Neural Rainfall-Runoff Modeling

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Fedor Scholz

aus Oldenburg

Tübingen

2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	03.02.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Martin Butz
2. Berichterstatter/-in:	PD Dr.-Ing. Uwe Ehret

Neural Rainfall-Runoff Modeling

Copyright © 2025 - Fedor Scholz, Eberhard Karls University of Tübingen.

This dissertation is original work, written solely for this purpose, and all the authors whose studies and publications contributed to it have been duly cited. Partial reproduction is allowed with acknowledgment of the author and reference to the degree, academic year, institution—*University of Tübingen*—and public defense date.

Large language models were used for editorial assistance, including grammar checking, rephrasing, and improving clarity.

Preparation of this work was facilitated by the use of the *IPLeiria-Thesis* template (José Areia, 2023).

Abstract

Hydrological modeling traditionally relies on mechanistic models that, while physically interpretable, struggle to capture the nonlinear dynamics of environmental systems. Data-driven approaches such as artificial neural networks (ANNs) can automatically discover patterns in observational data, often achieving superior performance, but typically operate as black boxes that may not adhere to established physical constraints. This dissertation introduces DRRAiNN (Distributed Rainfall-Runoff ArtIficial Neural Network), a fully differentiable and fully distributed ANN architecture for rainfall-runoff modeling. DRRAiNN consists of two components: a spatially distributed rainfall-runoff model operating on a regular grid, and a graph-based river discharge model that captures flow dynamics along the river network. The rainfall-runoff component uses specialized convolutional recurrent networks with physics-informed inductive biases to model lateral water propagation and evapotranspiration processes. The discharge component employs a graph neural network that respects the connectivity of gauging stations. This modular architecture enables end-to-end optimization using sparse discharge measurements while maintaining physical plausibility through architectural constraints. The model is evaluated on the Neckar river catchment in Southwest Germany using data from 17 gauging stations. DRRAiNN demonstrates improved performance compared to the European Flood Awareness System (EFAS) across multiple metrics. The fully differentiable architecture enables interpretability analysis through gradient-based attribution methods. These techniques demonstrate the ability to reconstruct physically meaningful catchment boundaries that exhibit reasonable correspondence with topographically derived watersheds, which shows that the model learns interpretable spatial patterns. Leave-one-out cross-validation experiments evaluate the model's ability to generalize to ungauged basins, which is a critical requirement for practical hydrological applications. Analysis reveals important insights about ANN behavior in environmental modeling. The model achieves strong performance partly by using elevation data as positional encoding rather than explicit flow routing, which explains both its effectiveness within the training domain and its spatial generalization challenges. A notable trade-off emerges between predictive accuracy and physical plausibility, with the most physically realistic model instances not corresponding to those with optimal forecast performance. Key limitations include dependency on accurate precipitation forecasts, evaluation restricted to a single river network, and resolution constraints. The daily data resolution limits the model's ability

to capture rapid hydrological responses, while the $4\,\mathrm{km} \times 4\,\mathrm{km}$ grid may inadequately represent fine-scale processes. Future work should focus on multi-catchment training, higher temporal resolution data, and satellite-based approaches for ungauged regions. Overall, this work advances the field of hydrological modeling by demonstrating that carefully designed ANN architectures can achieve both high predictive accuracy and physical interpretability, opening new possibilities for knowledge discovery in hydrology while highlighting the importance of balancing performance optimization with physical realism.

Keywords: rainfall-runoff modeling, river discharge prediction, artificial neural networks, data-driven modeling, differentiable modeling, hybrid modeling, interpretable machine learning, spatio-temporal modeling

Zusammenfassung

Die hydrologische Modellierung stützt sich traditionell auf mechanistische Modelle, die zwar physikalisch interpretierbar sind, jedoch Schwierigkeiten haben, die nichtlineare Dynamik von Umweltsystemen zu erfassen. Datengesteuerte Ansätze wie künstliche neuronale Netze können automatisch Muster in Beobachtungsdaten erkennen und erzielen oft eine überlegene Leistung, funktionieren jedoch in der Regel als Black Boxes, die physikalische Beschränkungen verletzen können. Diese Dissertation stellt DRRAiNN (Distributed Rainfall-Runoff ArtIficial Neural Network) vor, eine vollständig differenzierbare und vollständig verteilte künstliche neuronale Architektur für die Regen-Abfluss-Modellierung. DRRAiNN besteht aus zwei Komponenten: einem räumlich verteilten Regen-Abfluss-Modell, das auf einem regelmäßigen Gitter arbeitet, und einem graphbasierten Flussabflussmodell, das die Strömungsdynamik entlang des Flussnetzes erfasst. Die Niederschlags-Abfluss-Komponente verwendet spezielle konvolutionelle rekurrente Netzwerke mit physikalisch informierten induktiven Verzerrungen, um laterale Wasserausbreitungs- und Evapotranspirationsprozesse zu modellieren. Die Abflusskomponente verwendet ein graphbasiertes neuronales Netzwerk, das die Konnektivität der Messstationen berücksichtigt. Diese modulare Architektur ermöglicht eine Ende-zu-Ende-Optimierung unter Verwendung spärlicher Abflussmessungen, während die physikalische Plausibilität durch architektonische Einschränkungen gewahrt bleibt. Das Modell wird am Neckar-Einzugsgebiet in Südwestdeutschland unter Verwendung von Daten aus 17 Messstationen evaluiert. DRRAiNN zeigt im Vergleich zum European Flood Awareness System (EFAS) eine überlegene Leistung bei mehreren Metriken. Die vollständig differenzierbare Architektur ermöglicht eine Interpretierbarkeitsanalyse durch gradientenbasierte Attributionsmethoden. Diese Techniken rekonstruieren erfolgreich physikalisch sinnvolle Einzugsgebietsgrenzen, die mit topografisch abgeleiteten Wasserscheiden übereinstimmen, und zeigen, dass das Modell interpretierbare räumliche Muster lernt. Leave-one-out-Kreuzvalidierungsexperimente bewerten die Fähigkeit des Modells, auf nicht gemessene Einzugsgebiete zu verallgemeinern, was eine entscheidende Voraussetzung für praktische hydrologische Anwendungen ist. Die Analyse liefert wichtige Erkenntnisse über das Verhalten künstlicher neuronaler Netze in der Umweltmodellierung. Das Modell erzielt eine starke Leistung, unter anderem durch die Verwendung von Höhendaten als Positionskodierung anstelle einer expliziten Flussführung, was sowohl seine Wirksamkeit innerhalb des Trainingsbereichs als auch seine Herausforderungen bei der räumlichen Generalisierung erklärt. Es ergibt sich ein

bemerkenswerter Kompromiss zwischen Vorhersagegenauigkeit und physikalischer Plausibilität, wobei die physikalisch realistischsten Modellinstanzen nicht mit denen mit optimaler Vorhersageleistung übereinstimmen. Zu den wichtigsten Einschränkungen zählen die Abhängigkeit von genauen Niederschlagsvorhersagen, die Beschränkung der Bewertung auf ein einziges Flussnetz und Einschränkungen hinsichtlich der Auflösung. Die tägliche Datenauflösung schränkt die Erfassung schneller hydrologischer Reaktionen ein, während das Raster von $4\,\mathrm{km} \times 4\,\mathrm{km}$ möglicherweise keine feinmaschigen Prozesse angemessen darstellt. Zukünftige Arbeiten sollten sich auf das Training mit mehreren Einzugsgebieten, Daten mit höherer zeitlicher Auflösung und satellitengestützte Ansätze für nicht gemessene Regionen konzentrieren. Insgesamt bringt diese Arbeit das Gebiet der hydrologischen Modellierung voran, indem sie zeigt, dass sorgfältig entworfene künstliche neuronale Netzwerkarchitekturen sowohl eine hohe Vorhersagegenauigkeit als auch physikalische Interpretierbarkeit erreichen können. Damit eröffnet sie neue Möglichkeiten für die Gewinnung von Erkenntnissen in der Hydrologie und unterstreicht gleichzeitig die Bedeutung eines Gleichgewichts zwischen Leistungsoptimierung und physikalischem Realismus.

Schlüsselwörter: Niederschlag-Abfluss-Modellierung, künstliche neuronale Netze, datengetriebene Modellierung, differenzierbare Modellierung, hybride Modellierung, interpretierbares maschinelles Lernen, räumlich-zeitliche Modellierung

Contents

List of Figures

List of Tables

Glossary

autoregressive see autoregressive model. (*p. 34, 53, 70*)

autoregressive model Model that uses its own previous outputs as inputs for subsequent predictions. (*p. xxvii, 42, 49, 53*)

fully distributed see fully distributed model. (*p. xv, 13, 14, 19–21, 25, 26, 33, 75, 81, 82, 89, 90*)

fully distributed model Model that operates on a spatial grid, maintaining full spatial resolution without aggregation across the modeling domain. (*p. xxvii, 12, 25, 31*)

hypernetwork Neural network that generates weights or parameters for another neural network, enabling dynamic, context-dependent model behavior. (*p. xvi, 35–38, 78*)

inductive bias Set of assumptions or constraints built into a learning algorithm that guide it toward certain solutions and enable generalization beyond training data. (*p. xv, xxiii, 5, 7, 22, 23, 25–27, 31, 34, 36, 64, 65, 78, 87, 89, 90*)

lumped see lumped model. (*p. xv, 11–13, 15, 19, 23–26, 29*)

lumped model Model that processes spatially aggregated inputs at the basin-scale. (*p. xxvii, 11, 12, 25*)

semi-distributed see semi-distributed model. (*p. 13, 14, 19, 23, 24*)

semi-distributed model Model that divides a catchment into sub-basins or hydrological response units, processing spatially aggregated inputs within each unit. (*p. xxvii, 12, 14, 20, 25*)

Acronyms

%BiasFHV percent bias in flow duration curve high-segment volume. (*p. 44–46, 52, 53, 56, 70, 72*)

%BiasFLV percent bias in flow duration curve low-segment volume. (*p. 44, 45, 52, 53, 56, 70, 72*)

%BiasFMS percent bias in flow duration curve mid-segment slope. (*p. 44–46, 52–54, 56, 70, 72*)

%BiasRR percent bias in overall runoff ratio. (*p. 44, 45, 52–54, 56, 70, 72*)

ANN artificial neural network. (*p. i, ii, xvi, 4, 5, 7, 16–19, 21–25, 29–31, 33, 36–38, 43, 68, 75–78, 82–84, 87–90*)

CNN convolutional neural network. (*p. xvi, 20, 27, 30, 35–38, 65, 87*)

DEM digital elevation model. (*p. xvii, 11, 33, 39, 40, 46, 59–62, 64, 80*)

DISTANA distributed spatio-temporal artificial neural network architecture. (*p. 37, 38*)

DRRAiNN Distributed Rainfall-Runoff ArtIficial Neural Network. (*p. i, v, xv–xx, xxiii, 4–7, 10, 11, 25–27, 31, 33–35, 38, 39, 41–44, 46–56, 58–68, 70, 71, 73–90*)

DWConv depth-wise convolution. (*p. xvi, xvii, xix, 35–38, 59, 61, 65–67*)

EFAS European Flood Awareness System. (*p. i, v, xvi–xx, xxiii, 5, 7, 14, 15, 27, 33, 40, 43, 48–56, 59, 61–65, 67, 70–73, 75*)

ET evapotranspiration. (*p. xvi, 9, 14, 21, 22, 30, 34–38, 43, 47, 62–65, 77, 82*)

FDC flow duration curve. (*p. 45*)

GNN graph neural network. (*p. 20, 27*)

GRU gated recurrent unit. (*p. 18–20, 27, 37, 38*)

KGE Kling-Gupta efficiency. (*p. xvi–xix, 44, 45, 50, 52–54, 56, 58–65, 67, 68, 70, 72*)

LSTM long short-term memory. (*p. xvi, xviii, 18–20, 23, 25, 27, 29, 35, 65, 66, 80*)

MAE mean absolute error. (*p. 44, 45, 52–54, 56, 70, 72*)

ML machine learning. (*p. 11, 16, 18, 29, 31, 42, 82, 84, 87*)
MLP multi-layer perceptron. (*p. xvi, 17, 27, 35, 37, 38, 65*)
MSE mean squared error. (*p. 42, 44*)

NSE Nash-Sutcliffe efficiency. (*p. xvii–xix, 44, 45, 52–54, 56, 58–65, 67, 68, 70, 72*)

PBM process-based model. (*p. 12, 13, 15, 24, 29*)
PCC Pearson's correlation coefficient. (*p. 44, 45, 52–54, 56, 70, 72*)
PUB prediction in ungauged basins. (*p. 7, 28, 49, 62, 68, 70, 72, 74*)
PWConv position-wise convolution. (*p. xvi, xviii, xix, 35–38, 63–67*)
PWLSTM position-wise long short-term memory. (*p. xvi, xviii, 35, 36, 65*)

RNN recurrent neural network. (*p. 18, 23, 27*)
RUSLE revised universal soil loss equation. (*p. 84–86*)

SiLU sigmoid linear unit. (*p. 36, 66*)

TCN temporal convolution network. (*p. 20, 27*)

1 Introduction

Early warnings and action save lives —
UN Secretary-General António Guterres,
World Meteorological Day 2022

1.1 Motivation

Climate change represents one of the most pressing challenges humanity faces. One reason for this is the intensification of the water cycle (Milly et al., 2002; Oki et al., 2006; Huntington, 2006; Vargas Godoy et al., 2023). In 2024, global average temperature was $1.29\,^{\circ}\mathrm{C}$ above the 20th century average (NOAA National Centers for Environmental Information, 2025), with human activities as the primary driver of this warming (Lynas et al., 2021). According to the Clausius-Clapeyron relation, the atmosphere's water holding capacity increases by $7\,\%$ for every $1\,^{\circ}\mathrm{C}$ (Trenberth, 2011; Vargas Godoy et al., 2023). This additional atmospheric moisture increases the probability of precipitation (Vargas Godoy et al., 2023) and intensifies storm events, thereby substantially increasing flood risks (Milly et al., 2002; Trenberth, 2011). Simultaneously, higher temperatures shift precipitation from snow to rain and accelerate snowmelt, further amplifying flood risks (Trenberth, 2011). These processes collectively result in more intense precipitation during wet seasons and stronger extreme events (Vargas Godoy et al., 2023). Conversely, reduced snowpack availability during summer months, combined with increased evaporation rates and surface drying, exacerbates water scarcity (Trenberth, 2011; Vargas Godoy et al., 2023; OECD, 2025). These changes have manifested in a doubling of global drought-affected areas since 1900, accompanied by widespread groundwater depletion affecting $62\,\%$ of monitored aquifers and declining streamflow in rivers worldwide (OECD, 2025). Moreover, drought conditions create cascading risks: When intense rainfall follows extended dry periods, hardened soils with reduced infiltration capacity generate excessive surface runoff, thereby paradoxically increasing flood risks. Collectively, these hydrological shifts present unprecedented societal challenges.
Globally, 1-in-100-year floods threaten 1.81 billion people (Rentschler et al., 2022). It is estimated that $89\,\%$ of these people live in low- and middle-income countries,

which makes them especially vulnerable. In 2024, the International Disaster Database (Delforge et al., 2025) recorded 48.8 million people affected by floods globally, with 5883 deaths and 32.8 billion USD in economic losses. Similarly, 29.5 million people were affected by droughts with 13.3 billion USD in economic losses. Additionally, there were 45.8 million internal displacements worldwide that are related to disasters in general, with 19.1 million of those being related to floods and 387 000 being related to droughts (Internal Displacement Monitoring Centre, 2025). The associated economic risk totals approximately 9.8 trillion USD, which is roughly 12 % of the global GDP (Rentschler et al., 2022).
Therefore, effective flood prediction and early warning systems have become increasingly critical. There is broad scientific consensus that flood early warning systems reduce the risks posed by floods (Pilon, 2002; Basher, 2006; Perera et al., 2019; Perera et al., 2020). This includes the reduction of the number of fatalities (Hallegatte, 2012; Cools et al., 2016; World Meteorological Organization, 2023) and economic damage (Hallegatte, 2012; Thielen-del Pozo et al., 2015; Pappenberger et al., 2015; Cools et al., 2016). More concretely, it is estimated that these kinds of systems can avoid multiple hundred fatalities and reduce the costs of these disasters by 460 million to 2.7 billion EUR per year in Europe (Hallegatte, 2012). Another study estimates that for every Euro invested into flood early warning systems, the benefit is about 400 EUR (Pappenberger et al., 2015). The effectiveness of these systems fundamentally depends on accurate water flow forecasting, which plays a critical role in mitigating short-term flood impacts. For example, simulating river discharge is a prerequisite for flood inundation modeling (Hunter et al., 2007). These challenges necessitate advanced systems capable of accurate river discharge prediction.

1.2 Problem statement

However, predictive accuracy alone is insufficient for effective hydrological modeling. Physically-based hydrological models demonstrate generalization capabilities and provide valuable insights into water movement processes. These capabilities are essential for advancing scientific understanding and informing practical water management strategies (Shen et al., 2023) such as dam operations (Valeriano et al., 2010). 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), with recent evidence showing that such adaptation measures have successfully reduced flood vulnerability by 39 % to 63 % across Europe since 1950 (Paprotny et al., 2025). Despite the increasing modeling challenges posed by more frequent extreme events, physically-informed models remain essential for assessing future climate change impacts on ecosystem dynamics (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 flood damage. From a practical perspective, effective

discharge prediction models must enable efficient calibration and perform well with sparse observational data, a common constraint in river discharge monitoring. Furthermore, they should operate efficiently in real-time, and perform well across diverse hydrological conditions. Achieving all these requirements simultaneously presents significant challenges.

Current modeling approaches face a fundamental trade-off between these objectives, particularly regarding interpretability and accuracy (Núñez et al., 2023). Traditional hydrological models naturally provide high interpretability and respect established physical principles (Brutsaert, 2023). However, they suffer from expensive and complex calibration procedures (Shen et al., 2023), a limited ability to exploit large datasets effectively (Shen et al., 2023), and poor generalization capabilities across different basins (Hrachowitz et al., 2013; Nearing et al., 2020). In contrast, modern data-driven approaches demonstrate superior predictive performance (Kratzert et al., 2018; Shen, 2018; Nearing et al., 2020; Gauch et al., 2021) by automatically discovering complex patterns in large datasets. However, they are often criticized for operating as "black boxes" that provide limited physical insight and may violate fundamental physical principles (Núñez et al., 2023). Despite the abundance of environmental data, these approaches often fail to extract meaningful scientific understanding, creating a data-rich but insight-poor situation. This issue is amplified by spatial representation limitations: The vast majority of existing data-driven approaches employ spatially aggregated, basin-scale representations, which limits their ability to capture fine-scale and inter-basin processes.

These fundamental limitations mean that existing modeling approaches struggle to simultaneously meet all requirements for effective hydrological prediction and management. Traditional models fail to achieve their full potential, as evidenced by the consistently higher accuracy of data-driven alternatives (Kratzert et al., 2018; Shen, 2018; Kratzert et al., 2019b; Nearing et al., 2020; Gauch et al., 2021), limiting the quality of predictions available for operational applications. Their poor spatial generalization capabilities imply that they have to be calibrated to each basin separately (Hrachowitz et al., 2013; Nearing et al., 2020), with these limitations particularly impacting low- and middle-income regions where computational resources and observational data are scarce. Both traditional and data-driven models become vulnerable when encountering conditions that exceed their calibration ranges. This is a growing concern as climate change increases the frequency of unprecedented hydrological events. The complexity of data-driven approaches, on the other hand, makes it hard to interpret their inner workings. However, interpretability is important for trust, policy decisions, and scientific advancement (Başağaoğlu et al., 2022; Núñez et al., 2023). The lack of adequate constraints in data-driven models poses the risk that these models learn spurious relationships that do not relate to real-world subprocesses (Butz et al., 2025). Spatially aggregated representations disregard important spatial heterogeneity, rendering them less physically aligned with reality (Okiria et al., 2022). They also prevent understanding of water source origins and therefore limit the ability to identify flood-contributing

areas for targeted management interventions. This limitation is problematic for climate change adaptation, where understanding changing spatial patterns of water movement becomes critical for infrastructure planning and risk assessment.
To address these fundamental limitations, especially accuracy, interpretability, and the reliance on spatially aggregated representations, we develop a novel modeling approach that bridges the gap between data-driven performance and physical interpretability. Our approach combines the predictive power of neural networks with physics-inspired architectural choices in a fully distributed, fully differentiable framework that requires only sparse discharge observations for training.

1.3 Contribution

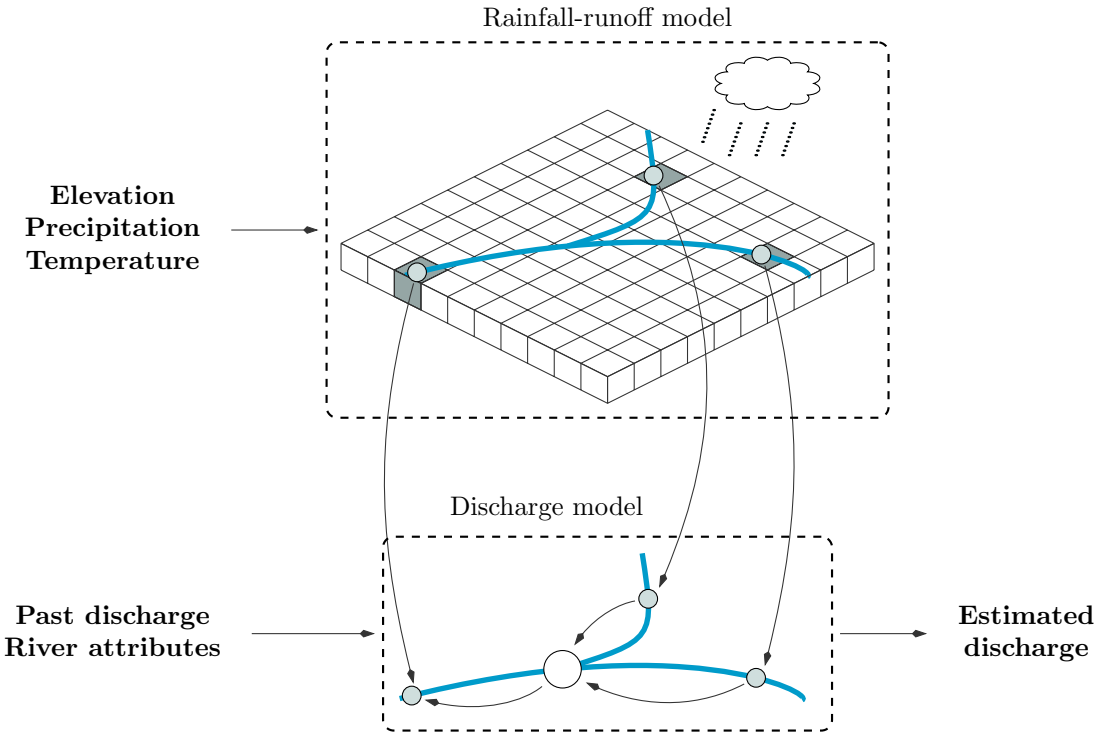


Figure 1.1: *An overview of DRRAiNN. The gridded rainfall-runoff model propagates precipitation across the landscape according to elevation and models evapotranspiration based on temperature. The graph-based discharge model then receives the rainfall-runoff model state at gauging station locations and processes it together with the previous (potentially estimated) discharge. This processing considers station adjacency, altitude differences, and inter-station river segment lengths. The output is discharge at each station.*

We present DRRAiNN (Distributed Rainfall-Runoff ArtIficial Neural Network), a fully differentiable, fully distributed rainfall-runoff model. As illustrated in Figure 1.1, DRRAiNN combines data-driven modeling with physics-inspired architectural choices to address the challenges of predictive accuracy, physical plausibility, interpretability, and latent variable inference. Our spatio-temporal artificial neural network (ANN) architecture estimates river discharge at gauging stations from gridded precipitation, temperature, elevation, and past discharge. DRRAiNN is fully distributed in the sense

that it internally operates on a grid. However, its outputs are point-wise river discharge measurements at given gauging station locations. Its full differentiability allows gradients to flow seamlessly through the entire system, enabling end-to-end optimization of all its components with sparse, point-wise discharge measurements being the only target variable. To avoid overfitting, and to improve interpretability and generalization, we incorporated several physics-inspired inductive biases (Gauch et al., 2020) into DRRAiNN. These include the modularization into a spatially fully distributed rainfall-runoff model and the utilization of a graph-based river discharge model. Additional architectural choices precondition DRRAiNN to encode distinct processes, such as lateral propagation of water across the landscape and local evapotranspiration.
Due to DRRAiNN's fully distributed and fully differentiable architecture, it is possible to answer spatially explicit questions, such as: Where is the true catchment area based on the observed hydrological dynamics? In other words, DRRAiNN enables source allocations using gradient-based attribution methods like integrated gradients (Sundararajan et al., 2017). These techniques can help to examine and understand internal model dynamics, enabling knowledge discovery. As a result, ANN approaches can transition from black-box models toward more interpretable, process-informed models. An appropriate model design can enforce the development of meaningful components that correspond to subprocesses of the overall real-world process. Their inner workings can be interrogated with interpretability methods to show what their estimations are based on.
Empirical evaluation on the Neckar river network demonstrates that DRRAiNN consistently outperforms the European Flood Awareness System (EFAS) across standard hydrological evaluation metrics. DRRAiNN maintains this superior performance over 50-day forecast horizons despite being trained to forecast autonomously for only 10 days, demonstrating the potential of physics-informed ANN architectures to achieve competitive performance with operational hydrological models while offering enhanced interpretability.

1.4 Research questions

Given DRRAiNN's unique combination of capabilities we investigate several questions that examine both DRRAiNN's practical performance and its potential for advancing hydrological understanding:

RQ1 Can DRRAiNN achieve performance comparable to state-of-the-art operational models, even if trained solely on sparse discharge data?

RQ2 Do patterns of discharge prediction difficulty vary systematically across gauging stations, and what factors might explain these spatial performance variations?

RQ3 Can gradient-based attribution methods applied to DRRAiNN's fully differentiable architecture reconstruct physically meaningful catchment boundaries without requiring precomputed watershed delineations?

RQ4 What is the relationship between predictive accuracy and physical plausibility in DRRAiNN?

RQ5 How do different input variables (precipitation, elevation, and temperature) contribute to DRRAiNN's performance and physical plausibility, as assessed through ablation studies and gradient-based attributions?

RQ6 Which physics-inspired architectural choices most effectively improve model performance and interpretability?

RQ7 How effectively does DRRAiNN generalize to ungauged basins?

1.5 Author's publications and research contributions

Publications:

- Fedor Scholz et al. (Feb. 2022). "Inference of Affordances and Active Motor Control in Simulated Agents". In: *Frontiers in Neurorobotics*. DOI: 10.3389/fnbot.2022.881673. arXiv: 2202.11532v3 [cs.AI]. URL: http://arxiv.org/abs/2202.11532v3
- Fedor Scholz et al. (2024d). "Quick and Accurate Affordance Learning". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 46. URL: https://escholarship.org/uc/item/21b6p6tt
- Fedor Scholz et al. (Mar. 2025a). "Fully differentiable, fully distributed Rainfall-Runoff Modeling". In: *EGUsphere* 2025, pp. 1–37. DOI: 10.5194/egusphere-2024-4119. URL: https://doi.org/10.5194/egusphere-2024-4119 (accepted)

Conference presentations:

- Manuel Traub et al. (Apr. 2024a). "High-Efficiency Rainfall Data Compression Using Binarized Convolutional Autoencoder". In: *EGU General Assembly 2024*. Vienna, Austria, EGU24–11768. DOI: 10.5194/egusphere-egu24-11768. URL: https://doi.org/10.5194/egusphere-egu24-11768
- Fedor Scholz et al. (Apr. 2024c). "Introducing a fully differentiable, fully distributed Rainfall-Runoff Model". In: *EGU General Assembly 2024*. Vienna, Austria, EGU24–5298. DOI: 10.5194/egusphere-egu24-5298. URL: https://doi.org/10.5194/egusphere-egu24-5298
- Fedor Scholz et al. (Apr. 2025b). "Inference of catchment areas from modeled discharge dynamics". In: *EGU General Assembly 2025*. Vienna, Austria, EGU25–19057. DOI: 10.5194/egusphere-egu25-19057. URL: https://doi.org/10.5194/egusphere-egu25-19057

Software and data:

- Fedor Scholz et al. (Oct. 2024a). *Fully differentiable, fully distributed River Discharge Prediction: code*. DOI: 10.5281/zenodo.13992583. URL: https://zenodo.org/records/15484058

- Fedor Scholz et al. (Oct. 2024b). *Fully differentiable, fully distributed River Discharge Prediction: data sets*. DOI: 10.5281/zenodo.13970575. URL: https://zenodo.org/records/15482198

This dissertation builds upon and extends the work presented in Scholz et al., 2025a, with portions of text adapted and revised from that publication. The current work advances the original approach through improved input feature selection, refined model architecture, and enhanced training procedures, resulting in better predictive performance and more robust model behavior. Additionally, the dissertation provides an extended analysis including additional evaluation metrics, prediction in ungauged basins (PUB), and in-depth discussion of the results. These improvements and extensions are detailed in Section 3.7.

1.6 Outline

The main contribution of this thesis is the development and evaluation of DRRAiNN, a fully differentiable, fully distributed ANN for river discharge modeling. While this work is largely technical, we also provide context within the broader landscape of rainfall-runoff modeling approaches. Chapter 2 provides a literature review positioning DRRAiNN within current modeling paradigms, covering hydrological fundamentals, lumped and distributed conceptual and process-based models, data-driven approaches, and hybrid/differentiable modeling to establish the theoretical foundation for DRRAiNN's architecture. Chapter 3 details DRRAiNN's technical architecture, which combines distributed rainfall-runoff modeling with graph-based discharge prediction. This chapter describes the physics-inspired inductive biases, implementation details, evaluation metrics, and gradient-based attribution methods for assessing physical plausibility. Chapter 4 evaluates DRRAiNN's performance against EFAS across 50-day lead times and analyzes station-specific variations. We demonstrate physically meaningful catchment reconstruction through attribution methods, examine spatial generalization through leave-one-out cross-validation, investigate effects of meteorological and static inputs, and present ablation studies examining key architectural components. Chapter 5 summarizes the results, addresses each research question individually, acknowledges model limitations, explores connections between cognitive science and environmental modeling, and outlines directions for future work.

2

Background

All models are approximations. Essentially, all models are wrong, but some are useful —
George E. P. Box

2.1 Hydrological modeling fundamentals

Hydrology concerns itself with the different natural processes that involve water and occur within the earth system (Brutsaert, 2023). Together, these processes make up the water cycle as illustrated in Figure 2.1. The water cycle represents a complex, interconnected system where water continuously moves between the atmosphere, land surface, and subsurface, undergoing various transformations across multiple spatial and temporal scales.

Water vapor situated in the atmosphere condenses and falls onto land surface or water bodies as precipitation in various forms including rain, snow, and sleet. Once precipitation reaches the land surface, it can follow multiple pathways: Some water may be intercepted by vegetation, while other portions infiltrate into soils, contribute to surface runoff, or accumulate as snow and ice. Water that infiltrates into soils may be stored temporarily as soil moisture, taken up by plant roots, or percolate deeper to recharge groundwater aquifers. Surface runoff collects in streams and rivers, eventually flowing toward lakes, wetlands, or the ocean. At any stage in these pathways, water can return to the atmosphere via evapotranspiration (ET), the combined process of evaporation from water and soil surfaces and transpiration from vegetation. Solar energy serves as the main driver of the water cycle, providing the energy necessary for ET, while gravity drives the movement of water through surface and subsurface pathways.

The global scale of these processes is remarkable: Worldwide there is $111\,000\,\mathrm{km}^3$ precipitation per year (Oki et al., 2006). Of that, $45\,500\,\mathrm{km}^3$ eventually reach the ocean via rivers, with a mean residence time of water in unmodified rivers of approximately 2.5 weeks. The remaining $65\,500\,\mathrm{km}^3$, and therefore more than half, undergo ET. Accounting for ET is therefore crucial for most hydrological models.

The different kinds of flow can be broadly categorized into surface and subsurface flows (Brutsaert, 2023). Surface flow includes overland flow occurring directly on the

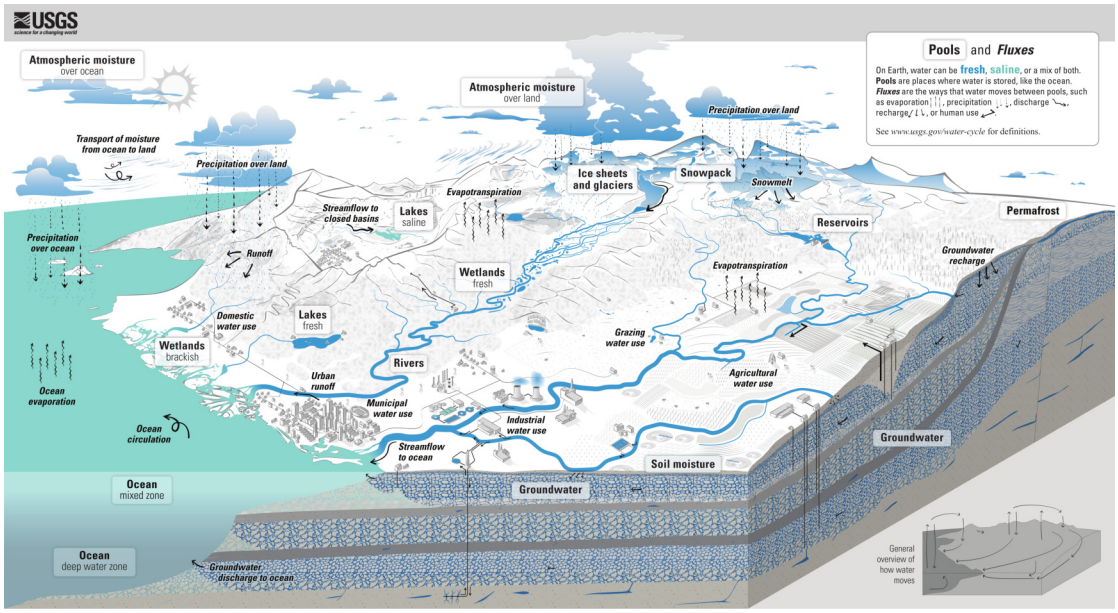


Figure 2.1: *The USGS water cycle diagram (United States Geological Survey, 2022). An online version with zoom functionality can be found at* `https://labs.waterdata.usgs.gov/visualizations/water-cycle/index.html#/`. *The diagram demonstrates the complexity of the terrestrial water cycle with its many storages and processes. In this work, we will focus on the subset of the water cycle that is situated on and below the ground, i.e., the subset between precipitation and river discharge, while considering evapotranspiration and ignoring human interventions. Precipitation will be given, i.e., we will not model atmospheric moisture.*

land surface as river, stream, or channel flow. Subsurface flow includes interflow occurring directly below the surface and groundwater occuring at greater depths. Of the $45\,500\,\mathrm{km}^3$ discharge per year, only around $15\,300\,\mathrm{km}^3$ flows upon the surface, while the remainder takes place in the form of subsurface runoff (Oki et al., 2006). $10\,000\,\mathrm{km}^3$ of the subsurface flow go into the ocean without ever seeing the surface. Since these flows are situated within different media, they exhibit different dynamics. River flow, for example, is almost one-dimensional and therefore easier to model than overland flow.

2.1.1 Terminology

In this study, the terms *discharge* and *streamflow* refer to the same concept, namely the amount of water that flows through the cross-section of a river or stream at a given point in a fixed time interval. Therefore, it is measured in $\mathrm{m}^3\,\mathrm{h}^{-1}$. Runoff, on the other hand, refers to the amount of water that is currently flowing inside a given catchment. DRRAiNN outputs river discharge only at specific points in space, namely where the gauging stations are located, making it a river discharge model. To achieve this, however, DRRAiNN also models how precipitation is converted into runoff and how runoff moves across the landscape. Therefore, we refer to DRRAiNN as a rainfall-runoff model, even though it is a combination of a rainfall-runoff model and a discharge model.

A (*river*) *catchment* (*area*), *drainage area*, or (*drainage*) *basin* refers to the area from which

precipitation eventually reaches a given point. Even though a catchment area exists for every point on the landscape, these points are usually located on a stream or river. In this study, we are interested in the catchment areas that belong to the gauging stations.
Regarding different modeling approaches, terminology can vary between different subgroups of the hydrological community. In this work, we use the term *mechanistic model* for any hydrological model that explicitly represents the underlying mechanisms of a hydrological system through components and relationships that correspond to real-world quantities and processes or abstractions thereof. This contrasts with data-driven models that learn patterns automatically without explicit mechanism representation. We regard *conceptual models*, which represent hydrological understanding through simplified structures such as storage reservoirs, and *process-based models*, which encode mechanisms through physical equations, as subsets of mechanistic models.
In the machine learning (ML) community, the term *prediction* is often used for the output of a model, regardless of time. In hydrology, on the other hand, *prediction* is reserved for model outputs that lie in the future (Gauch et al., 2020). If the output does not lie in the future, the term *simulation* is used. In hydrological terms, DRRAiNN therefore produces *simulations* since we provide observed meteorological forcings over the whole time series. According to Beven et al., 2013, DRRAiNN is an ex-post forecasting model, underlining the fact that DRRAiNN receives discharge observations up to some time step and from then on estimates discharge. This approach enables assessment of model performance under controlled conditions with known meteorological inputs, though operational forecasting would require coupling with weather prediction models. Our approach is described in more detail in Section 3.1.

2.1.2 Spatial scale: lumped vs. distributed models

Different hydrological models, whether mechanistic or data-driven, can act on various spatial scales. In lumped models, which represent the simplest case, inputs are aggregated over space at the basin scale before being fed into the model. This includes static attributes like elevation and meteorological forcings like precipitation and temperature. The fact that two spatial dimensions of the inputs to these models are removed makes these models computationally cheap, which can be a big advantage. To compute the spatially averaged quantities, the availability of the exact outline of the basin is a requirement. This boundary is usually inferred algorithmically from a digital elevation model (DEM).
The lumped modeling approach is based on the strong assumption that there is no lateral subsurface inflow into or outflow out of the basin. Considering that around $30\,200\,\mathrm{km}^3$ of the overall $45\,500\,\mathrm{km}^3$ discharge per year is situated below the ground, this can be a problem. Furthermore, lumped models cannot distinguish between different locations within the catchment, disregarding the basin's spatial heterogeneity. This can be a problem as well, especially in large basins and at fine time scales: Precipitation that occurs further upstream takes longer to reach the point of interest than

precipitation occuring right at the point of interest.
Semi-distributed models take one step toward spatial modeling by leveraging the river network topology. Here, an overall basin is divided into multiple subbasins which are connected via a tree-like graph structure that is determined by the subbasins' connectivity. Each subbasin is modeled separately and their interactions are modeled by a routing model that allows limited spatial interactions. The routing model usually aggregates the output stemming from the individual subbasins, again neglecting potential exchange of water between subbasins, however, more complex routines are possible. Semi-distributed models can be better at dealing with spatial heterogeneity, where the effectiveness largely depends on how fine-grained the division into different subbasins is.
Fully distributed models represent the most comprehensive spatial modeling approach. They operate without predetermined catchment boundaries on grid or grid-like structures to model both water content within cells and usually the dynamic exchange of water between cells. While they do not necessarily provide superior predictive performance compared to lumped models (Vansteenkiste et al., 2014), they offer unique advantages for spatial scenario simulations and provide explicit spatial patterns of state variables that can enhance process understanding (Tran et al., 2018). Rather than relying on fixed catchment delineations, these models have to implicitly infer catchment boundaries from observed dynamics, potentially capturing unobservable underground flows that the other approaches cannot represent. Since distributed models operate on a data structure with two additional spatial dimensions compared to lumped approaches, these models can be very expensive to calibrate and run.

2.2 Traditional approaches: mechanistic models

Traditionally, hydrological modeling has been addressed using mechanistic models that encode domain knowledge in some way. Usually, the different variables and equations in these models directly relate to real-world quantities and processes or abstractions thereof. This renders mechanistic models inherently interpretable, allowing researchers to ask specific questions by probing them. The gained information can then be used to advance our scientific understanding or, more concretely, allows modelers to assess how hypothetical situations would play out. We differentiate between conceptual models and process-based models (PBMs).
Conceptual models usually contain storages (often called buckets or reservoirs) that relate to real-world storages of water, such as the amount of snow or soil moisture. The storages are usually connected by simplified relationships that are loosely based on an intuitive, physical understanding. However, these relationships can also be based on more complex physical equations. Conceptual models typically require extensive calibration to achieve acceptable performance. This calibration process involves adjusting model parameters to optimize the fit between simulated and observed data. While necessary for model performance, calibration can introduce challenges: Parameters

may take on unrealistic values that compensate for errors in model structure, input data, or other parameter values (McMillan et al., 2016). This parameter fitting process can inadvertently introduce black-box elements into otherwise interpretable models, as calibrated parameters may not reflect actual physical properties.
In PBMs, encoded domain knowledge is explicitly represented by physical equations (often differential equations). These models describe physical processes with mathematical equations derived from physical laws and observations (Brutsaert, 2023). Some components of PBMs may be inferred from experiments in a laboratory, such as Darcy's law (Darcy, 1856). Others are based on simplifications of more general physics equations. For example, the shallow water equations can be derived by depth-integrating the more general Navier-Stokes equations (Vreugdenhil, 1994). These equations rely on the assumption that water depth is much smaller than the horizontal extent and that vertical velocities are very small. Due to their simpler form, they are easier to handle and can be used to model wave propagation in a river among others. Since PBMs are grounded in physical principles, they require little to no calibration, with parameters directly corresponding to measurable physical properties (Gauch et al., 2020).
Both conceptual models and PBMs can be implemented using either lumped or distributed spatial approaches. In lumped implementations, each type of storage or process is represented once for the entire basin, aggregating spatial variability into basin-scale parameters. In fully distributed implementations, each storage type and process exists for every grid cell, allowing explicit representation of spatial heterogeneity across the landscape. Distributed PBMs are computationally intensive, as they typically solve differential equations in each grid cell across potentially thousands of spatial locations. However, modern computing capabilities make fully distributed approaches increasingly feasible, enabling more detailed spatial process representation than was previously practical.

2.2.1 Examples

In the following, we provide an overview over notable mechanistic models. This list is by no means complete and its order is arbitrary.

Lumped models

The HBV model represents a classical conceptual approach, using simple storage reservoirs to represent snow accumulation and soil moisture and a routine that transforms runoff into discharge (Bergström, 1976). Despite this simplicity, HBV has demonstrated effectiveness across diverse conditions and remains widely used in operational forecasting systems throughout Scandinavia (Abebe et al., 2010). Some later versions of HBV were equipped with a routing routine for different subbasins, turning the model into a semi-distributed one (Bergström, 1992).
The Sacramento Soil Moisture Accounting (SAC-SMA) model employs a multi-layer

soil representation with separate routines for surface and subsurface flow, and ET (Burnash, 1973). This model was historically used in operational forecasting applications by the National Weather Service of the US (M. B. Smith et al., 2003) and remains an important benchmark model in hydrological research.
The GR4J model exemplifies the pursuit of model parsimony, achieving competitive performance with only four parameters representing production store capacity, groundwater exchange, routing store capacity, and unit hydrograph time base (Perrin et al., 2003).
HBV, SAC-SMA, and GR4J demonstrate that effective rainfall-runoff modeling can be achieved without complex process representations. However, since all of these models are conceptual models containing parameters that are not directly measurable, manual or automatic calibration of these models is essential.

Distributed models

TOPMODEL leverages topographic information to predict spatial patterns of soil moisture and runoff generation in subbasins that are connected via a routing model (Beven et al., 1979). This makes TOPMODEL a semi-distributed model that incorporates some degree of spatial heterogeneity while maintaining computational efficiency.
SWAT (Soil and Water Assessment Tool) was developed as an operational model that provides combined representations of water flow, sediment transport, and nutrient cycling with land use and management practices (Arnold et al., 1998). The model contains a routing routine that allows it to operate on natural subbasins or grid cells, rendering it semi-distributed or fully distributed.
SHE represents a fully distributed modeling approach that couples surface water flow, unsaturated zone processes, groundwater flow, and ET through rigorous physical equations (Abbott et al., 1986a; Abbott et al., 1986b). This comprehensive framework allows detailed investigation of water cycle interactions but requires extensive data and computational resources. SHE has been implemented in various modeling platforms such as MIKE SHE (Refshaard et al., 1995).
TOPMODEL, SWAT, and SHE are physics-based models, whose parameters can be measured or estimated from real-world observations, avoiding the need for calibration and allowing the models to be applied to ungauged basins.
The Variable Infiltration Capacity (VIC) is a fully distributed rainfall-runoff model that was designed for large-scale hydrology and climate studies, incorporating subgrid variability in soil moisture storage capacity and vegetation characteristics (Liang et al., 1994). VIC's emphasis on energy and water balance makes it well-suited for climate change impact studies and global hydrological modeling. Due to the presence of parameters in VIC that do not correspond to real-world quantities, calibration of this model is necessary.
The European Flood Awareness System (EFAS) provides continental-scale flood forecasting by combining meteorological forecasts with the LISFLOOD distributed hydro-

logical model (Thielen et al., 2009). EFAS demonstrates the application of a PBM at operational scales, providing early flood warnings across Europe. Since we will use EFAS as a benchmark model in this work, we provide a more detailed description in Section 3.5. GloFAS (Global Flood Awareness System) extends the EFAS approach to global coverage, coupling the HTESSEL land surface model with LISFLOOD to provide worldwide flood forecasting capabilities (Alfieri et al., 2013). This system illustrates how mechanistic modeling approaches can be scaled to address global water management challenges. In contrast to EFAS, GloFAS does not require calibration.

2.2.2 Limitations and challenges

Mechanistic hydrological models face several fundamental limitations that constrain their effectiveness in addressing modern modeling challenges. While mechanistic models offer valuable interpretability and process understanding, their practical application encounters significant barriers that limit their potential impact.

Hydrological processes are inherently complex, involving numerous interacting variables that create heterogeneous behavior across space and time (Marçais et al., 2017). The involved processes and their interactions are only partially understood in most cases (Hrachowitz et al., 2013), leading to high uncertainty and systematic biases in model predictions. This incomplete understanding is compounded by the fact that many key processes occur at scales that differ substantially from those observed under controlled laboratory conditions (Hrachowitz et al., 2013; Shen, 2018; Nearing et al., 2020).

While data assimilation techniques that incorporate observations into running models (Yuqiong Liu et al., 2012; Montzka et al., 2012; Camporese et al., 2022) have shown promise for reducing uncertainty and improving initialization in both lumped (Moradkhani et al., 2005; Yuqiong Liu et al., 2007; Yuqiong Liu et al., 2012) and distributed models (Rakovec et al., 2012), fundamental challenges remain.

Even when processes are well understood, critical input variables may be unobservable or difficult to measure directly. Underground topography, for example, significantly influences subsurface flow patterns but cannot be measured comprehensively across entire catchments. This creates fundamental limitations in model parameterization and validation, as key drivers of hydrological behavior remain hidden from direct observation.

Despite the rapidly increasing availability of hydrological data (Sit et al., 2020), mechanistic models struggle to fully exploit these rich datasets. Parameter calibration typically relies on limited data subsets and remains a lengthy, expert-driven process (Shen et al., 2023). This manual calibration approach not only limits the amount of information that can be effectively incorporated but also introduces human bias and subjectivity into the modeling process.

Distributed PBMs have significant computational demands (Ehret et al., 2020; Herrera et al., 2022) that limit their practical applicability (Vivoni et al., 2011). These models

must solve differential equations in each grid cell across potentially thousands of spatial locations and long time series, leading to substantial computational costs. The numerical solution of coupled partial differential equations, especially when accounting for complex boundary conditions and nonlinear processes, can become prohibitively expensive for real-time applications or large-scale studies. This computational burden is problematic for operational forecasting systems that require rapid model execution, and for ensemble simulations needed for uncertainty quantification (Vivoni et al., 2011).

Perhaps most critically, conceptual models typically require basin-specific calibration to achieve acceptable performance. Parameters calibrated for one basin often do not generalize well to other basins with different characteristics (Hrachowitz et al., 2013; Nearing et al., 2020). This limitation severely constrains the applicability of these models, particularly in data-sparse regions. The problem is especially acute for countries in the Global South, where detailed land surface and subsurface measurements are often unavailable, yet intelligent water management strategies may be most urgently needed.

These limitations collectively highlight the need for modeling approaches that can better exploit available data while maintaining physical interpretability and computational efficiency. This challenge motivates the data-driven and hybrid approaches explored in subsequent sections.

2.3 Data-driven approaches

Complementary to mechanistic models, ML and data-driven approaches have gained significant traction in hydrology in recent years, driven by the increasing availability of hydrological data (Sit et al., 2020). Unlike mechanistic models that encode domain knowledge explicitly, ML models automatically learn complex patterns and relationships directly from large datasets (Goodfellow et al., 2016). This fundamental difference in approach makes data quantity and quality crucial factors: Data-driven model performance typically scales with the quantity and quality of available training data. However, they struggle with inaccurate, incomplete, or biased datasets, which can lead to spurious correlations and poor generalization.

2.3.1 Artificial neural networks

Artificial neural networks (ANNs) represent a powerful class of data-driven models loosely inspired by biological neural networks (Goodfellow et al., 2016). ANNs consist of interconnected artificial neurons. These simple computational units receive weighted inputs, apply a nonlinear activation function, and produce outputs. Individual neurons are remarkably simple, yet when organized in layers and connected in networks, they possess the ability to approximate arbitrary continuous functions (Hornik et al., 1989). This universal approximation property makes ANNs suitable

for capturing the complex, nonlinear relationships inherent in hydrological systems.

An ANN typically consists of an input layer that receives data, one or more hidden layers that perform computations, and an output layer that produces predictions. Each connection between neurons has an associated weight parameter that determines the strength of the connection, and each neuron applies an activation function to introduce nonlinearity into the model. The combination of many such simple operations across multiple layers constitutes a multi-layer perceptron (MLP) and enables it to learn increasingly complex representations of the input data.

Training an ANN involves finding the optimal set of weights that minimize prediction errors on the training data. This optimization problem is typically solved through gradient descent, an iterative algorithm that adjusts weights in the direction of steepest decrease of a loss function (Goodfellow et al., 2016). The loss function quantifies the difference between the model's predictions and the true target values, providing a scalar measure of model performance that can be optimized.

For deep networks with many layers and potentially millions of parameters, computing gradients efficiently requires backpropagation. This algorithm applies the chain rule of calculus to propagate error gradients backwards through the network from output to weights and potentially inputs (Goodfellow et al., 2016). Backpropagation enables the calculation of gradients for all parameters in a single forward and backward pass through the network, making training of large ANNs computationally feasible.

Modern implementations rely heavily on automatic differentiation frameworks that can compute exact gradients of complex computational graphs automatically (Goodfellow et al., 2016). These frameworks track all mathematical operations performed during the forward pass and automatically generate the corresponding gradient computations for the backward pass. This capability has been crucial for the rapid development and adoption of increasingly sophisticated ANN architectures.

The training process for ANNs can be lengthy and computationally expensive, often requiring specialized hardware (such as GPUs) and substantial computational resources. Training time depends on factors including network size, dataset size, and convergence criteria, and may range from minutes for simple problems to days or weeks for large-scale applications. However, once an ANN is trained, inference, i.e., making predictions on new data, is typically fast and computationally inexpensive, making trained models suitable for real-time applications.

2.3.2 Artificial neural networks in hydrology

The application of ANNs in hydrology has gained significant momentum in recent years, driven by the steadily increasing availability of hydrological data (Sit et al., 2020). Data-driven models can be trained on vast amounts of observational data and automatically infer complex relationships that might be difficult to represent explicitly in mechanistic models. This capability has proven valuable in hydrological systems, where numerous interacting processes create complex, nonlinear relationships that are

challenging to capture through traditional modeling approaches.
While other Earth sciences have rapidly adopted data-driven techniques, hydrology has been somewhat slower to embrace these approaches (Shen, 2018). This slower adoption stems from several factors, including significant community skepticism toward non-physical models (Blöschl et al., 2019) and concerns about interpretability and physical plausibility of purely data-driven approaches. Additionally, the hydrological community faced practical challenges in model comparison and validation, as the field historically lacked standardized benchmarks enabling fair comparisons between different modeling approaches (Hrachowitz et al., 2013; Sit et al., 2020; Nearing et al., 2020), though recent efforts have begun to address this gap (Kratzert et al., 2023).
Despite these initial challenges, previous research has consistently demonstrated that ML approaches often outperform traditional methods in terms of predictive accuracy across diverse hydrological applications (Kratzert et al., 2018; Shen, 2018; Kratzert et al., 2019b; Nearing et al., 2020; Gauch et al., 2021; Gauch et al., 2020). These successes have established ANNs as powerful tools for hydrological modeling, capable of achieving superior performance while automatically discovering patterns in complex datasets. The growing recognition of these capabilities, combined with advances in addressing physical constraints and interpretability concerns, suggests that the full potential of data-driven approaches in hydrology remains largely untapped (Shen, 2018; Nearing et al., 2020).
For a comprehensive overview of ML applications in hydrology, we refer readers to Shen, 2018 and Sit et al., 2020.

2.3.3 Recurrent neural networks and storage equations

In order for ANNs to be able to model time series, they are equipped with a memory, called the hidden state. These so-called recurrent neural networks (RNNs) process time series in a sequential manner, updating their hidden state based on the last hidden state and the current input. This allows RNNs to store and memorize information across multiple time steps. A so-called long short-term memory (LSTM) is a special type of RNN that maintains two states, the hidden state and the cell state c (Hochreiter et al., 1997). The cell state is shielded from undesired updates by a gating mechanism, allowing the LSTM to memorize information across longer time intervals. Given the forget gate f, the input gate i, and a candidate hidden state g, the cell state c is updated in the following manner: $c_t = f \cdot c_{t-1} + i \cdot g$. A gated recurrent unit (GRU) is another type of RNN, which does not maintain a cell state, but updates its hidden state h in a similar manner via an update gate z and a candidate hidden state g (Cho et al., 2014): $h_t = z \cdot h_{t-1} + (1 - z) \cdot g$.
As described in Kratzert et al., 2018, there are striking parallels between conceptual, storage-based models and LSTMs and, for that matter, GRUs. Both are designed to process time series in a sequential manner by utilizing state variables. In a storage-based model, the update equation for the storage s usually has the following form:

$s_t = s_{t-1} \cdot p_1 + x \cdot p_2$, where x denotes the input and p_1 and p_2 represent parameters that determine how the state is updated, akin to the gates in LSTMs.
An important difference between these models is that in storage-based models, the parameters p_1 and p_2 are based on a physical understanding of the underlying process and tuned manually by the modeler. In LSTMs and GRUs, being data-driven models, the parameters that determine the operation of the gates are learned from observations. This means that their parameters do not necessarily reflect actual physical properties.

2.3.4 Examples

The application of data-driven models in hydrology has evolved across different spatial scales, from lumped approaches through semi-distributed to fully distributed methods.

Lumped models

In their seminal work, Kratzert et al. successfully applied an LSTM network for rainfall-runoff modeling at the basin scale (Kratzert et al., 2018), demonstrating that purely data-driven models can exceed the performance of traditional methods. This groundbreaking study established LSTMs as a viable alternative to conceptual models and sparked widespread interest in ANN applications for hydrology. Following this initial success, numerous studies have emerged applying largely similar LSTM-based architectures to various datasets and regions (Sit et al., 2020). However, significant methodological advancements to this framework have also been made. This includes the incorporation of physical constraints like mass conservation (Kratzert et al., 2019b; Hoedt et al., 2021; Y.-H. Wang et al., 2024), uncertainty estimation to quantify prediction confidence (Klotz et al., 2022; Nearing et al., 2023), and the extension of modeling to multiple timescales (Gauch et al., 2021; Acuña Espinoza et al., 2025).

Distributed models

Several semi-distributed strategies exist for incorporating connectivity between subbasins. One might argue that these models are hybrid models since connectivity is based on physics. Nevertheless, we will classify them as data-driven approaches since they incorporate only basic structural connectivity, whereas we reserve the term "hybrid" for models that integrate multiple physics-inspired constraints beyond network topology.
A simple method is to build a different model for each station in the river network, where each model receives information from all the upstream stations through separate input channels (Xiang et al., 2020). HydroNets enforces weight sharing between subbasins while maintaining some basin-specific model components, striking a balance between generalization and local adaptation (Moshe et al., 2020). This strategy revealed that incorporating river structure as prior knowledge reduces sample complexity and enables more accurate hydrological modeling even with limited training

data. A limitation of both of these approaches is that they cannot be readily applied to a different river network due to the station-specific components.

An alternative approach segments input features by Strahler river order, creating order-specific representations of upstream climate data that capture spatial variability while maintaining weight sharing through a unified model architecture (Muhebwa et al., 2024). Results show that as spatial resolution increases, model performance improves due to more granular hydrological information, supporting the value of semi-distributed models and motivating the use of fully distributed ones.

Graph neural networks (GNNs) offer a more flexible solution to this transferability challenge (Battaglia et al., 2018). GNNs represent river networks as graphs with nodes (representing subbasins or gauging stations) and edges (representing connectivity). This graph structure enables the same model architecture to be applied to different river networks without modification, as the connectivity information is encoded in the graph rather than hard-coded in the model architecture. To handle time series data, these GNNs are integrated with temporal modeling architectures such as GRUs (Sit et al., 2021), LSTMs (S. Chen et al., 2022), and temporal convolution networks (TCNs) (Sun et al., 2022; Wan et al., 2024).

Despite calls for more fully distributed data-driven models in rainfall-runoff modeling (Nearing et al., 2020), most existing methods face significant limitations that compromise their ability to represent hydrological processes realistically. A common class of distributed models employs CNN-LSTM architectures that process sequences of gridded input data through separated spatial and temporal processing stages (Anderson et al., 2022; P. Li et al., 2022; X. Li et al., 2022; Ueda et al., 2024; Pokharel et al., 2024). Here, the outputs of the CNN are flattened before being passed to an LSTM component. This architectural choice destroys spatial dependencies across time steps, preventing the model from maintaining spatial coherence in its temporal evolution.

Some researchers have attempted to address this limitation by using ConvLSTMs (Shi et al., 2015), which can jointly model spatial and temporal relationships by employing CNNs in the LSTM's gates. Recent work has applied ConvLSTMs to maintain spatiotemporal dependencies throughout the network (Oddo et al., 2024). However, even these improved methods often resort to global aggregation strategies before making final discharge predictions. The outputs of all grid cells are flattened into a single feature vector and passed through fully connected layers, eliminating the spatial structure at the crucial prediction stage. Similar global aggregation strategies appear throughout the literature (Xu et al., 2022; Zhu et al., 2023; Tyson et al., 2023; Pokharel et al., 2024; Börgel et al., 2025).

A notable exception that moves closer to physical plausibility is recent work that combined ConvLSTMs with ridge regression to learn which grid cells should contribute to discharge estimation at each gauging station (Longyang et al., 2024). This approach enabled the reconstruction of plausible underground flow paths between subbasins, demonstrating the potential for distributed models to yield both accurate predictions and physically interpretable insights when architectural design principles align with

hydrological processes.

2.3.5 Limitations and challenges

Despite achieving strong predictive performance in many hydrological applications, ANNs face important limitations when applied to physical systems. Due to their purely data-driven nature, ANN parameters and internal states typically do not correspond directly to measurable physical quantities or interpretable processes (Goodfellow et al., 2016). This black-box nature means that ANNs may learn spurious correlations in the training data rather than genuine physical relationships, potentially leading to poor performance when applied outside their training domain.

Moreover, ANNs do not inherently respect physical laws such as mass conservation, energy balance, or thermodynamic constraints unless explicitly designed to do so. This can result in physically implausible predictions, particularly in extrapolation scenarios or extreme conditions not well-represented in the training data. Unlike mechanistic models that can incorporate conservation equations directly, purely data-driven approaches may learn to predict discharge values that are physically impossible such as generating more water than was input through precipitation, or predicting negative ET rates. This limitation has motivated significant research into physics-informed ANNs and hybrid modeling approaches that combine the flexibility of ANNs with the physical consistency of mechanistic models.

ANNs excel at identifying statistical correlations but struggle to distinguish genuine causal relationships from spurious associations. In environmental modeling, this limitation is concerning because models may learn to rely on indirect indicators rather than underlying physical processes. For example, a model might learn that certain atmospheric patterns coincide with flooding without understanding the physical mechanisms linking precipitation, soil moisture, topography, and runoff generation. This correlation-based learning can lead to poor generalization when the statistical relationships change due to climate variability or anthropogenic influences.

The fundamental problem with most existing distributed approaches is that they aggregate spatial outputs globally, whether through simple summation or weighted combinations. This design choice eliminates any incentive for the model to propagate water across the landscape in a physically plausible manner. Instead of learning realistic flow patterns that respect topography and conservation principles, these models can achieve good predictive performance through statistical relationships that may not generalize beyond their training conditions. This limitation highlights a critical gap in current fully distributed data-driven approaches: While they possess the architectural capacity to represent complex spatial relationships, their design typically prevents them from learning the fundamental physical constraints that govern water movement across landscapes.

A related challenge stems from the uncritical transfer of ANN architectures from other domains to hydrological problems without sufficient consideration of the underly-

ing physical constraints. For instance, some studies have attempted to apply ConvLSTMs to hydrological forecasting using precipitation data from sparse point stations (Moishin et al., 2021), where the spatial convolution operations lack meaningful spatial structure to operate on. Such approaches demonstrate that successful application of data-driven methods in hydrology requires careful consideration of the physical nature of the problem and appropriate architectural design, rather than blind adoption of techniques that work well in other fields.
These limitations collectively demonstrate that while data-driven approaches offer powerful pattern recognition capabilities, their lack of physical grounding creates significant obstacles for reliable hydrological modeling. Addressing these challenges through hybrid approaches that combine data-driven flexibility with physics-based constraints represents a promising path forward for advancing the field.

2.4 Hybrid and differentiable modeling

Hybrid modeling approaches represent a promising strategy for addressing the fundamental limitations of both mechanistic and data-driven methods by combining their complementary strengths. Rather than viewing these paradigms as mutually exclusive alternatives, hybrid models integrate physics-based understanding with data-driven learning capabilities to achieve multiple desirable properties simultaneously: high predictive performance through automatic pattern discovery, physical interpretability through explicit process representation, robust generalization through physics-based constraints, and data efficiency through informed model structure.
This integration becomes particularly powerful when all model components are differentiable, enabling what is termed "differentiable modeling" (Shen et al., 2023). Differentiable modeling refers to a framework where gradient-based optimization can flow seamlessly through both data-driven and physics-based model components, allowing the entire hybrid system to be trained end-to-end using backpropagation and automatic differentiation.
Differentiable modeling can be approached from two complementary perspectives, each offering distinct advantages for addressing specific modeling challenges. From the physics-informed perspective, traditional mechanistic models serve as the foundation, with differentiable ANN components integrated to address specific knowledge gaps or computational bottlenecks. This approach allows researchers to maintain the interpretable structure and physical consistency of mechanistic models while leveraging ANNs to handle poorly understood processes, estimate difficult-to-measure parameters, or replace computationally expensive numerical solvers. For example, a traditional rainfall-runoff model might incorporate ANNs to estimate spatially variable soil parameters from limited observations, or to represent complex ET processes that are difficult to model explicitly. From the data-driven perspective, ANN architectures serve as the primary modeling framework, with physics-based knowledge incorporated as architectural constraints, loss function modifications, or inductive bi-

ases (Gauch et al., 2020; Roberts, 2021; Butz et al., 2025). Inductive biases encode prior assumptions about the underlying physical processes, effectively constraining the model's solution space to favor solutions that align with known physical principles. These biases can take various forms: architectural choices that reflect known system structure (such as conservation laws or spatial connectivity), regularization terms that penalize physically implausible solutions, or training procedures that incorporate physical constraints. By restricting the space of possible solutions to those consistent with physical understanding, these biases often improve generalization, enhance interpretability, and increase data efficiency.

2.4.1 Examples

Bias correction

Several approaches combine mechanistic models with machine learning by feeding mechanistic model outputs into data-driven models, which is sometimes called bias correction or residual prediction. Since the ANN processes the output of the mechanistic model rather than being integrated within it, the mechanistic components do not require differentiability for optimization. This is especially useful for older mechanistic models as there is no need to reimplement them in an automatic differentiation framework. One study demonstrated this approach by feeding outputs from the lumped GR4J conceptual model into various RNNs, including LSTMs, and additionally performed uncertainty quantification with the RNN component (Tian et al., 2018). The results showed that the integrated models achieved superior performance compared to either pure conceptual or pure data-driven models used individually. Similar improvements have been reported in other lumped (Konapala et al., 2020; Sezen et al., 2023) and distributed (Nimai et al., 2023) modeling cases, where this hybrid approach consistently outperforms both pure mechanistic and pure data-driven models.

Parameterization networks

The pipeline can also be reversed, where an ANN produces the parameters for a mechanistic model, an approach sometimes called a parameterization network. In this case, the mechanistic model needs to be differentiable and implemented in an automatic differentiation framework to allow error signals to flow through the mechanistic model and optimize the data-driven component. For example, recent work has developed a differentiable Muskingum-Cunge routing model that operates on given streamflow predictions in a semi-distributed setting (Bindas et al., 2024). The parameters of the routing model are predicted by a simple ANN that takes static attributes which describe the individual river reaches. Similarly, other studies have presented parameterization networks for semi-distributed modeling approaches that produce both dynamic and static parameters for conceptual runoff models and a differentiable Muskingum-Cunge routing models (Zhong et al., 2024a; Zhong et al., 2024b). However, this ap-

proach has shown mixed results, with some research suggesting that conceptual models do not serve as effective regularization for parameterizing ANNs, and that the ANN component primarily compensates for deficiencies in the hybrid model rather than providing synergistic benefits (Acuña Espinoza et al., 2024). While some degree of interpretability is retained compared to pure data-driven approaches, the addition of conceptual model components does not always lead to performance improvements.

Other approaches

Other hybrid modeling strategies include concurrent lumped approaches where ANNs receive both meteorological data and outputs from mechanistic models simultaneously, rather than in sequence (Lu et al., 2021). Such physics-informed hybrid models can present improved performance in terms of learning speed, prediction accuracy, and generalization capability, though their performance often depends heavily on the accuracy of the underlying physics-based component. Another lumped approach involves neural ordinary differential equations (Neural ODEs), where the differential equations of PBMs are replaced by ANNs while maintaining the overall model structure (Höge et al., 2022). This allows for more flexible representation of hydrological processes while preserving some physical interpretability. Some semi-distributed hybrid approaches combine differentiable ANN components for basin-scale modeling with non-differentiable routing models, requiring separate optimization strategies for the model components (Yu et al., 2024).

Physics-constrained distributed models

Some hybrid approaches focus on incorporating physical constraints into distributed ANN architectures. For example, certain models operate on grids but restrict cell-to-cell communication to the direction of steepest descent (Xiang et al., 2022; C. Wang et al., 2024; Huynh et al., 2024). This strong assumption effectively transforms the grid into a directed graph, excluding physically plausible underground flows in other directions. While this constraint incorporates topographic knowledge and may improve computational efficiency, it limits the model's ability to capture the full complexity of distributed water movement across the landscape.

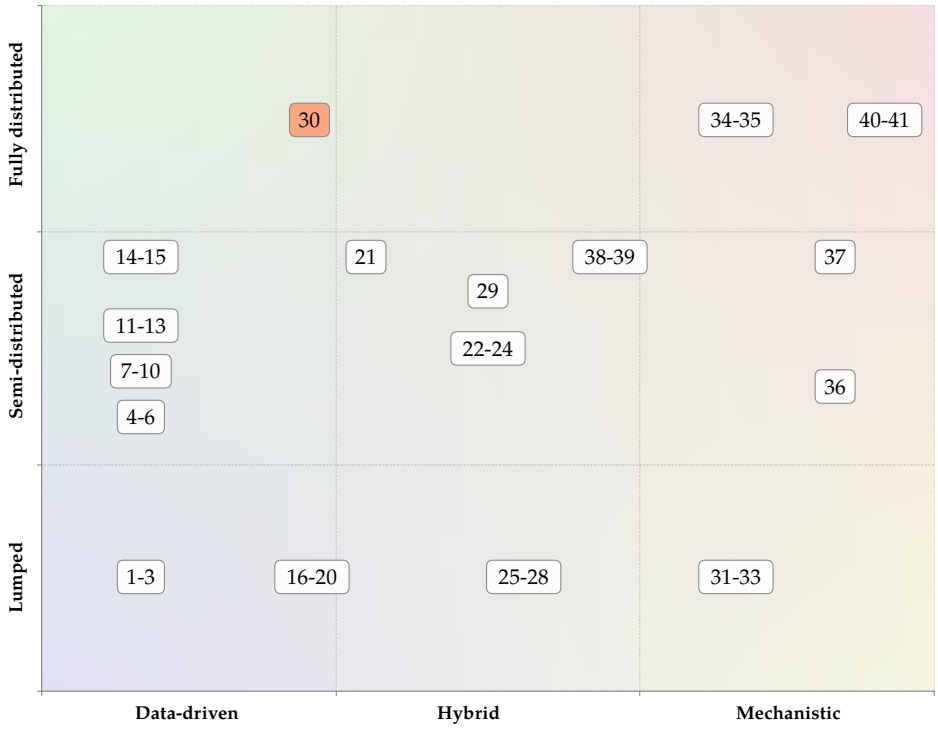
2.4.2 Limitations and challenges

The central challenge in hybrid model design lies in finding the optimal balance between physical constraints and data-driven flexibility. Modelers must incorporate sufficient physics-based structure to ensure physically meaningful behavior and robust generalization, while preserving enough degrees of freedom for the ANN components to discover novel relationships and adapt to local conditions. Overly restrictive physics-based constraints can prevent the model from learning important patterns present in the data, while insufficient constraints may allow the model to learn spurious relationships that lack physical meaning. Keeping this balance is challenging because it

requires domain expertise in both the underlying physical processes and modern machine learning techniques. Success depends on identifying which physical principles are most fundamental to preserve, which processes are sufficiently well understood to constrain explicitly, and which aspects of the system would benefit most from data-driven learning.

2.5 Hydrological model landscape

Having examined individual modeling paradigms, we now synthesize these approaches to understand their relationships and identify gaps in the current modeling landscape. Figure 2.2 positions existing rainfall-runoff modeling approaches along two fundamental dimensions: The horizontal axis represents the process representation spectrum, from pure data-driven methods (left) that learn patterns from observations, to mechanistic models (right) based on physical knowledge. The vertical axis reflects spatial discretization, ranging from lumped models that treat catchments as single units (bottom) to fully distributed models operating on regular grids (top).
Modern LSTM approaches (Kratzert et al., 2018) are positioned as data-driven and lumped, while conceptual models like HBV (Bergström, 1976) are mechanistic and lumped. Hybrid approaches (Höge et al., 2022; Bindas et al., 2024) occupy the center region by combining physical knowledge with ANN components.
Many ostensibly "distributed" approaches still impose significant spatial constraints. Semi-distributed models (Moshe et al., 2020; S. Chen et al., 2022) aggregate inputs within subbasins, while some grid-based approaches either restrict water movement to steepest descent directions (Xiang et al., 2022; C. Wang et al., 2024) or perform global spatial aggregation before discharge prediction (Ueda et al., 2024; Longyang et al., 2024). True fully distributed modeling without spatial aggregation or directional restrictions is primarily found in mechanistic models like EFAS (Thielen et al., 2009) and VIC (Liang et al., 1994).
DRRAiNN's positioning highlights a relative scarcity of approaches that combine ANN flexibility with fine-scale spatial representation. While maintaining data-driven learning capabilities, DRRAiNN incorporates physics-informed inductive biases throughout its fully distributed architecture, representing a novel contribution to this underexplored region of the modeling landscape.



Model References

1: LSTM Kratzert et al., 2018
2: LSTM+UE Klotz et al., 2022
3: LSTM+UE Nearing et al., 2023
4: GNN Moshe et al., 2020
5: GNN Per order Muhebwa et al., 2024
6: GNN Per station Xiang et al., 2020
7: GNN-GRU Sit et al., 2021
8: GNN-LSTM S. Chen et al., 2022
9: GNN-TCN Sun et al., 2022
10: GNN-TCN Wan et al., 2024
11: CNN-LSTM Ueda et al., 2024
12: CNN-LSTM Pokharel et al., 2024
13: CNN-LSTM P. Li et al., 2022
14: ConvLSTM → Reg Longyang et al., 2024
15: ConvLSTM → MLP S. Chen et al., 2022
16: LSTM+MC Kratzert et al., 2019b
17: LSTM+MC Hoedt et al., 2021
18: LSTM+MC Y.-H. Wang et al., 2024
19: LSTM+TS Gauch et al., 2021
20: LSTM+TS Acuña Espinoza et al., 2025
21: GNN → LSTM w/ descent Xiang et al., 2022
22: δ MC Bindas et al., 2024
23: Param + δ MC Zhong et al., 2024a
24: LSTM → Routing Yu et al., 2024
25: GR4J → RNN Tian et al., 2018
26: SAC-SMA → LSTM Konapala et al., 2020
27: PRMS → LSTM Lu et al., 2021
28: Neural ODEs Höge et al., 2022
29: BTOP → Lumped LSTM Nimai et al., 2023
30: **DRRAiNN** Scholz et al., 2025a
31: HBV Bergström, 1976
32: SAC-SMA Burnash, 1973
33: GR4J Perrin et al., 2003
34: VIC Liang et al., 1994
35: EFAS Thielen et al., 2009
36: TOPMODEL Beven et al., 1979
37: SWAT Arnold et al., 1998
38: Param w/ descent C. Wang et al., 2024
39: Param w/ descent Huynh et al., 2024
40: SHE Abbott et al., 1986a
41: GloFAS Alfieri et al., 2013

Figure 2.2: *Landscape of rainfall-runoff modeling approaches. Models are positioned by their degree of mechanistic vs. data-driven modeling (x-axis) and spatial representation from lumped to distributed (y-axis). DRRAiNN occupies a unique position as a fully distributed, data-driven approach with physics-informed inductive biases. Reg: Regression. Param: Parameterization. BTOP: Block-Wise Use of TOPMODEL, an extension of TOPMODEL with Muskingum-Cunge routing (Takeuchi et al., 2008). MC: Muskingum-Cunge routing. UE: Uncertainty estimation. MC: Mass conservation. TS: Time scales. PRMS: Precipitation-Runoff Modeling System, a mechanistic, semi-distributed model (Markstrom et al., 2015).*

Model Category	Examples	Spatial Representation	Process Representation	Interpretability	Calibration Approach	Computational Cost	Key Advantages	Key Limitations
Data-driven Lumped	LSTM (Kratzert et al., 2018) LSTM+UE (Klotz et al., 2022) LSTM+UE (Nearing et al., 2023) LSTM+MC (Kratzert et al., 2019b) LSTM+MC (Hoedt et al., 2021) LSTM+MC (Y.-H. Wang et al., 2024) LSTM+TS (Gauch et al., 2021) LSTM+TS (Acuña Espinoza et al., 2025)	Basin-aggregated	Learned patterns from data	Low	Automatic learning	Training: low Inference: low	Superior predictive performance	No spatial detail, no inter-basin processes, requires large datasets
Data-driven Semi-distributed	CNN-LSTM (Ueda et al., 2024) CNN-LSTM (Pokharel et al., 2024) CNN-LSTM (P. Li et al., 2022) GNN-GRU (Sit et al., 2021) GNN-LSTM (S. Chen et al., 2022) GNN-TCN (Sun et al., 2022) GNN-TCN (Wan et al., 2024) Hydronets (Moshe et al., 2020) Per order (Muhebwa et al., 2024) Per station (Xiang et al., 2020) ConvLSTM → Reg (Longyang et al., 2024), ConvLSTM → MLP (S. Chen et al., 2022)	Subbasin-aggregated or gridded inputs with spatial aggregation later in the pipeline	Learned with basic spatial constraints	Low-Medium	Automatic learning	Training: low-high Inference: low	Network topology integration, weight sharing across basins	Network-specific architectures, limited spatial flexibility, may require precomputed catchment boundaries
Hybrid Lumped	GR4J → RNN (Tian et al., 2018) SAC-SMA → LSTM (Konapala et al., 2020) PRMS → LSTM (Lu et al., 2021) Neural ODEs (Höge et al., 2022)	Basin-aggregated	Physics with learned components	Medium	Hybrid	Training: low-medium Inference: low-medium	Combines interpretability with performance, leverage physical understanding	Sequential processing limits synergy, performance gains not guaranteed
Hybrid Semi-distributed	δ MC (Bindas et al., 2024) Param + δ MC (Zhong et al., 2024a) LSTM → Routing (Yu et al., 2024) BTOP → Lumped LSTM (Nimai et al., 2023) GNN → LSTM /w descent (Xiang et al., 2022) Param w/ descent (C. Wang et al., 2024) Param w/ descent (Huynh et al., 2024)	Subbasin-aggregated	Learned parameterization and/or physics-based routing	Medium	Hybrid	Training: medium Inference: medium	Flexible routing, parameter learning, physical consistency	Requires precomputed catchment boundaries, model complexity
Hybrid Distributed	**DRRAiNN (this work)**	Gridded	Learned with physics-informed inductive biases	Medium	Automatic learning	Training: high Inference: low	No catchment boundary requirements	Limited spatial generalization, memory intensive
Mechanistic Lumped	HBV (Bergström, 1976) SAC-SMA (Burnash, 1973) GR4J (Perrin et al., 2003)	Basin-aggregated	Simplified physical relationships	High	Manual tuning	Training: low Inference: low	Interpretable parameters, computationally efficient	Poor performance, require precomputed catchment boundaries
Mechanistic Semi-distributed	TOPMODEL (Beven et al., 1979) SWAT (Arnold et al., 1998)	Subbasin-based	Physical relationships with routing	High	Manual tuning	Training: medium Inference: medium	Spatial heterogeneity representation	Predetermined subbasin boundaries, calibration complexity
Mechanistic Distributed	VIC (Liang et al., 1994) EFAS (Thielen et al., 2009) SHE (Abbott et al., 1986a) GloFAS (Alfieri et al., 2013)	Gridded	Physical equations (PDEs/ODEs)	High	Manual tuning	Training: high Inference: high	Strong physical basis	Computationally intensive

Table 2.1: *Comparison of rainfall-runoff modeling approaches across key characteristics. The table systematically compares different modeling paradigms that influence their applicability and performance. DRRAiNN occupies a unique position as a data-driven distributed approach with physics-inspired inductive biases. Reg: Regression. Param: Parameterization. BTOP: Block-Wise Use of TOPMODEL, an extension of TOPMODEL with Muskingum-Cunge routing (Takeuchi et al., 2008). (δ) MC: (differentiable) Muskingum-Cunge routing. UE: Uncertainty estimation. MC: Mass conservation. TS: Time scales. PRMS: Precipitation-Runoff Modeling System, a mechanistic, semi-distributed model (Markstrom et al., 2015).*

Table 2.1 provides a systematic comparison of these modeling approaches across key characteristics that influence their applicability and performance in different contexts.

2.6 Spatial generalization

One of the most challenging aspects of hydrological modeling concerns spatial generalization, i.e., the ability to apply models trained or calibrated in data-rich regions to areas where observations are scarce or entirely absent. Given that much of the world remains inadequately equipped with sensors for river discharge monitoring, this capability is crucial for global water management and scientific understanding. In the hydrological literature, this challenge is formally known as prediction in ungauged basins (PUB) (Sivapalan, 2003). The scope of this problem is vast: Earth's land surface remains largely ungauged or poorly gauged, especially in many developing countries where human impacts on hydrological systems are often greatest (Sivapalan, 2003).
However, the fundamental difficulty of PUB extends beyond data scarcity. Current understanding of basin responses remains inadequate for confident extrapolation from gauged to ungauged basins (Sivapalan, 2003), particularly given unprecedented human-induced land use and climatic changes. Traditional regionalization methods developed over several decades have met with limited success, with no single approach emerging as optimal despite extensive research efforts (He et al., 2011).
More recent efforts have attempted to address these challenges through physics-informed approaches. The Multiscale Parameter Regionalization framework (Samaniego et al., 2010) and scalable transfer function approaches (R. Imhoff et al., 2020) focus on improving parameterization and capturing spatial heterogeneity by deriving spatially distributed parameters from observable landscape characteristics through transfer functions. Such approaches aim to reduce basin-specific calibration needs while maintaining physical consistency. However, these physics-based regionalization methods remain constrained by the accuracy of underlying transfer functions and the completeness of encoded physical relationships.
This persistent challenge led to the realization that existing hydrological theories and models were largely inadequate for predictions in ungauged basins (Hrachowitz et al., 2013). The community recognized that meaningful progress required shifting focus from parameter fitting toward process understanding and model structural diagnostics, a paradigm shift emphasizing theoretical insights over empirical calibration approaches (Hrachowitz et al., 2013). Recent developments in machine learning have opened new avenues for addressing PUB challenges. Data-driven approaches can automatically extract patterns from large multi-catchment datasets, potentially identifying hydrological similarities that traditional regionalization methods miss. Early re-

sults have been encouraging: Out-of-sample LSTMs can outperform both conceptual models calibrated independently for each catchment and distributed PBMs when applied to ungauged basins (Kratzert et al., 2019a). This success has led some researchers to conclude that ML is especially effective at forecasting in ungauged basins (Nearing et al., 2023), suggesting that the pattern recognition capabilities of ANNs may be well-suited to identifying transferable hydrological relationships across different catchments.

Spatial information and hybrid modeling show particular promise for improving ungauged basin predictions. Training LSTMs with disaggregated, distributed features significantly improves model accuracy on held-out stations compared to lumped approaches (Muhebwa et al., 2024), while physics-informed hybrid LSTM models demonstrate superior out-of-distribution prediction capabilities (Lu et al., 2021). Advanced physics-informed ANN frameworks using spatial discretization with differentiable hydrological models have achieved remarkable results: When trained solely on downstream stations, they outperform traditional distributed models at both training and upstream ungauged locations (Zhong et al., 2024a).

These advances have important implications for distributed hydrological modeling. The success of data-driven approaches in ungauged basins shows that ANNs can learn transferable hydrological relationships that transcend specific catchment characteristics. This capability is particularly valuable for distributed models, which must represent spatial heterogeneity and flow connectivity across landscapes where detailed local calibration data are often unavailable. The integration of physics-based constraints with data-driven learning is especially promising, combining pattern recognition capabilities with fundamental physical principles that should govern hydrological behavior regardless of location.

2.7 Model interpretation and explainability

It is often criticized that developers of data-driven models do not put enough effort into the interpretation of their developed systems, thereby failing to gain a better understanding of the system's internal dynamics (Muñoz-Carpena et al., 2023). However, recent research has begun to demonstrate that ANNs can indeed provide meaningful insights into hydrological processes. For instance, studies have shown that LSTMs internally learn to represent patterns consistent with our qualitative understanding of hydrological systems (Kratzert et al., 2019b) and can develop representations of intermittent states like soil moisture despite never being explicitly trained on such variables (Lees et al., 2021). These findings suggest that data-driven approaches can not only achieve superior predictive performance but also contribute to scientific understanding by learning hydrological processes directly from observational data.

One promising avenue involves leveraging data-driven approaches to infer latent variables that are otherwise inaccessible to direct measurement. A considerable portion of total discharge originates from subsurface flow, yet it is not possible to directly

measure subsurface flow. This renders underground topography a latent driver of hydrological behavior (Shen, 2018). Recent continental-scale studies using physics-based distributed models have revealed the extent of these underground connections (Yang et al., 2025). They show that groundwater can travel hundreds of kilometers before emerging as streamflow and that deep groundwater from consolidated sediments contributes more than half of the baseflow in 56% of subbasins across the continental United States. These findings underscore both the complexity of subsurface flow patterns and the limitations of traditional water-balance approaches that may underestimate inter-basin groundwater flow due to concurrent import and export processes. These latent variables likely contribute to poor model generalization across basins. Data-driven approaches, especially ANNs, can support hydrological modeling in such cases because they allow latent variables to be inferred retrospectively from observation dynamics (Butz et al., 2019; Otte et al., 2020). This capability motivates a key question we address in this work: Given the observed dynamics, in which areas did precipitation contribute to the measured discharge? Similar to subsurface flow, ET cannot be directly measured and must be inferred indirectly. This renders interpretability analyses of ANNs valuable tools for extending our understanding of the water cycle (Sit et al., 2020).

These latent hydrological processes, particularly subsurface flow patterns and effective catchment boundaries, can be investigated through interpretability analyses of trained ANNs. Answering such spatial attribution questions requires interpretability methods that can reveal which spatial inputs most influence model predictions. Among existing interpretability approaches, gradient-based attribution methods are particularly well-suited for these spatial questions, as they can efficiently compute how model outputs respond to changes across spatially distributed inputs. These techniques include saliency maps, which compute the gradient of model outputs with respect to input features, and integrated gradients, which provide more stable attributions (Sundararajan et al., 2017). For distributed hydrological models processing gridded meteorological data, these methods can identify which geographic regions and variables most strongly influence streamflow predictions at specific locations. For example, Wunsch et al. (2022) demonstrated how spatial input sensitivity analysis of CNNs can identify approximate catchment locations for karst springs by revealing which grid cells the model focuses on for discharge predictions, with precipitation patterns showing the strongest spatial coherence with known catchment boundaries. This capability directly addresses questions about precipitation-discharge attribution and enables validation of whether models focus on hydrologically meaningful spatial patterns.

While gradient-based methods are especially useful for spatial analysis, they represent just one category within the broader landscape of interpretability approaches. Researchers have developed various interpretability approaches that can be categorized into model-agnostic and gradient-based methods. Model-agnostic approaches work by perturbing inputs or analyzing model outputs without requiring access to model gradients, including SHAP (Shapley Additive exPlanations), LIME (Local Interpretable

Model-agnostic Explanations), and permutation feature importance methods (Molnar, 2020). These are useful for legacy models not implemented in automatic differentiation frameworks, though they are generally less efficient for analyzing spatial patterns in distributed models.
These ML-based interpretability approaches complement traditional hydrological inference methods that tackle similar questions about latent processes. For example, studies analyze discharge measurements to infer flow direction, aquifer connectivity, and storage characteristics without ML (Thurber et al., 2024), addressing some of the same latent variable questions that motivate ANN attribution analyses. The integration of interpretability methods with such established domain knowledge represents a promising avenue for advancing both model understanding and hydrological science.

2.8 Summary

This review reveals a critical gap in the hydrological modeling landscape. While mechanistic models provide physical interpretability but struggle with calibration and generalization, data-driven approaches achieve superior predictive performance but lack physical grounding. Although fully distributed models exist, they fall into two categories: Mechanistic models that maintain spatial coherence but face the traditional limitations of physical models, and data-driven models that suffer from a fundamental architectural flaws like resorting to global spatial aggregation before discharge prediction, eliminating spatial dependencies, and removing incentives to learn physically plausible water movement patterns. This limitation prevents existing data-driven approaches from addressing spatially explicit questions about effective catchment boundaries, subsurface flow paths, and precipitation attribution. These questions are central to advancing hydrological understanding.
DRRAiNN addresses these limitations through a novel combination of full differentiability, full spatial distribution, and physics-informed architectural design. By operating on grids while preserving spatial dependencies throughout the network, DRRAiNN can learn to implicitly infer effective catchment areas from observed dynamics. Its modular architecture separates rainfall-runoff processes from discharge routing, incorporating physical understanding while maintaining flexibility to discover complex patterns. Physics-informed inductive biases guide the model toward physically meaningful solutions without overly constraining learning capacity.
DRRAiNN's full differentiability enables gradient-based attribution methods to answer spatially explicit questions: Where do the waters contributing to discharge actually originate? How do effective catchment boundaries differ from topographically derived watersheds when subsurface flows are considered? By combining neural network flexibility with physics-inspired constraints while maintaining full spatial resolution, DRRAiNN represents a promising step toward resolving tensions between accuracy, interpretability, and physical plausibility in hydrological modeling.

3

Method

The best thing about being a statistician is that you get to play in everyone's backyard —
John Tukey

The literature review revealed critical gaps in existing approaches: Fully distributed data-driven models either resort to global spatial aggregation or fail to maintain spatial coherence throughout computation. To address these limitations, we present DRRAiNN, a spatio-temporal ANN architecture that maintains spatial coherence throughout its computation while incorporating physics-informed constraints to estimate river discharge from static attributes and meteorological forcings in a distributed manner. We evaluate DRRAiNN's estimation abilities, physical plausibility, and some of its architectural design choices. We demonstrate its performance in a real-world setting on the Neckar River in Southwest Germany, comparing it to simulations from the European Flood Awareness System (EFAS, Mazzetti et al., 2023). DRRAiNN achieves better performance than EFAS for lead times of up to 50 days and provides interpretable source attributions that enable the reconstruction of effective catchment areas from modeled dynamics.

3.1 Model

DRRAiNN's structure is grounded in the following data and structural information sources. The locations $L_i = (x_i, y_i)$ for estimations of discharge in the river network are determined by discharge gauging stations that provide observed discharge $Q_{i,t}$ for time t in 24 h 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 extends beyond the DEM-derived catchment boundaries to ensure complete coverage of the hydrologically contributing area. 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 estimates discharge $Q_{i,t_s+1:t_s+T}$ over a temporal future horizon of T days via a function f, representing the learned spatio-temporal mapping implemented by the

model:

$$\tilde{Q}_{i,t_s+1:t_s+T} = f(S_{:,:}, F_{:,:,t_0:t_s+T}, Q_{i,t_0:t_s}) \tag{3.1}$$

3.1.1 Architectural Overview

The mathematical framework above is realized through a physics-informed architecture that distinguishes between fundamentally different hydrological processes. Surface and subsurface flow exhibit distributed lateral movement across landscapes, driven primarily by topographic gradients. In contrast, river discharge involves concentrated flow through established channels, governed by different temporal dynamics and connectivity patterns.
This physical understanding motivates DRRAiNN's modular design, which separates these processes into two coupled components: a rainfall-runoff model and a discharge model. The rainfall-runoff model operates on a regular grid to capture distributed catchment processes, while the discharge model uses a graph-based representation to simulate flow through the river network.
Several key inductive biases guide this architecture:

- **Process separation**: Distinct modeling of spatially extended (rainfall-runoff) versus concentrated (river channel) flow processes
- **Spatial coherence**: Maintenance of spatial relationships throughout computation without premature aggregation
- **Physical connectivity**: Incorporation of river network topology through graph structure
- **Dynamic parameterization**: Use of hypernetworks to adapt model behavior based on local environmental conditions

Figure 3.1 provides a schematic overview of this architecture, showing how these components interact within the overall modeling framework. At each time step, DRRAiNN processes the sequence in an autoregressive loop by first invoking the rainfall-runoff model, followed by the discharge model. The rainfall-runoff model receives gridded static maps S and meteorological forcings F as input to model the catchment on a grid. It is primed to distinguish between two important subprocesses, namely surface and subsurface flow, which is mainly driven by topography, and ET, which is mainly driven by temperature. It produces a latent representation, which we term *runoff embedding*, extracted at station locations and used as input to the discharge model. Despite being the main driver of discharge, it cannot be directly interpreted as runoff due to its self-organizing nature. The discharge model additionally receives an adjacency matrix A that describes the connectivity between stations, static river segment features, and the (potentially estimated) discharge $Q_{:,t-1}$ from the previous time step. It then estimates discharge $\tilde{Q}$ for each station, from which the training loss is computed.
In the following, we provide a more detailed description of DRRAiNN's components.

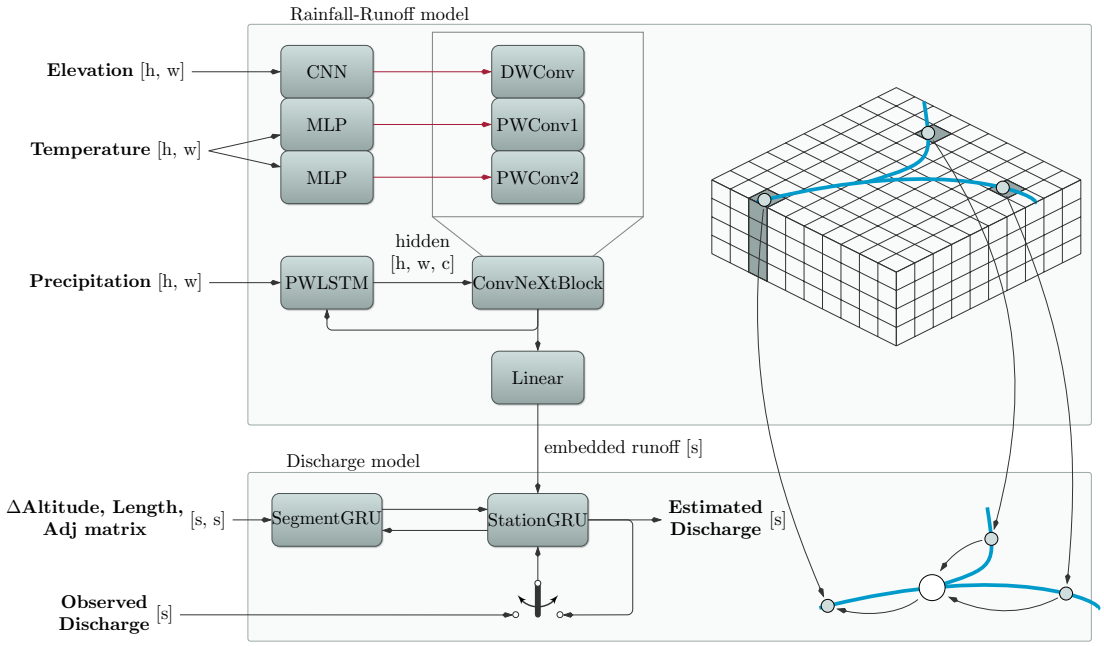


Figure 3.1: *Schematic overview of the DRRAiNN architecture. The gridded rainfall-runoff model has two main tasks: modeling precipitation redistribution across the landscape and modeling ET based on temperature. It receives precipitation as input to a position-wise long short-term memory (PWLSTM), whose hidden states (but not cell states) are updated using a ConvNeXtBlock. The ConvNeXtBlock weights are dynamically generated by hypernetworks (indicated by red arrows) rather than being fixed. The depth-wise convolution (DWConv) handles lateral water propagation and receives its weights from a CNN that takes elevation as input and has the same spatial extent as the DWConv kernel. The position-wise convolutions (PWConv1 and PWConv2) model local ET processes and receive their weights from an MLP that takes temperature as input. The LSTM hidden state is processed by a linear layer before being passed to the discharge model. This graph-based discharge model aggregates information at gauging stations, incorporating the previous (possibly inferred) discharge values, elevation differences between stations, and river segment lengths. The output is discharge at each station.*

3.1.2 Rainfall-Runoff model

The rainfall-runoff model consists of a position-wise long short-term memory (PWLSTM) and a CNN that are called in each time step. This renders the rainfall-runoff model local in space and time. Only spatially local and temporally previous information is used to update internal states.

The 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 regulates when and how the cell state is updated, allowing the model to retain information over extended time periods. This can be useful for implicitly modeling slow hydrological processes such as soil moisture or groundwater levels, which evolve more gradually than overland flow. The PWLSTM receives precipitation as input to update its hidden and cell states. It has a hidden size of 4. The weights of the PWLSTM are shared throughout the gridded area. As a result, while the PWLSTM at each grid cell maintains individual hidden and cell state values, the temporal processing principle is identical everywhere. The assumption is that the unfolding laws of physics are the same everywhere, although they may be locally parameterized.

The CNN models spatial relationships such as the propagation of water flow across the landscape and ET. It receives and updates the hidden state h of the PWLSTM to model spatial interactions, while leaving the PWLSTM's cell states untouched to preserve temporal memory. Surface and subsurface flow are spatially extended processes, whereas ET is primarily a local phenomenon, occurring independently at each grid cell. To reflect this distinction, we separate the CNN's treatment of these processes using different convolution types and input sources, which introduces an inductive bias into the architecture.

More precisely, the CNN is based on a modified ConvNeXt block (Z. Liu et al., 2022). ConvNeXt is a modern CNN variant that achieves better efficiency than standard CNNs while maintaining strong performance. ConvNeXt blocks use depth-wise convolutions (DWConvs), which process spatial patterns within each feature channel separately, followed by position-wise convolutions (PWConvs), which mix information across channels. Specifically, a ConvNeXt block consists of three layers: a DWConv with kernel size 7×7 followed by a position-wise inverted bottleneck given by two linear layers (PWConv1 and PWConv2). This design decouples spatial and channel-wise information flow, making the architecture well-suited for learning representations of spatially structured data. To stabilize training and enable effective learning, ConvNeXt incorporates layer normalization and residual connections that add the original input to the processed output. We apply the sigmoid linear unit (SiLU) activation function (Hendrycks et al., 2016) after the convolutional and between the linear layers.

In contrast to its original formulation, the weights of our ConvNeXt block are not static but location-dependent. They are parameterized by other ANNs, turning this network component into a hypernetwork (Ha et al., 2016). This means that the ConvNeXt block can behave differently at each location on the grid, breaking the translational invariance

assumption in CNNs. Calling DWConv results in the following operation:

$$y_{i,j,c} = \sum_{m=-3}^{3} \sum_{n=-3}^{3} w_{i,j,m,n,c} \cdot x_{i+m,j+n,c}, \tag{3.2}$$

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. We can still call this operation a convolution if we regard the input variables together with the weight-generating networks as the kernel. Calling PWConv1 and PWConv2 results in the following operation:

$$y_{i,j,c_{\text{out}}} = \sum_{c_{\text{in}}} w_{i,j,c_{\text{out}},c_{\text{in}}} \cdot x_{i,j,c_{\text{in}}} \tag{3.3}$$

Each layer of the ConvNeXt block is parameterized by a distinct hypernetwork, tailored to the type of process it represents. The weights of DWConv are produced by a CNN that has the same spatial extent (receptive field) as DWConv itself. The weights for PWConv1 and PWConv2 are produced by position-wise MLPs. By using different input variables for the different hypernetworks, we can distinguish between local and spatially extended processes. How water propagates across 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 to focus on local elevation gradients and improve generalization across different elevation ranges. Since this modifies the statistics, we multiply the resulting gradients with 2 to have a standard deviation of 1 again. ET, on the other hand, is a local process and is therefore best captured by the position-wise components. This is why we generate the weights for PWConv1 and PWConv2 from temperature. See Figure 3.2 for an illustration. Using a hypernetwork to specifically parameterize a ConvNeXt block was employed before for successful identity tracking (Traub et al., 2024b).
Lastly, the runoff embeddings are extracted at the station locations, fed through a single linear layer, and sent to the river discharge model. Aggregating the hidden states of all cells on the corresponding upstream river segment showed a tendency to overfit in preliminary experiments.

3.1.3 Discharge model

Our discharge model is a type of recurrent graph ANN called DISTANA (distributed spatio-temporal artificial neural network architecture, Karlbauer et al., 2019), with the graph structure defined by the actual river network and the stations. DISTANA maintains two types of recurrent units: station and segment kernels, both implemented as GRUs with a hidden size of 8. Station kernels are placed at the gauging stations, while segment kernels are located on segments between stations. These kernels communicate with each other via lateral connections with 4 channels. In each time step, the

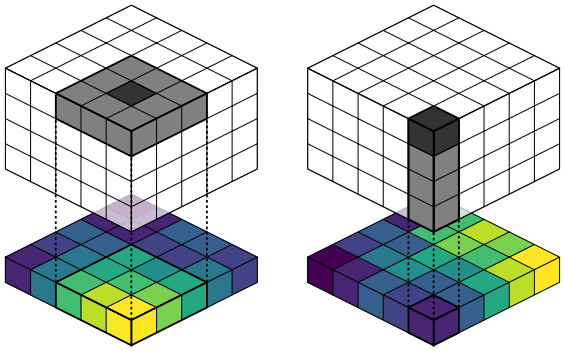


Figure 3.2: *Illustration of the hypernetworks used in DRRAiNN. In both panels, the dark gray cells represent locations whose hidden states are updated based on information from the light gray cells. The weights for these updates are generated by specialized ANNs that process different environmental variables. Left: A CNN takes elevation as input and produces weights for the depth-wise convolution (DWConv), which models lateral water propagation. The CNN has the same spatial extent as the DWConv kernel. Right: An MLP takes temperature as input and produces weights for the position-wise convolution (PWConv), which models localized ET.*

segment kernels are updated first, followed by the station kernels, which then estimate the discharge $\tilde{Q}$ at their respective locations. The segment kernels first concatenate the previous output of the upstream station kernels with static river segment attributes, specifically the altitude difference and segment length. After applying the GRU, the output is multiplied by the adjacency matrix, which is derived from the river network topology and station positions. The segment kernels thereby sum up information from upstream station kernels. The output of the segment kernels serves as input for the station kernels. The station kernels work similarly. They first concatenate the last output of the segment kernels with the last (potentially estimated) discharge and the output of the rainfall-runoff model. After applying the GRU, the output is split into the estimated discharge $\tilde{Q}$ and the input for the segment kernels in the next time step.

Although DRRAiNN receives hourly meteorological forcings F, it produces discharge estimates at a daily resolution. During the initial 10-day tune-in phase of each sequence, we feed the same observed discharge value Q into DRRAiNN for each hourly step within the day.

A key advantage of this modular architecture is its theoretical transferability: Because the runoff model operates on regular grids and DISTANA adapts to arbitrary river network topologies through the adjacency matrix, a trained DRRAiNN instance can theoretically be applied to any river network, regardless of domain size, number of gauging stations, or network connectivity. The model requires only the new river network topology, station coordinates, and gridded meteorological forcings for the target domain. However, the actual performance in new domains depends on the similarity of hydrological processes to those encountered during training.

3.2 Data

The input data for DRRAiNN consists of radar-based precipitation, elevation for above-ground topography, temperature, and river discharge data.
For precipitation, we use the radar-based precipitation product RADOLAN provided by the Deutsche Wetterdienst (*RADOLAN/RADVOR* 2016). 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 and a temporal resolution of $1\,\mathrm{h}$. RADOLAN data is log-standardized before being sent to the model due to its long-tail distribution. Specifically, we add 1 and take the logarithm, then compute the mean and standard deviation of the transformed data to standardize it. We replace missing values with 0s, which is the standardized mean.
For static topography information we use the digital elevation model (DEM) EU-DEM v1.1 provided by the Copernicus Land Monitoring Service of the European Environment Agency (*EU-DEM v1.1* 2016). We also use the DEM to compute the differences in altitudes between adjacent discharge gauging stations. Elevation values and derived differences are standardized before being sent to the model, i.e., we subtract their mean and divide by their standard deviation.
For temperature, we use the near-surface temperature (t2m) from the ERA5 data set (Hersbach et al., 2018). For solar radiation experiments, we use the surface short-wave downward radiation (SSRD) from the same ERA5 data set. Both variables come with a temporal resolution of $1\,\mathrm{h}$ and a relatively coarse spatial resolution of $0.25° \times 0.25°$. Temperature and solar radiation are standardized before being sent to the model, i.e., we subtract their respective means and divide by their standard deviations. We use rasterio (Gillies et al., 2013) to transform and reproject the DEM, temperature, and solar radiation data to match the RADOLAN coordinate reference system.
The topography of our river network is determined by the AWGN data set (*Amtliches Digitales Wasserwirtschaftliches Gewässernetz (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 from gauging stations to tune in the discharge model and as the only target variable to train, validate, and test our model. We use data collected and provided by the German Federal Institute of Hydrology via the Global Runoff Data Centre (*Global Runoff Data Centre* 2024). The data set contains observed daily river discharge from gauging stations worldwide, including those in Germany. Since the location information of the discharge gauging stations is partially wrong, we correct them manually. We then align the station locations to the nearest river segment (snapping). If the correction exceeds a predefined threshold, the station is excluded. When two stations are located too close to each other, one is removed to satisfy the Courant–Friedrichs–Lewy (CFL) convergence condition (Courant et al., 1967). This condition requires that temporal and spatial resolutions be adjusted together for the model to converge effectively. The finer the spatial resolution, the finer the temporal resolution must be: A wave or peak that travels from one simulated gaug-

ing station to the next in less than one timestep cannot be captured by the graph-based discharge model. In contrast, when it takes more than one timestep, this can be compensated to some extent through the memory component present in recurrent units. Due to its long-tail distribution, discharge data is log-standardized on a per-station basis before being sent to the model. We add 1 and take the logarithm, then standardize the data using station-wise means and standard deviations. We replace missing values with 0s, which is the standardized mean of the corresponding station.
Our choice of input datasets was guided by temporal resolution, data provenance, and practical availability. Although EFAS employs EMO-1 for precipitation input, we opted for RADOLAN due to important differences: EMO-1 offers a coarser $6\,\mathrm{h}$ resolution and is interpolated from sparse station data, in contrast to RADOLAN's direct radar-based observations. Although we expect only minor differences in performance in some settings, radar-derived datasets like RADOLAN provide finer spatial and temporal resolution, which is advantageous for distributed models. Similarly, we chose ERA5 for temperature data due to its gridded 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 grid format, or as gridded data aggregated monthly, which does not meet our temporal requirements. Daily datasets like EOBS may suffice if subdaily temporal patterns are encoded separately, but this would require additional preprocessing. A transition toward operation flood forecasting would place increased importance on the choice of precipitation forecast products (Ruben Olaf Imhoff et al., 2022). Ultimately, all data products entail inherent uncertainties and errors, and our choices reflect a balance between data availability, temporal resolution, and the specific requirements of our model.

3.3 Study site

The Neckar river network in Southwest Germany spans a catchment area of $14\,000\,\mathrm{km}^2$ with a mean elevation of $460\,\mathrm{m}$. According to ERA5, temperatures in this region ranged from $-25\,^\circ\mathrm{C}$ to $40\,^\circ\mathrm{C}$ during our training period, with a mean temperature of $0.95\,^\circ\mathrm{C}$ during winter and $17.95\,^\circ\mathrm{C}$ during summer. Our dataset includes measurements from 17 gauging stations distributed across the river network (see Figure 3.3). At the most downstream station in Rockenau, discharge during the training period ranged from $29.5\,\mathrm{m}^3\,\mathrm{s}^{-1}$ to $1690\,\mathrm{m}^3\,\mathrm{s}^{-1}$ with a mean of $133.3\,\mathrm{m}^3\,\mathrm{s}^{-1}$.
The catchment features a highly heterogeneous landscape, including narrow and wide valleys, diverse geology (e.g., limestone, sandstone), different soil textures (e.g., clay, marl), and subsurface structures such as karst systems and pore water aquifers (Ufrecht, 2002). 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 DEM alone. Latent underground topology routes the water in a different direction than elevation alone would

suggest.
By restricting the domain to the Neckar river network, we end up with an area of size $200\,\mathrm{km} \times 200\,\mathrm{km}$. Following the transformations described above, 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 grid covering the study area. We train our model on hydrological years 2006 to 2015, validate on 2016 to 2018, and test on 2019. Meteorological forcings F are provided at hourly resolution, while discharge is provided at daily resolution.

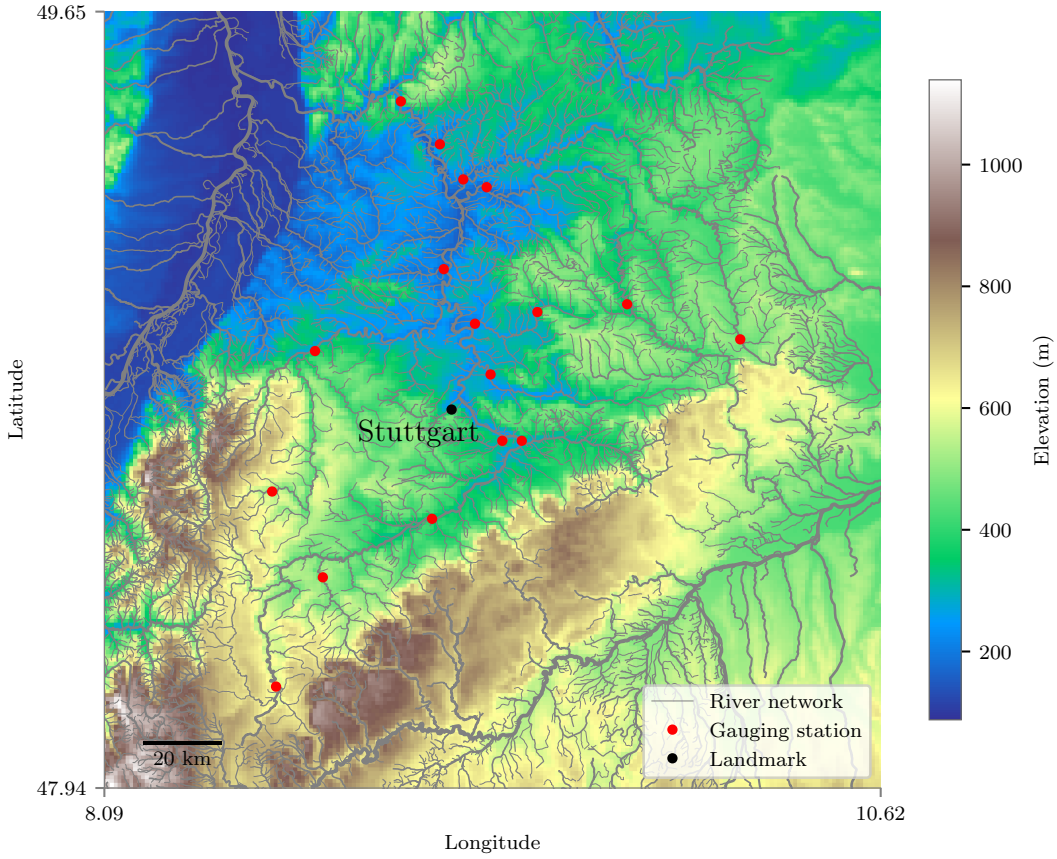


Figure 3.3: *The study area used in this work is the Neckar River catchment in Southwest Germany.*

3.4 Experimental setup

We train DRRAiNN on sequences of 20 days (480 hourly steps), using the first 10 days as a warm-up phase. During this phase, we feed the model observed discharge values to initialize and align its hidden states with the true system dynamics. This procedure resembles data assimilation in traditional hydrological models, where observa-

tions are used to update model states and reduce uncertainty. In ML terms, this corresponds to teacher forcing. The warm-up phase allows the rainfall-runoff component of DRRAiNN to infer latent hydrological states, such as soil moisture or groundwater storage, through its hidden state representations.

After the warm-up phase, DRRAiNN transitions into open-loop mode for the remaining 10 days of each sequence. In this predictive mode, the model does not have access to observed ground-truth discharge values. Instead, the discharge model feeds its own previous discharge estimates as inputs for subsequent time steps, which is why we call it an autoregressive model. The rainfall-runoff model, in contrast, continues to receive observed precipitation and temperature as inputs throughout the sequence. While informative, this setup does not reflect realistic operational conditions for discharge forecasting. Precipitation forecasting, in particular, remains a major challenge. Currently no algorithm can accurately predict precipitation 10 days ahead at a spatial resolution of $4\,\mathrm{km} \times 4\,\mathrm{km}$. However, this setup is well suited for knowledge discovery concerning hydrologic processes, which is the primary focus in this work. We leave the evaluation of DRRAiNN under realistic, forecast-based conditions for future work.

We use the mean squared error (MSE) computed on station-wise standardized discharge data as both the training and validation loss. Standardization ensures that stations with larger discharge values do not dominate the loss, promoting a balanced learning across all stations. Training is performed using truncated backpropagation through time (TBPTT), where the truncation length increases progressively over the course of training. Initially, we backpropagate the loss over 1-day sequences (24 time steps) to help DRRAiNN focus on short-term temporal relationships and stabilize learning. Over the course of training, we increase the truncation length, enabling the model to learn longer-term dependencies. The truncation length schedule is shown in Table 3.1. We adapt the batch size to fit the model within the memory constraints of a single NVIDIA GeForce GTX 1080 Ti.

Table 3.1: *Truncation length schedule in days for truncated backpropagation through time.*

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

To improve generalization and account for model variability due to random initialization, we train five independent instances of DRRAiNN per experiment, each initialized with a different seed. We use early stopping in the sense that we train each instance for 20 epochs but use the checkpoint with the lowest validation loss. This procedure is applied consistently to both the primary model and all ablation variants. We use the Ranger optimizer (Wright, 2019) with a learning rate of 0.0025 to optimize the 30 500 parameters in DRRAiNN. To stabilize training, we clip the gradient if its norm exceeds 1, thereby preventing large parameter updates in steep regions of the loss surface (Pas-

canu et al., 2012).
To increase the size of the training data set and improve generalization, we apply data augmentation. The symmetry group of the square contains eight elements: the identity, three rotations ($90°$, $180°$, and $270°$), and four reflections (horizontal, vertical, and two diagonals). For each training sequence, we apply a uniformly sampled symmetry to the spatial variables in each time step. We ensure physical consistency by tapping into the runoff embeddings at the transformed station locations. The river discharge model's graph structure remains unchanged by this augmentation.

3.5 Benchmark model: European Flood Awareness System

To provide context for DRRAiNN's performance, we compare it to the European Flood Awareness System (EFAS), an established and operational distributed process-based model (Thielen et al., 2009). We use publicly available EFAS reanalysis data, which eliminates the need to tune the benchmark model ourselves. This avoids potential biases that could arise from allocating unequal tuning effort to the benchmark model versus our own model. While DRRAiNN achieves higher performance than EFAS in many scenarios, our primary aim is to demonstrate the potential of distributed ANNs for river discharge estimation, rather than merely outperforming EFAS.
EFAS simulates runoff on an approximately $1.5\,\mathrm{km} \times 1.5\,\mathrm{km}$ grid with a temporal resolution of $6\,\mathrm{h}$. Among others, it receives as inputs static maps describing topography, river networks including lakes, reservoirs, and channel width, 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, which makes it an ex-post forecasting model (Beven et al., 2013). In contrast, EFAS does not use discharge measurements as input but relies on them for offline model calibration only, rendering it a simulation. Furthermore, DRRAiNN produces discharge estimates only at gauging station locations, whereas EFAS generates discharge predictions across the entire spatial grid. Therefore, it is necessary to extract outputs from EFAS grid cells that correspond to the station locations in order to make meaningful comparisons to DRRAiNN. EFAS also relies on additional input variables not used by DRRAiNN, such as soil type, vegetation, temperature, and potential ET. While this makes EFAS a powerful tool, it also limits its applicability in regions lacking such detailed input data. Another difference lies in the precipitation data used: EFAS relies on EMO-1, a $6\,\mathrm{h}$ product interpolated from weather station data, whereas DRRAiNN uses RADOLAN, a radar-based dataset offering higher spatial and temporal resolution. As a result, a direct comparison between EFAS and DRRAiNN is not valid. Nonetheless, EFAS serves as a baseline to contextualize the expected performance range of DRRAiNN. 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 visualizing hydrographs for selected gauging stations, we evaluate DRRAiNN using the following metrics commonly used in hydrology: Kling-Gupta efficiency (KGE, Gupta et al., 2009), Nash-Sutcliffe efficiency (NSE, Nash et al., 1970), Pearson's correlation coefficient (PCC), mean absolute error (MAE). Additionally, we will employ the following bias metrics (Yilmaz et al., 2008): percent bias in overall runoff ratio (%BiasRR), percent bias in flow duration curve mid-segment slope (%BiasFMS), percent bias in flow duration curve low-segment volume (%BiasFLV), and percent bias in flow duration curve high-segment volume (%BiasFHV). We report all of these metrics because each one highlights different aspects of model performance, and no single metric is free from limitations (Gupta et al., 2009).

MAE is particularly intuitive, as it is expressed in the same unit as discharge and directly quantifies the average deviation between predictions and observations:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Q_i - \tilde{Q}_i| \tag{3.4}$$

However, because it lacks normalization, stations with larger discharge magnitudes contribute disproportionately to the overall MAE. PCC quantifies the strength of linear association between the observed and estimated discharges:

$$\text{PCC} = \frac{\sum_{i=1}^{n}(\tilde{Q}_i - \mu_{\tilde{Q}})(Q_i - \mu_Q)}{\sqrt{\sum_{i=1}^{n}(\tilde{Q}_i - \mu_{\tilde{Q}})^2}\sqrt{\sum_{i=1}^{n}(Q_i - \mu_Q)^2}} \tag{3.5}$$

While it captures shared variability, it is insensitive to systematic differences in scale or bias. To also capture the scale, the NSE was developed, which can be seen as a MSE that is weighted by the variance of the observed discharge:

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(Q_i - \tilde{Q}_i)^2}{\sum_{i=1}^{n}(Q_i - \mu_Q)^2} \tag{3.6}$$

The NSE incorporates bias in normalized form scaled by the standard deviation of the target variable. This can mask individual contributions of different error components. Therefore, the KGE was developed to jointly evaluate correlation, bias, and variability as separate components:

$$\text{KGE} = 1 - \sqrt{(\text{PCC} - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}, \quad \alpha = \frac{\sigma_{\tilde{Q}}}{\sigma_Q}, \quad \beta = \frac{\mu_{\tilde{Q}}}{\mu_Q} \tag{3.7}$$

When computing KGE and NSE values, we use station-wise means and variances calculated from the training data set, following the approach in Kratzert et al., 2019b. For KGE, NSE, and PCC, higher values indicate better performance, with a maximum of 1 representing a perfect match. In contrast, lower values of MAE are better, with 0 indicating a perfect fit.

The bias metrics inform us about systematic biases in the model's predictions, i.e., whether they consistently under- or overestimate flow (Yilmaz et al., 2008). These metrics are based on flow duration curves (FDCs), which are fundamental diagnostic tools in hydrology that rank all discharge values from highest to lowest, providing a probability-based characterization of the entire flow regime at a given location (Vogel et al., 1994). By analyzing different segments of the FDC, these bias metrics enable systematic evaluation of model performance across distinct hydrological conditions. The closer these metrics are to 0, the lower the bias in the model. %BiasRR quantifies the overall relative bias across all data points, providing a general measure of whether the model systematically over- or underestimates discharge:

$$\%\text{BiasRR} = \frac{\sum_{i=1}^{n} \tilde{Q}_i - \sum_{i=1}^{n} Q_i}{\sum_{i=1}^{n} Q_i} \cdot 100 \tag{3.8}$$

%BiasFMS evaluates the bias in the slope of the flow duration curve midsegment (the midsegment between exceedance probabilities of 0.2 and 0.7), which serves as a signature index of the watershed's vertical redistribution of soil moisture and flow response characteristics:

$$\%\text{BiasFMS} = \frac{(\log(\tilde{Q}_{m1}) - \log(\tilde{Q}_{m2})) - (\log(Q_{m1}) - \log(Q_{m2}))}{\log(Q_{m1}) - \log(Q_{m2})} \cdot 100, \tag{3.9}$$

where $m1 = 0.2$ and $m2 = 0.7$ are exceedance probabilities. A steep midsegment slope indicates "flashy" watershed behavior with small soil storage capacity and predominantly overland flow, while flatter slopes are associated with slower, more sustained groundwater-dominated flow responses. %BiasFLV evaluates the bias in low-flow conditions by analyzing the lowest 30 % of flows (exceedance probabilities $0.7-1.0$), which is crucial for assessing model performance during drought conditions and base flow periods:

$$\%\text{BiasFLV} = -1 \cdot \frac{\sum_{l=1}^{L}(-\log(\tilde{Q}_l) - \log(\tilde{Q}_L)) - \sum_{l=1}^{L}(-\log(Q_l) - \log(Q_L))}{\sum_{l=1}^{L}[-\log(Q_l) - \log(Q_L)]} \cdot 100, \tag{3.10}$$

where l represents flows with exceedance probabilities $0.7 - 1.0$ and L is the minimum flow. %BiasFHV examines the bias in high-flow conditions by focusing on the highest 2 % of flows (exceedance probabilities < 0.02), which is essential for flood prediction

and extreme event modeling:

$$\%\text{BiasFHV} = \frac{\sum_{h=1}^{H} \tilde{Q}_h - \sum_{h=1}^{H} Q_h}{\sum_{h=1}^{H} Q_h} \cdot 100, \tag{3.11}$$

where h represents flows with exceedance probabilities < 0.02. This separation of bias evaluation across different flow regimes is hydrologically meaningful because models often exhibit different performance characteristics under various flow conditions. For instance, a model might accurately predict typical flows but systematically underestimate extreme floods or overestimate low flows during dry periods.

During open-loop inference, we evaluate metrics separately for each station and open-loop step, where the first step constitutes closed-loop estimation. Therefore, metrics are computed over the batch dimension only, allowing us to assess how model performance degrades with increasing lead times. This approach reduces the number of samples over which each metric is computed, which can lead to increased variability in the results, particularly for the bias metrics. %BiasFMS, which examines the slope of the midsegment, is particularly susceptible to this effect as it reduces each batch to two summary statistics before computing the slope. Although DRRAiNN was only trained on sequences that span 20 days, we evaluate it on 60-day sequences to investigate its ability to generalize beyond the training horizon. Additionally, we will plot the performance of the models on day-ahead predictions against the mean discharge of the different stations to identify potential systematic dependencies between flow magnitude and model accuracy. In all cases, we exclude 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 favorable metrics. With knowledge discovery being the main motivation of this work, we also test DRRAiNN for 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 lead to gains or losses of water not driven by meteorological forcings. By retrospectively inferring catchment areas from observed dynamics, we assess whether the rainfall-runoff model successfully propagates water across the landscape. The procedure is as follows: After a forward pass, we compute saliency maps by taking the gradient of the final discharge estimate with respect to all precipitation inputs. These maps tell us to which extent the model's output depends on the precipitation in each grid cell and time step. We multiply this gradient by the precipitation itself to focus the analysis on cells in which precipitation occurred. To examine how the attributions change over time, we split the sequence into subsequences of five days over which we take the mean. We do this for each station separately and visualize the resulting attributions to identify which areas contribute most to discharge estimation at each station. To reduce noise, we repeat this process across all test sequences and average the resulting attribution maps. In each station, we set the maximum value to be plotted to the 99% quantile of that station's values for

better visualization.
We compare the resulting attributions with catchment areas delineated from elevation data, which are widely used in the field. To evaluate their agreement quantitatively, we employ the following measure when comparing DRRAiNN to the ablated models: For each station, the attributions are zscore standardized, i.e., we subtract the mean and divide by the standard deviation. This way, each station has a similar effect on the overall metric. We then compute the Wasserstein distance between the attributions values inside the delineated catchment area and those outside 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 across the landscape in a physically plausible way. Specifically, it indicates that the model has implicitly learned the topographic structure of flow direction, i.e., that water generally flows downhill, solely from observed discharge dynamics. It is important to acknowledge that this evaluation approach assumes elevation-delineated catchment areas represent the true hydrological boundaries. Therefore, disagreement between model attributions and elevation-delineated boundaries does not necessarily indicate poor model performance, but could reveal hydrological processes not captured by topographic analysis alone.

3.7 Improvements over previous work

This work advances our previous publication Scholz et al., 2025a in several key areas:

Input feature selection We replaced solar radiation with temperature as an input variable, which provides better predictive information for ET modeling. While solar radiation represents the primary energy driver for both temperature and ET, temperature exhibits a more direct relationship to actual ET rates, which depend on atmospheric vapor pressure deficits and plant physiological responses that correlate more closely with air temperature than with solar energy input alone (Allen et al., 1998). Additionally, temperature offers superior spatial coverage and consistency compared to solar radiation measurements. The resulting differences in performance are examined in Section 4.5.2.

Model architecture The runoff embedding dimension was increased from 1 to 2, allowing the model to capture more complex relationships in the latent representation that connects the rainfall-runoff and discharge components. This expanded representational capacity enables DRRAiNN to encode richer spatial and temporal patterns in the transition from distributed hydrological processes to point-wise discharge predictions, improving the model's ability to distinguish between different flow generation mechanisms.

Training procedure We implemented an improved checkpoint selection strategy, using the best validation epoch for each random seed rather than the final training epoch. The previous approach relied on completing the full training schedule and evaluating based on final performance, which could be suboptimal if models overfit in later epochs. This modification ensures more consistent results across different random initializations and better generalization performance by preventing the selection of overfit model states, leading to more robust and reliable discharge predictions.

3.8 Implementation and reproducibility

To ensure reproducibility and facilitate future research, all code and data used in this work are publicly available. The complete DRRAiNN implementation, including model architecture, training procedures, and evaluation scripts, is available as open-source software (Scholz et al., 2024a). The preprocessed datasets for the Neckar catchment, including precipitation, elevation, and EFAS outputs, are published alongside the code (Scholz et al., 2024b), while discharge measurements must be obtained separately from the Global Runoff Data Centre (GRDC) following their data access procedures.
The implementation builds on several key software libraries: PyTorch (Paszke et al., 2019) and PyTorch Lightning (Falcon et al., 2019) for model development and training, Hydra (Yadan, 2019) for configuration management, Captum (Kokhlikyan et al., 2020) for gradient-based attribution methods, Rasterio (Gillies et al., 2013) for geospatial data processing, and PySheds (Bartos, 2020) for watershed delineation. Training requires approximately 22 h on a single NVIDIA GeForce GTX 1080 Ti or approximately 8 h on a single NVIDIA A100. A forward simulation of a 20-day sequence takes approximately 4 s.

4

Results

Prediction is very difficult, especially if it's about the future —
Niels Bohr

To evaluate DRRAiNN, we present hydrographs and compare performance with EFAS to contextualize DRRAiNN's results. We also show that DRRAiNN can retrospectively infer catchment-like structures, thus demonstrating how full differentiability supports physical interpretability. We then investigate how elevation and temperature data contribute to model performance. Through ablation studies, we examine the importance of separating local and spatially extended processes, as well as the role of hypernetworks in the architecture. Finally, we evaluate DRRAiNN's spatial generalization capabilities within the PUB framework.

4.1 Hydrographs

Both DRRAiNN and EFAS produce hydrographs that largely match the shape and magnitude of observed discharge across the range of discharge scales, demonstrating their strong performance (Figure 4.1). This includes both low flows (Figure 4.1a) and high flows (Figure 4.1b), with both models generally capturing the timing of discharge events well. However, the performance varies across stations, as illustrated by comparing each model's best (Figure 4.1c,d) and worst (Figure 4.1e,f) performing cases. Overall, the hydrographs suggest that EFAS tends to underestimate flows while DRRAiNN tends to overestimate low flows in particular. Since DRRAiNN is an autoregressive model, errors can accumulate over time leading to gradual decline in accuracy, as evident in its best-case performance at Untergriesheim (Figure 4.1c). Nonetheless, it is notable that DRRAiNN demonstrates robust extrapolation capabilities, accurately capturing peaks even after almost 50 days, despite being trained only on 20-day sequences.
Having examined the qualitative patterns in the hydrographs, we now turn to a comprehensive quantitative evaluation of DRRAiNN's performance across multiple metrics.

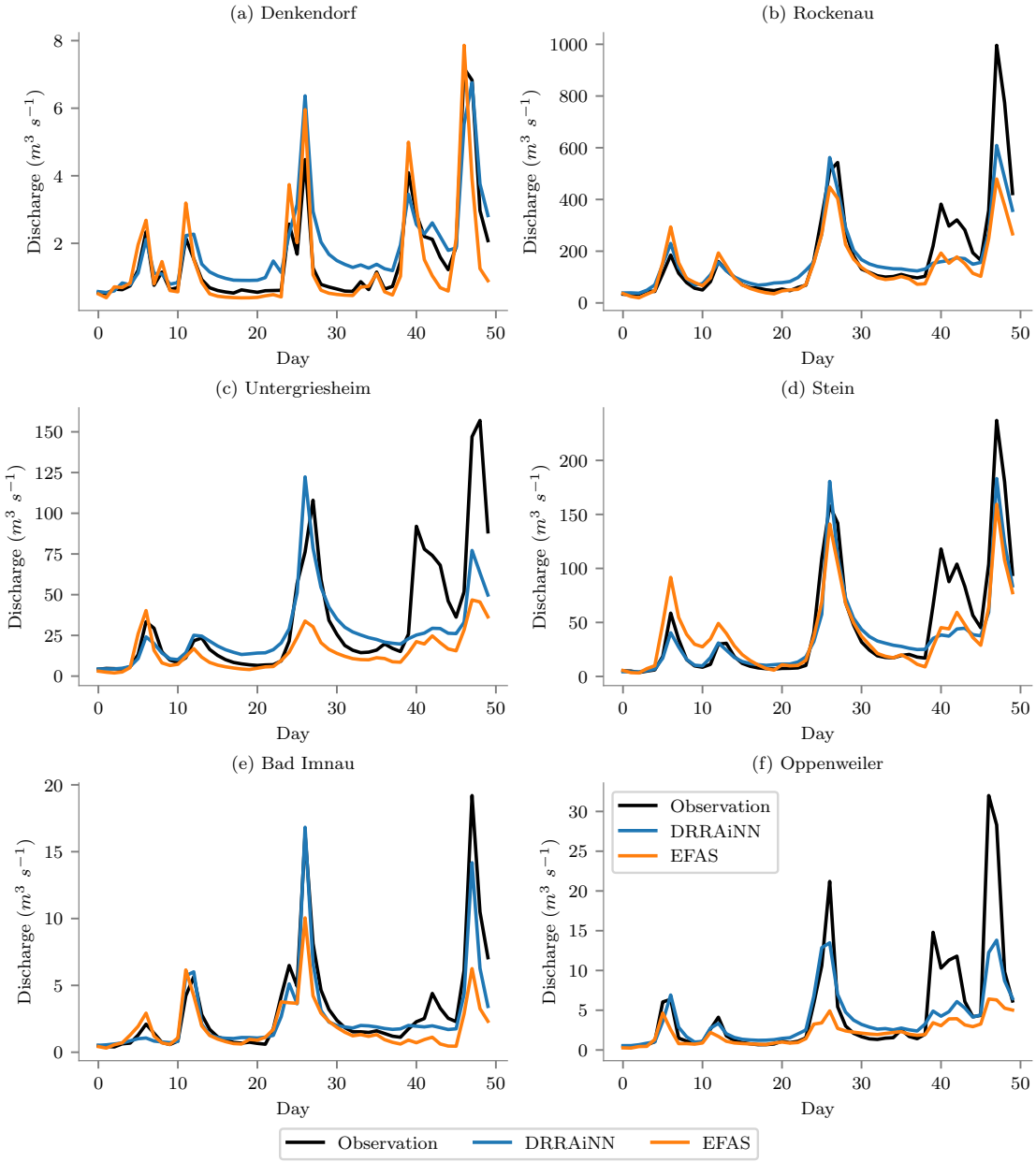


Figure 4.1: *Hydrographs showing observed discharge, DRRAiNN predictions for lead times up to 50 days, and EFAS simulations. The six panels show stations with the lowest (a) and highest (b) mean discharge, stations where DRRAiNN (c) and EFAS (d) achieve the best KGE performance, and stations where DRRAiNN (e) and EFAS (f) achieve the worst KGE performance on average. All results are from the test set, showing the sequence with the highest discharge variance to represent a challenging prediction scenario.*

4.2 Predictive performance

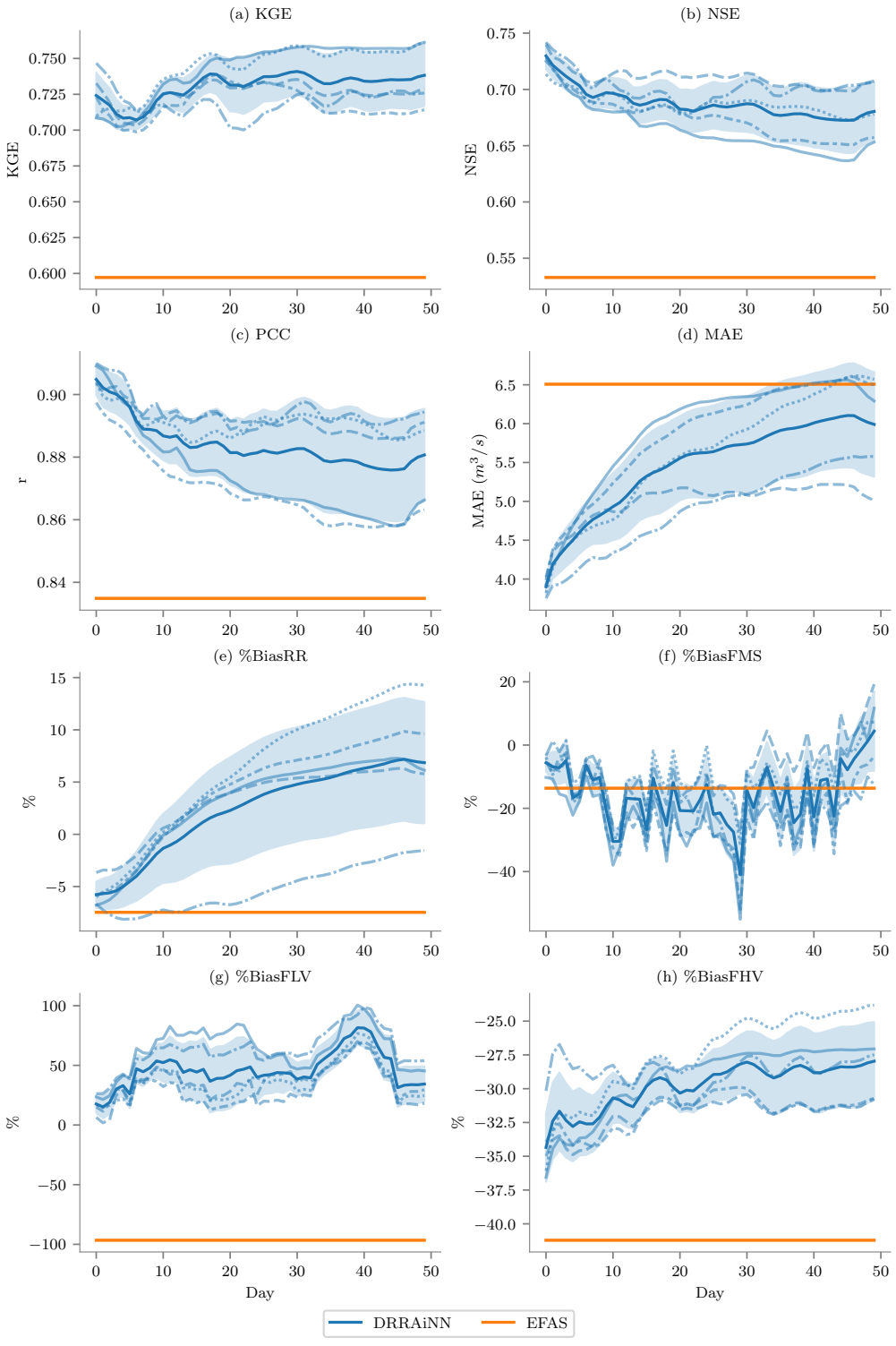


Figure 4.2: *Performance of five DRRAiNN instances compared to EFAS across eight hydrological performance metrics for lead times up to 50 days. Results are averaged across stations. For DRRAiNN, transparent lines show individual model instances, while the opaque line shows the ensemble mean and the shaded area shows the standard deviation across instances.*

Metric	Station Stats		KGE ↑		NSE ↑		PCC ↑		MAE ↓		%BiasRR → 0		%BiasFMS → 0		%BiasFLV → 0		%BiasFHV → 0	
Model	Mean	Std	DRRAiNN	EFAS	DRRAiNN	EFAS	DRRAiNN	EFAS	DRRAiNN	EFAS	DRRAiNN	EFAS	DRRAiNN	EFAS	DRRAiNN	EFAS	DRRAiNN	EFAS
Denkendorf	1.3758	1.3193	0.6179	**0.7430**	**0.7708**	0.7185	0.8475	**0.9034**	**0.2553**	0.3667	**0.1977**	0.8081	-54.3292	**16.2631**	**-24.6039**	-54.1243	-41.0673	**12.6537**
Mosbach	1.7058	2.4923	**0.7278**	0.5171	**0.6713**	0.4506	**0.9104**	0.8580	**0.4363**	0.7516	**-10.4070**	-24.3801	-36.9494	**3.0047**	**-43.6598**	-229.1235	**-43.0945**	-55.7115
Schwabsberg	1.7419	2.7898	**0.8258**	0.4294	**0.5052**	-0.0474	**0.9193**	0.8915	**0.6760**	1.1851	**-16.4297**	-32.5660	**27.2567**	-54.5031	94.2004	**17.1718**	**-50.3049**	-70.7381
Altensteig	2.2949	2.1843	**0.7777**	0.7163	**0.5586**	0.4899	**0.8892**	0.8535	**0.4741**	0.8143	-8.8433	**3.7689**	**7.4400**	28.8568	63.2123	**56.2601**	**-43.9464**	-44.3537
Oppenweiler	2.3922	3.2620	**0.5788**	0.2927	**0.5214**	0.1286	**0.8629**	0.6977	**0.6440**	1.1414	-11.2291	**-8.6735**	-13.3059	**5.9632**	**-7.5329**	-153.8356	**-55.8173**	-76.1367
Bad Imnau	3.6157	5.1527	**0.3979**	0.3164	**0.8868**	0.7796	**0.8928**	0.7468	**0.7091**	1.2445	**-5.7463**	-14.8709	**-26.1400**	-45.7181	**11.0712**	-115.8700	**-35.2789**	-50.2307
Rottweil	5.1150	6.1467	**0.6295**	0.5655	**0.7239**	0.6135	**0.8863**	0.8172	**1.1276**	1.6464	**-8.7966**	-10.7687	**12.9177**	20.8254	**40.2133**	-69.1205	**-37.9641**	-51.0631
Murr	5.6159	6.8477	**0.5640**	0.4529	**0.6189**	0.4240	**0.8676**	0.7925	**1.2976**	2.4352	**-5.3376**	12.2794	**21.7524**	24.9771	**-5.3538**	-238.4593	**-44.5794**	-60.7901
Neustadt	7.5300	7.7291	**0.7771**	0.5665	**0.5900**	0.2267	**0.8987**	0.7794	**1.8200**	2.7700	-9.1823	**-6.5674**	**-12.3432**	-32.7406	**-3.5763**	-147.4399	**-37.5558**	-57.4818
Gaildorf	9.4535	11.9829	**0.8848**	0.7469	**0.8030**	0.6186	**0.9363**	0.8706	**2.1232**	4.4693	**-4.0222**	10.8159	-12.7807	**2.7925**	**38.5754**	-71.2997	**-10.2133**	-16.9592
Untergriesheim	16.3501	17.6793	**0.8886**	0.4582	**0.7319**	0.2779	**0.9012**	0.7919	**3.4290**	6.8931	**-3.8040**	-30.5412	**-5.4842**	-10.0644	**32.1927**	-106.0822	**-35.2827**	-63.7038
Pforzheim	16.4702	12.9469	**0.7537**	0.6219	**0.8720**	0.7652	**0.9251**	0.8594	**2.1975**	3.6944	-0.6435	**0.2234**	**7.5965**	-37.9535	**10.4691**	-98.0902	**-27.1966**	-34.7401
Stein	22.1339	25.3836	**0.8715**	0.8107	**0.7540**	0.5656	**0.9148**	0.8349	**5.0265**	10.5352	-4.9470	**3.5797**	**0.7238**	-9.5424	**10.5086**	-72.9512	-27.9412	**-18.5758**
Kirchentellinsfurt	27.7958	25.3652	**0.7281**	0.6830	**0.9093**	0.8342	**0.9415**	0.8869	**3.7346**	6.8623	**-0.4703**	-8.4615	**-2.1579**	-39.2074	**7.3866**	-30.6062	**-12.2506**	-22.1817
Plochingen	49.0633	45.2066	**0.6868**	0.6865	**0.8450**	0.7366	**0.9396**	0.8616	**7.5476**	12.8698	**-4.2834**	-9.5777	**2.4933**	-20.8493	**32.2327**	-65.4190	**-27.4705**	-33.0006
Lauffen	83.0664	66.7827	0.7775	**0.7903**	**0.8628**	0.7621	**0.9465**	0.8800	**11.4504**	18.6305	**-6.2458**	-6.5792	**-26.3958**	-43.6868	**17.6262**	-155.5609	**-23.3820**	-23.7371
Rockenau	133.3257	113.6106	**0.8228**	0.7551	**0.7805**	0.7149	**0.9026**	0.8674	**23.3899**	34.3197	**1.9997**	-5.3975	**14.2726**	-40.2835	**26.8050**	-107.4998	**-30.5530**	-33.8700

Table 4.1: *Performance metrics of DRRAiNN and EFAS on each station with one-day lead times, including station statistics. Arrows indicate the direction of optimal performance: high values (↑), low values (↓), or values approaching zero (→ 0).*

Overall, DRRAiNN outperforms EFAS in most considered metrics across the majority of lead times (Figure 4.2). Since EFAS does not incorporate discharge values during inference, we report its performance as constant across lead times. In contrast, DRRAiNN's autoregressive nature causes errors to accumulate over time, leading to gradual performance decline at longer lead times. Table 4.1 provides precise metrics for one-day lead times.

The KGE plot (Figure 4.2a) reveals that DRRAiNN maintains strong performance over time at approximately 0.75. DRRAiNN consistently outperforms EFAS (KGE ≈ 0.60) throughout the entire 50-day horizon, despite being trained only on 20-day sequences. The NSE plot (Figure 4.2b) shows the expected autoregressive degradation, declining from 0.73 to 0.68, yet still substantially exceeding EFAS performance (0.53) even after 50 days. The PCC plot (Figure 4.2c) demonstrates strong linear relationships, starting at 0.90 and declining only slightly to 0.88, compared to EFAS' constant 0.83. The MAE plot (Figure 4.2d) reveals superior accuracy of DRRAiNN with an average error of $3.90\,\mathrm{m^3\,s^{-1}}$ in the beginning and $5.99\,\mathrm{m^3\,s^{-1}}$ after 50 days compared to EFAS' $6.51\,\mathrm{m^3\,s^{-1}}$. The bias metrics (Figure 4.2e-h) provide additional insights into model behavior across different flow regimes. The %BiasRR plot (Figure 4.2e) shows that both models initially exhibit similar, slightly negative bias in overall runoff ratio. However, DRRAiNN's bias increases steadily over time and eventually becomes positive, reflecting the characteristic error accumulation inherent in autoregressive models where prediction errors propagate and compound over longer forecast horizons. The %BiasFMS plot (Figure 4.2f) shows considerable variability for DRRAiNN, likely due to the limited data available for computing flow duration curve slopes. However, both models appear to exhibit a negative bias, suggesting that they predict flatter flow duration curve slopes than observed. This may indicate that the models overestimate the watershed's soil moisture storage capacity and predict more sustained, groundwater-dominated flow responses when the actual watershed behavior is more flashy with greater surface runoff components. For low flows, the %BiasFLV plot (Figure 4.2g) reveals that DRRAiNN generally maintains closer-to-zero bias compared to EFAS, indicating superior representation of low-flow conditions. While EFAS underestimates low flows, DRRAiNN overestimates them. Finally, the %BiasFHV plot (Figure 4.2h) shows that both models systematically underestimate high flows, though DRRAiNN exhibits less severe bias, suggesting better representation of flood peaks. This underestimation of extreme flows is a common limitation in hydrological modeling and highlights the ongoing challenge of accurately predicting rare, high-magnitude events that are crucial for flood risk assessment.

EFAS performance on the test year is poor compared to its validation set performance, where the difference between models was much smaller. The test year may represent challenging conditions for EFAS, and the performance gap observed here may not be

representative of typical performance differences between the approaches.

The performance plots reveal systematic patterns where all DRRAiNN instances tend to perform similarly at specific lead time steps. This behavior likely reflects sampling bias inherent in the single-year test set with 50-day sequences: Events occurring early in the test year (first 49 days) are underrepresented at longer lead times because they can only appear in sequences that started early enough to include them, while events occurring late in the test year (last 49 days) are underrepresented at shorter lead times because they only appear in sequences that started late in the year. Consequently, the meteorological characteristics of different parts of the test year systematically influence performance at specific lead times, creating apparent temporal dependencies.

Performance varies considerably across stations (Figure 4.3), which span nearly three orders of magnitude in discharge (from ≈ 1 to $> 100\,\mathrm{m^3\,s^{-1}}$). Which stations are harder to estimate, however, differs according to the different metrics, reflecting the distinct sensitivities each metric has, as discussed previously. For KGE (Figure 4.3a), both DRRAiNN and EFAS show similar variability across stations, though DRRAiNN generally achieves higher values at most locations. The NSE patterns (Figure 4.3b) reveal a striking difference: DRRAiNN exhibits much lower variability than EFAS and outperforms EFAS at all stations, while EFAS shows dramatic spread with some stations yielding very low or even negative NSE values. Similarly, PCC (Figure 4.3c) demonstrates that DRRAiNN maintains consistently high correlations with lower variability compared to EFAS, which exhibits greater spread across stations. For MAE (Figure 4.3d), both models show systematic increases with discharge magnitude, as expected for an absolute error metric, with both following similar scaling relationships though DRRAiNN generally maintains lower absolute errors.

Both DRRAiNN and EFAS show agreement on station difficulty rankings, which indicates that both models struggle with similar stations despite their different underlying approaches. However, the strength of this agreement varies considerably across metrics. The PCC correlation of 0.44 represents weak agreement, while the MAE correlation of 0.99 is unsurprising since absolute errors naturally scale with discharge magnitude, making both models exhibit similar error patterns across the discharge spectrum. More meaningful agreement is observed in KGE (correlation of 0.53, representing moderate agreement) and in NSE, which shows strong concordance at 0.91. Specifically, stations like Altensteig, Gaildorf, and Stein consistently yield high KGE performance, while Oppenweiler, Bad Imnau, and Murr represent challenges for both approaches. The standard deviation bars reveal that DRRAiNN shows relatively consistent variability across model instances.

The bias metrics (Figure 4.3e-h) demonstrate systematic patterns in how both models behave across different stations, with biases of DRRAiNN and EFAS showing correlation across the stations. For overall runoff ratio (%BiasRR), DRRAiNN maintains relatively consistent bias across stations, whereas EFAS displays more pronounced station-specific patterns, consistently underestimating discharge at some locations while overestimating at others (Figure 4.3e). Both models show substantial variability in %Bi-

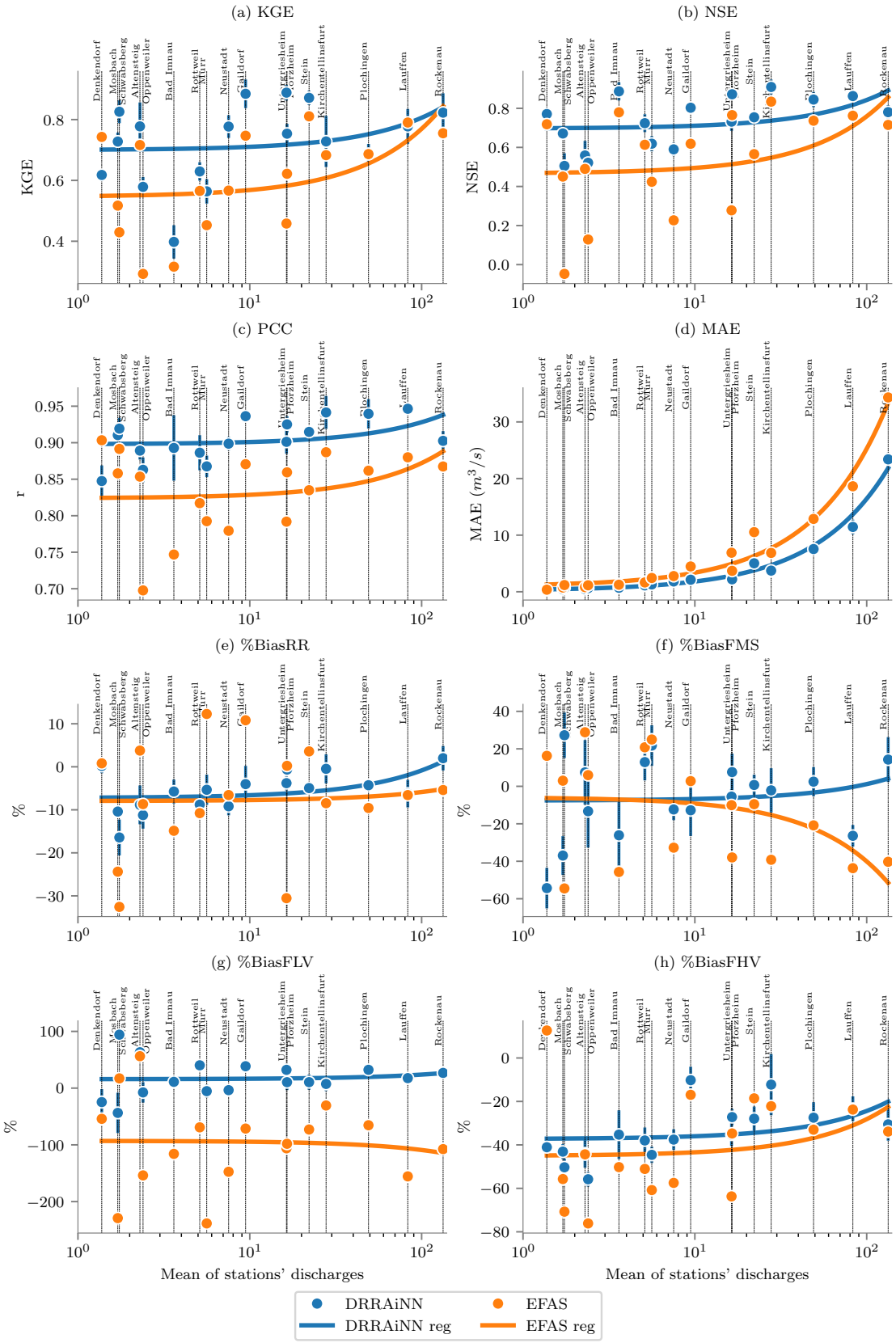


Figure 4.3: *Performance of DRRAiNN and EFAS at one day lead time across eight hydrological performance metrics and stations. Stations are ordered by mean discharge (log scale) and labeled with their names. Error bars show the standard deviation across DRRAiNN model instances. Solid lines represent linear regressions fitted to each model's performance against log-transformed mean discharge.*

asFMS, spanning both positive and negative biases across different stations (Figure 4.3f). For low flows (%BiasFLV), DRRAiNN demonstrates stable performance near zero bias, while EFAS exhibits consistent underestimation across most stations, with Altensteig and Mosbach being exceptions (Figure 4.3g). Note that the log-space calculation for %BiasFLV allows for values below $-100\,\%$. High flow bias (%BiasFHV) patterns show that both models systematically underestimate peak flows across nearly all stations, with Denkendorf being the sole exception where EFAS overestimates (Figure 4.3h). The correlation between DRRAiNN and EFAS bias patterns is strongest for high flow conditions.

The regression lines reveal strong systematic relationships between discharge magnitude and performance. Due to the logarithmic x-axis scaling, the linear regressions appear curved. For KGE, NSE, and PCC (Figure 4.3a-c), both models exhibit improved performance at higher-discharge stations, though DRRAiNN demonstrates a more gradual improvement slope, indicating more balanced performance across the discharge spectrum. Conversely, MAE (Figure 4.3d) increases substantially with discharge magnitude for both models, as expected given that MAE reflects absolute rather than relative errors. The %BiasRR reveals that DRRAiNN systematically underestimates discharge at stations with smaller catchments, while EFAS shows no clear relationship with catchment size (Figure 4.3e). For flow duration curve slope characteristics, DRRAiNN maintains relatively balanced %BiasFMS performance across discharge magnitudes, whereas EFAS exhibits increasingly negative bias at higher-discharge stations (Figure 4.3f). Although %BiasFLV remains generally stable across discharge magnitudes for both models (Figure 4.3g), both exhibit a systematic trend toward reduced negative bias (closer to zero) at higher-discharge stations compared to lower-discharge locations for %BiasFHV (Figure 4.3h).

The reasons for these systematic performance patterns, such as differences in catchment size, land cover, dam presence, upstream complexity, or the number of upstream gauging stations, could be analyzed in future work. These patterns have important practical implications: Smaller catchments may require different modeling approaches or additional input variables, while the strong performance at larger stations indicates that DRRAiNN is well-suited for major river forecasting applications. Such insights also suggest that station-specific calibration or bias correction could yield substantial improvements, particularly for the consistently challenging stations.

Beyond predictive accuracy, a key advantage of DRRAiNN's fully differentiable architecture is its potential for physical interpretability. We now examine whether the model can retrospectively infer meaningful spatial relationships.

4.3 Catchment area inference

We observe that DRRAiNN implicitly infers physically plausible catchment areas, as shown in Figure 4.4. Darker areas indicate regions with higher importance of precipitation for estimating discharge at the corresponding station. These attribution pat-

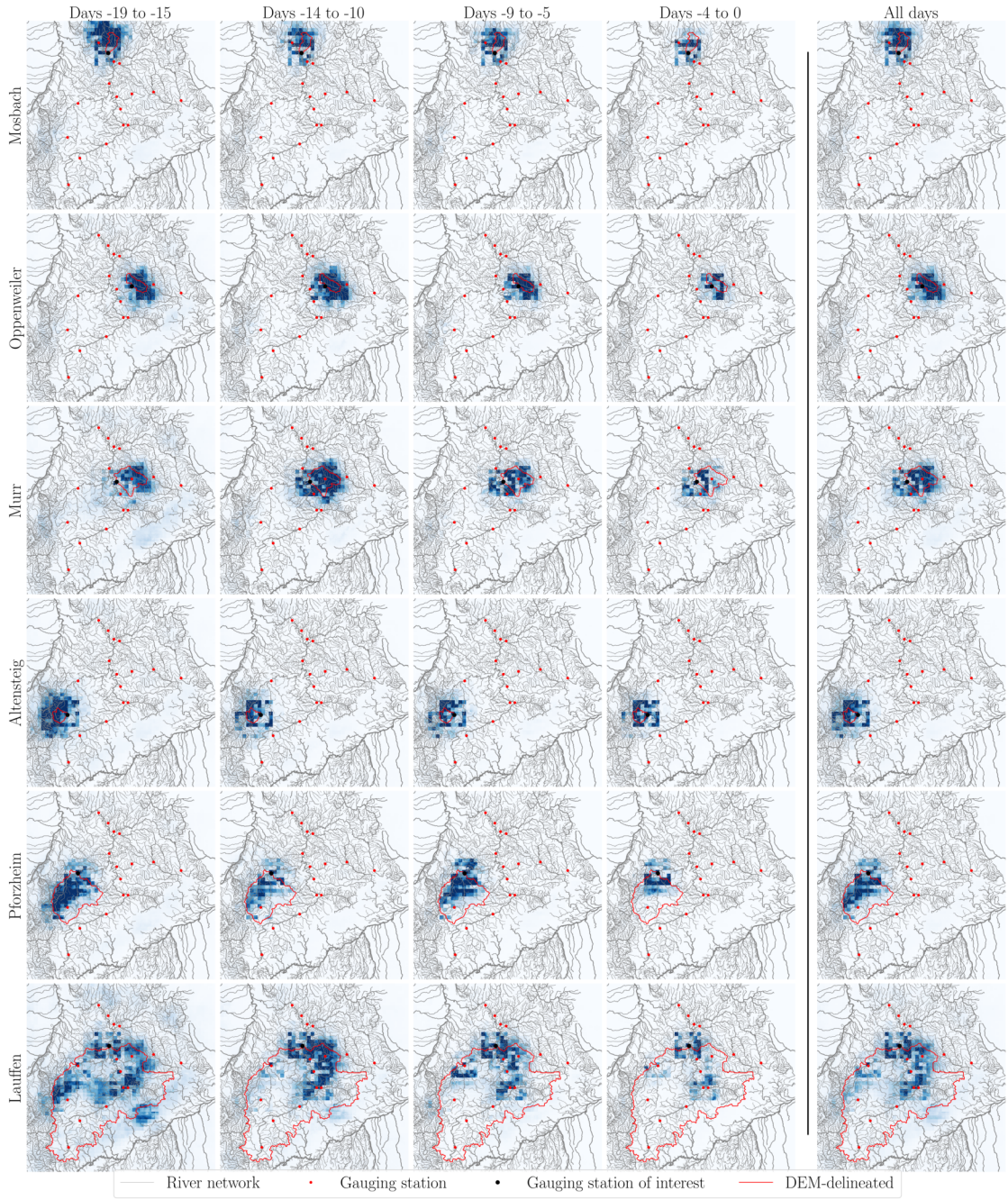


Figure 4.4: *Precipitation attribution maps showing the spatial influence of precipitation on discharge estimation at selected stations, averaged over 5-day intervals and all test set sequences. Darker colors indicate grid cells where precipitation has stronger influence on estimated discharge at the corresponding station. Traditional catchment areas delineated from elevation data are outlined in red for comparison.*

terns show spatial overlap with the catchment areas delineated from elevation alone (depicted in red), demonstrating that the model has learned physically meaningful spatial relationships from discharge observations.

The first four columns visualize attributions for subsequences of five days to illustrate temporal changes in spatial influence. The area of influence systematically increases when looking further into the past (leftward columns), which aligns with physical expectations: Precipitation from more distant locations requires longer travel times to reach the gauging station. This temporal expansion suggests that DRRAiNN implicitly accounts for flow routing and travel time distributions across the landscape. The rightmost column shows attributions averaged over complete 20-day sequences.

Most stations exhibit structures of 7×7 grid cells centered on the station location, exactly matching the ConvNeXt kernel size. These patterns reveal that DRRAiNN employs a hybrid approach: While it captures physically plausible catchment-scale influences, it also relies on precipitation in the immediate vicinity of gauging stations as a heuristic for discharge estimation, regardless of actual flow directions. This local dependence reflects statistical correlations in regional precipitation patterns rather than pure physical routing.

In the case of Pforzheim, DRRAiNN consistently assigns low importance to an area in the southeastern part, despite its inclusion in the elevation-delineated catchment area. This pattern could relate to documented underground flow paths near Pforzheim (Ufrecht, 2002). Due to the presence of these underground flow paths, water that would end up in Pforzheim according to the elevation instead moves towards the southeast, entering the Neckar River network via an alternative route. The results suggest that DRRAiNN may have detected these unobservable underground flows from precipitation and discharge dynamics. However, this interpretation remains speculative and would require dedicated hydrogeological investigation. It is also an example of how elevation-delineated catchments differ from real hydrological boundaries.

These results demonstrate the potential for physical interpretability in neural hydrological models. While attribution quality varies across model instances, the clearest case shows that DRRAiNN can infer physically meaningful spatial relationships. Interestingly, this model instance does not exhibit the highest KGE or NSE values, indicating a trade-off between accuracy and plausibility.

It is important to keep in mind that DRRAiNN is trained on daily discharge measurements. Learning sharp catchment delineations would require the data set to contain sequences in which it rained within the area, but not outside of it, over the extent of a $24\,\mathrm{h}$ period. As precipitation is very dynamic on this time scale, the chances for this to happen at all catchment boundaries are relatively low in a one year test set. In the future, we expect sharper results if we go from daily to hourly discharge data and use more data.

The catchment inference results raise fundamental questions about how DRRAiNN processes spatial information. To understand these mechanisms better, we now systematically examine the role of different input variables and architectural components.

For each variant, we train five independent model instances from scratch (different random seeds) following the training procedure described in Section 3.4. This ensures that each configuration is optimized specifically for its input and architectural constraints, rather than merely testing a single pre-trained model under different conditions.

4.4 Role of the elevation map

To better understand DRRAiNN's capabilities and limitations, it is crucial to examine how the model processes spatial information. Does the model learn transferable physical relationships that could enable spatial generalization, or does it primarily memorize location-specific statistical patterns?

4.4.1 Omitting the elevation map

As a first test, we remove the DEM entirely from the input. Without elevation data, the hypernetwork cannot produce spatially varying kernel weights for the depth-wise convolution, forcing the model to revert to translationally invariant processing, applying identical lateral flow patterns everywhere. This constraint eliminates the possibility of topographically-guided water routing, reducing the model to uniform spatial processing.

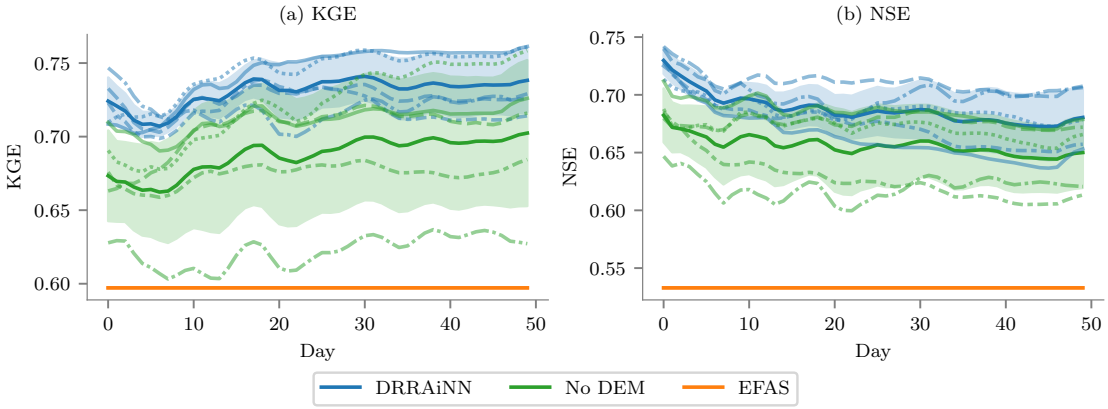


Figure 4.5: *Performance of five DRRAiNN instances compared to EFAS and DRRAiNN* ***without elevation*** *data across KGE and NSE metrics for lead times up to 50 days. The variant without elevation data does not receive DEM as input to the hypernetwork of the DWConv. Results are averaged across all stations. For DRRAiNN variants, transparent lines show individual model instances, while the opaque line shows the ensemble mean and the shaded area shows the standard deviation across instances.*

The performance impact is measurable but surprisingly modest (Figure 4.5). DRRAiNN without elevation shows larger performance spread in KGE (Figure 4.5a) than in NSE (Figure 4.5b), though both metrics remain consistently above EFAS levels throughout the 50-day forecast horizon. This relatively small degradation indicates that while elevation information is beneficial, DRRAiNN can compensate substantially through other mechanisms, such as learning precipitation patterns and temporal discharge dynamics.

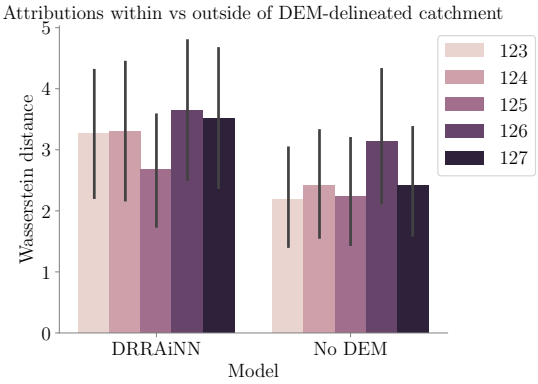


Figure 4.6: *Wasserstein distances between precipitation attributions inside and outside elevation-delineated catchment areas for DRRAiNN and DRRAiNN **without elevation** data. Higher distances indicate greater distinction between attributions within versus outside elevation-delineated catchment boundaries. Individual bars show results for all model instances, with standard deviations computed across gauging stations.*

However, the impact on physical interpretability is substantial. The Wasserstein distance analysis reveals that without elevation input, DRRAiNN shows considerably reduced ability to infer physically plausible catchment areas (Figure 4.6). The distance metric decreases notably, indicating that the learned precipitation attributions show weaker alignment with elevation-delineated catchment boundaries.

This contrast between modest performance loss and reduced interpretability suggests that elevation information serves dual roles: It provides a modest predictive advantage while contributing substantially to physically meaningful spatial attribution patterns. The model's ability to maintain reasonable discharge predictions without elevation reflects its capacity to learn statistical relationships from precipitation-discharge correlations, even without explicit flow routing guidance.

4.4.2 Providing a rotated elevation map

We now investigate the extent to which DRRAiNN uses elevation as positional encoding versus topographic flow routing. To examine this, we train and test DRRAiNN using elevation data rotated by 180°, which preserves the statistical distribution and spatial structure of elevation values while completely disrupting the actual topographic relationships.

The results reveal minimal performance degradation (Figure 4.7): KGE values show insignificant differences (Figure 4.7a), and NSE values are nearly identical (Figure 4.7b). Remarkably, one instance of the rotated DEM model even performs well above all original instances, demonstrating that correct topographic relationships are not essential for achieving high performance.

Even more remarkably, the model maintains its ability to reconstruct physically plausible catchment areas with rotated elevation input (Figure 4.8). The Wasserstein distances are statistically indistinguishable between original and rotated elevation cases,

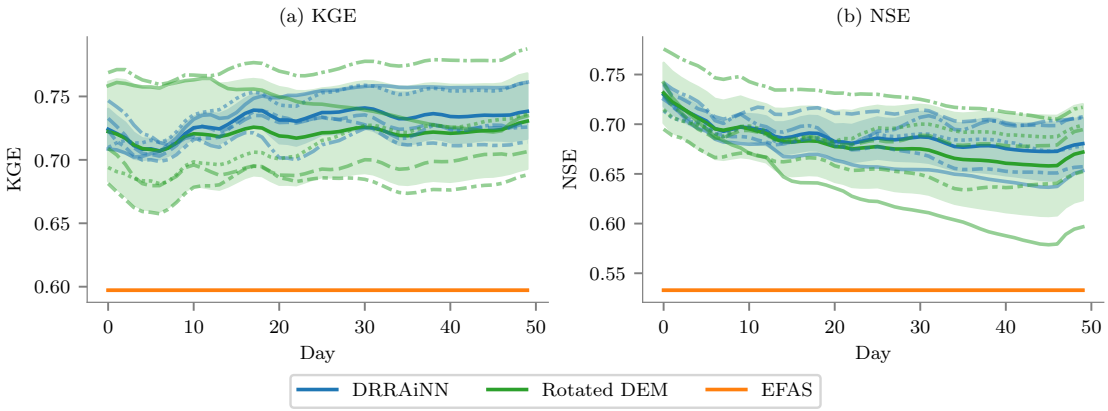


Figure 4.7: *Performance of five DRRAiNN instances compared to EFAS and DRRAiNN with **rotated elevation** data across KGE and NSE metrics for lead times up to 50 days. The rotated elevation variant receives a spatially rotated DEM as input to the hypernetwork of the DWConv. Results are averaged across all stations. For DRRAiNN variants, transparent lines show individual model instances, while the opaque line shows the ensemble mean and the shaded area shows the standard deviation across instances.*

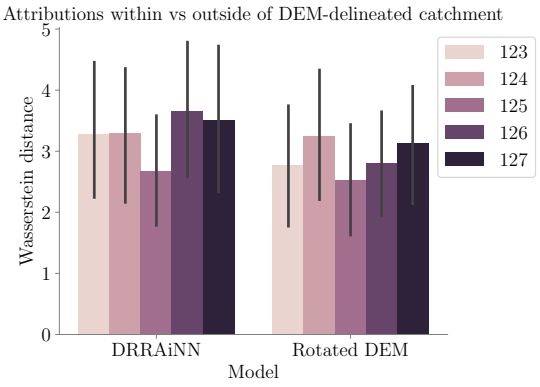


Figure 4.8: *Wasserstein distances between precipitation attributions inside and outside elevation-delineated catchment areas for DRRAiNN and DRRAiNN **with rotated elevation** data. Higher distances indicate greater distinction between attributions within versus outside elevation-delineated catchment boundaries. Individual bars show results for all model instances, with standard deviations computed across gauging stations.*

indicating that topographic correctness is not essential for learning appropriate spatial attribution patterns.
These counterintuitive findings suggest that DRRAiNN has learned to infer catchment structure primarily from the statistical relationships between distributed precipitation patterns and observed discharge responses, rather than from explicit topographic routing. The DEM appears to function more as a spatial coordinate system that enables location-specific parameter adaptation than as a guide for physical water movement.
This finding has profound implications for spatial generalization and PUB: It suggests that accurate topographic data may be less critical than previously assumed, provided sufficient precipitation-discharge training data is available to learn the underlying spatial relationships. However, this also raises questions about the model's physical interpretability. While the learned patterns appear physically plausible, they may reflect statistical rather than mechanistic understanding of hydrological processes.
Having examined how DRRAiNN uses topographic information, we next investigate the role of temperature.

4.5 Role of temperature

Temperature and solar radiation inputs are crucial for representing evapotranspiration (ET) processes, which significantly affect water balance in hydrological systems. ET removes water from the catchment, reducing the amount available for runoff generation. By incorporating temperature data, we enable DRRAiNN to implicitly account for these losses and better predict discharge variations, particularly during warm periods when ET rates are highest.
In our previously published work, DRRAiNN used solar radiation instead of temperature, based on the rationale that solar radiation is the primary driver of both temperature and ET (Allen et al., 1998). Here, we systematically examine whether temperature provides superior information content and investigate the effects of combining both inputs.

4.5.1 Omitting temperature

Removing both temperature and solar radiation inputs results in substantial performance degradation (Figure 4.9). However, the model maintains competitive performance with EFAS for the first five days in KGE (Figure 4.9a) and up to 30 days in NSE (Figure 4.9b), suggesting that precipitation patterns alone contain substantial predictive information for short-term forecasting.
The performance degradation is most pronounced at longer lead times, likely because ET effects accumulate over time. During extended dry periods or warm seasons, the inability to account for water losses becomes increasingly problematic for discharge prediction accuracy. Future work could investigate DRRAiNN's sensitivity to temperature input during different seasons.

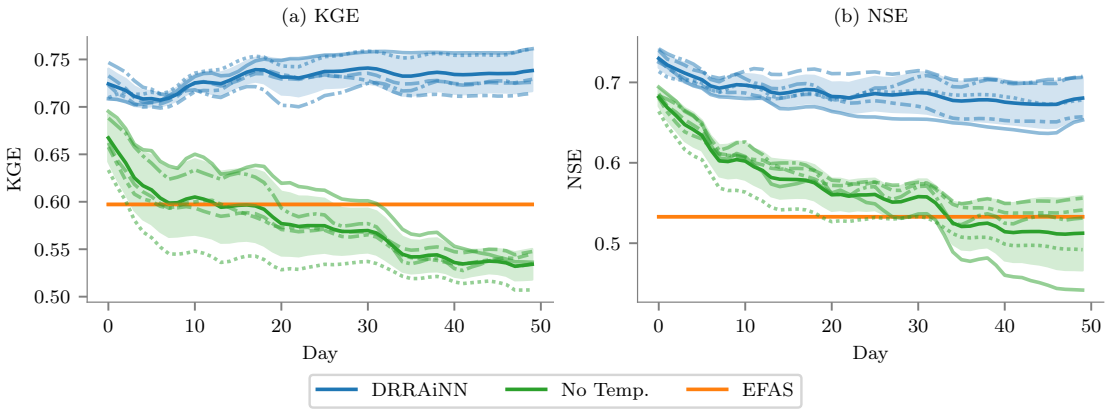


Figure 4.9: *Performance of five DRRAiNN instances compared to EFAS and DRRAiNN **without temperature** data across KGE and NSE metrics for lead times up to 50 days. The variant without temperature data does not receive temperature as input to the hypernetworks of the PWConvs. Results are averaged across all stations. For DRRAiNN variants, transparent lines show individual model instances, while the opaque line shows the ensemble mean and the shaded area shows the standard deviation across instances.*

4.5.2 Providing solar radiation instead of temperature

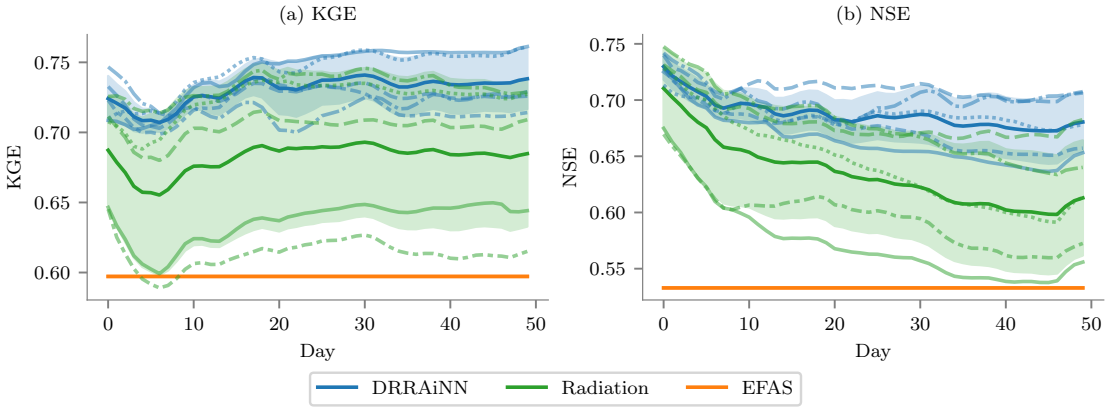


Figure 4.10: *Performance of five DRRAiNN instances compared to EFAS and DRRAiNN with **radiation instead of temperature** data across KGE and NSE metrics for lead times up to 50 days. The variant with radiation data receives radiation instead of temperature as input to the hypernetworks of the PWConvs. Results are averaged across all stations. For DRRAiNN variants, transparent lines show individual model instances, while the opaque line shows the ensemble mean and the shaded area shows the standard deviation across instances.*

To determine the optimal meteorological input, we compare temperature against the solar radiation used in our previous work. Direct comparison reveals that temperature provides superior predictive information in terms of KGE (Figure 4.10a) and NSE (Figure 4.10b) compared to solar radiation. While performance remains well above EFAS levels, this finding demonstrates that our previously published model configuration was suboptimal.

The superior performance of temperature reflects its more direct relationship to actual ET rates, which depend not only on solar energy input but also on atmospheric vapor pressure deficits and plant physiological responses that correlate more closely with air

temperature (Allen et al., 1998).

4.5.3 Providing temperature and solar radiation

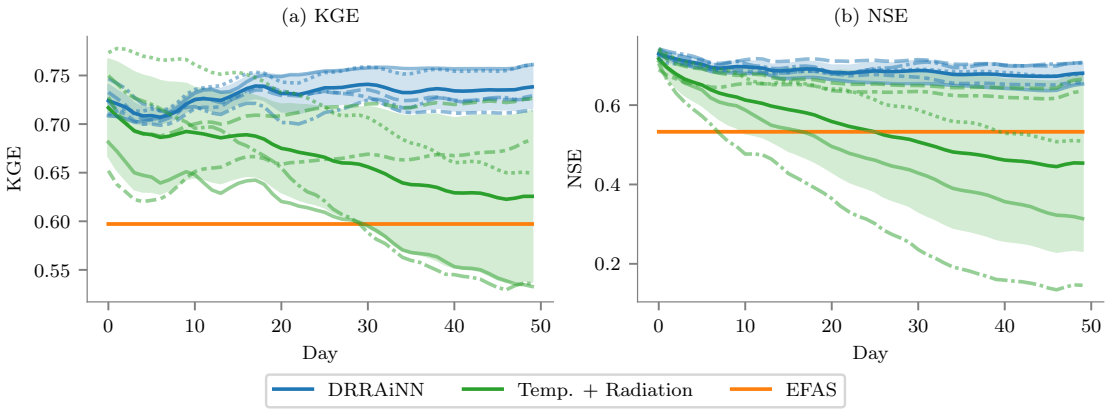


Figure 4.11: *Performance of five DRRAiNN instances compared to EFAS and DRRAiNN with **radiation and temperature** data across KGE and NSE metrics for lead times up to 50 days. The variant with radiation and temperature data receives radiation and temperature as input to the hypernetworks of the PWConvs. Results are averaged across all stations. For DRRAiNN variants, transparent lines show individual model instances, while the opaque line shows the ensemble mean and the shaded area shows the standard deviation across instances.*

Providing both temperature and solar radiation simultaneously yields mixed results for both KGE (Figure 4.11a) and NSE (Figure 4.11b). While some model instances achieve performance comparable to the temperature-only version, others perform substantially worse than EFAS. The pattern suggests that providing both temperature and solar radiation simultaneously impairs robustness, as evidenced by the large performance variability across different random initializations.

These findings highlight the importance of careful input selection in neural hydrological models: Using more input features is not always better, and redundant information can degrade rather than improve performance. For practical applications, temperature alone provides the optimal balance of ET information without introducing harmful redundancy.

Our investigation of input variables reveals their importance for model performance and, in the case of the DEM, physical plausibility. We now examine how specific architectural design choices contribute to DRRAiNN's capabilities.

4.6 Architectural design choices

To assess the contributions of specific inductive biases and architectural choices, we conducted a series of ablations on DRRAiNN. These experiments help determine which design decisions are crucial for both predictive performance and physical interpretability.

4.6.1 Separation of local and spatially extended processes

A key inductive bias in DRRAiNN is the explicit separation between spatially extended processes and local processes. Lateral water movement across the landscape is a spatially extended process primarily driven by elevation gradients. ET, conversely, is a local process largely influenced by local meteorological conditions, particularly temperature.

We encode this distinction by assigning these processes to different components of the ConvNeXt block: The DWConv is parameterized by a CNN that receives elevation as input, while PWConv1 and PWConv2 are parameterized by a MLPs that receive temperature. To test this design choice, we remove this separation by feeding elevation and temperature, together with precipitation, directly into the PWLSTM. Consequently, 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 also removed.

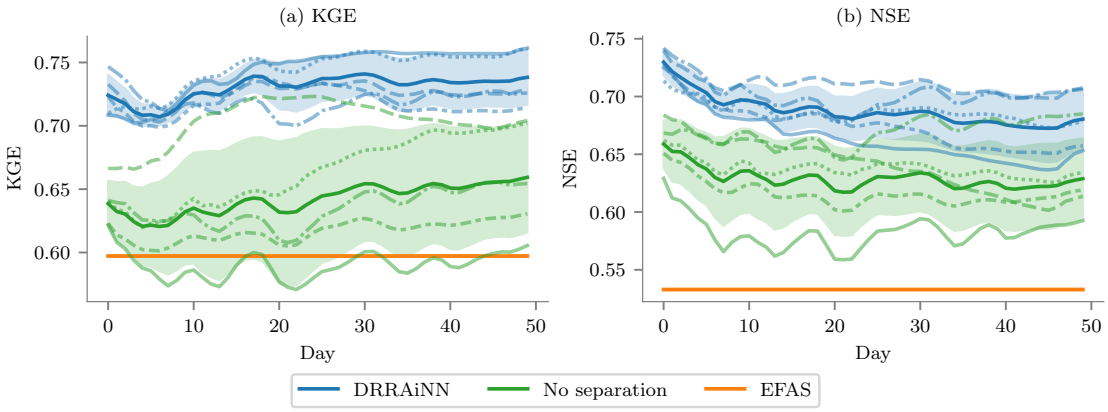


Figure 4.12: *Performance of five DRRAiNN instances compared to EFAS and DRRAiNN with **all inputs fed to the point-wise LSTM** across KGE and NSE metrics for lead times up to 50 days. The variant with all inputs fed to the point-wise LSTM receives precipitation, temperature, and elevation data as direct input to the PWLSTM instead of using hypernetworks. Results are averaged across all stations. For DRRAiNN variants, transparent lines show individual model instances, while the opaque line shows the ensemble mean and the shaded area shows the standard deviation across instances.*

Removing process separation results in substantial performance degradation (Figure 4.12a,b). Physical interpretability suffers dramatically, with Wasserstein distances dropping substantially (Figure 4.13), indicating that the model loses its ability to infer plausible catchment boundaries.

These results demonstrate that explicitly distinguishing between spatially extended and local processes is fundamental to DRRAiNN's success. The architectural separation not only improves predictive accuracy but is essential for maintaining physical interpretability, suggesting that this inductive bias captures genuine hydrological principles.

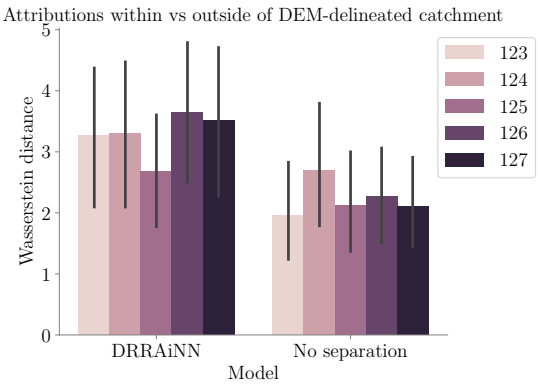


Figure 4.13: *Wasserstein distances between precipitation attributions inside and outside elevation-delineated catchment areas for DRRAiNN and DRRAiNN* ***with all inputs fed to the pointwise LSTM****. Higher distances indicate greater distinction between attributions within versus outside elevation-delineated catchment boundaries. Individual bars show results for all model instances, with standard deviations computed across gauging stations.*

4.6.2 Role of hypernetworks

Beyond process separation, another key component of DRRAiNN is its use of hypernetworks for location-specific adaptation. Hypernetworks enable DRRAiNN to dynamically adapt its processing based on local conditions. To assess their contribution, we train DRRAiNN without hypernetworks while preserving the separation between spatially extended and local processes.

In this ablation, elevation is concatenated with the hidden state, passed through a position-wise linear layer with SiLU activation, and fed into the DWConv. This preprocessing is necessary because DWConv requires equal input and output channel dimensions. Temperature is concatenated with the hidden state and fed directly into PWConv1, with input channels adjusted accordingly. The relativity bias is removed since it depends on the hypernetwork architecture.

Removing hypernetworks results in substantial performance decreases (Figure 4.14a,b). The impact on physical interpretability is measurable but less severe than process separation (Figure 4.15).

These results indicate that hypernetworks provide meaningful benefits by enabling location-specific parameter adaptation, though they are less critical than the fundamental separation of hydrological processes. The hypernetworks appear to enhance the model's ability to adapt its processing to local conditions, contributing to both performance and physical plausibility.

The architectural ablations confirm the importance of DRRAiNN's key design components for accuracy and physical plausibility. However, the earlier elevation experiments (Section 4.4) revealed that the model may rely more on statistical patterns than physical processes. This raises important questions about spatial generalization capabilities, which we now examine through ungauged basin prediction.

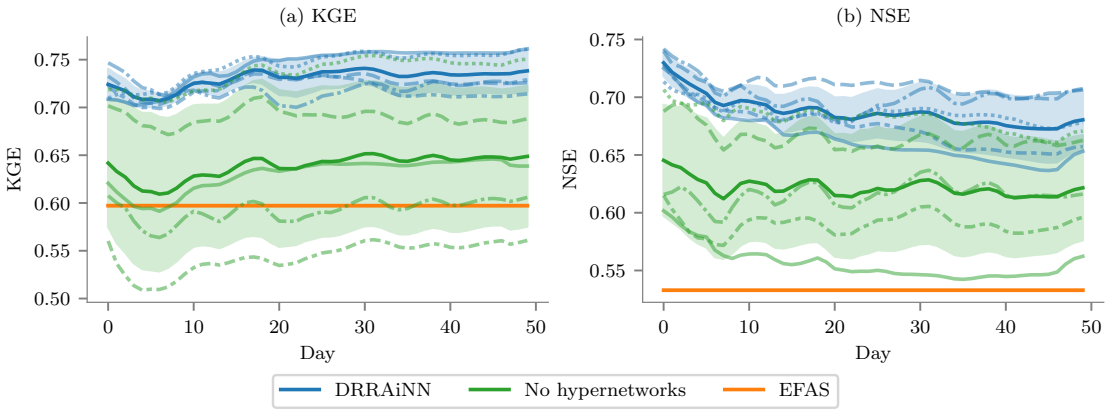


Figure 4.14: *Performance of five DRRAiNN instances compared to EFAS and DRRAiNN* ***without hypernetworks*** *across KGE and NSE metrics for lead times up to 50 days. The variant without hypernetworks directly feeds elevation data into the DWConv and temperature data into the PWConvs instead of using hypernetworks to generate convolution weights. Results are averaged across all stations. For DRRAiNN variants, transparent lines show individual model instances, while the opaque line shows the ensemble mean and the shaded area shows the standard deviation across instances.*

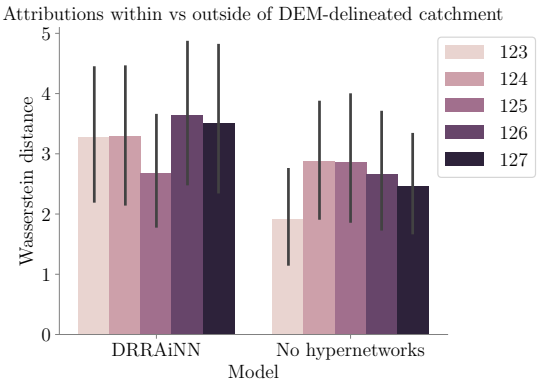


Figure 4.15: *Wasserstein distances between precipitation attributions inside and outside elevation-delineated catchment areas for DRRAiNN and DRRAiNN* ***without hypernetworks****. Higher distances indicate greater distinction between attributions within versus outside elevation-delineated catchment boundaries. Individual bars show results for all model instances, with standard deviations computed across gauging stations.*

4.7 Spatial generalization

In this section, we explore DRRAiNN's generalization capabilities for prediction in ungauged basins (PUB) within the Neckar river network. We employ leave-one-out cross-validation: For each station, we train five instances of DRRAiNN (different random seeds) while treating that station as ungauged during training. Ungauged stations never receive historical discharge data as input nor contribute to the loss function during training. During testing, however, we evaluate model performance at these withheld stations to assess spatial generalization capabilities within the same catchment system. All results are combined such that each station's prediction comes from model instances that treated it as ungauged.

This approach represents an intermediate step toward true ungauged basin prediction: While the stations are treated as ungauged during training, they remain within the same river network where other stations provide training data. This setup tests the model's ability to interpolate hydrological behavior to ungauged locations within a known system, rather than extrapolating to completely independent river networks.

Standardization is crucial for ANN training stability, particularly when dealing with variables spanning multiple orders of magnitude like river discharge. Our original DRRAiNN implementation standardizes each station's discharge separately using that station's log-transformed mean and standard deviation. However, this station-specific information would be unavailable for truly ungauged basins. Before proceeding to the actual PUB evaluation, we first test an alternative standardization approach on the full dataset (with all stations gauged) to determine its feasibility. This gives us an indication of the relative contribution of standardization changes versus ungauged basin prediction to overall performance losses.

4.7.1 Catchment size-based discharge standardization

For ungauged basin applications, we ideally need a proxy for discharge magnitude that can be determined without historical flow measurements. The elevation-delineated catchment area of a station provides such a proxy, as it can be computed a priori even for ungauged stations. Although one motivation for this work is that actual catchment areas cannot always be reliably inferred from elevation alone, catchment size still serves as a reasonable heuristic for a station's mean discharge magnitude (Figure 4.16).

In our catchment size-based standardization approach, we standardize each station's discharge by first dividing it by the elevation-delineated catchment area size, then applying log-transformed mean and standard deviation computed across all gauged stations. The catchment size-based standardization reduces mean performance in KGE (Figure 4.17a) and NSE (Figure 4.17b). While some individual model instances achieve performance levels comparable to the original station-specific DRRAiNN, others perform substantially worse, creating a large spread in outcomes across different random seeds. Despite this performance degradation and increased variability, we proceed

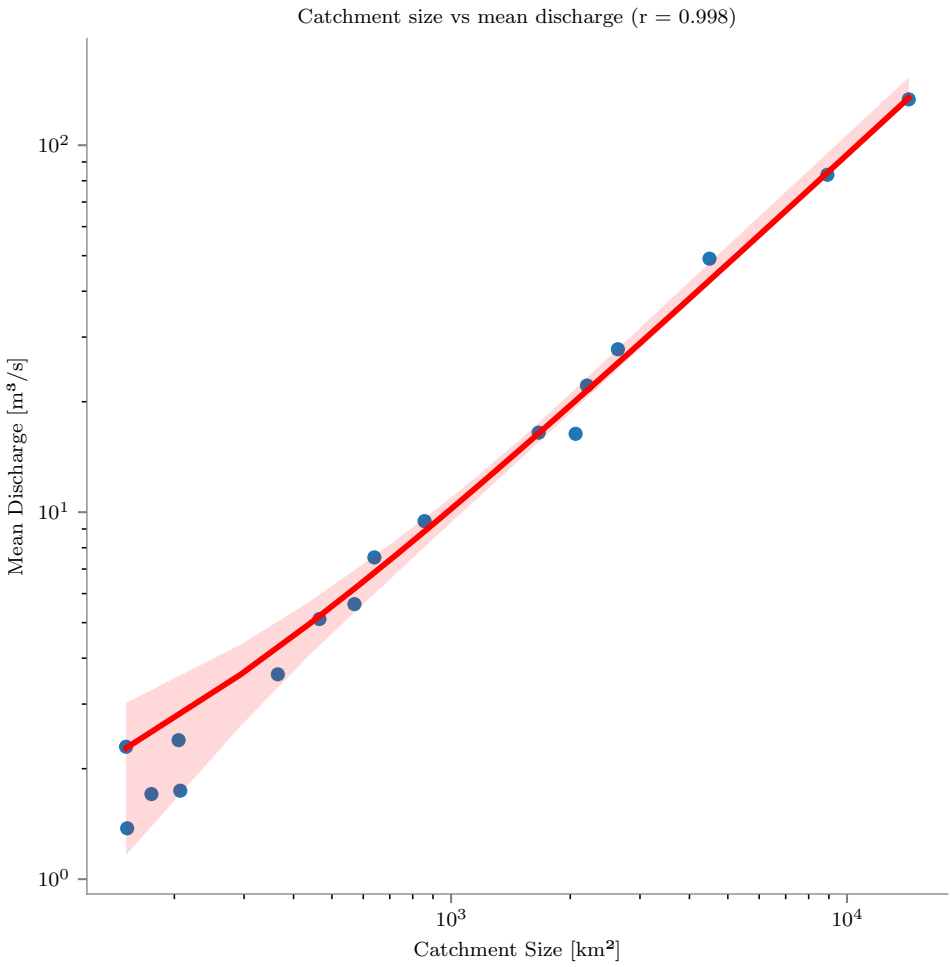


Figure 4.16: *Relationship between elevation-delineated catchment area and mean observed discharge across all 17 gauging stations in the Neckar river network. Each point represents one gauging station. The strong positive correlation demonstrates that catchment area serves as a reasonable proxy for discharge magnitude, justifying its use for standardization in ungauged basin applications.*

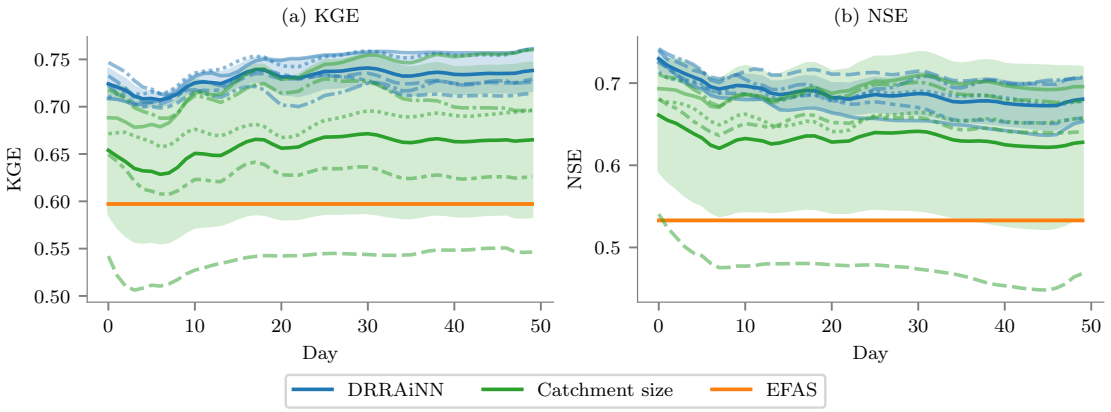


Figure 4.17: *Performance of five DRRAiNN instances compared to EFAS and DRRAiNN with **catchment size-standardized discharge** across KGE and NSE metrics for lead times up to 50 days. The catchment-standardized variant divides discharge by the station's catchment size before applying standardization using log-transformed mean and standard deviation computed across all gauged stations. Results are averaged across all stations. For DRRAiNN variants, transparent lines show individual model instances, while the opaque line shows the ensemble mean and the shaded area shows the standard deviation across instances.*

with this standardization approach for the ungauged basin evaluation because it represents the only viable alternative that does not require station-specific discharge statistics.

4.7.2 Prediction in ungauged basins

The transition to PUB results in substantial performance degradation across most metrics (Figure 4.18). DRRAiNN maintains competitive KGE (Figure 4.18a), albeit slightly worse performance relative to EFAS in the ungauged setting, but shows larger deficits in NSE, PCC and MAE (Figure 4.18b-d). The autoregressive performance degradation observed in the gauged setting is less pronounced here, with some metrics even showing slight improvement over time (Figure 4.18b, c). This may indicate that longer tune-in periods could improve performance by providing more information for the model to adapt to the ungauged location's hydrological characteristics before beginning autonomous prediction.

The bias metrics reveal additional challenges in ungauged basin prediction (Figure 4.18). %BiasRR shows considerable spread across model instances, making it difficult to determine systematic patterns compared to the original model and EFAS (Figure 4.18e). %BiasFMS reveals a larger negative bias compared to both the original model and EFAS, suggesting underestimation of flow duration curve slopes (Figure 4.18f). For low flows (%BiasFLV), the model shows larger positive bias than the original, indicating systematic overestimation of low flow volumes (Figure 4.18g). High flow bias (%BiasFHV) reveals larger negative bias with substantial spread across instances, suggesting inconsistent underestimation of peak flows (Figure 4.18h). These bias patterns indicate that ungauged basin prediction not only reduces overall accuracy but also introduces systematic distortions across different flow regimes.

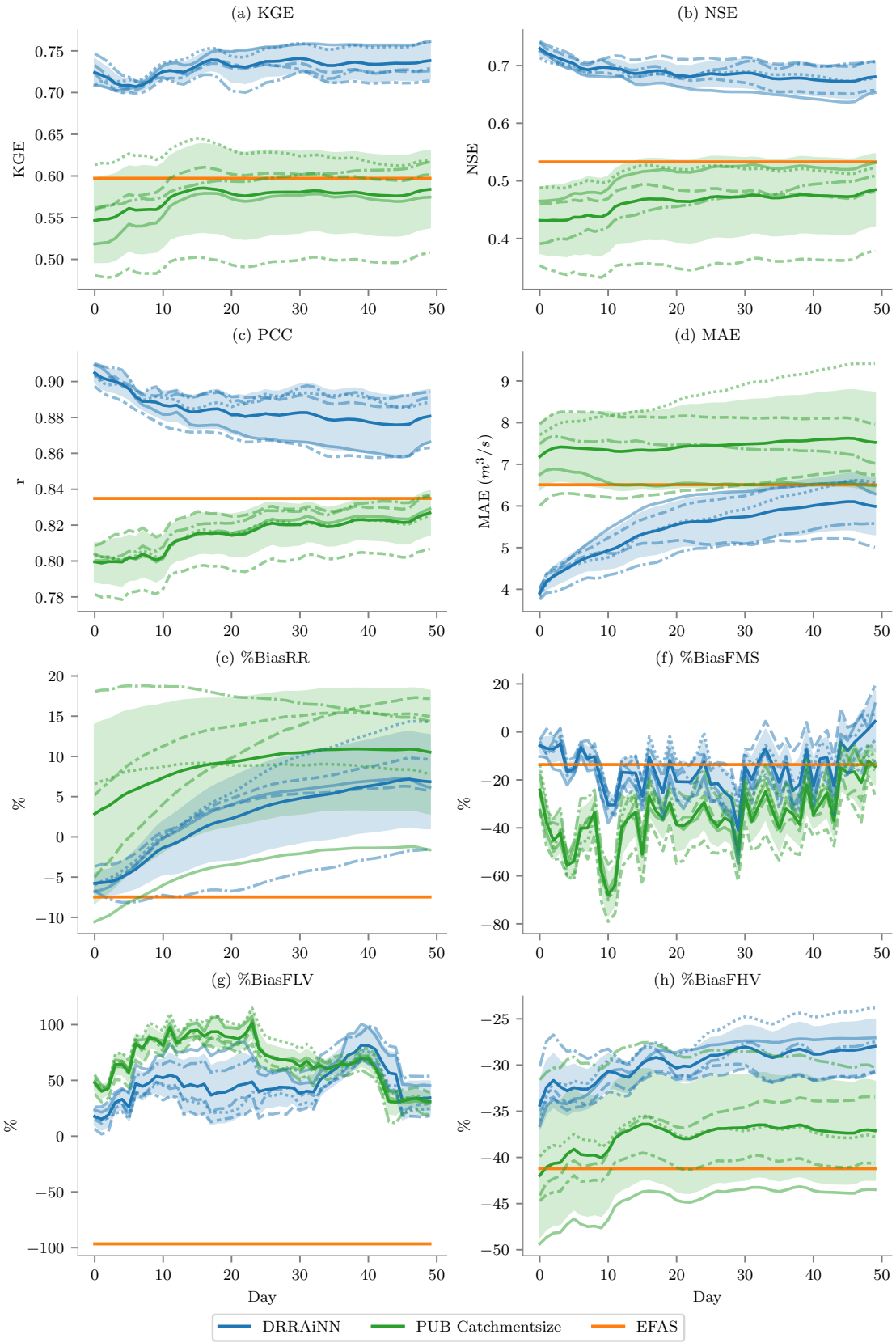


Figure 4.18: *Performance of five DRRAiNN instances compared to EFAS and DRRAiNN for* ***ungauged basins*** *across eight hydrological performance metrics for lead times up to 50 days. For ungauged basin prediction, DRRAiNN is trained using leave-one-out cross-validation, where each station is treated as ungauged during training and never receives historical discharge data as input. Results are averaged across all stations. For DRRAiNN variants, transparent lines show individual model instances, while the opaque line shows the ensemble mean and the shaded area shows the standard deviation across instances.*

The larger spread in performance across all metrics indicates high variability between model instances, suggesting that ungauged basin performance is highly sensitive to random initialization.

Station-level analysis reveals highly heterogeneous spatial generalization performance (Figure 4.19). Performance degradation varies dramatically across locations: Some stations experience catastrophic losses (e.g., Schwabsberg and Untergriesheim), while others show minimal impact or even improvements. Intriguingly, Murr exhibits higher KGE values in the ungauged setting, though this improvement does not extend to other metrics (Figure 4.19a). For NSE, the ungauged setting shows more dramatic performance losses, with several stations (particularly Schwabsberg and Altensteig) dropping to near-zero or negative values, indicating poor variance explanation (Figure 4.19b). PCC shows similar dramatic performance losses to NSE, with several stations experiencing substantial drops in correlation values (Figure 4.19c). MAE patterns reveal that absolute errors increase substantially in the ungauged setting, with some stations showing errors comparable to or exceeding EFAS levels (Figure 4.19d).

The bias metrics reveal additional complexities in ungauged basin prediction across stations. %BiasRR shows larger spread across stations compared to the original model, with Denkendorf, Bad Imnau, Murr, Untergriesheim, and Pforzheim showing large positive bias, indicating systematic overestimation of runoff at these locations (Figure 4.19e). %BiasFMS shows larger spread across stations, with more extreme negative and positive values compared to the gauged setting, indicating inconsistent estimation of flow duration curve slopes (Figure 4.19f). %BiasFLV is slightly larger than in the original model, mostly due to Altensteig and Gaildorf having large positive bias, suggesting systematic overestimation of low flow volumes at these specific stations (Figure 4.19g). %BiasFHV shows similar spread across stations as the original model, with varying degrees of underestimation that differ substantially between model instances (Figure 4.19h). These bias patterns suggest that ungauged basin prediction not only reduces overall accuracy but also introduces highly variable systematic distortions that depend strongly on both station location and model initialization.

The regression analysis indicates that PUB performance degradation is not uniformly distributed across discharge magnitudes (Figure 4.19). However, the relative improvement in stations with higher discharges likely reflect the number of upstream gauged stations available during training rather than simply catchment size or discharge magnitude. Stations with fewer upstream neighbors in the training network appear particularly vulnerable to ungauged treatment, as the model has limited opportunity to learn upstream-downstream flow relationships that could inform predictions at the withheld location. Conversely, stations with multiple upstream training stations benefit from the model's ability to propagate information through the river network, enabling better spatial interpolation even when the target station itself is ungauged. This finding suggests that PUB performance depends critically on the spatial configuration of the gauging network and the connectivity between training and target locations, highlighting the importance of network topology for spatial generalization in distributed

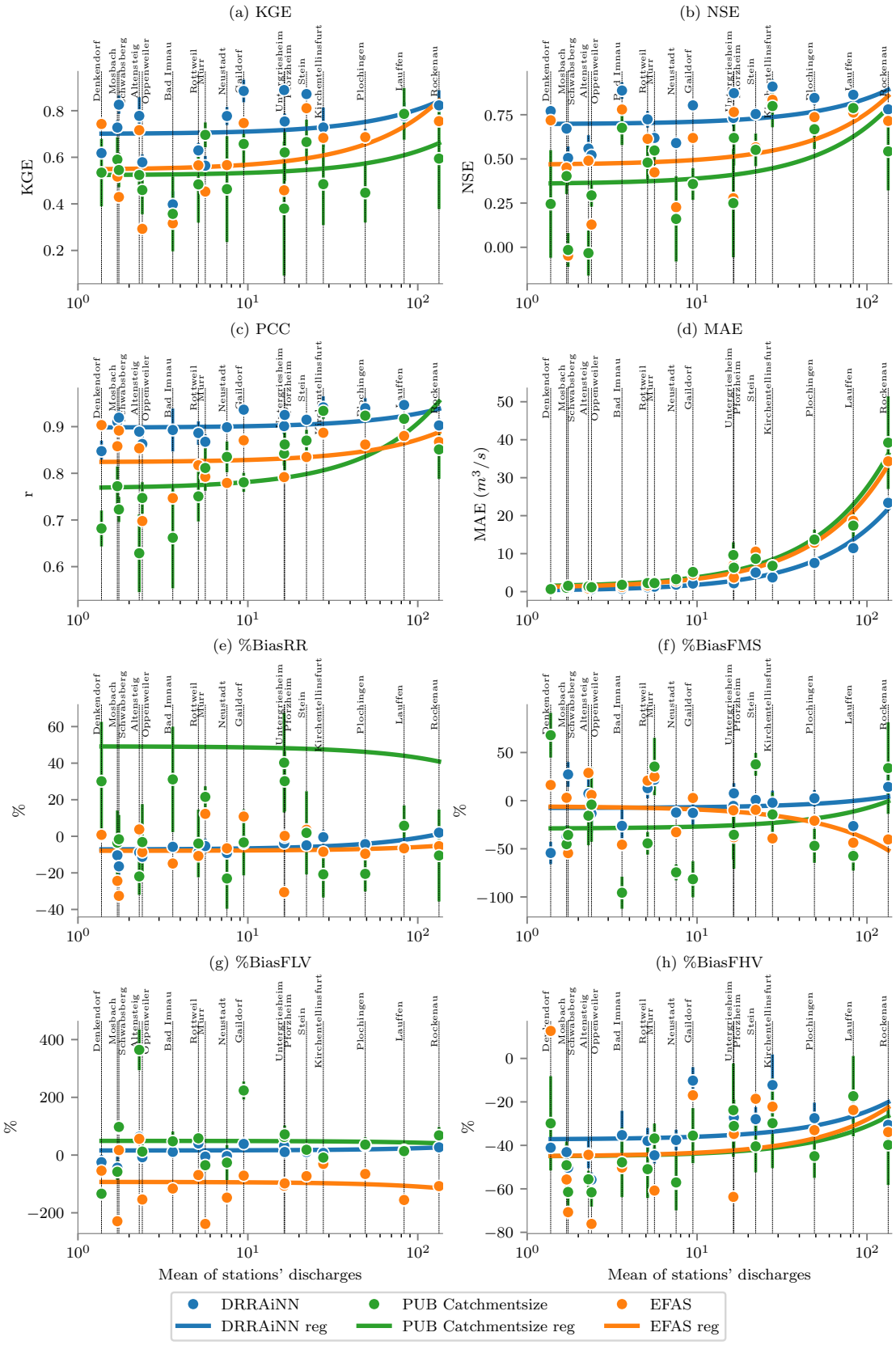


Figure 4.19: *Performance of DRRAiNN and EFAS at one-day lead time across eight hydrological performance metrics and stations in* ***ungauged basin*** *prediction. Each station's DRRAiNN performance is evaluated using models trained with that station treated as ungauged. Stations are ordered by mean discharge (log scale) and labeled with their names. Error bars show the standard deviation across DRRAiNN model instances. Solid lines represent linear regressions fitted to each model's performance against log-transformed mean discharge.*

hydrological models.
These spatial generalization challenges may be fundamentally linked to how DRRAiNN processes spatial information. The model's design explicitly breaks translational invariance through hypernetworks, enabling it to learn location-specific relationships between precipitation patterns and discharge responses. As we demonstrate in Section 4.4, the model uses elevation data as positional encoding: DRRAiNN learns location-specific statistical relationships between precipitation patterns and discharge responses. Such location-specific learning would naturally struggle when applied to ungauged locations where the statistical relationships have not been observed during training.
These results highlight both the promise and limitations of DRRAiNN for ungauged basin prediction. While the model demonstrates some spatial generalization capability, maintaining reasonable performance at several stations, the substantial variability in outcomes suggests that additional information could improve ungauged basin performance. DRRAiNN was neither built nor optimized with PUB in mind, leaving room for improvement.

5 Discussion

What we observe is not nature itself, but nature exposed to our method of questioning —
Werner Heisenberg

Our Distributed Rainfall-Runoff ArtIficial Neural Network (DRRAiNN) represents a fully distributed artificial neural network (ANN) architecture that estimates river discharge from past discharge observations, gridded elevation maps, and gridded meteorological forcings (precipitation and temperature). The results provide answers to the research questions posed in Section 1.4. DRRAiNN achieves superior performance compared to the European Flood Awareness System (EFAS) for lead times extending up to 50 days (**RQ1**). DRRAiNN was trained exclusively on 20-day sequences, including a 10-day warm-up period, demonstrating robust generalization to much longer prediction horizons.

5.1 Station-specific performance variability

We observe heterogeneity in discharge estimation difficulty across gauging stations (**RQ2**). DRRAiNN and EFAS tend to perform poorly at the same subset of stations, suggesting that modeling difficulties are station-specific rather than model-specific.
Several factors likely contribute to this variability across stations. Unobservable subsurface quantities, such as soil moisture and groundwater, can significantly influence discharge without being captured in the available input data. Additionally, anthropogenic interventions such as dam operations, water diversions, or urban infrastructure create dynamics that are challenging to model. Moreover, spatial heterogeneity may not be adequately resolved at the $4\,\mathrm{km} \times 4\,\mathrm{km}$ grid resolution employed by DRRAiNN. Input data quality represents another factor, as inconsistencies and biases in precipitation and discharge measurements could disproportionately affect certain stations.
Future research employing gradient-based attribution techniques could provide valuable insights into these station-specific patterns, identifying which meteorological and topographic features most strongly influence discharge at problematic stations. If attribution analysis reveals that certain stations are systematically affected by spatial pro-

cesses not adequately captured in the current $4\,\mathrm{km} \times 4\,\mathrm{km}$ grid resolution, this could motivate incorporating higher-resolution satellite-derived environmental data. Additionally, comprehensive databases of anthropogenic modifications such as dam locations and characteristics (Lehner et al., 2024) could provide valuable context for understanding station-specific performance differences and improving model interpretability in human-modified watersheds.

5.2 Catchment area inference

DRRAiNN can reconstruct physically meaningful catchment boundaries through gradient-based attribution (**RQ3**), demonstrating that distributed ANNs can learn spatial relationships from modeled discharge dynamics. The progressive expansion of attribution patterns as we look further back in time reveals that the model implicitly captures flow routing and travel time effects.

These attribution patterns, however, reflect DRRAiNN's hybrid approach that combines physical understanding with data-driven exploitation of regional precipitation correlations (see Section 5.4). This hybrid nature becomes apparent in the attribution maps, where 7×7 grid patterns emerge, directly corresponding to the convolutional kernel size. Training DRRAiNN on multiple river systems could reduce these artificial spatial patterns by encouraging the model to learn more generalizable, globally valid relationships rather than region-specific correlations.

Our evaluation of attribution quality relies on elevation-delineated catchment areas as ground truth. However, this reference standard has inherent limitations: Topographic delineation cannot capture subsurface flows, karst systems, or anthropogenic modifications to natural drainage patterns. Consequently, the attribution metric requires careful interpretation. While high similarity to elevation-delineated catchments may indicate physical plausibility, low similarity does not necessarily indicate poor model performance. Instead, it may reveal processes invisible to surface topography.

The Pforzheim case exemplifies this complexity, where attributions diverge from the elevation-delineated catchment. This divergence reflects the model's ability to detect known subsurface flow processes in the region, suggesting that data-driven attribution may sometimes provide more realistic representations of effective catchment areas than purely topographic approaches. However, this hypothesis requires further investigation to establish causality. Verification could be achieved through targeted tracer studies using artificial tracers or natural isotopic signatures, which have proven effective for delineating subsurface flow pathways and confirming model-inferred flow connections (Cook et al., 2012; Kendall et al., 2012).

5.3 Accuracy-plausibility trade-off

A striking finding from our interpretability analysis is that the model instance producing the most physically plausible attribution maps does not correspond to the one with

optimal predictive performance. This observation reveals tension between maximizing forecast accuracy and ensuring physical realism in model behavior (**RQ4**), which is a trade-off with significant implications for operational hydrological modeling.
This finding challenges the conventional assumption that predictive accuracy serves as a sufficient proxy for model quality in hydrological applications. While standard performance metrics effectively quantify forecast skill, they may inadequately capture whether models adhere to underlying physical principles. This limitation becomes concerning when models are deployed beyond their training domains or used to inform management decisions that depend on understanding causal relationships rather than mere predictive associations.
The implications extend beyond research, as operational flood forecasting and water resource management require models that do not only predict accurately but also provide physically meaningful insights into hydrological processes. Model evaluation should therefore consider multi-objective approaches that explicitly balance predictive skill with physical plausibility.

5.4 Role of input variables

The input variable experiments yield insights about how ANN models process environmental information (**RQ5**). Particularly revealing is the discovery that elevation functions primarily as positional encoding rather than explicit flow routing: DRRAiNN leverages elevation data to break translational invariance and learn location-specific adaptations, rather than learning universal hydrological principles that could transfer across space.
This statistical learning approach creates a trade-off: It enables accurate predictions and interpretable attributions within the training domain, but fundamentally limits spatial generalization. The model overfits to the spatial characteristics of the training catchment, learning location-specific statistical relationships rather than generalizable hydrological principles. While this approach can capture complex processes like underground flows (as potentially demonstrated at Pforzheim), it does so through empirical relationships rather than transferable physical understanding.
This limitation has important implications for operational applications, as it means that DRRAiNN would need to be trained separately for each new region before deployment, requiring local discharge observations. Training on multiple river networks across diverse regions (as discussed in Section 5.7.2) would likely reduce this overfitting behavior, forcing the model to rely less on elevation as positional encoding and develop more transferable physical relationships. Other approaches that could nudge the model towards operating in a physically more plausible manner are discussed in Section 5.6.
The temperature experiments reinforce fundamental hydrological principles: ET processes are critical for accurate water balance modeling and temperature provides superior information compared to solar radiation. The instability introduced by redundant

inputs (temperature + solar radiation) demonstrates that using more input features does not automatically improve ANNs. Careful input selection is crucial, though appropriate architectural constraints and inductive biases could alternatively guide how redundant information is processed within the model. This could enable stable learning from correlated environmental variables without manual feature curation.
These findings underscore the importance of carefully examining how ANNs utilize different input variables. Assumptions about physical processing may not always hold, with direct consequences for model generalization and interpretability.

5.5 Architectural design choices

Our ablation studies provide clear evidence for the importance of distinguishing between spatially extended and local hydrological processes within the model architecture. The incorporation of hypernetworks proves beneficial, suggesting that dynamic parameterization enhances the model's ability to adapt to varying hydrological conditions. When these architectural components are removed, models not only exhibit degraded predictive performance but also fail to generate physically realistic catchment boundaries in attribution analyses. This failure indicates that these components encode essential hydrological processes, particularly the complex patterns of water movement across heterogeneous topographies (**RQ6**).
These findings demonstrate the critical importance of incorporating domain-informed inductive biases into ANNs, particularly when learning from sparse target data. DRRAiNN learns meaningful spatial patterns across the gridded domain despite being trained on only point-wise discharge observations from 17 gauging stations. Rather than relying solely on data-driven pattern recognition, these physics-inspired architectural constraints guide the model toward physically meaningful relationships while reducing susceptibility to spurious correlations. This success connects to fundamental principles about how effective learning systems develop, which reflects broader insights from cognitive science and evolutionary biology.

5.6 Spatial generalization

The leave-one-out cross-validation within the Neckar river network reveals both promise and limitations for ungauged basin prediction (**RQ7**). DRRAiNN demonstrates meaningful spatial interpolation capabilities within a known hydrological system, but the highly heterogeneous performance across stations underscores the complexity of even this intermediate generalization task.
We also found that generalization difficulty likely depends on the availability of upstream training stations. Stations with fewer upstream gauged neighbors during training appear more vulnerable to performance degradation, suggesting that the model benefits from learning upstream-downstream relationships within the river network.

This finding has implications for operational applications, as it suggests that strategic placement of stations could improve ungauged basin performance.
Improving spatial generalization remains a fundamental challenge in hydrology. With regards to DRRAiNN, expanding training to multiple river systems (as discussed Section 5.7.2) would improve spatial generalization by reducing the model's dependence on catchment-specific positional biases and encouraging the learning of universal hydrological principles. However, a fundamental constraint must be acknowledged: Without discharge data from a specific catchment, the model cannot infer subsurface flow paths and local underground topology. This creates an inherent trade-off between two modeling strategies. Training on a single catchment allows the model to learn and encode catchment-specific subsurface structures, albeit at the cost of overfitting to local conditions. Conversely, training for spatial generalization across multiple catchments requires avoiding such overfitting, but necessarily accepts that the model cannot achieve perfect predictions in new locations without local calibration.
Several technical approaches can help navigate this trade-off. These include incorporating explicit physical constraints, either as hard constraints such as mass conservation (Hoedt et al., 2021; Harder et al., 2023; Wi et al., 2023) or as soft constraints through regularization techniques that penalize violations of energy conservation or unrealistic flow velocities. Semantically partitioning the rainfall-runoff model's hidden state into surface and subsurface components could further encode physical structure. Auxiliary prediction tasks during training, such as predicting intermediate state variables like soil moisture or evapotranspiration, could encourage the development of physically meaningful internal representations. Transfer learning approaches could leverage pretrained components from well-monitored catchments to bootstrap learning in data-sparse regions, reducing the training data requirements for new applications. While discharge measurements will always be necessary for training data-driven models, advances in remote sensing technology may enable applications to previously ungauged river networks (Gigi et al., 2019).

5.7 Technical considerations

The following sections examine the technical constraints and opportunities that influence DRRAiNN's current capabilities and future development pathways.

5.7.1 Temporal and spatial resolution

The daily temporal resolution of discharge data constrains DRRAiNN's ability to capture rapid hydrological responses and limits the sharpness of catchment boundary inference. Learning precise spatial attributions would require precipitation events that are spatially confined to specific areas within $24\,\mathrm{h}$ periods. The relative rarity of these events reduces the signal available for learning catchment delineations.
Moving to hourly discharge data therefore presents a natural next step for improving

DRRAiNN's capabilities for short-term predictions. Higher temporal resolution could enhance both predictive performance and attribution quality, potentially enabling the model to trace the origins of individual discharge peaks. This improvement would be valuable for flood forecasting applications where rapid response dynamics are critical. Higher temporal resolution would also enable detection of subdaily hydrological signatures such as diurnal evapotranspiration cycles, urban runoff patterns, and snowmelt dynamics that are smoothed out in daily aggregations. Event-based analysis would become possible, allowing the model to distinguish between different storm types and their characteristic hydrological responses.

However, transitioning to hourly data presents practical challenges: While the Landesanstalt für Umwelt Baden-Württemberg provides hourly measurements for stations primarily on smaller streams in the Neckar catchment, the overlap between stations with both hourly and daily data is limited. This constraint would require DRRAiNN to handle mixed temporal resolutions within a single training framework. Two architectural approaches could address this challenge: deploying separate LSTM components for each temporal scale (Gauch et al., 2021), or utilizing a single LSTM architecture with temporal resolution flags that indicate the current time scale (Acuña Espinoza et al., 2025). The latter approach offers the advantage of shared representations across temporal scales, enabling the model to learn consistent hydrological relationships that transfer between hourly and daily dynamics. Developing such a flexible system would enable training on substantially larger discharge datasets by combining measurements from different institutions that operate at varying temporal resolutions, ultimately improving model robustness and generalization capabilities.

The $4\,\mathrm{km} \times 4\,\mathrm{km}$ grid resolution may inadequately capture fine-scale hydrological processes, particularly in smaller subcatchments where local variability becomes important. Higher-resolution data sources are readily available for enhancing DRRAiNN's spatial representation. RADOLAN precipitation data is available at $1\,\mathrm{km} \times 1\,\mathrm{km}$ resolution, providing four times the spatial detail of current inputs. DEMs are available at much finer resolutions, including the free Copernicus DEM at $30\,\mathrm{m} \times 30\,\mathrm{m}$ (European Space Agency et al., 2019) and even higher resolution LiDAR-derived DEMs for specific regions.

While higher spatial resolution would be desirable, increasing grid density poses substantial computational challenges, particularly regarding GPU memory constraints that limit practical model deployment. See Section 5.7.5 for further information on computational considerations.

5.7.2 Spatial and temporal scope

The 10-year training period, while substantial, may not capture the full range of hydrological variability, particularly extreme events that occur at longer return periods. This limitation reflects data availability constraints, as consistent meteorological forcing data becomes increasingly sparse for earlier periods. Extending the training pe-

riod would require careful preprocessing to handle data gaps and non-stationarity in climate patterns.

Our evaluation is restricted to the Neckar river network in southwestern Germany, representing a single climate regime and hydrological setting. Training across diverse catchments would presumably encourage more generalizable physical relationships and reduce the model's dependence on elevation as positional encoding. The current single-network approach prevents assessment of generalizability across different climatic conditions, geological settings, or catchment characteristics.

Expanding to multi-regional training requires systematic selection criteria to ensure representative coverage of hydrological diversity. Key steps include selecting catchments that span different conditions regarding climate, topography, geology, land use; ensuring adequate temporal overlap in high-quality discharge and meteorological data; and developing boundary condition strategies for handling edge effects where catchments extend beyond the gridded domain. Nested modeling approaches could address boundary conditions by using coarser-resolution models to provide boundary fluxes for higher-resolution regional implementations. Standardization protocols might be necessary to harmonize data from different monitoring networks, addressing variations in measurement methods, data quality, and temporal alignment.

5.7.3 Data sparsity and infrastructure requirements

Despite its fully distributed architecture, DRRAiNN requires point-wise discharge observations for training, limiting applicability to regions with established monitoring networks. Although the model's performance with only 17 training stations is encouraging, this station density within the Neckar catchment far exceeds what is typically available in many global regions. The challenge becomes more acute in developing countries, where hydrological monitoring networks are often sparse, discontinuous, or entirely absent. Transfer learning strategies could address this limitation by pretraining DRRAiNN on data-rich regions and fine-tuning on sparse observations from target catchments.

However, DRRAiNN's architecture offers advantages for data-sparse scenarios: The model requires only discharge-correlated quantities rather than precise physical measurements as training targets. Multi-fidelity approaches could combine high-quality gauge data with lower-quality proxy measurements, such as citizen science observations or opportunistic measurements from transportation infrastructure. Recent advances in satellite-based discharge estimation (Gigi et al., 2019) could provide additional data, which would enable DRRAiNN deployment in ungauged regions where multispectral satellite imagery is available but ground measurements are not. This flexibility stems from the model's ability to learn relationships between any input signals that correlate with discharge dynamics, rather than requiring adherence to strict physical units.

5.7.4 Additional input variables

DRRAiNN incorporates fewer input variables compared to traditional models, potentially missing important drivers of hydrological variability. Incorporating additional input variables such as land cover, parent material, soil texture, vegetation indices, and potential ET data would provide richer environmental context. However, incorporating multiple correlated inputs requires careful consideration, as demonstrated by the instability observed when combining temperature and solar radiation. Strategic input selection or architectural modifications that handle redundant information through appropriate regularization would be necessary to prevent degraded performance from feature correlation.

DRRAiNN's fully distributed architecture enables spatial sensitivity analysis through interpretability methods, revealing when and where different variables influence predictions. This capability could uncover links between model representations and real-world hydrological processes, advancing scientific understanding alongside predictive performance. For example, attribution maps could show how soil texture influences discharge during drought periods versus how land cover affects flood response.

Currently, DRRAiNN requires a 10-day warm-up period for hidden states to adapt to catchment dynamics. The rainfall-runoff component likely uses this period to estimate soil moisture conditions, which significantly influence infiltration processes. Incorporating soil moisture as an explicit input variable could eliminate this warm-up requirement, reducing computational costs and training time. Recent advances in ML-based soil moisture estimation demonstrate the feasibility of this approach. Studies have shown that ANNs can effectively predict soil moisture from meteorological variables and remote sensing data (Shokati et al., 2024; Shokati et al., 2025a), providing the soil moisture inputs needed to eliminate DRRAiNN's warm-up period. Alternatively, compressed precipitation histories spanning days or weeks could provide similar benefits (Traub et al., 2024a; Ehret et al., 2025).

5.7.5 Computational efficiency and scalability

DRRAiNN's computational requirements present both advantages and limitations for operational deployment. Although DRRAiNN comprises only ~33 500 parameters, the fully distributed architecture requires substantial GPU memory, particularly for the gridded rainfall-runoff component. Training the model on the Neckar catchment requires approximately 22 h on a single NVIDIA GeForce GTX 1080 Ti or approximately 8 h on a single NVIDIA A100. Applying DRRAiNN to a larger region with more stations would not increase the number of parameters. It would, however, increase the number of activations. Gradient checkpointing, which we already employed during training, significantly reduces memory requirements by trading computation for memory. Here, intermediate activations are recomputed during the backward pass rather than stored. This technique could be further optimized through selective checkpointing strategies that identify which intermediate layers provide the best memory-

computation trade-offs (Feng et al., 2021).
Once trained, inference is computationally efficient, requiring only 4 s for a 20-day sequence, making it suitable for operational flood forecasting applications. This computational efficiency represents a significant advantage over process-based models, which typically require much longer execution times due to numerical solution of differential equations and iterative calibration procedures. While traditional hydrological models may require hours to days for complex catchment simulations, DRRAiNN's inference speed enables real-time operational deployment with minimal computational overhead. This efficiency gain becomes valuable for ensemble forecasting scenarios, where hundreds of model runs may be required to quantify prediction uncertainties, and for scenario analysis applications where rapid evaluation of multiple management alternatives is needed.

5.8 Future work

The findings presented in this work establish DRRAiNN as a viable approach for distributed rainfall-runoff modeling, yet they also reveal fundamental questions and limitations that warrant further investigation. Addressing these challenges will require advances across multiple dimensions: from improving input data quality and expanding spatial coverage to enhancing physical interpretability and developing operational deployment strategies. The subsequent sections outline these research directions, progressing from immediate practical considerations to longer-term architectural innovations.

5.8.1 Toward operational use

A key limitation for operational applications is the inherent difficulty of obtaining sufficiently accurate, high-resolution precipitation forecasts over multi-day lead times. Throughout this work, we assumed perfect precipitation forecasts by using historical observational data, focusing on water dynamics after precipitation reaches the surface. Numerical weather prediction models are limited in predicting localized extreme precipitation events and reducing forecast uncertainty, particularly for the lead times where DRRAiNN demonstrates superior performance. The model's practical utility therefore depends heavily on precipitation forecast quality. This dependency remains largely unexplored.
Several strategies could mitigate precipitation forecast uncertainty. Ensemble precipitation forecasting could be integrated with DRRAiNN to generate probabilistic discharge predictions, propagating meteorological uncertainties through the hydrological model. Hybrid forecast approaches could use ANN-based post-processing to bias-correct numerical weather predictions based on historical forecast errors in the specific catchment (Rasp et al., 2018). Adaptive forecasting strategies could dynamically adjust prediction confidence based on precipitation forecast skill, providing high-confidence

discharge predictions when precipitation forecasts are reliable and appropriately uncertain predictions when meteorological inputs are questionable. Real-time data assimilation could continuously update model states using observed precipitation and discharge to minimize the impact of forecast errors in earlier time steps.

Integration with flood inundation modeling represents a natural extension of DRRAiNN's capabilities. The model's discharge predictions could serve as boundary conditions for hydraulic models that simulate flood extent and depth across floodplains (Hunter et al., 2007). Real-time flood mapping could be achieved by coupling DRRAiNN's rapid inference capabilities with simplified hydraulic models or ML-based inundation models. This integration would enable comprehensive flood forecasting systems that provide not only discharge predictions but also spatial maps of flood risk, supporting emergency management decisions and public warning systems.

However, these operational applications share a fundamental requirement: robust uncertainty quantification. Whether handling precipitation forecast uncertainty or supporting emergency management decisions, reliable confidence estimates are essential for safe deployment. Quantifying prediction uncertainties represents a critical advancement for hydrological applications, as emphasized in previous research (Hrachowitz et al., 2013; Nearing et al., 2020). Equipping DRRAiNN with uncertainty estimation capabilities would enable the provision of confidence intervals alongside discharge predictions, essential for informed decision-making in safety-critical contexts.

Several promising approaches need investigation. Distributional parameter estimation could be implemented by extending the architecture to produce additional outputs interpreted as standard deviations in a negative log-likelihood loss function. Alternative methodologies include Bayesian ANNs (Neal, 2012; Lu et al., 2021), Monte Carlo dropout (Gal et al., 2016), and variational methods (Graves, 2011). Each approach has different trade-offs between computational efficiency and uncertainty quality, requiring systematic evaluation within the hydrological modeling context.

5.8.2 Applications beyond river discharge

DRRAiNN's architecture enables opportunities beyond traditional river discharge prediction. The model's capacity for gradient-based source attribution, combined with its distributed spatial processing, provides new opportunities in environmental monitoring and watershed management, enabling targeted interventions based on source identification.

Current approaches to erosion and sediment transport modeling face significant challenges in capturing the complex relationships between meteorological forcings, topography, soil types, and land use (Shokati et al., 2025b). Traditional empirical models like RUSLE (Revised universal soil loss equation, Renard et al., 1994) provide static estimates but miss temporal dynamics, while recent ML advances (Shokati et al., 2025b) show promise but typically treat distributed monitoring stations as isolated problems. This approach misses the spatial connectivity inherent in sediment transport processes,

as demonstrated by global sediment budgets that reveal complex interactions between erosion, transport, and deposition processes across river networks (Zarfl et al., 2022). These network-scale dynamics emphasize the need for distributed approaches that can capture upstream-downstream relationships rather than treating individual monitoring points in isolation.

DRRAiNN's architecture presents promising opportunities for erosion modeling through several adaptation pathways. The model could be modified to predict turbidity or sediment concentration by incorporating RUSLE outputs as additional input variables to enhance erosion potential estimation. This approach would leverage DRRAiNN's fully distributed grid-based architecture, eliminating spatial lumping and enabling the model to capture subcatchment heterogeneity that traditional models miss. Each grid cell would process local erosion dynamics while the graph neural network component would model sediment transport through the river network, capturing upstream-downstream relationships.

The model's gradient-based attribution capabilities could transform erosion management from reactive monitoring to proactive hotspot identification. During storm events, the system could backpropagate through the network to identify specific contributing areas responsible for sediment loading at downstream monitoring points. This source attribution capability would be valuable for distinguishing between different erosion processes and their temporal dynamics.

Studies in the Ammer catchment, a tributary of the Neckar River, demonstrate how sediment sources shift with seasons and discharge conditions (Y. Liu et al., 2018): Urban areas and in-stream processes contribute more during summer months, while bed and bank erosion become significant only above specific flow thresholds. These time- and flow-dependent source transitions highlight the importance of dynamic attribution methods that DRRAiNN's architecture can provide.

The model's temporal processing capabilities could enable real-time identification of erosion hotspots that change with flow and seasonal conditions. By incorporating dynamic land use data, the system could distinguish between natural erosion processes and anthropogenic sources from agricultural practices or construction activities. The model's long-term memory could support seasonal erosion analysis, incorporating vegetation changes and climate patterns that affect erosion prediction. However, these applications remain hypothetical and would require substantial model development and validation before implementation.

Beyond erosion modeling, DRRAiNN's architecture could be adapted for various other water quality applications that would benefit from spatially explicit source attribution. Nutrient modeling represents a critical application area, where the system could track nitrogen and phosphorus concentrations to prevent eutrophication events while distinguishing between agricultural fertilizer and wastewater treatment plant sources. Similarly, the model could monitor industrial contamination from heavy metals and pesticides, microplastics (Souza Machado et al., 2018), pharmaceutical compounds (Glaser et al., 2020), and bacterial contamination. Physical and chemical parameters

such as water temperature, pH, conductivity, and chlorophyll-a concentrations could also be predicted with source identification capabilities. The model's distributed architecture would enable simultaneous multi-parameter monitoring with parallel source attribution for multiple variables, supporting real-time pollution event detection and regulatory compliance monitoring.

In practice, this multi-parameter approach could be implemented by extending DRRAiNN's architecture to predict multiple output variables alongside discharge. The model would maintain a shared representation throughout most of the network, since water serves as the common transport medium for all contaminants, before branching into parameter-specific output heads in the final layers. This shared representation leverages the underlying hydrological processes that govern the transport of all waterborne substances while allowing parameter-specific dynamics to be captured in the final prediction steps. The training process would require careful loss balancing between discharge and water quality parameters, potentially using weighted multi-task learning approaches such as GradNorm (Z. Chen et al., 2018) to account for different measurement scales, frequencies, and uncertainties across parameters.

Successful implementation of these applications would require comprehensive, high-quality datasets that extend beyond inputs currently used by DRRAiNN. Essential data inputs include dynamic land use and land cover information, detailed soil characteristics and geological data, and erosion potential estimates from established frameworks like RUSLE. For water quality applications, point source locations and discharge characteristics from wastewater treatment plants and industrial facilities would be crucial for accurate source attribution. While the Landesanstalt für Umwelt Baden-Württemberg provides valuable water quality measurements such as conductivity, oxygen, temperature, turbidity, and pH data from monitoring stations, with temperature data available from 99 online measurement stations and chemical measurements from 442 stations, the temporal resolution of chemical measurements remains relatively low for real-time applications. Given these data limitations, initial development and validation efforts would benefit from starting with process-based simulation studies, where synthetic datasets can provide the dense spatio-temporal coverage needed to train and evaluate multi-parameter DRRAiNN models before transitioning to real-world deployment.

Beyond monitoring applications, DRRAiNN's differentiable architecture enables direct inference of optimal management actions through gradient-based optimization of land use inputs. For example, to minimize peak discharge or reduce erosion potential at specific monitoring locations, the trained model could backpropagate gradients through the network to identify which land use modifications would most effectively achieve these goals. The model's gradient-based action inference capabilities, building on established frameworks for goal-directed behavior (Otte et al., 2017), would enable systematic exploration of management scenarios in simulation before real-world implementation. Initial applications would likely focus on simulated environments where the model can safely explore different land use scenarios and quantify their effects on water quality parameters, providing evidence-based recommendations for

targeted interventions that address specific pollution sources or erosion hotspots.

5.9 Connections to evolutionary cognition and affordances

The development of effective ANN architectures for environmental modeling can draw upon principles discovered through evolutionary cognition. Since biological systems have undergone billions of years of evolutionary optimization to model and predict environmental dynamics, cognitive science provides a rich source of architectural inspiration for simulation science. From an information-theoretic perspective, evolution can be viewed as a extensive search process for optimal inductive biases that enable organisms to predict future world states given environmental forcings. The resulting architectures embody solutions to fundamental modeling challenges that simulation science also faces.

The success of biologically-inspired ANN architectures demonstrates the value of this approach. ANNs in general draw inspiration from the hierarchical structure of biological neural networks (Goodfellow et al., 2016), while CNNs are specifically inspired by the hierarchical processing of the human visual system and have revolutionized computer vision tasks (Krizhevsky et al., 2012). Similarly, variational methods in ML align with free energy principles that may underlie biological information processing (Friston, 2010; Scholz et al., 2022). These examples illustrate that cognitive principles can guide the design of effective computational architectures.

Contemporary ML research essentially recapitulates evolution's search for effective inductive biases, albeit through engineered rather than natural selection. Both processes favor architectural constraints that facilitate learning while preventing overfitting to spurious patterns. In DRRAiNN's case, physics-inspired biases, such as modular separation of rainfall-runoff and discharge processes, spatial convolution patterns, and graph-based river network representations, serve analogous functions to evolutionary cognitive adaptations.

This architectural philosophy connects most directly to DRRAiNN's hypernetwork components and the concept of affordances from cognitive science. Affordances describe how organisms perceive and interact with environmental structures based on their functional possibilities. DRRAiNN's hypernetworks embody this principle through second-order parameter adaptation: They dynamically modify the local routing and evapotranspiration behaviors based on environmental context. The hypernetworks learn to perceive the "flow affordances" of different topographic features (where elevation gradients afford downhill water movement) and "evapotranspiration affordances" of different temperature conditions (where local thermal conditions afford specific water loss rates). This creates context-dependent representations that capture how specific environmental configurations afford particular hydrological processes, paralleling how organisms evolved to detect environmental features that afford specific actions.

This conceptual framework aligns with our previous work on affordance learning, where we demonstrated how artificial agents can develop affordance maps that en-

code action possibilities in local environmental contexts (Scholz et al., 2022). Just as biological systems learn to navigate by detecting environmental features that afford movement opportunities in specific contexts, DRRAiNN learns to route water flow by detecting topographic features that afford hydrological processes. Both systems develop internal representations that capture the functional relationship between environmental structure and behavioral outcomes, suggesting that affordance-based learning principles may provide a general framework for developing physically meaningful ANN architectures.

Another crucial insight from cognitive science concerns the role of embodiment in learning world models. The "good regulator theorem" (Conant et al., 1970) states that any effective regulator of a system must contain or have access to a model of that system. This implies that biological systems, insofar as they successfully regulate their survival, must form internal models of their environment. This principle suggests that embodied interaction with the environment is fundamental to how biological systems develop and refine these internal world models.

However, simulation science faces a fundamental constraint: Researchers cannot freely choose environmental "actions" in the way that biological agents can. Instead, environmental forcings, such as precipitation patterns, temperature variations, or flow rates, are determined by uncontrollable natural processes. This distinction highlights a key difference between cognitive modeling and environmental simulation.

However, in the context of surrogate modeling, which is increasingly popular in hydrological science, a model could theoretically decide which environmental forcings should be fed into simulations to generate training data that maximizes information gain. This approach could adaptively focus on specific situations, locations, or land cover types that provide the most informative environmental patterns. Such selective data generation could be understood through the lens of affordances, i.e., the action possibilities that environments offer to agents (Gibson et al., 1986). We have previously demonstrated how agents can learn affordances through exploration based on information gain (Scholz et al., 2024d), suggesting pathways for incorporating such mechanisms into environmental modeling frameworks.

5.10 Key insights for different communities

The hydrological community might find the gradient-based attribution results most surprising. We can extract meaningful catchment boundaries directly from a trained neural network, where no watershed delineation algorithm is required. The model learns spatial connectivity from sparse discharge observations alone. However, this only works when we sacrifice some predictive accuracy. The best-performing model does not produce the most physically plausible attribution maps, which suggests we need to rethink how we evaluate models. Are we optimizing for the right thing?

The use of elevation as positional encoding reveals a fundamental tension in distributed hydrological modeling. DRRAiNN learns location-specific relationships rather than

transferable physical principles. This is not simply a limitation, but rather a choice with real consequences. When the model overfits to a specific catchment's spatial characteristics, it can potentially capture complex, hidden processes like subsurface flow pathways that would not show up in a more general model. The Pforzheim attribution patterns might be an example of this. But the flip side is obvious: Such location-specific learning does not transfer elsewhere. We cannot have both perfect spatial generalization and the ability to learn catchment-specific underground connections.

For machine learning researchers, this work makes a case for domain-specific inductive biases. Physics-informed constraints, such as process separation, spatial connectivity, mass conservation, reduce the model's flexibility. Nevertheless, these constraints do not just improve performance, they fundamentally change what the model learns. Maintaining spatial coherence throughout computation rather than collapsing to lumped representations preserves important information. The architectural insight here is that end-to-end differentiability lets us train fully distributed models when equipped with adequate inductive biases. Encoding domain knowledge into the architecture can be a better approach than increasing model size.

Water resource managers should know that neural approaches can match operational model quality for discharge forecasting. Real-time deployment is feasible. However, the elephant in the room is precipitation forecast quality: garbage in, garbage out still applies. The leave-one-out validation results show systematic spatial generalization problems that matter for ungauged locations. Models should not be deployed to new catchments without understanding these limitations.

5.11 Conclusion

This work shows that fully distributed ANN architectures can achieve competitive performance with operational hydrological models while providing novel insights into environmental processes. DRRAiNN successfully learns meaningful spatial relationships from sparse discharge observations and generalizes effectively across lead times that substantially exceed its training sequences.

DRRAiNN's hybrid approach combines genuine physical understanding with statistical pattern recognition. The model's capacity to infer realistic catchment boundaries and capture travel time effects reveals meaningful process representation, while its reliance on precipitation-discharge correlations highlights how ANNs can extract hydrological understanding from limited observational data.

The architectural ablations underscore that physics-inspired inductive biases become essential when learning from sparse target data. DRRAiNN's success in extracting gridded spatial patterns from only point-wise discharge observations demonstrates how domain-informed constraints guide ANNs toward physically meaningful solutions while avoiding spurious correlations. This success with inductive biases also illustrates how principles from cognitive science can inform environmental modeling architectures.

However, our results reveal important nuances in how ANNs process environmental information. The role of elevation as positional encoding rather than explicit flow routing suggests that successful performance may emerge through different mechanisms than initially anticipated. This behavior reflects training on a single river network. More diverse training data would presumably encourage the development of generalizable physical relationships rather than local positional biases. The path toward operational deployment requires addressing precipitation forecast dependency, expanding spatial generalization beyond single river networks, and integrating robust uncertainty quantification.

Despite these challenges, DRRAiNN represents a promising advance in data-driven hydrological modeling, demonstrating that fully distributed architectures can achieve both predictive accuracy and physical interpretability when guided by appropriate inductive biases.

Bibliography

Abbott, Michael B et al. (Oct. 1986a). "An introduction to the European Hydrological System—Systeme Hydrologique Europeen,"SHE", 1: History and philosophy of a physically-based, distributed modelling system". en. In: *Journal of hydrology* 87.1-2 (1-2), pp. 45–59. ISSN: 0022-1694. DOI: 10.1016/0022-1694(86)90114-9. URL: https://doi.org/10.1016/0022-1694(86)90114-9.

Abbott, Michael B et al. (Oct. 1986b). "An introduction to the European Hydrological System—Systeme Hydrologique Europeen,"SHE", 2: Structure of a physically-based, distributed modelling system". In: *Journal of hydrology* 87.1-2, pp. 61–77. ISSN: 0022-1694. DOI: https://doi.org/10.1016/0022-1694(86)90115-0. URL: https://www.sciencedirect.com/science/article/pii/0022169486901150.

Abebe, Nibret A., Fred L. Ogden, and Nawa R. Pradhan (Aug. 2010). "Sensitivity and uncertainty analysis of the conceptual HBV rainfall–runoff model: Implications for parameter estimation". en. In: *Journal of Hydrology* 389.3-4 (3-4), pp. 301–310. DOI: 10.1016/j.jhydrol.2010.06.007. URL: https://doi.org/10.1016/j.jhydrol.2010.06.007.

Acuña Espinoza, Eduardo et al. (June 2024). "To bucket or not to bucket? Analyzing the performance and interpretability of hybrid hydrological models with dynamic parameterization". en. In: *Hydrology and Earth System Sciences* 28.12 (12), pp. 2705–2719. DOI: 10.5194/hess-28-2705-2024. URL: https://doi.org/10.5194/hess-28-2705-2024.

Acuña Espinoza, Eduardo et al. (Mar. 2025). "Technical note: An approach for handling multiple temporal frequencies with different input dimensions using a single LSTM cell". en. In: *Hydrology and Earth System Sciences* 29.6 (6), pp. 1749–1758. DOI: 10.5194/hess-29-1749-2025. URL: https://doi.org/10.5194/hess-29-1749-2025.

Al Hossain, Bhuiya Md Tamim et al. (2015). "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.

Alfieri, L. et al. (2013). "GloFAS - global ensemble streamflow forecasting and flood early warning". In: *Hydrology and Earth System Sciences* 17.3, pp. 1161–1175. DOI: 10.5194/hess-17-1161-2013. URL: https://hess.copernicus.org/articles/17/1161/2013/.

Allen, Richard G et al. (1998). "Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56". In: *Fao, Rome* 300.9, p. D05109.

Anderson, S. and V. Radić (2022). "Evaluation and interpretation of convolutional long short-term memory networks for regional hydrological modelling". In: *Hydrology and Earth System Sciences* 26.3, pp. 795–825. DOI: 10.5194/hess-26-795-2022. URL: https://hess.copernicus.org/articles/26/795/2022/.

Arnold, Jeffrey G et al. (1998). "Large area hydrologic modeling and assessment part I: model development 1". In: *JAWRA Journal of the American Water Resources Association* 34.1, pp. 73–89.

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

Bartos, Matt (2020). *pysheds: simple and fast watershed delineation in python*. DOI: 10.5281/zenodo.3822494. URL: https://github.com/mdbartos/pysheds.

Başağaoğlu, Hakan et al. (2022). "A Review on Interpretable and Explainable Artificial Intelligence in Hydroclimatic Applications". In: *Water* 14.8. ISSN: 2073-4441. DOI: 10.3390/w14081230. URL: https://www.mdpi.com/2073-4441/14/8/1230.

Basher, Reid (2006). "Global early warning systems for natural hazards: systematic and people-centred". In: *Philosophical transactions of the royal society a: mathematical, physical and engineering sciences* 364.1845, pp. 2167–2182.

Battaglia, Peter W. et al. (June 2018). "Relational inductive biases, deep learning, and graph networks". In: *arXiv preprint arXiv:1806.01261*. arXiv: 1806.01261v3 [cs.LG]. URL: http://arxiv.org/abs/1806.01261v3.

Bergström, Sten (1976). *Development and application of a conceptual runoff model for Scandinavian catchments*.

Bergström, Sten (1992). *The HBV model–its structure and applications*.

Beven, Keith J and Michael J Kirkby (1979). "A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant". In: *Hydrological sciences journal* 24.1, pp. 43–69.

Beven, Keith J and Peter Young (2013). "A guide to good practice in modeling semantics for authors and referees". In: *Water Resources Research* 49.8, pp. 5092–5098.

Bharati, Luna et al. (2011). *The impacts of water infrastructure and climate change on the hydrology of the Upper Ganges River Basin*. Vol. 142.

Bindas, Tadd et al. (2024). "Improving River Routing Using a Differentiable Muskingum-Cunge Model and Physics-Informed Machine Learning". In: *Water Resources Research* 60.1. e2023WR035337 2023WR035337, e2023WR035337. DOI: https://doi.org/10.

1029/2023WR035337. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023WR035337. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023WR035337.

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

Börgel, F. et al. (2025). "From weather data to river runoff: using spatiotemporal convolutional networks for discharge forecasting". In: *Geoscientific Model Development* 18.6, pp. 2005–2019. DOI: 10.5194/gmd-18-2005-2025. URL: https://gmd.copernicus.org/articles/18/2005/2025/.

Brutsaert, Wilfried (2023). *Hydrology*.

Burnash, Robert JC (1973). *A generalized streamflow simulation system: Conceptual modeling for digital computers*.

Butz, Martin V. et al. (Sept. 2019). "Learning, planning, and control in a monolithic neural event inference architecture". en. In: *Neural Networks* 117. arXiv: 1809.07412, pp. 135–144. DOI: 10.1016/j.neunet.2019.05.001. URL: http://dx.doi.org/10.1016/j.neunet.2019.05.001 (visited on 09/16/2020).

Butz, Martin V. et al. (Jan. 2025). "Contextualizing predictive minds". en. In: *Neuroscience & Biobehavioral Reviews* 168, p. 105948. DOI: 10.1016/j.neubiorev.2024.105948. URL: https://doi.org/10.1016/j.neubiorev.2024.105948.

Camporese, Matteo and Manuela Girotto (2022). "Recent advances and opportunities in data assimilation for physics-based hydrological modeling". In: *Frontiers in Water* 4, p. 948832.

Chen, Shengyu, Jacob A. Zwart, and Xiaowei Jia (Aug. 2022). "Physics-Guided Graph Meta Learning for Predicting Water Temperature and Streamflow in Stream Networks". In: *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC USA). KDD '22. Washington DC, USA, pp. 2752–2761. ISBN: 9781450393850. DOI: 10.1145/3534678.3539115. URL: https://doi.org/10.1145/3534678.3539115.

Chen, Zhao et al. (2018). "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks". In: *International conference on machine learning*. PMLR, pp. 794–803.

Cho, Kyunghyun et al. (Oct. 2014). "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* (Doha, Qatar). DOI: 10.3115/v1/w14-4012. arXiv: 1409.1259v2 [cs.CL]. URL: https://doi.org/10.3115/v1/w14-4012.

Conant, Roger C and W Ross Ashby (1970). "Every good regulator of a system must be a model of that system". In: *International journal of systems science* 1.2, pp. 89–97.

Cook, Peter G and Andrew L Herczeg (2012). "Environmental tracers in subsurface hydrology". In.

Cools, Jan, Demetrio Innocenti, and Sarah O'Brien (2016). "Lessons from flood early warning systems". In: *Environmental science & policy* 58, pp. 117–122.

Courant, Richard, Kurt Friedrichs, and Hans Lewy (1967). "On the partial difference equations of mathematical physics". In: *IBM journal of Research and Development* 11.2, pp. 215–234.

Darcy, Henry (1856). *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.

Delforge, Damien et al. (2025). "EM-DAT: the emergency events database". In: *International Journal of Disaster Risk Reduction*, p. 105509.

Ehret, Uwe, Jingyang Chen, and Sebastian Lerch (Apr. 2025). "A comparative study of algorithms for lossy compression of 2-d meteorological gridded fields". In: *EGU General Assembly 2025*. EGU25-5977. Vienna, Austria. DOI: 10.5194/egusphere-egu25-5977. URL: https://meetingorganizer.copernicus.org/EGU25/EGU25-5977.html.

Ehret, Uwe et al. (2020). "Adaptive clustering: reducing the computational costs of distributed (hydrological) modelling by exploiting time-variable similarity among model elements". In: *Hydrology and Earth System Sciences* 24.9, pp. 4389–4411. DOI: 10.5194/hess-24-4389-2020. URL: https://hess.copernicus.org/articles/24/4389/2020/.

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

European Space Agency and European Union (2019). *Copernicus DEM - Global and European Digital Elevation Model*. Copernicus Dataspace. Data acquired 2011-2015 via TanDEM-X mission, available in three instances: EEA-10 (10m, Europe), GLO-30 (30m, global), GLO-90 (90m, global). URL: https://dataspace.copernicus.eu/explore-data/data-collections/copernicus-contributing-missions/collections-description/COP-DEM.

Falcon, William and The PyTorch Lightning team (2019). *PyTorch Lightning*. The lightweight PyTorch wrapper for high-performance AI research. DOI: 10.5281/zenodo.3828935. URL: https://www.pytorchlightning.ai.

Feng, Jianwei and Dong Huang (June 2021). "Optimal Gradient Checkpoint Search for Arbitrary Computation Graphs". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11433–11442.

Friston, Karl (Feb. 2010). "The free-energy principle: a unified brain theory?" en. In: *Nature Reviews Neuroscience* 11.2 (2). Publisher: Nature publishing group, pp. 127–138. DOI: 10.1038/nrn2787. URL: http://dx.doi.org/10.1038/nrn2787.

Gal, Yarin and Zoubin Ghahramani (June 2016). "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA, pp. 1050–1059. DOI: 10.1117/12.2551649.6144131789001. arXiv: 1506.02142v6 [stat.ML]. URL: https://proceedings.mlr.press/v48/gal16.html.

Gauch, Martin et al. (2020). "A machine learner's guide to streamflow prediction". In: *AI for Earth Sciences Workshop at NeurIPS*.

Gauch, Martin et al. (Apr. 2021). "Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network". en. In: *Hydrology and Earth System Sciences* 25.4 (4), pp. 2045–2062. DOI: 10.5194/hess-25-2045-2021. URL: http://dx.doi.org/10.5194/hess-25-2045-2021.

Gibson, James Jerome and James J Gibson (1986). *The ecological approach to visual perception*. Vol. 1.

Gigi, Yotam et al. (Jan. 2019). "Towards global remote discharge estimation: Using the few to estimate the many". In: *arXiv preprint arXiv:1901.00786*. arXiv: 1901.00786v1 [cs.LG]. URL: http://arxiv.org/abs/1901.00786v1.

Gillies, Sean et al. (2013). *Rasterio: geospatial raster I/O for Python programmers*. Mapbox. URL: https://github.com/rasterio/rasterio.

Glaser, Clarissa et al. (2020). "Temporal and spatial variable in-stream attenuation of selected pharmaceuticals". In: *Science of The Total Environment* 741, p. 139514. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2020.139514. URL: https://www.sciencedirect.com/science/article/pii/S004896972033031X.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. Vol. 1. 2.

Graves, Alex (2011). "Practical variational inference for neural networks". In: *Advances in neural information processing systems*, pp. 2348–2356.

Global Runoff Data Centre (2024). 56068 Koblenz, Germany. URL: https://grdc.bafg.de/.

Gupta, Hoshin V. et al. (2009). "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling". In: *Journal of Hydrology* 377.1, pp. 80–91. ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2009.08.003. URL: https://www.sciencedirect.com/science/article/pii/S0022169409004843.

Ha, David, Andrew Dai, and V Le Quoc (2016). "HyperNetworks. arXiv preprint, page". In: *arXiv preprint arXiv:1609.09106*.

Hallegatte, Stéphane (2012). "A cost effective solution to reduce disaster losses in developing countries: hydro-meteorological services, early warning, and evacuation". In: *Policy research working paper* 6058.

Harder, Paula et al. (2023). *Physics-Constrained Deep Learning for Downscaling*. Tech. rep. Copernicus Meetings.

He, Y., A. Bárdossy, and E. Zehe (2011). "A review of regionalisation for continuous streamflow simulation". In: *Hydrology and Earth System Sciences* 15.11, pp. 3539–3553. DOI: 10.5194/hess-15-3539-2011. URL: https://hess.copernicus.org/articles/15/3539/2011/.

Hendrycks, Dan and Kevin Gimpel (June 2016). "Gaussian Error Linear Units (GELUs)". In: *arXiv preprint arXiv:1606.08415*. arXiv: 1606.08415v5 [cs.LG]. URL: http://arxiv.org/abs/1606.08415v5.

Herrera, Paulo A, Miguel Angel Marazuela, and Thilo Hofmann (2022). "Parameter estimation and uncertainty analysis in hydrological modeling". In: *Wiley Interdisciplinary Reviews: Water* 9.1, e1569.

Hersbach, H et al. (2018). *ERA5 hourly data on single levels from 1940 to present*.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". en. In: *Neural Computation* 9 (8), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hoedt, Pieter-Jan et al. (Jan. 2021). "MC-LSTM: Mass-Conserving LSTM". In: *Proceedings of Machine Learning Research*. arXiv: 2101.05186v3 [cs.LG]. URL: http://arxiv.org/abs/2101.05186v3.

Höge, M. et al. (2022). "Improving hydrologic models for predictions and process understanding using neural ODEs". In: *Hydrology and Earth System Sciences* 26.19, pp. 5085–5102. DOI: 10.5194/hess-26-5085-2022. URL: https://hess.copernicus.org/articles/26/5085/2022/.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5, pp. 359–366.

Hrachowitz, M. et al. (Aug. 2013). "A decade of Predictions in Ungauged Basins (PUB)—a review". en. In: *Hydrological Sciences Journal* 58.6 (6), pp. 1198–1255. DOI: 10.1080/02626667.2013.803183. URL: http://dx.doi.org/10.1080/02626667.2013.803183.

Hunter, Neil M et al. (2007). "Simple spatially-distributed models for predicting flood inundation: A review". In: *Geomorphology* 90.3-4, pp. 208–225.

Huntington, Thomas G. (2006). "Evidence for intensification of the global water cycle: Review and synthesis". In: *Journal of Hydrology* 319.1, pp. 83–95. ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2005.07.003. URL: https://www.sciencedirect.com/science/article/pii/S0022169405003215.

Huynh, Ngo Nghi Truyen et al. (2024). "Multiscale Learnable Physical Modeling and Data Assimilation Framework: Application to High-Resolution Regionalized Hydrological Simulation of Flash Floods". In: *Authorea Preprints*.

Imhoff, RO et al. (2020). "Scaling point-scale (pedo) transfer functions to seamless large-domain parameter estimates for high-resolution distributed hydrologic modeling: An example for the Rhine River". In: *Water Resources Research* 56.4, e2019WR026807.

Imhoff, Ruben Olaf et al. (2022). "Large-sample evaluation of radar rainfall nowcasting for flood early warning". In: *Water Resources Research* 58.3, e2021WR031591.

Internal Displacement Monitoring Centre (May 2025). *Global Report on Internal Displacement 2025 (GRID 2025)*. Internal Displacement Monitoring Centre. URL: https://www.internal-displacement.org/global-report/grid2025/.

José Areia (Dec. 2023). *Polytechnic University of Leiria: LaTeX Thesis Template*. URL: https://github.com/joseareia/ipleiria-thesis.

Karlbauer, Matthias et al. (Dec. 2019). "A distributed neural network architecture for robust non-linear spatio-temporal prediction". In: *arXiv preprint arXiv:1912.11141*. arXiv: 1912.11141v1 [cs.LG]. URL: http://arxiv.org/abs/1912.11141v1.

Kendall, Carol and Jeffrey J McDonnell (2012). *Isotope tracers in catchment hydrology*.

Klotz, Daniel et al. (2022). "Uncertainty estimation with deep learning for rainfall–runoff modeling". In: *Hydrology and Earth System Sciences* 26.6, pp. 1673–1693.

Kokhlikyan, Narine et al. (2020). "Captum: A unified and generic model interpretability library for pytorch". In: *arXiv preprint arXiv:2009.07896*.

Konapala, Goutam et al. (2020). "Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US". In: *Environmental Research Letters* 15.10, p. 104022.

Kratzert, Frederik et al. (2018). "Rainfall–runoff modelling using long short-term memory (LSTM) networks". In: *Hydrology and Earth System Sciences* 22.11, pp. 6005–6022.

Kratzert, Frederik et al. (2019a). "Toward improved predictions in ungauged basins: Exploiting the power of machine learning". In: *Water Resources Research* 55.12, pp. 11344–11354.

Kratzert, Frederik et al. (Dec. 2019b). "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets". en. In:

Hydrology and Earth System Sciences 23.12 (12), pp. 5089–5110. DOI: 10.5194/hess-23-5089-2019. URL: http://dx.doi.org/10.5194/hess-23-5089-2019.

Kratzert, Frederik et al. (May 2023). "Caravan - A global community dataset for large-sample hydrology". In: *Scientific Data* 10.1, p. 61. DOI: 10.5194/egusphere-egu23-5256. URL: https://doi.org/10.5194/egusphere-egu23-5256.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Lees, Thomas et al. (2021). "Hydrological concept formation inside long short-term memory (LSTM) networks". In: *Hydrology and Earth System Sciences Discussions* 2021, pp. 1–37.

Lehner, Bernhard et al. (2024). "The Global Dam Watch database of river barrier and reservoir information for large-scale applications". In: *Scientific Data* 11.1, p. 1069.

Li, Peifeng, Jin Zhang, and Peter Krebs (2022). "Prediction of flow based on a CNN-LSTM combined deep learning approach". In: *Water* 14.6, p. 993.

Li, Xia et al. (2022). "Hybrid CNN-LSTM models for river flow prediction". In: *Water Supply* 22.5, pp. 4902–4919.

Liang, Xu et al. (1994). "A simple hydrologically based model of land surface water and energy fluxes for general circulation models". In: *Journal of Geophysical Research: Atmospheres* 99.D7, pp. 14415–14428.

Liu, Y. et al. (2018). "Contributions of catchment and in-stream processes to suspended sediment transport in a dominantly groundwater-fed catchment". In: *Hydrology and Earth System Sciences* 22.7, pp. 3903–3921. DOI: 10.5194/hess-22-3903-2018. URL: https://hess.copernicus.org/articles/22/3903/2018/.

Liu, Yuqiong and Hoshin V Gupta (2007). "Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework". In: *Water resources research* 43.7.

Liu, Yuqiong et al. (2012). "Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities". In: *Hydrology and earth system sciences* 16.10, pp. 3863–3887.

Liu, Zhuang et al. (June 2022). "A ConvNet for the 2020s". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA). DOI: 10.1109/cvpr52688.2022.01167. arXiv: 2201.03545v2 [cs.CV]. URL: http://dx.doi.org/10.1109/cvpr52688.2022.01167.

Longyang, Qianqiu et al. (2024). "Explainable Spatially Distributed Hydrologic Modeling of a Snow Dominated Mountainous Karst Watershed Using Attention". In: *Authorea Preprints*.

Lu, Dan et al. (2021). "Streamflow simulation in data-scarce basins using Bayesian and physics-informed machine learning models". In: *Journal of Hydrometeorology* 22.6, pp. 1421–1438.

Lynas, Mark, Benjamin Z Houlton, and Simon Perry (2021). "Greater than 99% consensus on human caused climate change in the peer-reviewed scientific literature". In: *Environmental Research Letters* 16.11, p. 114005.

Marçais, Jean and Jean-Raynald de Dreuzy (2017). "Prospective interest of deep learning for hydrological inference". In: *Groundwater* 55.5, pp. 688–692.

Markstrom, Steven L et al. (2015). *PRMS-IV, the precipitation-runoff modeling system, version 4*. Tech. rep. US Geological Survey.

Mazzetti, C et al. (2023). *River discharge and related historical data from the European Flood Awareness System, v5.0, European Commission, Joint Research Centre (JRC)*. URL: https://cds.climate.copernicus.eu/cdsapp#!/dataset/efas-historical.

McMillan, H.K., D.J. Booker, and C. Cattoën (2016). "Validation of a national hydrological model". In: *Journal of Hydrology* 541, pp. 800–815. ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2016.07.043. URL: https://www.sciencedirect.com/science/article/pii/S0022169416304735.

Milly, P Christopher D et al. (2002). "Increasing risk of great floods in a changing climate". In: *Nature* 415.6871, pp. 514–517.

Moishin, Mohammed et al. (2021). "Designing deep-based learning flood forecast model with ConvLSTM hybrid algorithm". In: *IEEE Access* 9, pp. 50982–50993.

Molnar, Christoph (2020). *Interpretable machine learning*.

Montzka, Carsten et al. (2012). "Multivariate and multiscale data assimilation in terrestrial systems: A review". In: *Sensors* 12.12, pp. 16291–16333.

Moradkhani, Hamid et al. (2005). "Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter". In: *Water resources research* 41.5.

Moshe, Zach et al. (Mar. 2020). *HydroNets: Leveraging River Network Structure and Deep Neural Networks for Hydrologic Modeling*. DOI: 10.5194/egusphere-egu2020-4135. arXiv: 2007.00595v1 [cs.LG]. URL: http://dx.doi.org/10.5194/egusphere-egu2020-4135.

Muhebwa, Aggrey et al. (2024). "Improving discharge predictions in ungauged basins: Harnessing the power of disaggregated data modeling and machine learning". In: *Water Resources Research* 60.9, e2024WR037122.

Muñoz-Carpena, Rafael et al. (2023). "Convergence of mechanistic modeling and artificial intelligence in hydrologic science and engineering". In: *PLOS Water* 2.8, e0000059.

Nash, J Eamonn and Jonh V Sutcliffe (1970). "River flow forecasting through conceptual models part I—A discussion of principles". In: *Journal of hydrology* 10.3, pp. 282–290.

Neal, Radford M (2012). *Bayesian learning for neural networks*. Vol. 118. DOI: 10.2139/ssrn.5363231. URL: https://doi.org/10.2139/ssrn.5363231.

Nearing, Grey et al. (Feb. 2020). "What Role Does Hydrological Science Play in the Age of Machine Learning?" In: *Water Resources Research* 57.3, e2020WR028091. DOI: 10.31223/osf.io/3sx6g. URL: https://doi.org/10.31223/osf.io/3sx6g.

Nearing, Grey et al. (July 2023). "AI Increases Global Access to Reliable Flood Forecasts". In: *arXiv preprint arXiv:2307.16104*. arXiv: 2307.16104v3 [cs.LG]. URL: http://arxiv.org/abs/2307.16104v3.

Nimai, Silang et al. (2023). "Enhancing runoff simulation using BTOP-LSTM hybrid model in the Shinano River basin". In: *Water* 15.21, p. 3758.

NOAA National Centers for Environmental Information (Jan. 2025). *Monthly Global Climate Report for Annual 2024*. Online. Retrieved July 25, 2025. DOI: https://www.ncei.noaa.ncdc:C00672. URL: https://www.ncei.noaa.gov/access/monitoring/monthly-report/global/202413.

Núñez, Jorge, Catalina B. Cortés, and Marjorie A. Yáñez (2023). "Explainable Artificial Intelligence in Hydrology: Interpreting Black-Box Snowmelt-Driven Streamflow Predictions in an Arid Andean Basin of North-Central Chile". In: *Water* 15.19. ISSN: 2073-4441. DOI: 10.3390/w15193369. URL: https://www.mdpi.com/2073-4441/15/19/3369.

Oddo, Perry C et al. (2024). "Deep Convolutional LSTM for improved flash flood prediction". In: *Frontiers in Water* 6, p. 1346104.

OECD (June 2025). *Global Drought Outlook*. Organisation for Economic Co-operation and Development. URL: https://www.oecd.org/en/publications/global-drought-outlook_d492583a-en/full-report/executive-summary_4133c25e.html.

Oki, Taikan and Shinjiro Kanae (2006). "Global hydrological cycles and world water resources". In: *science* 313.5790, pp. 1068–1072.

Okiria, Emmanuel et al. (2022). "A comparative evaluation of lumped and semi-distributed conceptual hydrological models: does model complexity enhance hydrograph prediction?" In: *Hydrology* 9.5, p. 89.

Otte, Sebastian, Matthias Karlbauer, and Martin V. Butz (Oct. 2020). "Active Tuning". In: *arXiv:2010.03958 [cs]*. arXiv: 2010.03958. URL: http://arxiv.org/abs/2010.03958 (visited on 10/12/2020).

Otte, Sebastian et al. (2017). "Inferring Adaptive Goal-Directed Behavior Within Recurrent Neural Networks". en. In: *Artificial Neural Networks and Machine Learning –ICANN 2017*. Ed. by Alessandra Lintas et al. Vol. 10613. Series Title: Lecture Notes in Com-

puter Science. Cham, pp. 227–235. DOI: 10.1007/978-3-319-68600-4_27. URL: http://link.springer.com/10.1007/978-3-319-68600-4_27 (visited on 09/16/2020).

Palmer, Margaret A et al. (2008). "Climate change and the world's river basins: anticipating management options". In: *Frontiers in Ecology and the Environment* 6.2, pp. 81–89.

Pappenberger, Florian et al. (2015). "The monetary benefit of early flood warnings in Europe". In: *Environmental Science & Policy* 51, pp. 278–291. ISSN: 1462-9011. DOI: https://doi.org/10.1016/j.envsci.2015.04.016. URL: https://www.sciencedirect.com/science/article/pii/S1462901115000891.

Paprotny, Dominik et al. (2025). "Attribution of flood impacts shows strong benefits of adaptation in Europe since 1950". In: *Science Advances* 11.33, eadt7068.

Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio (Nov. 2012). "On the difficulty of training Recurrent Neural Networks". In: arXiv: 1211.5063v2 [cs.LG]. URL: http://arxiv.org/abs/1211.5063v2.

Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". en. In: *Advances in neural information processing systems*, p. 12.

Perera, Duminda et al. (2019). "Flood early warning systems: a review of benefits, challenges and prospects". In: *UNU-INWEH, Hamilton*.

Perera, Duminda et al. (2020). "Identifying societal challenges in flood early warning systems". In: *International Journal of Disaster Risk Reduction* 51, p. 101794. ISSN: 2212-4209. DOI: https://doi.org/10.1016/j.ijdrr.2020.101794. URL: https://www.sciencedirect.com/science/article/pii/S2212420920312966.

Perrin, Charles, Claude Michel, and Vazken Andréassian (2003). "Improvement of a parsimonious model for streamflow simulation". In: *Journal of hydrology* 279.1-4, pp. 275–289.

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

Pokharel, Sudan and Tirthankar Roy (2024). "A parsimonious setup for streamflow forecasting using CNN-LSTM". In: *Journal of Hydroinformatics*, jh2024114.

RADOLAN/RADVOR (2016). URL: https://opendata.dwd.de/climate_environment/CDC/grids_germany/hourly/radolan/.

Rakovec, O et al. (2012). "State updating of a distributed hydrological model with Ensemble Kalman Filtering: effects of updating frequency and observation network density on forecast accuracy". In: *Hydrology and Earth System Sciences* 16.9, pp. 3435–3449.

Rasp, Stephan and Sebastian Lerch (2018). "Neural networks for postprocessing ensemble weather forecasts". In: *Monthly Weather Review* 146.11, pp. 3885–3900.

Refshaard, J. C. and B. Storm (1995). "MIKE SHE." English. In: Colorado, pp. 809–846. ISBN: 9780918334916.

Renard, Kenneth G et al. (1994). *The revised universal soil loss equation*, pp. 105–126.

Rentschler, Jun, Melda Salhab, and Bramka Arga Jafino (2022). "Flood exposure and poverty in 188 countries". In: *Nature communications* 13.1, p. 3527.

Roberts, Daniel A. (Mar. 2021). "Why is AI hard and Physics simple?" In: *arXiv preprint arXiv:2104.00008*. arXiv: 2104.00008v1 [hep-th]. URL: http://arxiv.org/abs/2104.00008v1.

Samaniego, Luis, Rohini Kumar, and Sabine Attinger (2010). "Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale". In: *Water Resources Research* 46.5.

Scholz, Fedor et al. (Feb. 2022). "Inference of Affordances and Active Motor Control in Simulated Agents". In: *Frontiers in Neurorobotics*. DOI: 10.3389/fnbot.2022.881673. arXiv: 2202.11532v3 [cs.AI]. URL: http://arxiv.org/abs/2202.11532v3.

Scholz, Fedor et al. (Oct. 2024a). *Fully differentiable, fully distributed River Discharge Prediction: code*. DOI: 10.5281/zenodo.13992583. URL: https://zenodo.org/records/15484058.

Scholz, Fedor et al. (Oct. 2024b). *Fully differentiable, fully distributed River Discharge Prediction: data sets*. DOI: 10.5281/zenodo.13970575. URL: https://zenodo.org/records/15482198.

Scholz, Fedor et al. (Apr. 2024c). "Introducing a fully differentiable, fully distributed Rainfall-Runoff Model". In: *EGU General Assembly 2024*. Vienna, Austria, EGU24–5298. DOI: 10.5194/egusphere-egu24-5298. URL: https://doi.org/10.5194/egusphere-egu24-5298.

Scholz, Fedor et al. (2024d). "Quick and Accurate Affordance Learning". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 46. URL: https://escholarship.org/uc/item/21b6p6tt.

Scholz, Fedor et al. (Mar. 2025a). "Fully differentiable, fully distributed Rainfall-Runoff Modeling". In: *EGUsphere* 2025, pp. 1–37. DOI: 10.5194/egusphere-2024-4119. URL: https://doi.org/10.5194/egusphere-2024-4119.

Scholz, Fedor et al. (Apr. 2025b). "Inference of catchment areas from modeled discharge dynamics". In: *EGU General Assembly 2025*. Vienna, Austria, EGU25–19057. DOI: 10.5194/egusphere-egu25-19057. URL: https://doi.org/10.5194/egusphere-egu25-19057.

Sezen, Cenk and Mojca Šraj (2023). "Hourly rainfall-runoff modelling by combining the conceptual model with machine learning models in mostly karst Ljubljanica River

catchment in Slovenia". In: *Stochastic Environmental Research and Risk Assessment*, pp. 1–25.

Shen, Chaopeng (2018). "A transdisciplinary review of deep learning research and its relevance for water resources scientists". In: *Water Resources Research* 54.11, pp. 8558–8593.

Shen, Chaopeng et al. (July 2023). "Differentiable modelling to unify machine learning and physical models for geosciences". en. In: *Nature Reviews Earth & Environment* 4 (8), pp. 552–567. DOI: 10.1038/s43017-023-00450-9. URL: http://dx.doi.org/10.1038/s43017-023-00450-9.

Shi, Xingjian et al. (June 2015). "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting". In: *Advances in neural information processing systems*. arXiv: 1506.04214v2 [cs.CV]. URL: http://arxiv.org/abs/1506.04214v2.

Shokati, Hadi et al. (2024). "Random Forest-Based Soil Moisture Estimation Using Sentinel-2, Landsat-8/9, and UAV-Based Hyperspectral Data". In: *Remote Sensing* 16.11. ISSN: 2072-4292. DOI: 10.3390/rs16111962. URL: https://www.mdpi.com/2072-4292/16/11/1962.

Shokati, Hadi et al. (2025a). "Comparing UAV-Based Hyperspectral and Satellite-Based Multispectral Data for Soil Moisture Estimation Using Machine Learning". In: *Water* 17.11. ISSN: 2073-4441. DOI: 10.3390/w17111715. URL: https://www.mdpi.com/2073-4441/17/11/1715.

Shokati, Hadi et al. (2025b). "Erosion-SAM: Semantic segmentation of soil erosion by water". In: *CATENA* 254, p. 108954. ISSN: 0341-8162. DOI: https://doi.org/10.1016/j.catena.2025.108954. URL: https://www.sciencedirect.com/science/article/pii/S0341816225002565.

Sit, Muhammed, Bekir Demiray, and Ibrahim Demir (July 2021). "Short-term Hourly Streamflow Prediction with Graph Convolutional GRU Networks". In: *arXiv preprint arXiv:2107.07039*. arXiv: 2107.07039v1 [cs.LG]. URL: http://arxiv.org/abs/2107.07039v1.

Sit, Muhammed et al. (June 2020). *A Comprehensive Review of Deep Learning Applications in Hydrology and Water Resources*. DOI: 10.31223/osf.io/xs36g. arXiv: 2007.12269v1 [physics.geo-ph]. URL: http://dx.doi.org/10.31223/osf.io/xs36g.

Sivapalan, Murugesu (2003). "Prediction in ungauged basins: a grand challenge for theoretical hydrology." In.

Smith, Michael B. et al. (2003). "Hydrologic Model calibration in the National Weather Service". In: vol. 6, pp. 133–152. DOI: 10.1029/ws006p0133. URL: https://doi.org/10.1029/ws006p0133.

Souza Machado, Anderson Abel de et al. (2018). "Microplastics as an emerging threat to terrestrial ecosystems". In: *Global Change Biology* 24.4, pp. 1405–1416. DOI: https://doi.org/10.1111/gcb.14020. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.14020. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.14020.

Sun, A. Y. et al. (2022). "A graph neural network (GNN) approach to basin-scale river network learning: the role of physics-based connectivity and data fusion". In: *Hydrology and Earth System Sciences* 26.19, pp. 5163–5184. DOI: 10.5194/hess-26-5163-2022. URL: https://hess.copernicus.org/articles/26/5163/2022/.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR, pp. 3319–3328.

Takeuchi, Kuniyoshi et al. (2008). "A BTOP model to extend TOPMODEL for distributed hydrological simulation of large basins". In: *Hydrological Processes: An International Journal* 22.17, pp. 3236–3251.

Thielen, J. et al. (2009). "The European Flood Alert System –Part 1: Concept and development". In: *Hydrology and Earth System Sciences* 13.2, pp. 125–140. DOI: 10.5194/hess-13-125-2009. URL: https://hess.copernicus.org/articles/13/125/2009/.

Thielen-del Pozo, J et al. (2015). "The benefit of continental flood early warning systems to reduce the impact of flood disasters". In: *EUR Sci. Tech. Res. Rep.*

Thurber, Daniel et al. (2024). "Dissolving the mystery of subsurface controls on snowmelt–discharge dynamics in karst mountain watersheds using hydrologic timeseries". In: *Hydrological Processes* 38.5, e15170.

Tian, Ye et al. (2018). "Integration of a parsimonious hydrological model with recurrent neural networks for improved streamflow forecasting". In: *Water* 10.11, p. 1655.

Tran, Quoc Quan, J. De Niel, and P. Willems (2018). "Spatially Distributed Conceptual Hydrological Model Building: A Generic Top-Down Approach Starting From Lumped Models". In: *Water Resources Research* 54.10, pp. 8064–8085. DOI: https://doi.org/10.1029/2018WR023566. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018WR023566. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR023566.

Traub, Manuel et al. (Apr. 2024a). "High-Efficiency Rainfall Data Compression Using Binarized Convolutional Autoencoder". In: *EGU General Assembly 2024*. Vienna, Austria, EGU24–11768. DOI: 10.5194/egusphere-egu24-11768. URL: https://doi.org/10.5194/egusphere-egu24-11768.

Traub, Manuel et al. (2024b). "Loci-segmented: improving scene segmentation learning". In: *International Conference on Artificial Neural Networks*. Springer, pp. 45–61.

Trenberth, Kevin E (2011). "Changes in precipitation with climate change". In: *Climate research* 47.1-2, pp. 123–138.

Tyson, Conor et al. (2023). "Effects of meteorological forcing uncertainty on high-resolution snow modeling and streamflow prediction in a mountainous karst watershed". In: *Journal of Hydrology* 619, p. 129304.

Ueda, Futo et al. (2024). "A Transfer Learning Approach Based on Radar Rainfall for River Water-Level Prediction". In: *Water* 16.4, p. 607.

Ufrecht, Wolfgang (2002). "Ein Hydrogeologisches Modell für den Karst-und Mineralwasseraquifer Muschelkalk im Großraum Stuttgart". In: *Hydrogeologische Modelle–ein Leitfaden mit Fallbeispielen, Schriftenreihe der Deutschen Geologischen Gesellschaft* 24.10.

United States Geological Survey (Oct. 2022). *Water Cycle Diagram*. Public domain. Updated diagram released October 13, 2022. U.S. Geological Survey, Water Science School. URL: https://www.usgs.gov/special-topics/water-science-school/science/water-cycle-diagrams#overview (visited on 07/17/2025).

Valeriano, Oliver Cristian Saavedra et al. (2010). "Optimal dam operation during flood season using a distributed hydrological model and a heuristic algorithm". In: *Journal of Hydrologic Engineering* 15.7, pp. 580–586.

Van Vliet, Michelle TH et al. (2013). "Global river discharge and water temperature under climate change". In: *Global Environmental Change* 23.2, pp. 450–464.

Vansteenkiste, Thomas et al. (2014). "Intercomparison of five lumped and distributed models for catchment runoff and extreme flow simulation". In: *Journal of Hydrology* 511, pp. 335–349. ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2014.01.050. URL: https://www.sciencedirect.com/science/article/pii/S0022169414000729.

Vargas Godoy, Mijael Rodrigo and Yannis Markonis (May 2023). *Water Cycle Changes in Reanalyses*. DOI: 10.5194/egusphere-egu23-259. URL: https://doi.org/10.5194/egusphere-egu23-259.

Vivoni, Enrique R et al. (2011). "Real-world hydrologic assessment of a fully-distributed hydrological model in a parallel computing environment". In: *Journal of Hydrology* 409.1-2, pp. 483–496.

Vogel, Richard M and Neil M Fennessey (1994). "Flow-duration curves. I: New interpretation and confidence intervals". In: *Journal of Water Resources Planning and Management* 120.4, pp. 485–504.

Vreugdenhil, Cornelis Boudewijn (1994). *Numerical methods for shallow-water flow*. Vol. 13.

Wan, Shu et al. (2024). "Spatio-temporal Causal Learning for Streamflow Forecasting". In: *2024 IEEE International Conference on Big Data (BigData)*. IEEE, pp. 6161–6170.

Wang, Chao et al. (2024). "Distributed Hydrological Modeling With Physics-Encoded Deep Learning: A General Framework and Its Application in the Amazon". In: *Water Resources Research* 60.4. e2023WR036170 2023WR036170, e2023WR036170. DOI: https://doi.org/10.1029/2023WR036170. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023WR036170. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023WR036170.

Wang, Yuan-Heng and Hoshin V. Gupta (Apr. 2024). "A Mass-Conserving-Perceptron for Machine-Learning-Based Modeling of Geoscientific Systems". en. In: *Water Resources Research* 60 (4). DOI: 10.1029/2023wr036461. arXiv: 2310.08644v1 [cs.LG]. URL: https://doi.org/10.1029/2023wr036461.

Wi, Sungwook and Scott Steinschneider (2023). "On the need for physical constraints in deep learning rainfall-runoff projections under climate change". In: *EGUsphere* 2023, pp. 1–46.

World Meteorological Organization (2023). *Flood Forecasting and Early Warning*. Tech. rep. 19. World Meteorological Organization. URL: https://library.wmo.int/records/item/37081-flood-forecasting-and-early-warning.

Wright, Less (2019). *Ranger - a synergistic optimizer*. URL: https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer.

Wunsch, A. et al. (2022). "Karst spring discharge modeling based on deep learning using spatially distributed input data". In: *Hydrology and Earth System Sciences* 26.9, pp. 2405–2430. DOI: 10.5194/hess-26-2405-2022. URL: https://hess.copernicus.org/articles/26/2405/2022/.

Xiang, Zhongrun and Ibrahim Demir (Sept. 2020). "Distributed long-term hourly streamflow predictions using deep learning –A case study for State of Iowa". en. In: *Environmental Modelling & Software* 131, p. 104761. DOI: 10.1016/j.envsoft.2020.104761. URL: http://dx.doi.org/10.1016/j.envsoft.2020.104761.

Xiang, Zhongrun and Ibrahim Demir (2022). "Fully distributed rainfall-runoff modeling using spatial-temporal graph neural network". In.

Xu, Tianfang et al. (2022). "Hybrid physically based and deep learning modeling of a snow dominated, mountainous, karst watershed". In: *Water Resources Research* 58.3, e2021WR030993.

Yadan, Omry (2019). *Hydra - A framework for elegantly configuring complex applications*. Github. URL: https://github.com/facebookresearch/hydra.

Yang, Chen, Laura E Condon, and Reed M Maxwell (2025). "Unravelling groundwater–stream connections over the continental United States". In: *Nature Water* 3.1, pp. 70–79.

Yilmaz, Koray K, Hoshin V Gupta, and Thorsten Wagener (2008). "A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model". In: *Water Resources Research* 44.9.

Yu, Q. et al. (2024). "Enhancing long short-term memory (LSTM)-based streamflow prediction with a spatially distributed approach". In: *Hydrology and Earth System Sciences* 28.9, pp. 2107–2122. DOI: 10.5194/hess-28-2107-2024. URL: https://hess.copernicus.org/articles/28/2107/2024/.

Zarfl, Christiane and Frances E. Dunn (2022). "The delicate balance of river sediments". In: *Science* 376.6600, pp. 1385–1386. DOI: 10.1126/science.abq6986. eprint: https://www.science.org/doi/pdf/10.1126/science.abq6986. URL: https://www.science.org/doi/abs/10.1126/science.abq6986.

Zhong, Liangjin, Huimin Lei, and Jingjing Yang (2024a). "Development of a distributed physics-informed deep learning hydrological model for data-scarce regions". In: *Water Resources Research* 60.6, e2023WR036333.

Zhong, Liangjin et al. (2024b). "Advancing streamflow prediction in data-scarce regions through vegetation-constrained distributed hybrid ecohydrological models". In: *Journal of Hydrology* 645, p. 132165. ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2024.132165. URL: https://www.sciencedirect.com/science/article/pii/S0022169424015610.

Zhu, Shuang et al. (2023). "Spatiotemporal deep learning rainfall-runoff forecasting combined with remote sensing precipitation products in large scale basins". In: *Journal of Hydrology* 616, p. 128727.

Acknowledgements

This dissertation would not have been possible without the support, guidance, and encouragement of many individuals and institutions. I am deeply grateful to all who contributed to this journey. First and foremost, I extend my heartfelt gratitude to my primary supervisor, Prof. Dr. Martin V. Butz, who taught me how to be a scientist and provided an intellectually stimulating yet peaceful working environment. His unwavering belief in my work, even before I believed in it myself, was instrumental in bringing this research to realization. His mentorship has shaped not only this dissertation but also my approach to scientific inquiry. I am equally grateful to my co-supervisor Prof. Dr. Christiane Zarfl for her supervision, thoughtful discussions, and constructive feedback throughout this research. Her expertise in hydrology and environmental sciences provided essential guidance in bridging the gap between machine learning and hydrology. My sincere appreciation goes to Prof. Dr. Thomas Scholten for his insightful discussions and valuable feedback that enriched and motivated this work. I also thank Prof. Dr. Georg Martius and Prof. Dr. Bedartha Goswami for serving on my Thesis Advisory Committee, providing crucial feedback, and engaging in stimulating discussions that helped refine the research direction. I am grateful for my colleagues in the Cognitive Modeling group, who fostered an environment of intellectual curiosity and genuine collegiality. Special thanks go to Prof. Dr. Sebastian Otte, Dr. Matthias Karlbauer, Manuel Traub, Jannik Thümmel, and Jan Prosi for countless discussions in our reading group, entertaining lunch breaks, and their willingness to share their expertise. My gratitude extends to Manuela Di Paolo, who helped navigate the often complex world of academic bureaucracy, making administrative processes significantly more manageable. I would like to acknowledge the generous collaboration and support from colleagues at the Karlsruhe Institute of Technology (KIT). Dr. Ralf Loritz and PD Dr. Uwe Ehret welcomed me into their research groups, showed genuine appreciation for my work, and provided invaluable assistance with future endeavors. Alexander Dolich contributed through meaningful discussions that broadened my understanding of hydrological modeling. This research was made possible through funding from the Cluster of Excellence "Machine Learning: New Perspectives for Science" at the University of Tübingen. I am also grateful to the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for providing excellent workshops and fostering an international research community. Finally, I owe the deepest gratitude to my family and friends, who believed in me during times when I struggled to believe in

myself. Their unwavering support, patience, and encouragement provided the emotional foundation that made this journey possible. To all who contributed to this work in ways both large and small—thank you.