https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

Fully differentiable, fully distributed Rainfall-Runoff Modeling

Fedor Scholz[1], Manuel Traub[1], Christiane Zarfl[2], Thomas Scholten[3], and Martin V. Butz[1]

[1]Neuro-Cognitive Modeling Group, University of Tübingen, Tübingen, Germany
[2]Environmental Systems Analysis, University of Tübingen, Tübingen, Germany
[3]Soil Science and Geomorphology, University of Tübingen, Tübingen, Germany

Correspondence: Fedor Scholz (fedor.scholz@uni-tuebingen.de)

Abstract. Traditional hydrological modeling simulates rainfall-runoff dynamics using process-based models (PBMs), which are grounded in physical laws and therefore highly interpretable. Due to environmental systems being highly complex, though, sub-processes are sometimes hard or even impossible to identify and quantify. Alternatively, data-driven approaches, like deep neural networks (DNNs), can automatically discover relationships within the data, which often leads to superior performance. Due to DNNs' complexity, however, these relationships are hard to investigate and often fail to respect physical laws. *Differentiable modeling* calls for knowledge discovery by combining both approaches to benefit from their respective advantages. In this work, we present DRRAiNN (Distributed Rainfall-Runoff ArtIficial Neural Network), a targeted neural network architecture that successfully estimates river discharge based on meteorological forcings and elevation in the Neckar river basin, relying on daily water discharge measurements from only 17 stations. We evaluate our model against the European Flood Awareness System (EFAS) reanalysis on the Neckar river catchment in Southwest Germany, where some instances of our model outperform EFAS at lead times of over 100 days. Our model architecture is physically inspired, fully differentiable, and fully distributed. This combination enables the use of efficient source allocation algorithms, which help us identify the water sources responsible for the water discharge dynamics at specific gauging stations. In the future, this approach could be utilized to, e.g., infer erosion sites from turbidity data when integrated with an appropriate erosion model.

1 Introduction

Accurate water flow forecasting plays a critical role in mitigating short-term flood impacts, such as preventing loss of life and reducing economic damage (Pilon, 2002). For example, simulating river discharge empowers us to make informed decisions in water management such as dam operations (Valeriano et al., 2010). Accuracy is not everything though. Good hydrological models should respect physical laws to be able to generalize well to new situations, and to offer insights into the underlying processes that govern water movement. A solid understanding of the dynamics of water systems is necessary to estimate the impacts of environmental planning and to improve infrastructure design (Palmer et al., 2008; Bharati et al., 2011). It also enables a better assessment of how climate change might alter existing ecosystems in the future (Palmer et al., 2008; Van Vliet et al., 2013; Al Hossain et al., 2015). Additionally, models that respect physical laws can be used to infer the origins of observed discharge, thereby further facilitating the development of policies that mitigate the damages caused by floods. On a practical level, a good model should be easily tuned and perform well even if data is sparse, which is often the case for river discharge.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

To address some of these challenges, hydrologists traditionally employ process-based models (PBMs) that describe physical processes with mathematical equations derived from physical laws and observations (Brutsaert, 2023). This renders PBMs inherently interpretable, allowing researchers to ask specific questions by probing them. Some components of PBMs might be inferred from experiments in a laboratory, such as Darcy's law (Darcy, 1856). Others are based on simplifications of more general physics equations. To simplify models further, sometimes lumped approaches are employed. Here, meteorological forcings like precipitation are averaged over time and space on the basin scale. Therefore, the outline of the basin must be available a priori for this approach to be feasible. The outline is usually inferred from a digital elevation model, thereby potentially disregarding underground flows and pipes.

Real-world processes can be very complex with lots of variables that make the overall process highly heterogeneous (Marçais and de Dreuzy, 2017). Recent advances, such as the Multiscale Parameter Regionalization (MPR) framework (Samaniego et al., 2010) and scalable transfer function approaches (Imhoff et al., 2020) have focused on improving parameterization and capturing spatial heterogeneity in distributed hydrological models to alleviate these issues. To reduce uncertainty and initialize PBMs adequately, data assimilation incorporates concrete observations into running models (Liu et al., 2012; Camporese and Girotto, 2022; Montzka et al., 2012). Advancements in data assimilation can improve performance both in lumped (Moradkhani et al., 2005; Liu and Gupta, 2007; Liu et al., 2012) as well as distributed models (Rakovec et al., 2012). Nevertheless, usually, not all involved sub-processes and their interactions are known (Hrachowitz et al., 2013), leading to high uncertainty and biases. Even if a sub-process is known, though, certain input variables may simply be unobservable, such as underground topography. Additionally, the scales of real-world processes may be very different compared to lab conditions (Hrachowitz et al., 2013; Shen, 2018; Nearing et al., 2021).

As is the case for many sciences, the amount of available data for hydrology is increasing significantly (Sit et al., 2020). But even with abundant data, PBMs struggle to fully exploit them: Only little data is used to adapt their parameters, which nevertheless is a lengthy and costly process since it is often done manually by experts (Shen et al., 2023). Since the parameters are usually calibrated for a specific basin only, these tuned models do not generalize well to other basins with different properties (Hrachowitz et al., 2013; Nearing et al., 2021). This is especially problematic for countries of the Global South for which no detailed land surface and subsurface measurements are available but where intelligent water management might be even more important.

The discovery of relationships in huge amounts of data is a challenging task for which ANNs can be a remedy. In contrast to PBMs, ANN approaches are mainly data-driven, allowing them to automatically find relationships in the training data. Given the superior performance of early data-driven approaches in hydrology, there is reason to assume that the potential of data-driven approaches has by no means been fully exploited, yet (Shen, 2018; Nearing et al., 2021). However, the relationships found by ANNs often remain latent due to their high complexity. ANNs contain huge amounts of parameters and states that usually do not directly relate to real-world quantities. Training them is feasible only due to gradient descent via automatic differentiation. This process is lengthy and costly. Once an ANN is trained, though, inference is very fast and cheap. Their high complexity also often leads to neural networks not respecting physical laws despite very good performance, in turn leading to bad generalization capabilities. This calls for measures that guide data-driven approaches toward physical plausibility.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

While other Earth sciences started to adopt machine learning (ML) techniques, hydrology seems to lag behind (Shen, 2018). There is a significant amount of distrust in the community toward non-physical models (Blöschl et al., 2019). But this is not the only reason: In line with the above-mentioned lack of coordination, the field is missing benchmarks that are easy to access and enable a fair comparison between models (Hrachowitz et al., 2013; Sit et al., 2020; Nearing et al., 2021), although there are some recent efforts (Kratzert et al., 2023). Nevertheless, the amount of available data is steadily increasing. This is exactly where ML techniques offer great potential. They can be trained on vast amounts of data and infer relationships in the data automatically. It was already shown that ML often outperforms traditional approaches with regards to accuracy (Shen, 2018).

It is often criticized, that ML modelers do not put enough effort into the interpretation of their developed systems to gain a better understanding of their internal dynamics (Muñoz-Carpena et al., 2023). As mentioned above, a considerable amount of runoff is situated below the ground and therefore not observable. It is not yet possible to see through the ground, rendering the underground topography latent (Shen, 2018). We believe that these latent variables are the reason that models often generalize poorly to other basins. However, this is another problem where ML and especially ANNs can help since latent variables can be inferred retrospectively from observations (Butz et al., 2019; Otte et al., 2020). One question we will ask in this paper is: *Given the observed dynamics, in which areas did precipitation contribute to the measured discharge?* A similar argument can be made for evapotranspiration, which is not directly observable as well. This kind of model inversion (Sit et al., 2020) therefore renders another possibility to extend our understanding of the water cycle with ML. For a comprehensive review of modern ML in hydrology, we refer the reader to Shen (2018).

A combination of the above-mentioned approaches could leverage the advantages of both worlds. If pursued with the goal of knowledge discovery, this combination was recently coined “differentiable modeling” (Shen et al., 2023). It could result in well-performing interpretable models that automatically find new relationships in the data, respect physical laws and therefore generalize well and need comparatively little data. It offers two different perspectives: Coming from traditional modeling approaches, the incorporation of differentiable model parts allows to automatically close knowledge gaps. This could manifest in various ways, such as estimating unknown parameters or representing entire sub-processes using approaches like artificial neural networks.

Coming from the ML side, relationships that are known to be true can be incorporated into already differentiable models as constraints, or inductive biases. These inductive biases introduce prior assumptions about the data-generating process, effectively constraining the model’s solution space. By doing so, they can improve performance, enhance generalization, and make learning more efficient. Furthermore, they help guide the model towards discovering meaningful structures in the data, aligning its behavior with established principles (Butz et al., 2024).

Our work builds on differentiable modeling, presenting DRRAiNN (Distributed Rainfall-Runoff ArtIficial Neural Network), a physics-inspired, fully differentiable, fully distributed rainfall-runoff model. Our targeted spatio-temporal artificial neural network (ANN) architecture estimates river discharge at given measurement stations from gridded precipitation, solar radiation, elevation, and past discharge. This approach poses a severe challenge to ANN-based learning approaches, though: the targeted river discharge data is very sparse. To avoid overfitting and to improve interpretability and generalization, we had to incorporate several physics-inspired inductive biases into DRRAiNN. One of these biases is the modularization into a spatially fully

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

distributed rainfall-runoff model and a graph-based river discharge model. Another one preconditions DRRAiNN to represent the sub-processes of lateral propagation of water over the landscape and evapotranspiration in certain sub-components. With this, we hope to show that deep learning approaches must not necessarily be perceived as purely black boxes. Instead, they can be designed to contain meaningful components that correspond to sub-processes of the overall real-world process. To a certain extent, their inner workings can be interrogated with interpretability methods to show what their estimations are based on.

We thus focus on successfully designing and training such a fully differentiable model. We evaluate the DRRAiNN's estimation abilities, physical plausibility, and the need for our main design choices. We showcase its performance in a real-world setting on the Neckar River in Southwest Germany, comparing it with simulations from the European Flood Awareness System (EFAS, Mazzetti et al. (2023)). DRRAiNN outperforms EFAS with lead times of up to 45 days. Due to DRRAiNN being fully distributed and fully differentiable, our approach allows us to answer 'where' questions, such as: What is the true catchment area, including underground flows? That is, DRRAiNN opens up the possibility for performing source allocations using gradient-based techniques like integrated gradients (Sundararajan et al., 2017). These techniques can help in examining and understanding internal model dynamics, potentially leading to knowledge discovery and thereby further repealing the black box perspective. We show reconstructed catchment areas from observed dynamics, demonstrating the feasibility of attribution methods within DRRAiNN.

2 Related Work

In their seminal work, Kratzert et al. successfully use LSTMs (Hochreiter and Schmidhuber, 1997) for rainfall-runoff modeling on a daily scale (Kratzert et al., 2018). Since then, numerous studies have emerged, applying basically the same model to various data sets (Sit et al., 2020). Notably, significant advancements to the model have also been made, including the incorporation of physical constraints (Kratzert et al., 2019; Hoedt et al., 2021), uncertainty estimation (Klotz et al., 2022), and the extension of modeling to multiple timescales (Gauch et al., 2021). All of the above-mentioned models have in common that they are lumped models, i.e., forcings are spatially aggregated over the catchment area which was inferred from a digital elevation model. This practice is so prevalent that it is often not even mentioned in the papers. In semi-distributed modeling, models make limited use of the river topology (Xiang and Demir, 2020; Moshe et al., 2020; Sit et al., 2021; Kratzert et al., 2021). Here, the catchment area is divided into multiple sub-basins that can communicate with each other. Within each sub-basin, however, forcing are again spatially aggregated. In contrast, fully distributed models operate on a grid without any spatial aggregation. Even though there is a call for more fully distributed data-driven models for rainfall-runoff modeling (Nearing et al., 2021), not many approaches exist in the literature.

The model presented in (Xiang and Demir, 2022) indeed operates on a grid, but communication between neighboring cells is only possible in the direction of the steepest descent. This strong assumption essentially turns the grid into a sparser graph. The CNN-LSTMs presented in (Ueda et al., 2024; Pokharel and Roy, 2024a; Li et al., 2022) operate on a grid without any assumption of flow directions. However, the LSTMs are not applied in every grid cell but instead receive the flattened outputs of the CNNs, which renders the modeling of space and time completely separate.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

In (Schmidt et al., 2020), the authors applied a ConvLSTM (Shi et al., 2015) and found that it does not make use of spatial patterns, which made it perform worse than a lumped approach. In contrast, (Oddo et al., 2024) used a ConvLSTM to estimate river discharge with a one-hour lead time by flattening the outputs of all cells before feeding them into a linear layer. Similarly, (Longyang et al., 2024) used a modified ConvLSTM architecture combined with ridge regression to determine from which grid pixels the output should be aggregated to estimate discharge at the station. More examples of this kind can be found in the literature (Zhu et al., 2023; Tyson et al., 2023; Pokharel and Roy, 2024b; Xu et al., 2022). Since all of these models aggregate the outputs of the spatial component globally over space, the model lacks the incentive to propagate water across the landscape. We therefore assume that these models behave physically unrealistic and generalize poorly to other basins.

For a comprehensive list of applications and publications regarding machine learning and hydrology in general, we refer the reader to (Sit et al., 2020).

3 Methods

3.1 Study site

The Neckar river network in Southwest Germany has a catchment size of $14\,000\ \mathrm{km}^2$ and exhibits a heterogeneous landscape: It encompasses narrow and wide valleys, different kinds of rocks like limestone and sandstone, different types of soils like clay and marl, underground topographies like karst, and formations like aquifers. This makes the modeling of the Neckar River network a challenging endeavor. To give a concrete example, there are underground flows south of Pforzheim that route water toward the east, while the elevation model suggests a different flow direction. (Ufrecht, 2002). This relationship cannot be inferred from a digital elevation model alone: Latent underground structures route the water in a different direction than the elevation model alone would suggest.

The Neckar and its sub-catchments were subject to multiple distributed hydrological studies in the past (Imhoff et al., 2020; Samaniego et al., 2010; Schalge et al., 2021).

3.2 Data

We use the following data as input for DRRAiNN: radar-based precipitation, elevation for above-ground topography, solar radiation, and river discharge data. Preliminary experiments showed no improvement when including temperature, therefore we exclude it following Occam's razor. By restricting the domain to the Neckar River, we end up with an area of size $200\ \mathrm{km}^2$. After the transformations described in the paragraphs below, all gridded data is reduced from a $1\ \mathrm{km} \times 1\ \mathrm{km}$ grid to a $4\ \mathrm{km} \times 4\ \mathrm{km}$ grid by taking the mean. This results in a 50×50 pixel grid. We use the hydrological years 2006 to 2015 for training and 2016 to 2018 for validation.

For precipitation, we use the estimated product RADOLAN by the Deutsche Wetterdienst (RADOLAN, 2016), which is collected from radar stations that are distributed over Germany. The data domain is a $900\mathrm{km} \times 900\mathrm{km}$ pixel grid with a resolution of $1\mathrm{km} \times 1\mathrm{km}$ that covers all of Germany. This grid forms the basis of the grid our model operates on. RADOLAN

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

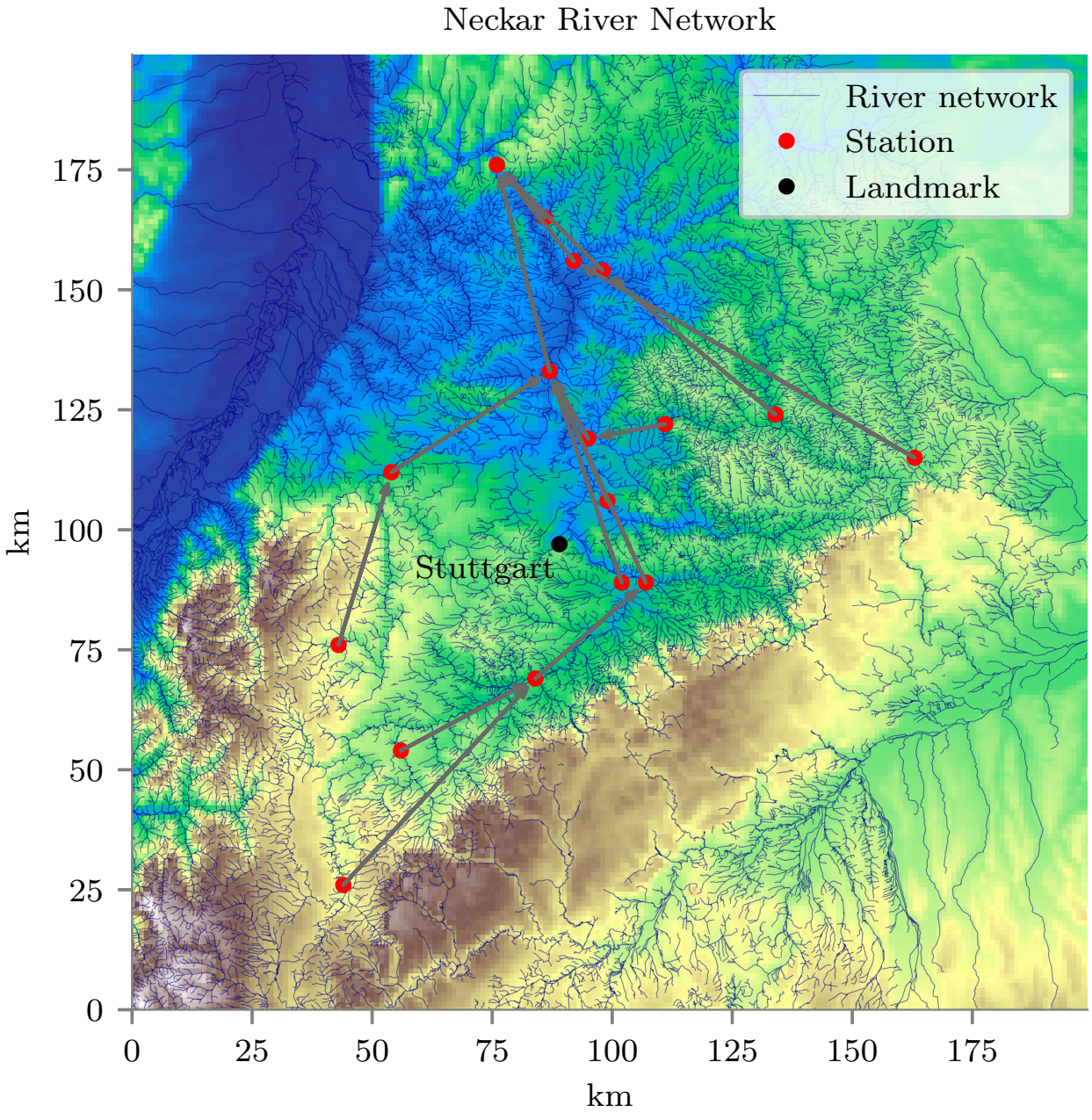

Figure 1. The study area used in this work is the Neckar River catchment in Southwest Germany.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

data is log-standardized before being sent to the model due to its long-tail distribution. This means we add 1, take the logarithm, subtract the mean, and divide by the standard deviation. We replace missing values with 0s, which is the log-standardized mean.

We derive static topography features from the digital elevation model (DEM) EU-DEM v1.1 by the European Union Copernicus Land Monitoring Service European Environment Agency (EU-DEM, 2016). We use rasterio (Gillies et al., 2013) to transform and reproject the data into the RADOLAN coordinate reference system. We also use the DEM to compute the differences in altitudes between adjacent discharge measurement stations. All these variables are standardized before being sent to the model, i.e., we subtract their mean and divide by their standard deviation.

For solar radiation, we use surface short-wave (solar) radiation downwards (SSRD) from the ERA5 data set (Hersbach et al., 2018). We use rasterio (Gillies et al., 2013) to transform and reproject the data into the RADOLAN coordinate reference system. Like the precipitation data, solar radiation data is log-standardized.

The topography of our river network is determined by the AWGN data set (AWGN, 2023). We use it to compute the adjacency matrix that describes which stations are connected via river segments and the corresponding river segment lengths.

Finally, we use discharge measurement data to tune in the discharge model and, more importantly, as the only target variable to train and validate our model. We use data collected and provided by the German Federal Institute of Hydrology via the Global Runoff Data Centre (GRDC, 2024). The data set contains observed daily river discharge from measurement stations across the world. Since the location information of the discharge measurement stations is partially wrong, we corrected them manually. We then try to snap the station locations to the river network. If this correction is larger than a certain threshold, the station is discarded. If two stations are very close to each other, one of them is discarded. Discharge data is log-standardized station-wise before being sent to the model due to its long-tail distribution. This means we add 1, take the logarithm, subtract the station-wise mean, and divide by the station-wise standard deviation. We replace missing values with 0s, which is the log-standardized mean of the corresponding station.

3.3 Model

We present DRRAiNN, a spatio-temporal artificial neural network architecture that estimates river discharge from static attributes and meteorological forcings in a distributed manner. The locations $L_i = (x_i, y_i)$ for estimations of discharge in the river network are determined by discharge measurement stations that provide observed discharge $Q_{i,t}$ for time t in 24h periods. The connectivity of stations, determined by the river network, is encoded in an adjacency matrix $A_{i,j}$. Static maps $S_{x,y}$ and meteorological forcings $F_{x,y,t}$ for hourly time points t are encoded on a grid that spans the whole catchment area of the river network. Given static maps $S_{:,:}$, meteorological forcings $F_{:,:,t_0:t_s+T}$ over the whole duration $(t_0 \dots t_s + T)$ in hours, and past discharge $Q_{i,t_0:t_s}$ over the tune-in period $(t_0 \dots t_s)$ in days, DRRAiNN f estimates future discharge $Q_{i,t_s+1:t_s+T}$ over a temporal future horizon of T days:

$$\tilde{Q}_{i,t_s+1:t_s+T} = f(S_{:,:}, F_{:,:,t_0:t_s+T}, Q_{i,t_0:t_s}) \quad (1)$$

,

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

In contrast to most other neural networks in hydrology, DRRAiNN includes a rainfall-runoff model that is fully spatially distributed: We do not lump variables across space over basins, but our model operates on a grid instead. Since surface/subsurface and river flow dynamics behave differently as described above, we model these sub-processes separately. Therefore, DRRAiNN consists of two components, the rainfall-runoff model and the discharge model. The rainfall-runoff model operates recurrently on a grid, rendering it fully distributed. It is supposed to model surface/subsurface flow and evapotranspiration. The discharge model operates recurrently on a graph and is supposed to model river flow inside of channels.

DRRAiNN processes a time series in the following manner: Over the whole sequence, we alternately call the rainfall-runoff model, which is implemented by a recurrent convolutional ANN, and the discharge model, which is implemented by a recurrent graph ANN. The rainfall-runoff model receives static landscape features and meteorological forcings as input to estimate runoff on a grid. It is primed to model two important sub-processes separately, namely surface/subsurface flow, which is mainly driven by topography, and evapotranspiration, which is mainly driven by solar radiation. Even though it cannot directly be interpreted as such, we call its output *runoff* since this is the main driver for the discharge model: The estimated runoff is collected at station locations and sent to the discharge model. The discharge model additionally receives an adjacency matrix that describes the connectivity between stations, static river segment features, and the (potentially estimated) discharge from the previous time step. It estimates discharge for each station, from which the loss is computed.

We implement DRRAiNN in pytorch (Paszke et al., 2019). In the following, we provide a more detailed description of DRRAiNN's components. See Fig. 2 for a depiction of the overall model.

3.3.1 Rainfall-Runoff Model

The rainfall-runoff model consists of a position-wise long short-term memory (LSTM) and a convolutional neural network (CNN) that are called in alternation. This renders the rainfall-runoff model local in space and time: Only neighboring and past information is used to update internal states.

3.3.2 Modeling temporal dynamics

The position-wise LSTM (PWLSTM) is responsible for modeling the temporal relationships in the data and therefore maintains a hidden and a cell state for each grid cell. The gating mechanism of the LSTM can shield the cell states from unwanted updates. It thus allows to maintain information over long periods. This can be particularly useful to implicitly model, e.g., soil moisture or groundwater levels, which exhibit slower dynamics than overland flow. The LSTM receives precipitation as input to update its hidden and cell states. It has a hidden size of 4. Importantly, the weights of the LSTM are shared throughout the gridded area. As a result, while the LSTM at each grid cell maintains individual hidden state values, the temporal processing principle is identical everywhere. The assumption is that the unfolding physics is the same everywhere, although they may be locally parameterized.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

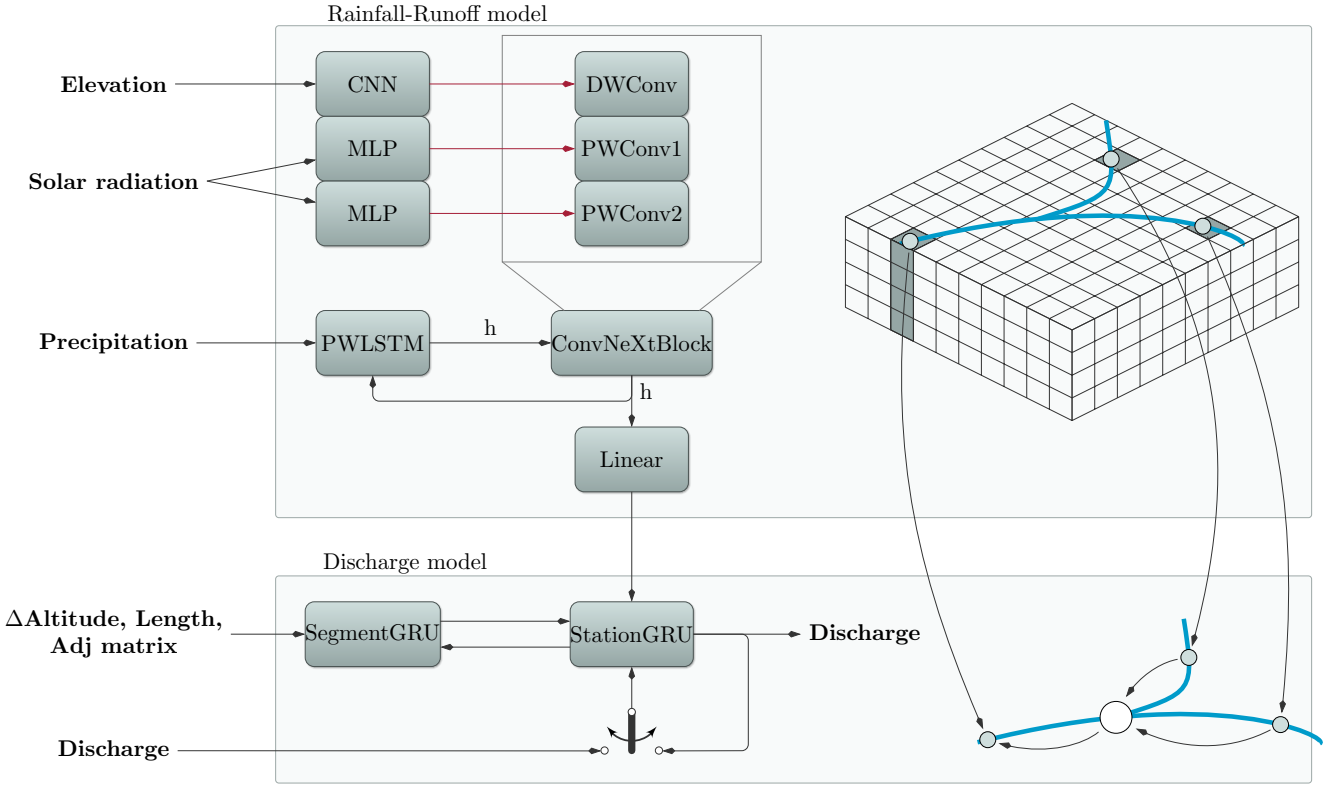

Figure 2. A detailed view of DRRAiNN. The gridded rainfall-runoff model's task is to propagate the received precipitation over the landscape according to the elevation and to model evapotranspiration based on solar radiation. It receives precipitation as its main input to the point-wise LSTM, whose hidden state is modified by the ConvNeXtBlock. The ConvNeXtBlock's weights are not static but produced by other neural networks. The depth-wise convolution's weights are produced by a convolutional neural network that has the same receptive field but receives elevation as input. Its main purpose is to model lateral propagation of water over the landscape. The point-wise convolutions' weights are produced by a multi-layer perceptron that receives solar radiation as input. Its main purpose is to model evapotranspiration, a process that is local in space. Before the hidden state is sent to the discharge model, it is processed by a simple linear layer. The graph-based discharge model then receives the processed state of the rainfall-runoff model at the measurement stations' locations and processes it together with the last (potentially inferred) discharge according to the adjacency of the stations, their differences in altitudes, and the river segment lengths between the stations. Its output is the discharge at each station.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

3.3.3 Modeling spatial dynamics

The CNN is responsible for modeling spatial relationships, i.e., the propagation of water flow over the landscape and evapotranspiration. It receives and updates the hidden state h of the PWLSTM, leaving the PWLSTM's cell states untouched. Note that surface/subsurface flow is a spatially extended process, while evapotranspiration is a local (though vertical) process that happens mostly independently of neighboring cells. We incorporate this into our architecture as an additional inductive bias.

More precisely, the CNN is given by a modified ConvNeXt block (Liu et al., 2022). A ConvNeXt block consists of three layers, namely a depth-wise convolutional layer (DWConv) with kernel size 7×7 followed by a position-wise inverted bottleneck given by two linear layers (PWConv1 and PWConv2). This way, ConvNeXt disentangles horizontal and vertical information flow. We use the SiLU activation function between all layers (Hendrycks and Gimpel, 2016). In contrast to its original formulation, the weights of our ConvNeXt block are not static. Rather, they are parameterized by other neural networks, turning this network component into a *hypernetwork* (Traub et al., 2024a). This means that the ConvNeXt block can behave differently at each location on the grid. Calling DWConv results in the following operation:

$$y_{i,j,c} = \sum_{m=-3}^{3} \sum_{n=-3}^{3} w_{i,j,m,n,c} \cdot x_{i+m,j+n,c}, \tag{2}$$

where y is the output, x the input, w are the weights produced by the hypernetwork, c is the considered channel, and i and j are coordinates. Note that we can still call this operation a convolution if we regard the input variables together with the weight-generating networks as the kernel.

We parameterize the different layers of the ConvNeXt block with different weight-generating networks that receive different inputs. The weights of DWConv are produced by a simple CNN that has the same kernel size as DWConv itself. The weights for PWConv1 and PWConv2 are produced by simple position-wise MLPs. By using different input variables for the different hypernetworks, we can distinguish between local and spatially extended processes. How water propagates over the landscape depends mainly on the topography, which is why we generate the weights of DWConv from elevation. Before feeding the elevation into the hypernetwork, we subtract the elevation of the center cell from the elevations of all other cells within each receptive field, since we are interested in slopes and not absolute elevation. Evapotranspiration, on the other hand, is a very local process and should therefore be modeled by the position-wise components. This is why we generate the weights for PWConv1 and PWConv2 from solar radiation. See Fig. 3 for an illustration.

3.3.4 Adapter

Lastly, the hidden states at the station locations are collected, fed through a single linear layer, and sent the the river discharge model. Collecting and summing up the hidden states of all cells on the corresponding upstream river segment showed a tendency to overfitting in preliminary experiments.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

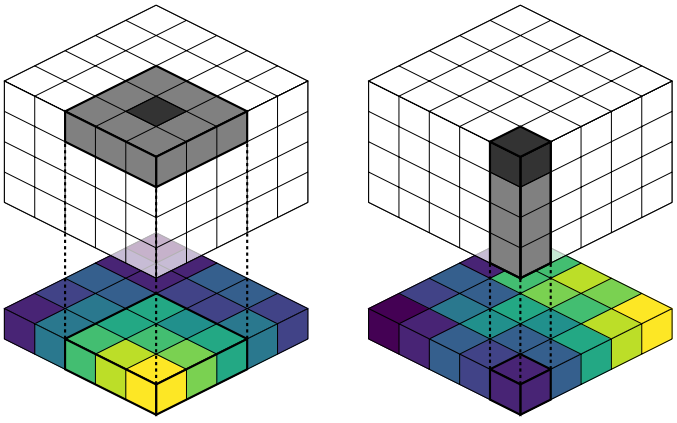

Figure 3. Illustration of the hypernetworks. The dark cells are updated based on the values of the light cells. The weights for these updates are produced by other neural networks that have the same receptive field but look at another type of data. Left: The weights for the depth-wise convolution are produced by a convolutional neural network that receives elevation as input. Right: The weights for the point-wise convolution are produced by a multi-layer perceptron that receives solar radiation as input.

3.3.5 Discharge Model

Our discharge model is a recurrent graph neural network, with the graph structure determined by the actual river network and the stations. It maintains two types of kernels, station and segment kernels, both of which are Gated Recurrent Units (GRUs) (Cho et al., 2014) with a hidden size of 8. Station kernels sit on the discharge measurement stations, segment kernels sit on the segments between those stations. They communicate with each other via lateral connections that have 4 channels (Fig. 2). The station kernel additionally receives the output of the rainfall-runoff model, the last (potentially inferred) discharge, and, based on that, estimates the discharge at that station. The segment kernel additionally receives the difference in altitude between the corresponding stations as well as the length of the river segment that it models. In general, both types of kernels work similarly except that the transition kernel receives an adjacency matrix according to which it aggregates data from upstream stations. The adjacency matrix is determined by the station locations and the river network topology. In a single time step, first, the transition kernels and subsequently the station kernels are called. Each kernel first concatenates its static, dynamic, and lateral inputs and then applies the GRU. In the case of the transition kernel, the output of the GRU is multiplied by the adjacency matrix, thereby summing up incoming information from upstream station kernels. Afterward, the tensor is split into dynamic and lateral outputs.

Even though we feed hourly meteorological forcings into DRRAiNN, we only produce daily discharge estimates. During the initial 10 days tune-in phase of each sequence, we therefore feed the same observed discharge value into DRRAiNN over one day. Additionally, we feed a daily marker into the station kernel (not depicted in the figure), which informs DRRAiNN about when a new day begins.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

Table 1. Truncation length schedule in days for TBPTT

#Epochs	Truncation length	Batch size
10	1	256
4	2	128
2	4	64
1	10	32
1	20	32

3.4 Experimental setup

We train DRRAiNN on sequences of 20 days, with the first 10 days serving as a warm-up phase. During this phase, observed discharge values are assimilated into the model to initialize its hidden states and align them with the system's dynamics. This approach is akin to data assimilation in traditional hydrological models, where observations are used to update model states and reduce uncertainty. In machine learning, this closed-loop setup is called teacher forcing. The warm-up phase allows the rainfall-runoff component of DRRAiNN to potentially estimate quantities like soil moisture or aquifer recharge within its hidden states. It therefore allows the model to align the hidden states with the physical state of the system before transitioning to predictive, open-loop mode.

Following the warm-up phase, DRRAiNN transitions into an open-loop mode for the remaining 10 days of the sequence. In this predictive mode, the discharge model feeds its own discharge estimations into subsequent time steps. The rainfall-runoff model on the other hand continues to be provided with historical observed precipitation and solar radiation. This is not a realistic setting for operational discharge forecasting. Especially precipitation forecasting is a hard problem and currently no algorithm exists that could accurately predict precipitation 10 days into the future on a $4\ \mathrm{km} \times 4\ \mathrm{km}$ scale. However, this setup is useful for knowledge discovery concerning hydrologic processes, which is the main focus of this paper. We leave the operational evaluation of our model on historical precipitation forecasts for future work.

We use the mean squared error (MSE) on the station-wise standardized discharge data as the training and validation loss. This ensures that the training process does not disproportionately favor stations with higher discharge magnitudes. Training is conducted with truncated backpropagation through time (TBPTT), employing a scheduled truncation length. At the beginning of training, we let DRRAiNN learn relationships in the data that are more local in time by backpropagating the loss over subsequences of 1 day. Note that our model operates on an hourly time scale, which means that a 1 day sequence consists of 24 time steps. Throughout training, we increase the truncation length, thereby allowing the model to learn relationships that are increasingly distant in time. The truncation length schedule, which is shown in 1, was determined empirically. We adjust the batch size such that training can take place on a single A100 graphics card.

To improve generalization and account for model variability due to random initialization, we train five instances of DRRAiNN for each experiment, each with a different seed for the random number generator. Results are reported based on the three seeds with the lowest validation loss, a practice we consistently apply to both the primary model and its ablations. We

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

use Ranger (Wright, 2019) with the learning rate set to 0.0025 to optimize the 30.600 parameters in DRRAiNN, which takes about 7 hours. We clip the gradient if its norm exceeds 1 to avoid large jumps at steep regions in the loss surface. We use hydra to manage our configurations (Yadan, 2019).

To increase the size of the training data set and improve generalization, we perform data augmentation. The symmetry group of the square contains eight symmetries, namely: identity, rotation by 90, 180, and 270 degrees, and reflection in the x, y, and the two diagonal axes. For each sequence, we apply a uniformly sampled symmetry to the following variables in each time step: elevation, precipitation, solar radiation, and the mask that is used to translate from grid to graph.

3.5 Benchmark model: European Flood Awareness System

To provide context for DRRAiNNs performance, we compare it to the European Flood Awareness System (EFAS), an established and operational distributed process-based model. Since EFAS reanalysis data is readily available for download, we do not have to tune EFAS ourselves. This avoids potential biases arising from unequal effort in tuning the benchmark model versus the self-developed model. While DRRAiNN achieves higher performance than EFAS in many scenarios, our focus is not solely on outperforming EFAS but on demonstrating the potential of distributed neural networks for river discharge estimation.

EFAS simulates runoff on an approximately $1.5\ \mathrm{km} \times 1.5\ \mathrm{km}$ grid with a temporal resolution of $6\ \mathrm{h}$, which is similar to our setup. It receives as inputs static maps describing topography, river networks, soil, and vegetation, as well as meteorological forcings such as precipitation, temperature, and potential evaporation.

While EFAS serves as a useful benchmark, the comparison to DRRAiNN is not perfectly fair due to fundamental differences in the input and output variables. Both models receive gridded meteorological forcings, but DRRAiNN additionally receives discharge measurements during the tune-in period. In contrast, EFAS does not use discharge measurements as input. Instead, these are used exclusively for model calibration. Furthermore, DRRAiNN estimates discharge only at station locations where observations are available, while EFAS estimates discharge in all grid cells. EFAS also relies on additional input variables not used by DRRAiNN, such as soil type, vegetation, temperature, and potential evapotranspiration. This makes EFAS particularly powerful but also less transferable to regions where detailed input data of this kind might be unavailable. Another difference lies in the precipitation data used: EFAS uses EMO-1, a 6-hourly product interpolated from weather station data, whereas DRRAiNN uses RADOLAN, a radar-derived dataset with finer spatial and temporal resolution. As a result, a direct comparison between EFAS and DRRAiNN is not valid. Nonetheless, the EFAS data can serve as a baseline and an orientation for the performance regime we should be able to match. We thus emphasize that our goal is not to directly compare performance but to provide a baseline that allows us to place the principled quality of DRRAiNN's performance with respect to alternative state-of-the-art forecasting approaches.

3.6 Evaluation

Besides the depiction of hydrographs at some of the modeled stations, we employ the following evaluation metrics to assess the performance of DRRAiNN: Kling-Gupta efficiency (KGE, (Gupta et al., 2009)), Nash-Sutcliffe efficiency (NSE, (Nash and Sutcliffe, 1970)), Pearson's correlation coefficient (PCC), and the mean absolute error (MAE). We report all of these

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

metrics because they are widely used in the hydrological sciences and because there is no single metric that does not have any disadvantages (Gupta et al., 2009). One advantage of the MAE is that it provides a direct and intuitive measure that shows to which extent the models' estimations are off as it has the same unit as the measured quantity. As no normalization takes place in its computation, though, this metric is disproportionately influenced by stations with larger discharges. The PCC shows how much variation is shared between the observed and estimated discharges, however, it does not account for systematic differences in scale or bias. To also capture the scale, the NSE was developed, which can be seen as a mean squared error that is weighted by the variance of the observed discharge. The NSE also does not account for bias, though, which is why the KGE was introduced to capture correlation, bias, and variance. When computing KGE and NSE values, we use station-wise means and variances from the training data set as done in (Kratzert et al., 2019). For KGE, NSE, and PCC, higher values are better with 1 corresponding to a perfect fit. For MAE, lower values are better with 0 corresponding to a perfect fit.

When performing open-loop inference, we also evaluate metrics separately for the different number of open-loop steps performed (where the first one should be similar to closed-loop estimation). This way we can see to which extent performance drops with increasing lead times. Even though DRRAiNN was only trained on sequences that span 20 days, we always evaluate on 100 day sequences to see whether our model can generalize with regards to lead time. Additionally, we will plot the performance of the models against the mean discharge of the different stations to see whether we find systematic relationships between these quantities. In all cases, we remove the initial 10 days tune-in period before calculating metrics and producing plots.

As discussed above, we are interested in more than just good performance in terms of matching hydrographs and good metrics. With knowledge discovery being the main motivation of this work, we will also test DRRAiNN on physical plausibility. A physically implausible model might learn spurious relationships in the data. It could, for example, exploit the DEM to encode local biases that let water spawn or disappear without this process being driven by the meteorological forcings. By retrospectively inferring catchment areas from observed dynamics, we assess whether the rainfall-runoff model successfully propagates water over the landscape. The procedure is as follows: After a forward pass, we compute the gradient of the last time step output with respect to the precipitation input. The result is a so-called saliency map which tells us to which extent the model's output depends on the precipitation in each grid cell. We multiply this gradient by the precipitation itself to focus the analysis on cells in which precipitation occurred. By doing this for each station separately and visualizing the resulting attributions, we can see which areas on the map contribute to the discharge estimation at the corresponding station. To reduce noise, we do this for every sequence in our validation data set and average the outcomes.

We compare the resulting attributions with catchment areas delineated from elevation data, as those are commonly used in hydrology. To evaluate their agreement quantitatively, we employ the following measure when comparing DRRAiNN to the ablated models: For each station, the attributions are first standardized to lie between 0 and 1. We then compute the Wasserstein distance between the attributions within the delineated catchment area and those outside of it. A higher Wasserstein distance indicates better alignment between the attributions and the catchment areas delineated from elevation data. This quantitative measure complements the qualitative comparison, providing stronger evidence for our model's ability to propagate water over

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

the landscape in a physically plausible way. Specifically, it suggests that the model has learned from the observed dynamics alone that water flows downward.

4 Results

For evaluating DRRAiNN, we first provide hydrographs and compare performance with EFAS. We furthermore show that DRRAiNN has the potential to infer catchment areas, thus highlighting the system's potential due to its full differentiability.

4.1 Hydrographs

First, EFAS produces well-matching and plausible hydrographs, rendering it a strong contestant (Fig. 4). As EFAS produces gridded outputs, it is necessary to pick the correct grid cells to compare the model outputs at the specific stations. We likely chose the correct cells, since the reanalysis hydrographs produced by EFAS match the historical observations well for all considered stations. Within the low flow regime (Fig. 4a), it seems that EFAS tends to underestimate discharge.

Second, the results show that DRRAiNN can produce plausible hydrographs that match the observed discharges well, too, especially during the first few days of the estimation. This includes low flows (Fig. 4a) as well as high flows (Fig. 4d) with no apparent systematic difference in performance. Throughout the 100 days, however, the hydrographs tend to match the observed discharge less and less, which is expected as DRRAiNN operates autoregressively: After the closed-loop tune-in phase (which is not shown here), DRRAiNN receives its last inferred discharge as input in the next time step. Therefore, the error accumulates over time. However, considering that DRRAiNN was trained on 20 day sequences only, it is surprising to see that it is in general able to hit peaks even after 80 days. The large peak on day 80 in Lauffen and Rockenau (Fig. 4b and 4d) is underestimated by both models, indicating a bias towards lower values.

4.2 Performance

Overall, DRRAiNN can outperform EFAS in the initial days of the estimation horizon in all considered metrics (Fig. 5). Please note that, since discharge values are never fed into EFAS but only used for calibration, the performance of EFAS is constant. As expected, the performance of DRRAiNN decreases over time, since its autoregressive nature leads to error accumulation as described above.

The KGE plot (Fig. 5a) shows that DRRAiNN produces significantly better results during the initial days. Averaged over the seeds, starting with a KGE of about 0.76, it takes about 48 days before our model's estimations become worse than those of EFAS on average, even though DRRAiNN was only ever trained on 20 day sequences. Two instances of our model can keep up with EFAS even after 100 days. The NSE plot (Fig. 5b) shows our model starting with a value of about 0.81, again with EFAS beating it after about 45 days on average. The fact that one of DRRAiNN's instances intersects EFAS' line earlier in the KGE plot than in the NSE plot points to a larger systematic bias in this instance compared to EFAS, as this is the main difference between those two metrics. The PCC plot (Fig. 5c) shows a strong linear relationship between the observed and inferred discharge values with a value of about 0.9 on average at the beginning. Here, DRRAiNN captures this relationship

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

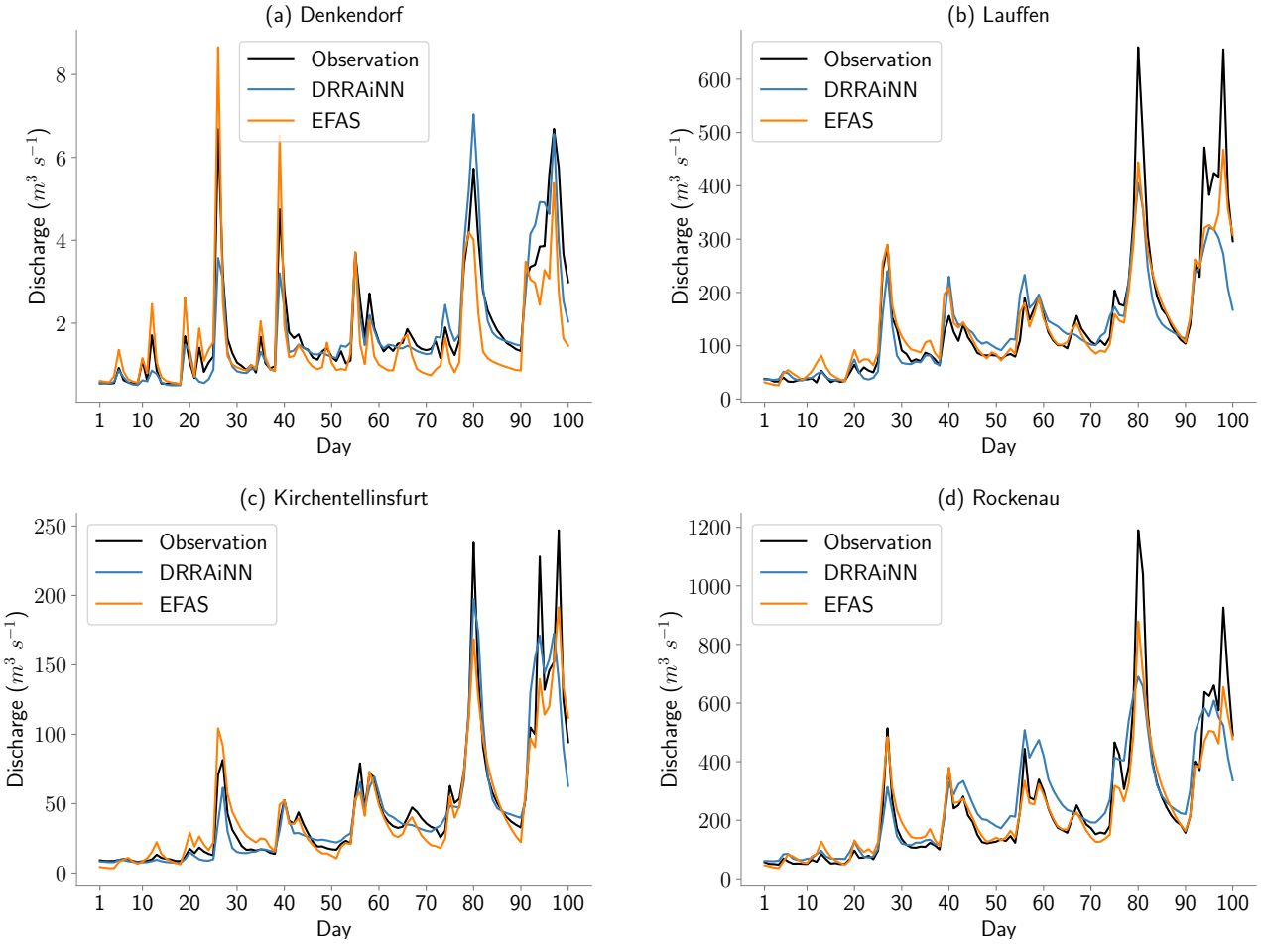

Figure 4. Hydrographs of observed discharge, discharge simulated by EFAS, and discharge inferred by one of our model instances out of five with lead times up to 100 days. They correspond to the stations with the lowest (a), and highest (d), as well as to those stations with the best KGE performance of EFAS (b), and our model on average (c), respectively. We chose those sequences from our validation data set that have the largest variance in discharge as variance likely acts as a proxy for difficulty.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

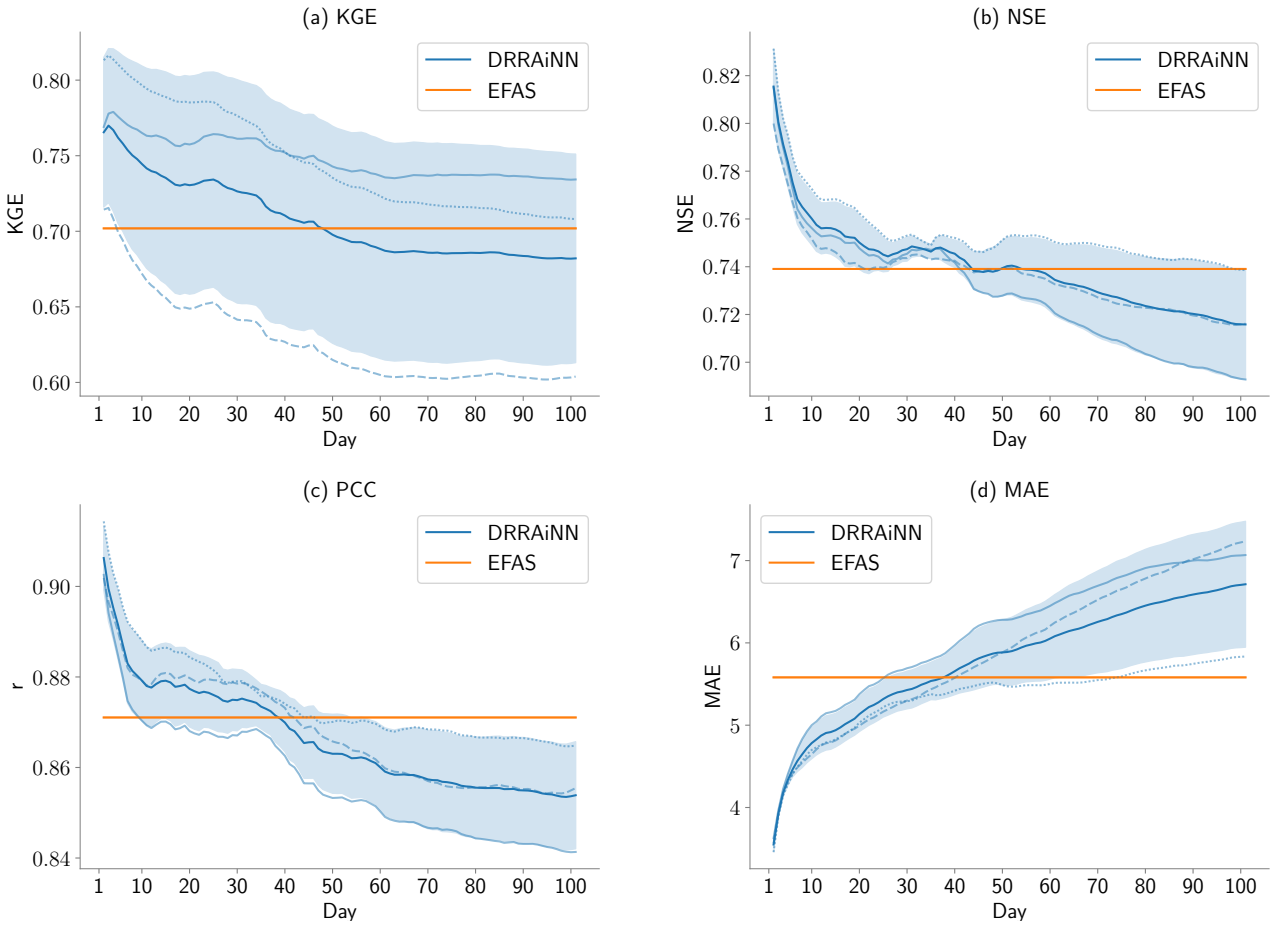

Figure 5. Performances based on different metrics of EFAS and DRRAiNN with lead times up to 100 days. The results are averaged over the stations, and the different seeds of DRRAiNN are depicted with different line styles.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

better than EFAS during the first 40 days. Note that the linear correlation is also part of KGE and NSE. As the MAE allows direct interpretation, its plot (Fig. 5d) shows that EFAS is off by about $5.7\mathrm{m}^3\ \mathrm{s}^{-1}$ on average, while DRRAiNN with $3.3\mathrm{m}^3\ \mathrm{s}^{-1}$ on average on the first day produces a considerable smaller error. After about 40 days, EFAS produces better results on average.

All considered metrics show that the individual model instances perform differently. The order, however, is not fixed but depends on the considered metric and even more so on the considered lead time. Some seeds perform better during the initial days, while others are better with greater lead times: The best two instances, for example, switch after about 40 days in KGE. The seeds differ in weight initialization only, meaning that some instances start the training with a larger bias towards capturing short-term and others with a larger bias towards capturing long-term relationships in the data.

The erratic lines in Fig. 6 show that stations vary in difficulty: Discharge of some stations is harder to estimate than that of others, regardless of which metric is considered. Which stations are harder to estimate, however, is different across the metrics since the metrics focus on different aspects as discussed above. The different seeds of DRRAiNN, and more interestingly, also EFAS, agree on which stations are harder to estimate to some extent: The KGE values in Fig. 6a, e.g., show that Altensteig, Rottweil, and Kirchentellinsfurt consistently belong to the easier ones, while Oppenweiler, Bad Imnau, and Murr belong to the harder ones. We assume that this is related to unobservable underground flows and pipes, however, this could be further analyzed in future work.

The regression lines help us to see whether there is a systematic relationship between a station's mean discharge and its predictability. We performed linear regression here, and the regression lines are only exponential due to the logarithmic x-axis. The KGE plot (Fig. 6a) shows that both models tend to perform better at stations with higher mean discharges. This effect is more pronounced in EFAS, while our model exhibits a more balanced behavior. This is even more the case if we consider the NSE (Fig. 6b) and the PCC (Fig. 6c). Here, DRRAiNN's performance barely depends on the station's mean discharge at all, which cannot be said about EFAS. Differences in the patterns of the KGE (Fig. 6a) and NSE plots (Fig. 6b) show that the models have different biases for the different stations, since this is the main difference between KGE and NSE, as discussed above. Both, DRRAiNN and EFAS, produce significantly larger MAEs with increased mean discharge (Fig. 6d). This is expected, though, as MAE does not account for the stations' mean discharges or their variability in discharge, unlike the other metrics.

4.3 Catchment area inference

We can successfully reconstruct physically plausible catchment areas that DRRAiNN must have inferred implicitly (Fig. 7). Lighter areas show higher importance for estimating discharge at the corresponding station. These areas correlate with the catchment areas depicted in red, which are delineated from elevation alone. The results are not perfect. One should keep in mind, however, that DRRAiNN is trained on daily discharge measurements. This means for sharp delineations, the training data set ideally needs to contain sequences in which it rained within the area, but not outside, over the extent of a 24 h period. As precipitation is very dynamic on this time scale, the chances for this are relatively low. In the future, we expect much better results if we go from daily to hourly discharge data.

In the case of Pforzheim (Fig. 7b), DRRAiNN missed an area in the lower right part that is considered part of the delineated catchment area. This could be explained by underground flows that were found in previous studies near Pforzheim (Ufrecht,

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

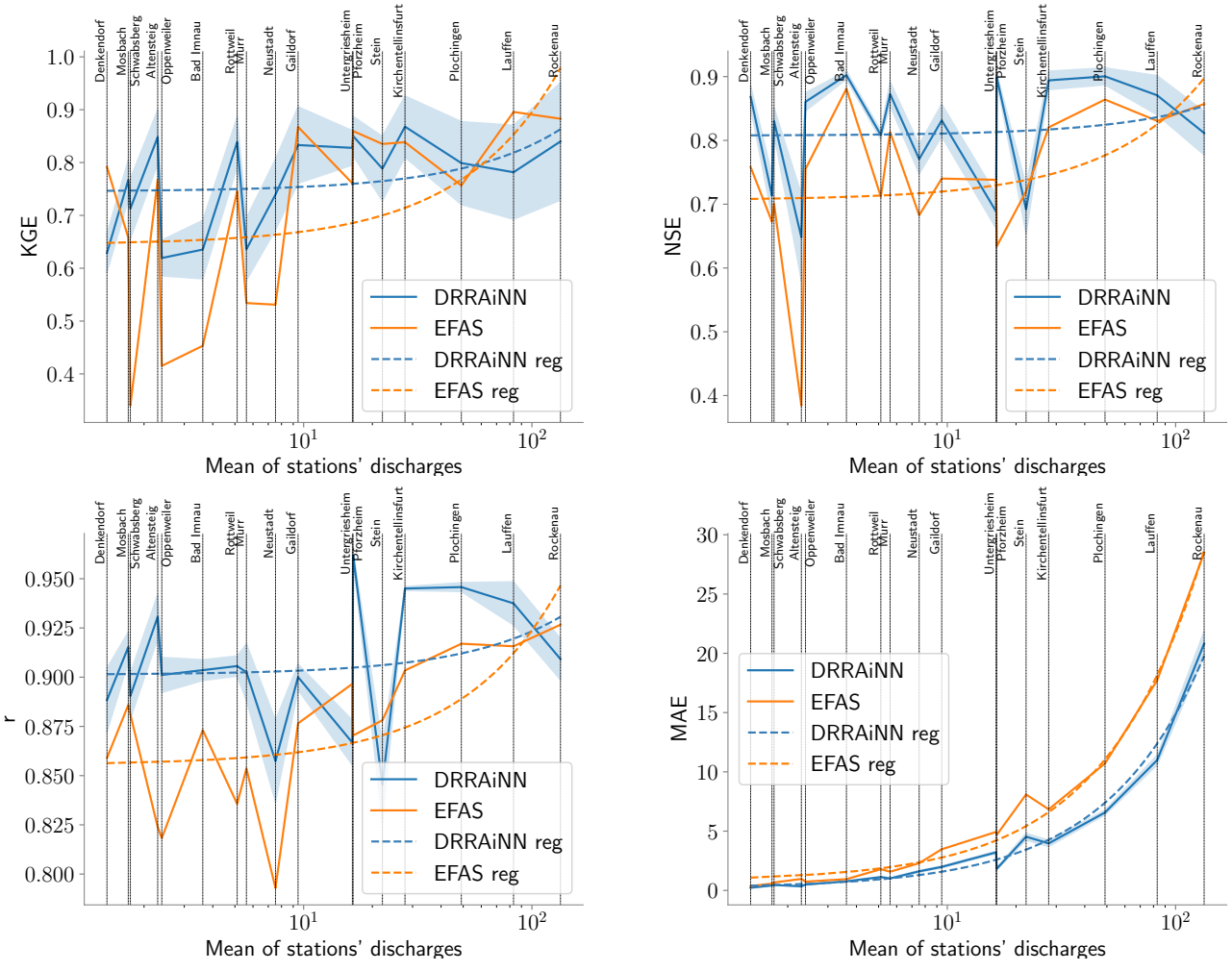

Figure 6. Performances of DRRAiNN and EFAS on 1 day lead time based on different metrics at the different stations. The x-axis denotes the logarithmic means of the stations' discharges. The blue shadow depicts the standard deviation over the different seeds. The dashed lines represent a linear regression on the logarithmic stations' means and the metric.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

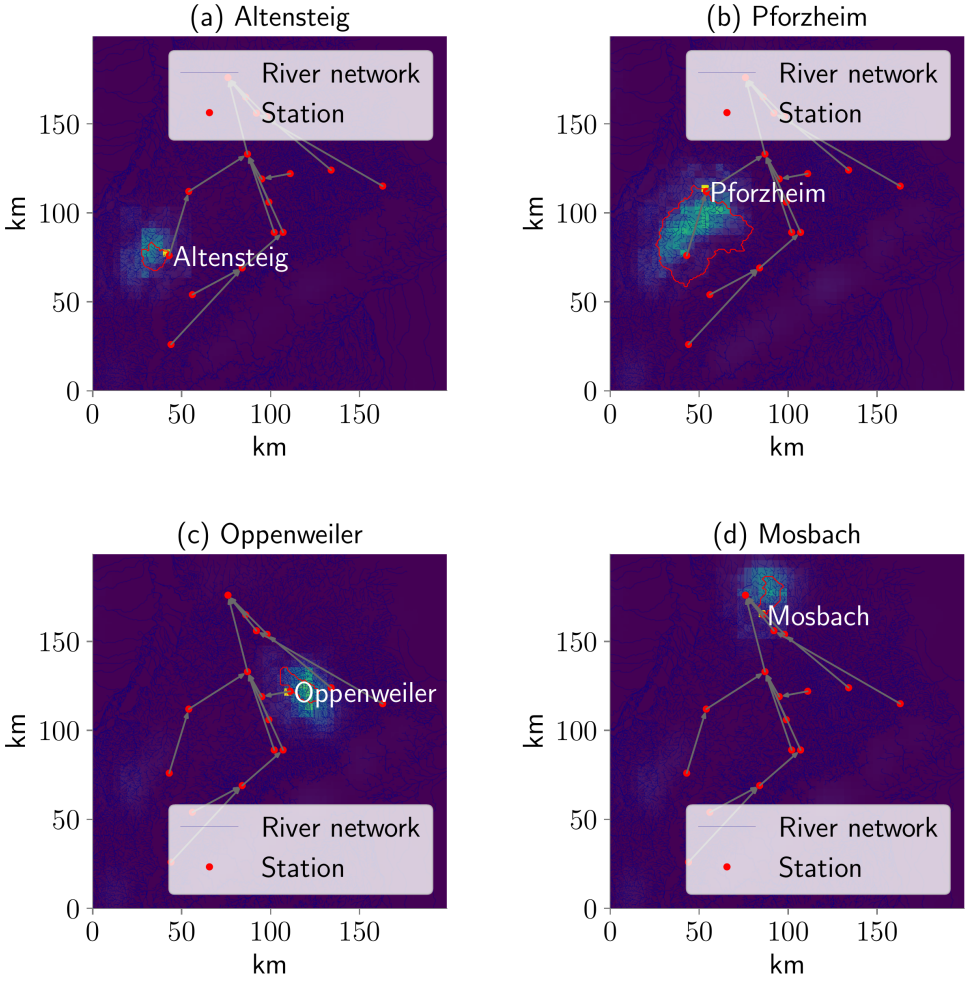

Figure 7. Attribution maps of precipitation for discharge estimation at selected stations, averaged over all validation sequences. Brighter colors indicate grid cells where precipitation has a stronger influence on the estimated discharge at the corresponding station. For comparison, the traditional catchment areas delineated from elevation data are outlined in red. This juxtaposition highlights the agreement between data-driven attributions and physically derived catchment boundaries. The method used to compute these attributions is described in detail in Subsection 3.6 of the main text.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

2002). Water that would pass through Pforzheim if no underground flows existed, instead flows towards the southeast, entering the Neckar River network in a different channel. Our results might be evidence that DRRAiNN detected these unobservable underground flows from precipitation and discharge dynamics, however, this hypothesis arguably needs more investigation in the future.

Note, that these results mainly serve as a proof of principle: We only present the results of the best seed here. This is valid as we performed temporal validation on all seeds beforehand to check for their temporal generalization capabilities. This could also be done if one was interested in building an operational model.

5 Discussion

In this work, we present DRRAiNN, a fully differentiable, fully distributed neural network architecture that successfully estimates river discharge from past discharge, an elevation map, gridded precipitation, and gridded solar radiation. Individual instances of DRRAiNN can produce better KGE performance than EFAS with lead times of up to 100 days. This shows that DRRAiNN can produce reasonable estimations far into the future even though it was only trained on 20 day sequences.

Our analysis reveals that discharge estimation at various stations exhibits differing levels of difficulty. Notably, DRRAiNN and EFAS tend to identify the same stations as challenging to predict, suggesting that the difficulty is intrinsic to the stations and their associated data rather than specific to the model architecture. This variability in prediction difficulty likely arises from several factors. Stations influenced by unobserved variables like complex subsurface topography, land cover heterogeneity, or anthropogenic factors (e.g., dam operations) may pose greater challenges for both models. Furthermore, spatial variations in the quality of input data could contribute to discrepancies in performance. Future investigations employing attribution techniques could offer deeper insights into these station-specific variations and guide the development of architectural modifications or regularizations to address these challenges effectively.

We performed several ablations on our model (Appendix A). First, we showed that DRRAiNN can exploit the DEM as a positional encoding by training and validating it on a rotated DEM. While this did not lead to worse performance, it resulted in significantly less physically plausible behavior (Appendix A1). Then, we checked whether the inductive bias that makes the model distinguish between spatially extended and local processes is useful (Appendix A2). Last, we removed the hypernetworks to examine their impact (Appendix A3). For some combinations of performed ablation, metric, and lead time, there is no significant differences in terms of performance compared to the original DRRAiNN model. However, none of the ablated models is able to produce physically realistic catchment areas, which we showed qualitatively and underlined quantitatively.

Our ablation studies highlight the importance of distinguishing between spatially extended and local processes, as well as the incorporation of hypernetworks. The inability of ablated models to produce realistic catchment areas suggests that these components encode crucial hydrological processes, such as water movement over complex topographies. This finding indicates that certain inductive biases not only improve model interpretability but also prevent spurious correlations.

Interestingly, the model instance achieving the best attribution maps does not correspond to the one achieving the best performance metrics. This highlights a trade-off between optimizing for predictive accuracy and ensuring the model behaves

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

in a physically meaningful way. It suggests that while the metrics measure how well the model captures patterns in the training data, they may not fully capture the alignment with physical principles.

Our choice of input datasets was guided by considerations of temporal resolution, data sources, and practical availability, all of which impact model performance. Although EFAS uses EMO-1 for precipitation data, we opted for RADOLAN due to key differences: EMO-1 provides 6-hourly resolution and is interpolated from station data rather than derived directly from radar observations. While we expect minor differences in performance between RADOLAN and EMO-1, radar-derived datasets like RADOLAN generally offer finer spatial and temporal detail, which is advantageous for distributed models. Similarly, for solar radiation data, ERA5 was chosen due to its raster format and hourly resolution. Alternative datasets, such as those provided by DWD, are either available only as station-wise hourly data, which lack the required raster format, or as raster data aggregated monthly, which does not meet our temporal requirements. Daily datasets like EOBS could suffice if temporal patterns are encoded separately, but this would require additional preprocessing steps. If one aims to transition toward operational flood forecasting in the future, the choice of precipitation forecast will become critically important (Imhoff et al., 2022). Ultimately, all data products come with inherent uncertainties and errors, and our choices reflect a balance between data availability, temporal resolution, and model needs.

An increase in the amount of training data is always beneficial in machine learning. Currently, DDRAiNN is not designed for scalability, as its application is expected to require retraining for each specific context. A first step towards improving its adaptability would be to train DDRAiNN on hourly discharge data. We anticipate this would yield performance improvements and qualitatively better attributions, potentially even capturing the origins of individual peaks in the hydrographs. To explore the model's spatial generalization capabilities, we aim to apply DDRAiNN to diverse catchments across Germany, Europe, or even globally. By validating it on catchments that are not part of the training data, we can systematically assess its ability to generalize to unseen regions. Improving this spatial generalization remains a key challenge and likely requires additional constraints or inductive biases in the model. Promising candidates are the incorporation of physical constraints like mass conservation (Hoedt et al., 2021; Harder et al., 2023; Wi and Steinschneider, 2023) or semantically splitting the hidden state of the rainfall-runoff model into surface and subsurface components. These enhancements could pave the way for future scalability and broader applicability.

As traditional process-based models make use of many more input variables, feeding them as additional inputs could lead to performance improvements in DRRAiNN as well. This includes land cover, geology, soil, vegetation, temperature, and potential evapotranspiration among others. Interpretability methods can then be used to perform a sensitivity analysis, revealing which input variables are important when and, due to our model being fully distributed, where.

Currently, DRRAiNN uses a warm-up period of 10 days for the hidden states to tune into the dynamics. The rainfall-runoff model potentially captures precipitation during this time to estimate soil moisture, which has a huge impact on infiltration. Therefore, soil moisture as an additional input variable is of special interest as it might allow us to get rid of the initial 10 days, thereby reducing training costs. An alternative would be to feed in the compressed precipitation history of the days or even weeks before (Traub et al., 2024b).

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

Concerning output variables, DRRAiNN could also be used to estimate quantities other than discharge. For some measurement stations, additional water-related information, like turbidity, is available. To estimate turbidity, information about potential erosion can be helpful, as is provided by the RUSLE model (Renard et al., 1994), for example. Similarly to the catchment area inference performed in this study, a trained instance of such a model could be interrogated to infer the origins of measured turbidity, potentially informing us about sites of actual erosion. This information could be used to create policies for soil protection. Other variables of interest include the concentration of toxins and oxygen for similar applications.

Operational flood forecasting would be a safety-critical application of DRRAiNN. Therefore, it is important to quantify uncertainties as was suggested elsewhere before (Hrachowitz et al., 2013; Nearing et al., 2021). Equipping our model with this ability would allow us to provide confidence intervals when reporting inferred discharge values. In this regard, distributional parameter estimation is a technique where our architecture would produce an additional output that is interpreted as the standard deviation in a negative log-likelihood loss. Other techniques include Bayesian neural networks (Neal, 2012), Monte-Carlo dropouts (Gal and Ghahramani, 2015), and variational methods (Graves, 2011).

Another hurdle for operational flood forecasting is the inherent difficulty of obtaining sufficiently accurate, high-resolution precipitation forecasts with lead times of several days. Even though numerical weather prediction models, such as those provided by the DWD, are readily available, they are limited in predicting localized extreme precipitation events and reducing forecast uncertainty. In this work, we always assumed perfect forecasts by using historical observational precipitation data since we focused on the dynamics of water once it reaches the earth's surface. Therefore, further examination of DRRAiNN's predictive abilities when provided with precipitation forecasts is needed to see whether it would be suitable for operational flood forecasting.

6 Conclusions

In this study, we introduced DRRAiNN, a fully distributed neural network architecture that estimates river discharge from precipitation, solar radiation, elevation maps, and past discharge measurements from gauging stations. Despite being trained on sparse target data, namely daily discharge observations from only 17 stations, DRRAiNN outperforms the operational benchmark model EFAS in terms of KGE and NSE across various lead times. Beyond its predictive accuracy, DRRAiNN provides physically interpretable attributions, enabling the identification of precipitation sources contributing to discharge at specific stations. Our analyses highlight the importance of incorporating hydrologically meaningful constraints, or inductive biases. These biases not only enhance interpretability but also ensure the model adheres to physical principles, as evidenced by its ability to delineate realistic catchment areas. With its predictive performance, interpretability, and physical consistency, DRRAiNN represents a promising step forward in the application of neural networks to distributed hydrological modeling.

Code and data availability. The preprocessed data sets can be found at Scholz et al. (2024a). The code can be found at Scholz et al. (2024b).

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

Appendix A: Ablations

A1 Rotated elevation map

We want to check whether DRRAiNN makes plausible use of the elevation map to propagate water downwards over the landscape. An alternative would be that DRRAiNN uses the elevation to orient itself in the landscape, exploiting it as a positional encoding. This way DRRAiNN can learn local biases at the different positions in the map. Most likely, one will always observe a combination of both effects.

To examine this, we here train and validate DRRAiNN with the same elevation map as before but rotated by 180 degrees. This has the advantage of the elevation map having the same statistics as before, making this comparison fair.

For most metrics and lead times, we do not find a significantly better performance of DRRAiNN if trained and validated on the original elevation map in contrast to the rotated one (Fig. A1). This confirms our suspicion that the model can exploit the elevation map as a positional encoding. However, in this case, we are not able to reconstruct plausible catchment areas (Fig. A2), which is underlined by our quantitative measure (Fig. A3). This is evidence that our original model makes suitable use of the elevation map that goes beyond positional encoding.

A2 All LSTM

One key inductive bias in DRRAiNN is the explicit distinction between spatially extended processes and local processes. The lateral propagation of water over the landscape is a spatially extended process that is mainly driven by elevation. Evapotranspiration, on the other hand, is a local process that is mainly driven by solar radiation. We incorporate this into DRRAiNN by mapping these processes on the DWConv and the PWConv components within the ConvNeXt block. DWConv is parameterized by a CNN that receives elevation as input, while PWConv1 and PWConv2 are parameterized by an MLP that receives solar radiation as input. In this ablation, we discard this bias by feeding the elevation and solar radiation together with the precipitation into the PWLSTM. Therefore, the relativity bias, realized by subtracting the elevation of the center cell from the elevations of all other cells within each receptive field of the hypernetwork, is discarded here as well.

We find a significant performance drop in earlier lead times for all metrics except MAE (Fig. A4). Furthermore, the inferred catchment areas do not look physically plausible (Fig. A5), which is underlined quantitatively (Fig. A6). This shows that the explicit distinction between these sub-processes is advantageous for DRRAiNN, both in terms of accuracy and plausibility.

A3 No hypernetworks

Here, we train DRRAiNN without hypernetworks to examine their usefulness. To stay close to the original architecture, we want to maintain the inductive bias that distinguishes the spatially extended process of propagating water over the landscape and the local process of evapotranspiration. Therefore, the elevation map is concatenated with the hidden state, fed into a position-wise linear layer, and only then fed into the DWConv. This is necessary as DWConv requires the number of input and output channels to be the same. Therefore, the relativity bias, realized by subtracting the elevation of the center cell from the

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

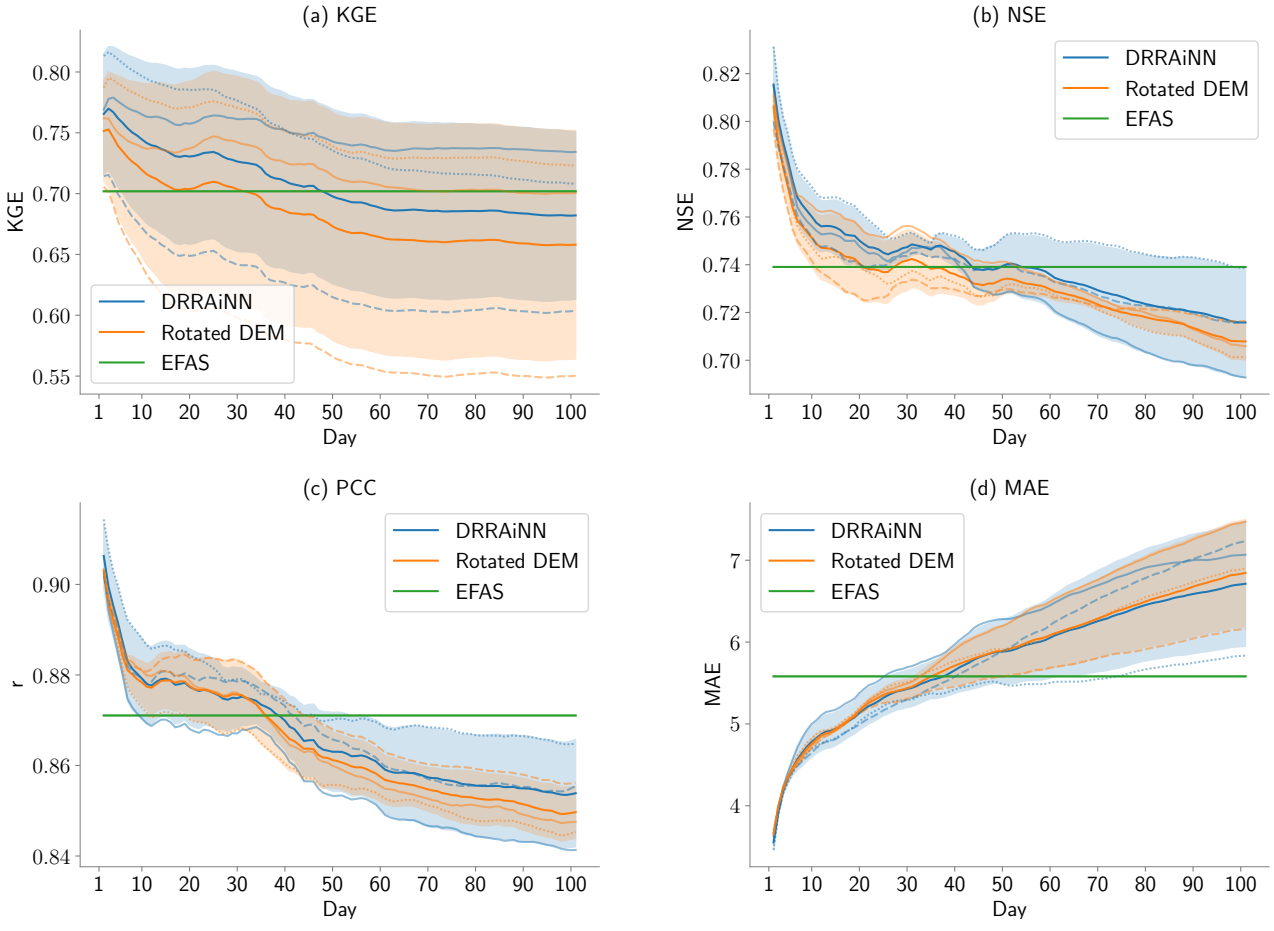

Figure A1. Performances based on different metrics of EFAS, DRRAiNN, and DRRAiNN on a rotated elevation map with lead time up to 100 days. The results are averaged over the stations, and the different seeds of our model are depicted with different line styles.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

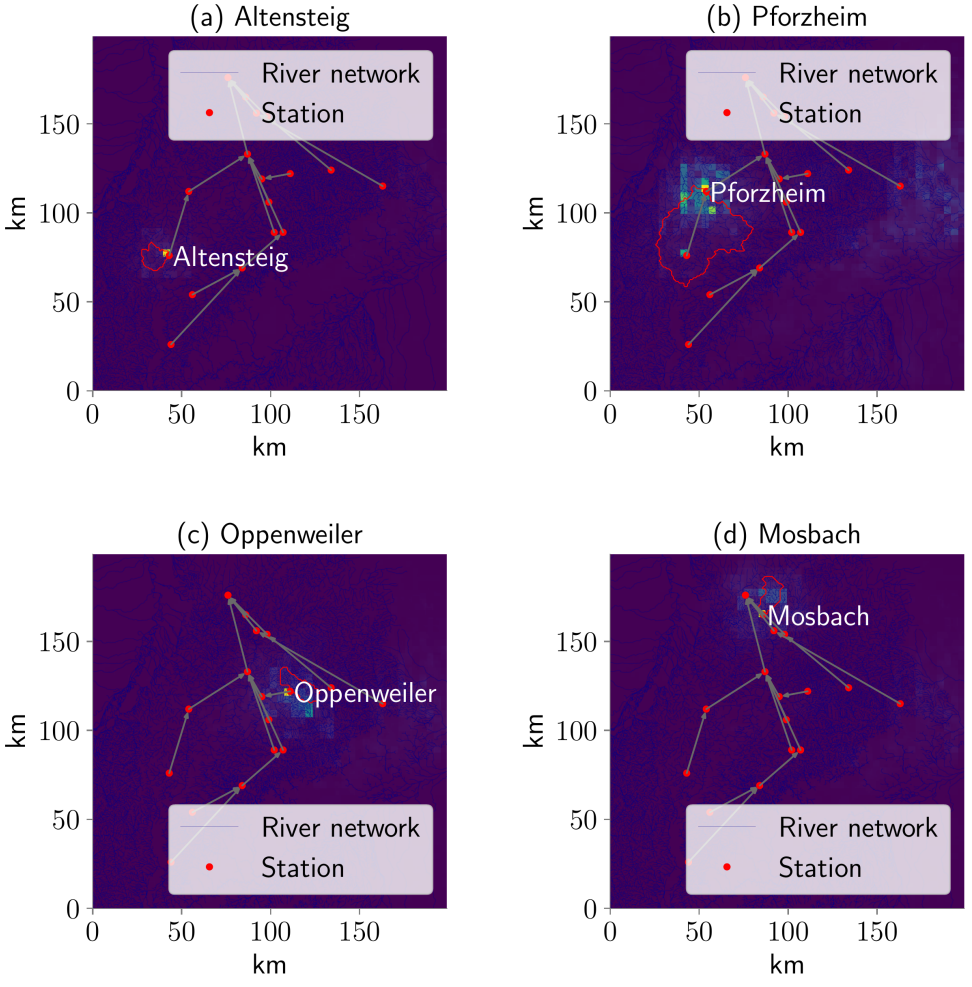

Figure A2. Attribution maps of precipitation for discharge estimation at selected stations aggregated over all validation sequences with a rotated elevation map. The brighter the color of a pixel, the more important is precipitation in that grid cell for discharge estimation at the corresponding station. The catchment areas inferred from elevation alone are shown in red.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

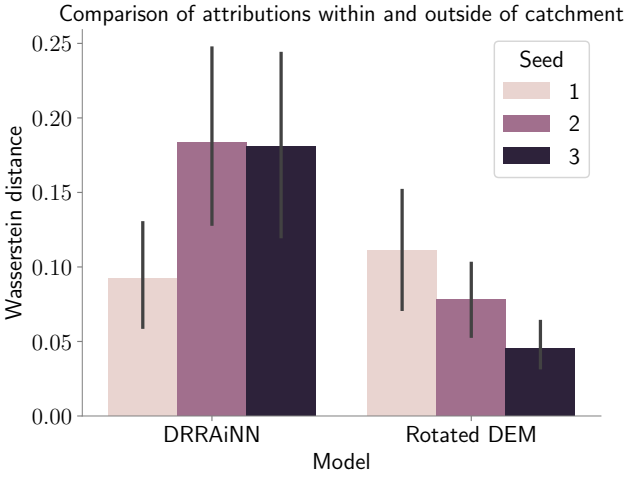

Figure A3. Wasserstein distance between normalized attributions within and outside of the catchment areas delineated from the digital elevation model. A higher distance points towards a better agreement between inferred and delineated catchment areas, and therefore a more physically realistic behavior of the model. The depicted standard deviations are computed over the different gauging stations.

elevations of all other cells within each receptive field of the hypernetwork, is discarded here as well. Solar radiation, on the other hand, is concatenated with the hidden state and directly fed into PWConv1.

Removing the hypernetworks from DRRAiNN leads to a significant decrease in performance for KGE, especially during the first days (Fig. A7a). For NSE this effect is less pronounced (Fig. A7b), while we do not observe a systematic difference in PCC and MAE (Fig. A7c and A7d). The ablated model does not produce plausible attribution maps (Fig. A8), which is underlined quantitatively (Fig. A9).

Author contributions. All authors contributed to the conceptualization of the paper. FS, MT, and MB designed the model architecture. FS developed the code and performed the experiments. FS prepared the manuscript with contributions from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. Additional support came from the Open Access Publishing Fund of the University of Tübingen. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Fedor Scholz and Manuel Traub. ChatGPT was used to improve the writing style of the manuscript.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

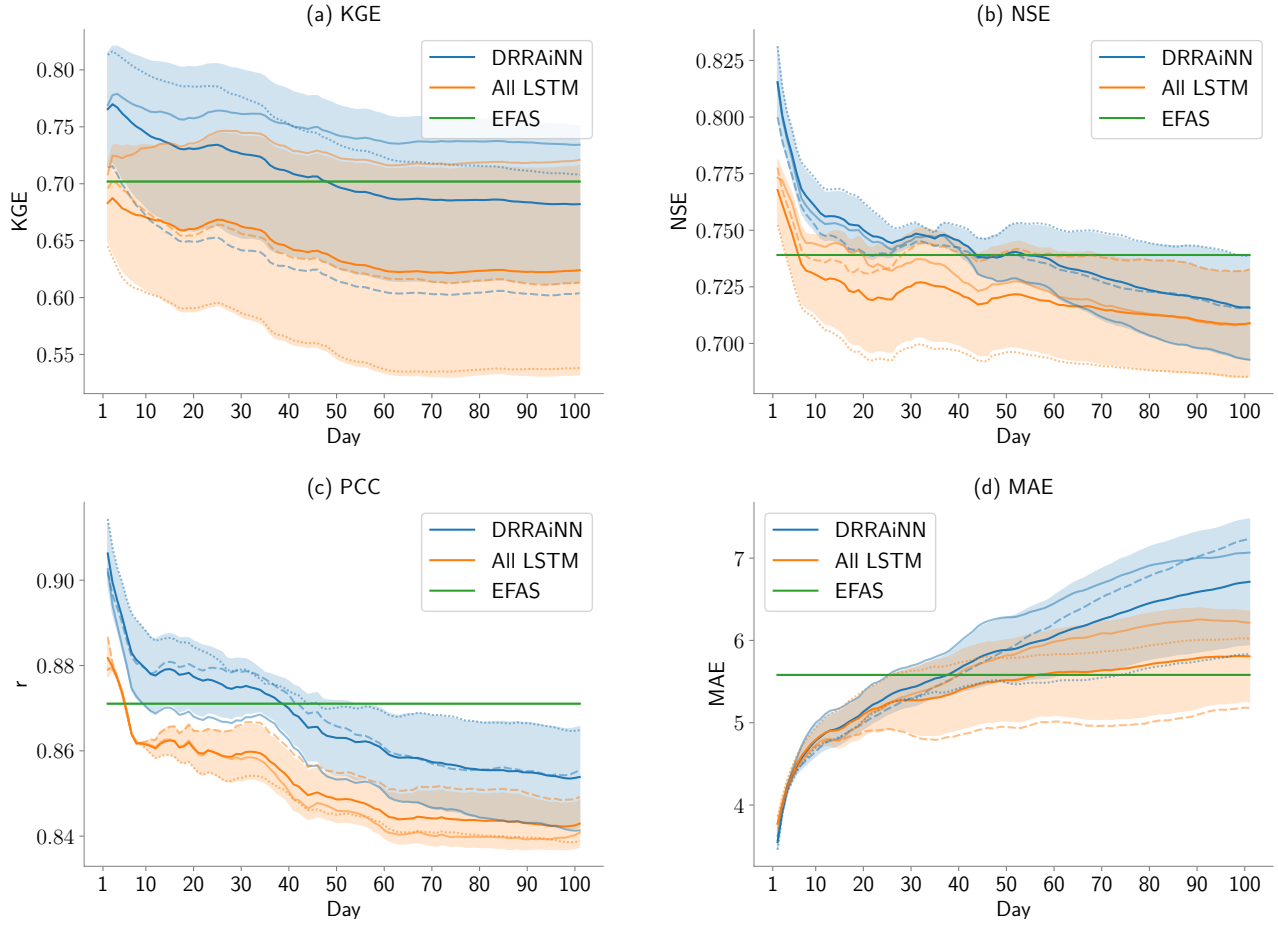

Figure A4. Performances based on different metrics of EFAS, DRRAiNN, and ablated DRRAiNN where all forcings are fed into the PWLSTM with lead up to 100 days. The results are averaged over the stations, and the different seeds of our model are depicted with different line styles.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

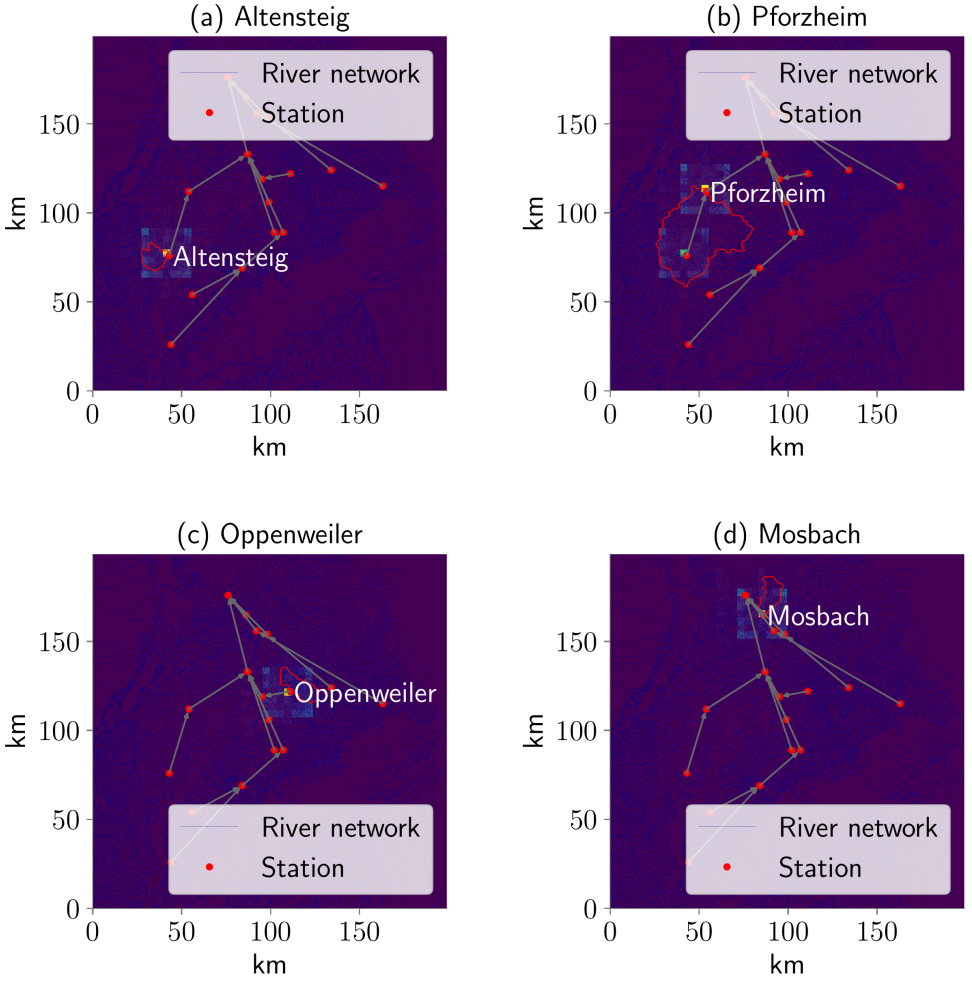

Figure A5. Attribution maps of precipitation for discharge estimation at selected stations aggregated over all validation sequences when all forcings are fed into the PWLSTM. The brighter the color of a pixel, the more important is precipitation in that grid cell for discharge estimation at the corresponding station. The catchment areas inferred from elevation alone are shown in red.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

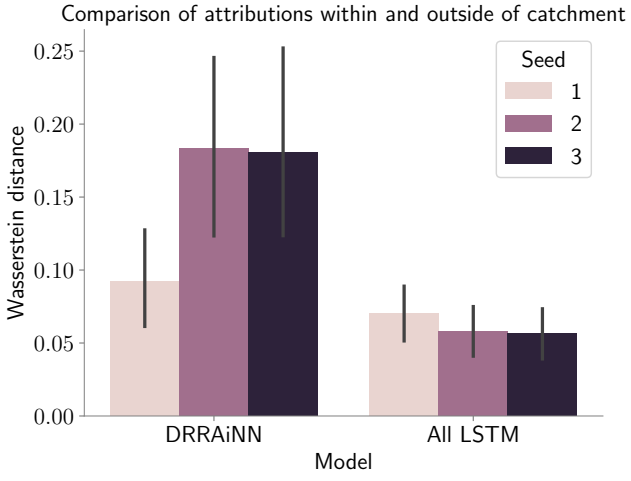

Figure A6. Wasserstein distance between normalized attributions within and outside of the catchment areas delineated from the digital elevation model. A higher distance points towards a better agreement between inferred and delineated catchment areas, and therefore a more physically realistic behavior of the model. The depicted standard deviations are computed over the different gauging stations.

References

Al Hossain, B. M. T., Ahmed, T., Aktar, M. N., Fida, M., Khan, A., Islam, A., Yazdan, M. M. S., Noor, F., and Rahaman, A. Z.: Climate Change Impacts on Water Availability in the Meghna Basin, in: Proceedings of the 5th International Conference on Water and Flood Management (ICWFM-2015), Dhaka, Bangladesh, pp. 6–8, 2015.

AWGN, 2023: Amtliches Digitales Wasserwirtschaftliches Gewässernetz (AWGN), https://www.lubw.baden-wuerttemberg.de/wasser/awgn, 2023.

Bharati, L., Lacombe, G., Gurung, P., Jayakody, P., Hoanh, C. T., and Smakhtin, V.: The impacts of water infrastructure and climate change on the hydrology of the Upper Ganges River Basin, vol. 142, IWMI, 2011.

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., Allen, S. T., Amin, A., Andréassian, V., Arheimer, B., Aryal, S. K., Baker, V., Bardsley, E., Barendrecht, M. H., Bartosova, A., Batelaan, O., Berghuijs, W. R., Beven, K., Blume, T., Bogaard, T., de Amorim, P. B., Böttcher, M. E., Boulet, G., Breinl, K., Brilly, M., Brocca, L., Buytaert, W., Castellarin, A., Castelletti, A., Chen, X., Chen, Y., Chen, Y., Chifflard, P., Claps, P., Clark, M. P., Collins, A. L., Croke, B., Dathe, A., David, P. C., de Barros, F. P. J., de Rooij, G., Baldassarre, G. D., Driscoll, J. M., Duethmann, D., Dwivedi, R., Eris, E., Farmer, W. H., Feiccabrino, J., Ferguson, G., Ferrari, E., Ferraris, S., Fersch, B., Finger, D., Foglia, L., Fowler, K., Gartsman, B., Gascoin, S., Gaume, E., Gelfan, A., Geris, J., Gharari, S., Gleeson, T., Glendell, M., Bevacqua, A. G., González-Dugo, M. P., Grimaldi, S., Gupta, A. B., Guse, B., Han, D., Hannah, D., Harpold, A., Haun, S., Heal, K., Helfricht, K., Herrnegger, M., Hipsey, M., Hlaváčiková, H., Hohmann, C., Holko, L., Hopkinson, C., Hrachowitz, M., Illangasekare, T. H., Inam, A., Innocente, C., Istanbulluoglu, E., Jarihani, B., Kalantari, Z., Kalvans, A., Khanal, S., Khatami, S., Kiesel, J., Kirkby, M., Knoben, W., Kochanek, K., Kohnová, S., Kolechkina, A., Krause, S., Kreamer, D., Kreibich, H., Kunstmann, H., Lange, H., Liberato, M. L. R., Lindquist, E., Link, T., Liu, J., Loucks, D. P., Luce, C., Mahé, G., Makarieva,

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

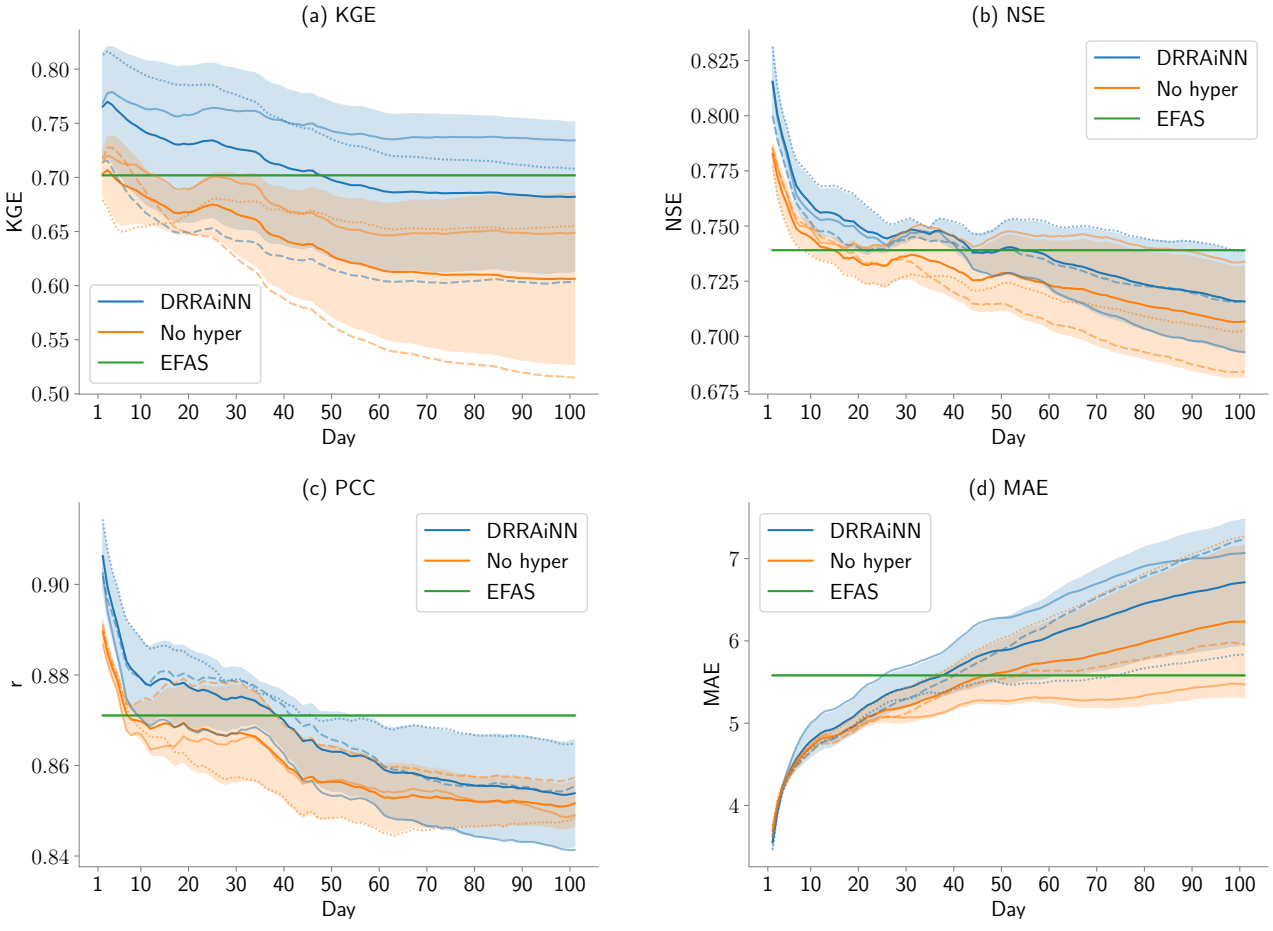

Figure A7. Performances based on different metrics of EFAS, DRRAiNN, and DRRAiNN without the hypernetworks with lead time up to 100 days. The results are averaged over the stations, and the different seeds of our model are depicted with different line styles.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

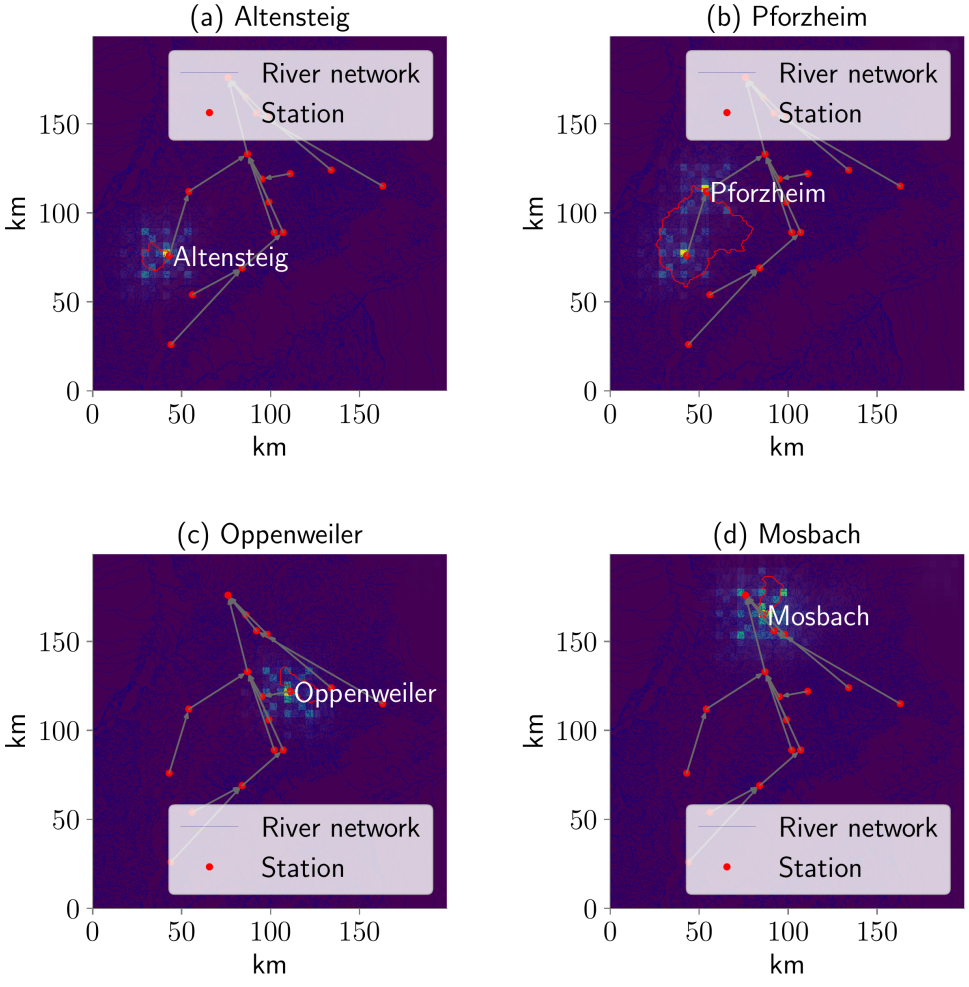

Figure A8. Attribution maps of precipitation for discharge estimation at selected stations aggregated over all validation sequences without hypernetworks. The brighter the color of a pixel, the more important is precipitation in that grid cell for discharge estimation at the corresponding station. The catchment areas inferred from elevation alone are shown in red.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

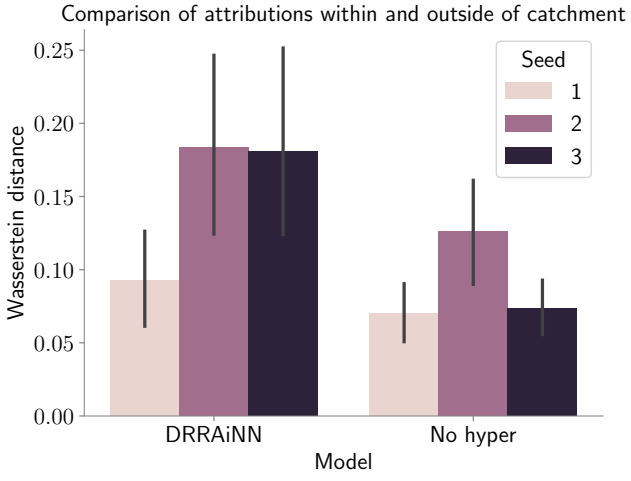

Figure A9. Wasserstein distance between normalized attributions within and outside of the catchment areas delineated from the digital elevation model. A higher distance points towards a better agreement between inferred and delineated catchment areas, and therefore a more physically realistic behavior of the model. The depicted standard deviations are computed over the different gauging stations.

O., Malard, J., Mashtayeva, S., Maskey, S., Mas-Pla, J., Mavrova-Guirguinova, M., Mazzoleni, M., Mernild, S., Misstear, B. D., Montanari, A., Müller-Thomy, H., Nabizadeh, A., Nardi, F., Neale, C., Nesterova, N., Nurtaev, B., Odongo, V. O., Panda, S., Pande, S., Pang, Z., Papacharalampous, G., Perrin, C., Pfister, L., Pimentel, R., Polo, M. J., Post, D., Sierra, C. P., Ramos, M.-H., Renner, M., Reynolds, J. E., Ridolfi, E., Rigon, R., Riva, M., Robertson, D. E., Rosso, R., Roy, T., Sá, J. H., Salvadori, G., Sandells, M., Schaefli, B., Schumann, A., Scolobig, A., Seibert, J., Servat, E., Shafiei, M., Sharma, A., Sidibe, M., Sidle, R. C., Skaugen, T., Smith, H., Spiessl, S. M., Stein, L., Steinsland, I., Strasser, U., Su, B., Szolgay, J., Tarboton, D., Tauro, F., Thirel, G., Tian, F., Tong, R., Tussupova, K., Tyralis, H., Uijlenhoet, R., van Beek, R., van der Ent, R. J., van der Ploeg, M., Loon, A. F. V., van Meerveld, I., van Nooijen, R., van Oel, P. R., Vidal, J.-P., von Freyberg, J., Vorogushyn, S., Wachniew, P., Wade, A. J., Ward, P., Westerberg, I. K., White, C., Wood, E. F., Woods, R., Xu, Z., Yilmaz, K. K., and Zhang, Y.: Twenty-three unsolved problems in hydrology (UPH) –a community perspective, Hydrological Sciences Journal, 64, 1141–1158, https://doi.org/10.1080/02626667.2019.1620507, 2019.

Brutsaert, W.: Hydrology, Cambridge university press, 2023.

Butz, M. V., Bilkey, D., Humaidan, D., Knott, A., and Otte, S.: Learning, planning, and control in a monolithic neural event inference architecture, Neural Networks, 117, 135–144, https://doi.org/10.1016/j.neunet.2019.05.001, arXiv: 1809.07412, 2019.

Butz, M. V., Mittenbühler, M., Schwöbel, S., Achimova, A., Gumbsch, C., Otte, S., and Kiebel, S.: Contextualizing predictive minds, Neuroscience \& Biobehavioral Reviews, p. 105948, 2024.

Camporese, M. and Girotto, M.: Recent advances and opportunities in data assimilation for physics-based hydrological modeling, Frontiers in Water, 4, 948 832, 2022.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y.: On the Properties of Neural Machine Translation: Encoder–Decoder Approaches, in: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Association for Computational Linguistics, https://doi.org/10.3115/v1/w14-4012, 2014.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

Darcy, H.: Les fontaines publiques de la ville de Dijon: exposition et application des principes à suivre et des formules à employer dans les questions de distribution d'eau, vol. 1, Victor dalmont, 1856.

EU-DEM, 2016: EU-DEM v1.1, Dataset, https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1, 2016.

Gal, Y. and Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, http://arxiv.org/abs/1506.02142v6, 2015.

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

Gillies, S. et al.: Rasterio: geospatial raster I/O for Python programmers, https://github.com/rasterio/rasterio, 2013.

Graves, A.: Practical variational inference for neural networks, in: Advances in neural information processing systems, pp. 2348–2356, 2011.

GRDC, 2024: Global Runoff Data Centre, https://grdc.bafg.de/, 2024.

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

Harder, P., Ramesh, V., Hernandez-Garcia, A., Yang, Q., Sattigeri, P., Szwarcman, D., Watson, C., and Rolnick, D.: Physics-Constrained Deep Learning for Downscaling, Tech. rep., Copernicus Meetings, 2023.

Hendrycks, D. and Gimpel, K.: Gaussian Error Linear Units (GELUs), arXiv preprint arXiv:1606.08415, http://arxiv.org/abs/1606.08415v5, 2016.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., et al.: ERA5 hourly data on single levels from 1940 to present, 2018.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., Hochreiter, S., and Klambauer, G.: MC-LSTM: Mass-Conserving LSTM, Proceedings of Machine Learning Research, http://arxiv.org/abs/2101.05186v3, 2021.

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., Fenicia, F., Freer, J., Gelfan, A., Gupta, H., Hughes, D., Hut, R., Montanari, A., Pande, S., Tetzlaff, D., Troch, P., Uhlenbrook, S., Wagener, T., Winsemius, H., Woods, R., Zehe, E., and Cudennec, C.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological Sciences Journal, 58, 1198–1255, https://doi.org/10.1080/02626667.2013.803183, 2013.

Imhoff, R., Van Verseveld, W., Van Osnabrugge, B., and Weerts, A.: Scaling point-scale (pedo) transfer functions to seamless large-domain parameter estimates for high-resolution distributed hydrologic modeling: An example for the Rhine River, Water Resources Research, 56, e2019WR026 807, 2020.

Imhoff, R. O., Brauer, C. C., van Heeringen, K.-J., Uijlenhoet, R., and Weerts, A. H.: Large-sample evaluation of radar rainfall nowcasting for flood early warning, Water Resources Research, 58, e2021WR031 591, 2022.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty estimation with deep learning for rainfall–runoff modeling, Hydrology and Earth System Sciences, 26, 1673–1693, 2022.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using long short-term memory (LSTM) networks, Hydrology and Earth System Sciences, 22, 6005–6022, 2018.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019.

Kratzert, F., Klotz, D., Gauch, M., Klingler, C., Nearing, G., and Hochreiter, S.: Large-scale river network modeling using Graph Neural Networks, in: EGU General Assembly Conference Abstracts, pp. EGU21–13 375, 2021.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, https://doi.org/10.5194/egusphere-egu23-5256, 2023.

Li, P., Zhang, J., and Krebs, P.: Prediction of flow based on a CNN-LSTM combined deep learning approach, Water, 14, 993, 2022.

Liu, Y. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water resources research, 43, 2007.

Liu, Y., Weerts, A. H., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., Seo, D.-J., Schwanenberg, D., Smith, P., van Dijk, A. I. J. M., van Velzen, N., He, M., Lee, H., Noh, S. J., Rakovec, O., and Restrepo, P.: Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities, Hydrology and earth system sciences, 16, 3863–3887, 2012.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S.: A ConvNet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, https://doi.org/10.1109/cvpr52688.2022.01167, 2022.

Longyang, Q., Choi, S., Tennant, H., Hill, D., Ashmead, N., Neilson, B. T., Newell, D. L., McNamara, J. P., and Xu, T.: Explainable Spatially Distributed Hydrologic Modeling of a Snow Dominated Mountainous Karst Watershed Using Attention, Authorea Preprints, 2024.

Marçais, J. and de Dreuzy, J.-R.: Prospective interest of deep learning for hydrological inference, Groundwater, 55, 688–692, 2017.

Mazzetti, C., Carton de Wiart, C., Gomes, G., Russo, C., Decremer D Ramos, A., Grimaldi, S., Disperati, J., Ziese, M., Schweim, C., Sanchez Garcia, R., Jacobson, T., Salamon, P., and Prudhomme, C.: River discharge and related historical data from the European Flood Awareness System, v5.0, European Commission, Joint Research Centre (JRC), https://cds.climate.copernicus.eu/cdsapp#!/dataset/efas-historical, 2023.

Montzka, C., Pauwels, V. R., Franssen, H.-J. H., Han, X., and Vereecken, H.: Multivariate and multiscale data assimilation in terrestrial systems: A review, Sensors, 12, 16 291–16 333, 2012.

Moradkhani, H., Hsu, K.-L., Gupta, H., and Sorooshian, S.: Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, Water resources research, 41, 2005.

Moshe, Z., Metzger, A., Kratzert, F., Morin, E., Nevo, S., Elidan, G., and Elyaniv, R.: HydroNets: Leveraging River Network Structure and Deep Neural Networks for Hydrologic Modeling, https://doi.org/10.5194/egusphere-egu2020-4135, 2020.

Muñoz-Carpena, R., Carmona-Cabrero, A., Yu, Z., Fox, G., and Batelaan, O.: Convergence of mechanistic modeling and artificial intelligence in hydrologic science and engineering, PLOS Water, 2, e0000 059, 2023.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 10, 282–290, 1970.

Neal, R. M.: Bayesian learning for neural networks, vol. 118, Springer Science \& Business Media, 2012.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What role does hydrological science play in the age of machine learning?, Water Resources Research, 57, e2020WR028 091, 2021.

Oddo, P. C., Bolten, J. D., Kumar, S. V., and Cleary, B.: Deep Convolutional LSTM for improved flash flood prediction, Frontiers in Water, 6, 1346 104, 2024.

Otte, S., Karlbauer, M., and Butz, M. V.: Active Tuning, arXiv:2010.03958 [cs], http://arxiv.org/abs/2010.03958, arXiv: 2010.03958, 2020.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

Palmer, M. A., Reidy Liermann, C. A., Nilsson, C., Flörke, M., Alcamo, J., Lake, P. S., and Bond, N.: Climate change and the world's river basins: anticipating management options, Frontiers in Ecology and the Environment, 6, 81–89, 2008.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, p. 12, 2019.

Pilon, P. J.: Guidelines for reducing flood losses, Tech. rep., United Nations International Strategy for Disaster Reduction (UNISDR), 2002.

Pokharel, S. and Roy, T.: A Parsimonious Setup for Streamflow Forecasting using CNN-LSTM, arXiv preprint arXiv:2404.07924, 2024a.

Pokharel, S. and Roy, T.: A parsimonious setup for streamflow forecasting using CNN-LSTM, Journal of Hydroinformatics, p. jh2024114, 2024b.

RADOLAN, 2016: RADOLAN/RADVOR, https://opendata.dwd.de/climate_environment/CDC/grids_germany/hourly/radolan/, 2016.

Rakovec, O., Weerts, A., Hazenberg, P., Torfs, P., and Uijlenhoet, R.: State updating of a distributed hydrological model with Ensemble Kalman Filtering: effects of updating frequency and observation network density on forecast accuracy, Hydrology and Earth System Sciences, 16, 3435–3449, 2012.

Renard, K. G., Laflen, J., Foster, G., and McCool, D.: The revised universal soil loss equation, Routledge, 1994.

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resources Research, 46, 2010.

Schalge, B., Baroni, G., Haese, B., Erdal, D., Geppert, G., Saavedra, P., Haefliger, V., Vereecken, H., Attinger, S., Kunstmann, H., Cirpka, O. A., Ament, F., Kollet, S., Neuweiler, I., Hendricks Franssen, H.-J., and Simmer, C.: Presentation and discussion of the high-resolution atmosphere–land-surface–subsurface simulation dataset of the simulated Neckar catchment for the period 2007–2015, Earth System Science Data, 13, 4437–4464, https://doi.org/10.5194/essd-13-4437-2021, 2021.

Schmidt, L., Gusho, E., de Back, W., Vinogradova, K., Kumar, R., Rakovec, O., Attinger, S., and Bumberger, J.: Spatially-distributed Deep Learning for rainfall-runoff modelling and system understanding, in: EGU General Assembly Conference Abstracts, p. 20736, 2020.

Scholz, F., Traub, M., Zarfl, C., Scholten, T., and Butz, M. V.: Fully differentiable, fully distributed River Discharge Prediction: data sets, https://doi.org/10.5281/zenodo.13970576, 2024a.

Scholz, F., Traub, M., Zarfl, C., Scholten, T., and Butz, M. V.: Fully differentiable, fully distributed River Discharge Prediction: code, https://doi.org/10.5281/zenodo.13992584, 2024b.

Shen, C.: A transdisciplinary review of deep learning research and its relevance for water resources scientists, Water Resources Research, 54, 8558–8593, 2018.

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., Rahmani, F., Song, Y., Beck, H. E., Bindas, T., Dwivedi, D., Fang, K., Höge, M., Rackauckas, C., Mohanty, B., Roy, T., Xu, C., and Lawson, K.: Differentiable modelling to unify machine learning and physical models for geosciences, Nature Reviews Earth & Environment, 4, 552–567, https://doi.org/10.1038/s43017-023-00450-9, 2023.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, Advances in neural information processing systems, http://arxiv.org/abs/1506.04214v2, 2015.

Sit, M., Demiray, B., Xiang, Z., Ewing, G., Sermet, Y., and Demir, I.: A Comprehensive Review of Deep Learning Applications in Hydrology and Water Resources, https://doi.org/10.31223/osf.io/xs36g, 2020.

https://doi.org/10.5194/egusphere-2024-4119
Preprint. Discussion started: 7 March 2025

Sit, M., Demiray, B., and Demir, I.: Short-term Hourly Streamflow Prediction with Graph Convolutional GRU Networks, arXiv preprint arXiv:2107.07039, http://arxiv.org/abs/2107.07039v1, 2021.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic attribution for deep networks, in: International conference on machine learning, pp. 3319–3328, PMLR, 2017.

Traub, M., Becker, F., Sauter, A., Otte, S., and Butz, M. V.: Loci-segmented: improving scene segmentation learning, in: International Conference on Artificial Neural Networks, pp. 45–61, Springer, 2024a.

Traub, M., Scholz, F., Scholten, T., Zarfl, C., and Butz, M. V.: High-Efficiency Rainfall Data Compression Using Binarized Convolutional Autoencoder, Tech. rep., Copernicus Meetings, 2024b.

Tyson, C., Longyang, Q., Neilson, B. T., Zeng, R., and Xu, T.: Effects of meteorological forcing uncertainty on high-resolution snow modeling and streamflow prediction in a mountainous karst watershed, Journal of Hydrology, 619, 129 304, 2023.

Ueda, F., Tanouchi, H., Egusa, N., and Yoshihiro, T.: A Transfer Learning Approach Based on Radar Rainfall for River Water-Level Prediction, Water, 16, 607, 2024.

Ufrecht, W.: Ein Hydrogeologisches Modell für den Karst-und Mineralwasseraquifer Muschelkalk im Großraum Stuttgart, Hydrogeologische Modelle–ein Leitfaden mit Fallbeispielen, Schriftenreihe der Deutschen Geologischen Gesellschaft, 24, 2002.

Valeriano, O. C. S., Koike, T., Yang, K., and Yang, D.: Optimal dam operation during flood season using a distributed hydrological model and a heuristic algorithm, Journal of Hydrologic Engineering, 15, 580–586, 2010.

Van Vliet, M. T., Franssen, W. H., Yearsley, J. R., Ludwig, F., Haddeland, I., Lettenmaier, D. P., and Kabat, P.: Global river discharge and water temperature under climate change, Global Environmental Change, 23, 450–464, 2013.

Wi, S. and Steinschneider, S.: On the need for physical constraints in deep learning rainfall-runoff projections under climate change, EGUsphere, 2023, 1–46, 2023.

Wright, L.: Ranger - a synergistic optimizer., \urlhttps://github.com/lessw2020/Ranger-Deep-Learning-Optimizer, 2019.

Xiang, Z. and Demir, I.: Distributed long-term hourly streamflow predictions using deep learning –A case study for State of Iowa, Environmental Modelling & Software, 131, 104 761, https://doi.org/10.1016/j.envsoft.2020.104761, 2020.

Xiang, Z. and Demir, I.: Fully distributed rainfall-runoff modeling using spatial-temporal graph neural network, 2022.

Xu, T., Longyang, Q., Tyson, C., Zeng, R., and Neilson, B. T.: Hybrid physically based and deep learning modeling of a snow dominated, mountainous, karst watershed, Water Resources Research, 58, e2021WR030 993, 2022.

Yadan, O.: Hydra - A framework for elegantly configuring complex applications, Github, https://github.com/facebookresearch/hydra, 2019.

Zhu, S., Wei, J., Zhang, H., Xu, Y., and Qin, H.: Spatiotemporal deep learning rainfall-runoff forecasting combined with remote sensing precipitation products in large scale basins, Journal of Hydrology, 616, 128 727, 2023.