of Hydrology* 377.1, pp. 80–91. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2009.08.003.

Hrachowitz, M. and M. P. Clark (2017). "HESS Opinions: The complementary merits of competing modelling philosophies in hydrology". In: *Hydrology and Earth System Sciences* 21.8, pp. 3953–3973. DOI: 10.5194/hess-21-3953-2017.

Hrachowitz, M. et al. (2013). "A decade of Predictions in Ungauged Basins (PUB)—a review". In: *Hydrological Sciences Journal* 58.6, pp. 1198–1255. DOI: 10.1080/02626667.2013.803183.

Imhoff, R. O. et al. (2020). "Scaling Point-Scale (Pedo)transfer Functions to Seamless Large-Domain Parameter Estimates for High-Resolution Distributed Hydrologic Modeling: An Example for the Rhine River". In: *Water Resources Research* 56.4. DOI: 10.1029/2019WR026807.

Jones, J. A. et al. (Apr. 2012). "Ecosystem Processes and Human Influences Regulate Streamflow Response to Climate Change at Long-Term Ecological Research Sites". In: *BioScience* 62.4, pp. 390–404. ISSN: 0006-3568. DOI: 10.1525/bio.2012.62.4.10.

Karim, R. (2018). *Animated RNN, LSTM and GRU*. URL: https://towardsdatascience.com/animated-rnn-lstm-and-gru-ef124d06cf45.

Konikov, L.F. (2013). "Groundwater depletion in the United States (1900-2008): U.S. Geological Survey Scientific Investigations Report 2013-5079". In: URL: http://pubs.usgs.gov/sir/2013/5079.

Kratzert, F. et al. (2018). "Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks". In: *Hydrology and Earth System Sciences* 22.11, pp. 6005–6022. DOI: 10.5194/hess-22-6005-2018.

Kratzert, F. et al. (2019a). "NeuralHydrology – Interpreting LSTMs in Hydrology". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, pp. 347–362.

Kratzert, F. et al. (2019b). "Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning". In: *Water Resources Research* 55.12, pp. 11344–11354. DOI: 10.1029/2019WR026065.

Kratzert, F. et al. (2019c). "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets". In: *Hydrology and Earth System Sciences* 23.12, pp. 5089–5110. DOI: 10.5194/hess-23-5089-2019.

LeCun, Y., Y. Bengio, and G. Hinton (2015). "Deep learning". In: *Nature* 521, pp. 436–44. DOI: 10.1038/nature14539.

Luce, C. (2014). "Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales". In: *Eos, Transactions American Geophysical Union* 95.2, pp. 22–22. DOI: 10.1002/2014EO020025.

Nash, J.E. and J.V. Sutcliffe (1970). "River flow forecasting through conceptual models part I — A discussion of principles". In: *Journal of Hydrology* 10.3, pp. 282–290. ISSN: 0022-1694. DOI: 10.1016/0022-1694(70)90255-6.

NeuralHydrology (2020). *neuralhydrology*. https://github.com/neuralhydrology.

Nielsen, M. (2015). *Neural Networks and Deep Learning*. Determination Press. URL: http://neuralnetworksanddeeplearning.com/.

Razavi, T. and P. Coulibaly (2013). "Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods". In: *Journal of Hydrologic Engineering* 18.8, pp. 958–975. DOI: 10.1061/(ASCE)HE.1943-5584.0000690.

Richey, A. S. et al. (2015). "Quantifying renewable groundwater stress with GRACE". In: *Water Resources Research* 51.7, pp. 5217–5238. DOI: 10.1002/2015WR017349.

Searcy, J. K. (1959). *Flow-duration curves*. 1542. US Government Printing Office. DOI: 10.3133/wsp1542A.

Sharma, S. (2017). "Activation Functions in Neural Networks". In: *towards data science*. URL: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6.

Shen, C. (2018). "A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists". In: *Water Resources Research* 54.11, pp. 8558–8593. DOI: 10.1029/2018WR022643.

Verseveld, W. van et al. (2020). "Wflow.jl". In: DOI: 10.5281/zenodo.5679039. URL: https://github.com/Deltares/Wflow.jl.

Vertessy, R. A. and H. Elsenbeer (1999). "Distributed modeling of storm flow generation in an Amazonian rain forest catchment: Effects of model parameterization". In: *Water Resources Research* 35.7, pp. 2173–2187. DOI: 10.1029/1999WR900051.

Xia, Y. et al. (2012). "Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products". In: *Journal of Geophysical Research: Atmospheres* 117.D3. DOI: 10.1029/2011JD016048.

Yamazaki, D. et al. (Dec. 2017). "MERIT DEM: A new high-accuracy global digital elevation model and its merit to global hydrodynamic modeling". In: *AGU Fall Meeting Abstracts*. Vol. 2017, H12C–04. URL: https://ui.adsabs.harvard.edu/abs/2017AGUFM.H12C..04Y.

Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008). "A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model". In: *Water Resources Research* 44.9. DOI: 10.1029/2007WR006716.