Improved Streamflow Simulation through Ensemble and Stochastic Conceptual Data-driven Approaches

by

Kyung Whan Hah

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Civil Engineering (Water)

Waterloo, Ontario, Canada, 2022

Author’s Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis consists of a manuscript written for publication. This research was conducted at the University of Waterloo by Kyung Whan (David) Hah under the supervision of Dr. John Quilty and with assistance from Dr. Anna Sikorska-Senoner. I developed, documented the methodology, implemented the method within the software, and conducted the numerical evaluations. The co-authors provided guidance during each step of the research and provided feedback on draft manuscripts.

Abstract

To better understand the complexities of water movement on earth, hydrologists have developed process-based hydrological models (HMs) and data-driven models (DDMs), both of which have been applied to a host of water resources applications (e.g., flood forecasting, reservoir operations, drought monitoring, hydraulic design). HMs attempt to simplify hydrological processes of interest (e.g., snowmelt, sub-surface flow), while DDMs estimate statistical relationships between explanatory/input and response/target variables using historical data. Traditionally, HMs and DDMs have been developed independently, however, there has been growing interest in using DDMs to improve HM simulations. Among various approaches for combining process-based theory with DDMs, the conceptual data-driven approach (CDDA) was recently proposed, where DDMs are used to correct the residuals (errors) stemming from ensemble HMs. The CDDA was shown to substantially reduce the simulation uncertainty. Since the CDDA only accounts for the HM parameter uncertainty, a subsequent study introduced the stochastic CDDA (SCDDA) to account for various sources of uncertainty (i.e., input data, input variable selection, parameters, and model output). However, the (original) SCDDA used HMs as input to the DDMs within a stochastic framework, thus, estimating the uncertainty of the DDMs, not the CDDA. Here, a new SCDDA is introduced where the CDDA uncertainty is estimated instead of the DDM uncertainty (as in the original SCDDA) by taking advantage of the multiple parameter sets generated by the CDDA through a stochastic framework. Hence, the new SCDDA serves as the second stage in post-processing HMs, where the stochastic framework can be used to improve the CDDA simulations. The new SCDDA is tested in a daily streamflow simulation case study using three Swiss catchments where it is benchmarked against the CDDA as well as ensemble and stochastic HMs. In total, nine HM-DDM combinations (variants) are explored within the CDDA and SCDDA based on three popular HMs and three state-of-the-art DDMs. The ensemble and stochastic HMs are based on the same three HMs used in the CDDA and SCDDA. A total of 34 years of daily streamflow, precipitation, maximum, minimum, and mean air temperatures, and potential evapotranspiration time series were partitioned into warm-up, calibration/training, validation, and test sets for model development and assessment. Several deterministic (mean absolute error, root mean squared error, Nash Sutcliffe Efficiency, Kling Gupta Efficiency (KGE), and percent bias) and probabilistic (mean continuous ranked probability score (CRPS), alpha index (α_R), and average width) performance metrics, as well as graphical tools (e.g., time series plots, raincloud plots, coverage probability plots (CPP)), were used to assess the simulations and compare the various models. The CDDA improved the CRPS of the ensemble HM by 18-69%, and the new SCDDA further improved the CRPS of the CDDA by up to 15%. However, it was found that the SCDDA could not improve the reliability of any CDDA variants that had an α_R above 0.85. Since the computational requirements of the CDDA and SCDDA can be significant, the effect of ensemble size on model performance was analyzed and revealed that approximately 100 (ensemble) members, for both ensemble and stochastic models, could be used without sacrificing performance. Finally, to test whether the CDDA and SCDDA can account for important processes missing within an HM, both approaches adopt an HM with and without a snow routine and are tested in a snow-driven catchment. It is found that both cases (with and without snow) had negligible difference in performance suggesting that the CDDA and SCDDA may account for missing processes in HMs.

Acknowledgements

This endeavour would not have been possible without Prof. John Quilty for his support and mentorship over the last two years. I am extremely grateful to have a supervisor who is dedicated to his students. His devotion to learning inspired me to study beyond my comfort zone and became one of the most enjoyable moments during my graduate studies.

I would also like to thank the Hydrology group and my peers in the Collaborative Water Program for the many discussions over the last two years. Special thanks to Sina and John for satisfying my itch whenever I had a question. I am grateful to have peers I can reach out to whenever I have a problem.

I want to extend my gratitude to Dr. Sikorska-Senoner, who provided insights for the thesis, and financial support from the Queen Elizabeth II Graduate Scholarship in Science and Technology.

Finally, I would like to express my deepest appreciation to my family. Stella, your love and support over the last eight years gave me the strength to extend beyond my limits. You’ve taught me how to appreciate the little things in life and helped me overcome my many weaknesses in the last two years. Mom, Dad, and John, thank you for supporting me in any decision I make. I am incredibly blessed to have such an amazing family.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ANN	Artificial neural network
AW	Average width
CDDA	Conceptual-data-driven approach
CPP	Coverage probability plot
CRPS	Continuous ranked probability score
DDM	Data-driven model
HM	Process-based hydrological model
KGE	Kling Gupta Efficiency
LSTM	Long-short term memory networks
MAE	Mean absolute error
NSE	Nash Sutcliffe Efficiency
PBIAS	Percent bias
Q22	Quilty et al. (2022)
RF	Random forests
RMSE	Root mean squared error
RNN	Recurrent neural network
SCDDA	Stochastic conceptual-data-driven approach
SSQ21	Sikorska-Senoner & Quilty (2021)
XGB	Extreme gradient boosting

Chapter 1

Introduction

Hydrologists and water resources practitioners are commonly tasked with estimating the uncertainty of hydrological variables for various applications such as flood forecasting and water management, which have benefited from process-based (physical or conceptual) hydrological models (HMs) (Anand et al., 2018; Jain et al., 2018). Although statistical methods have been used in the hydrology domain for decades (Singh, 2018), the accelerated advancements in technology and computer science, especially data-driven models (DDMs), have given a new perspective to the hydrological modelling paradigm by producing accurate models that have similar or better predictive capability than process-based models (Beven, 2020; Nearing et al., 2021). There has been a large body of research for both model types as one promotes understanding of the hydrological system (process-based) and the other focuses on predictive accuracy (data-driven), both of which are used for many hydrological applications. Hence, it is essential to differentiate the two model types as they can dictate the modelling process and guide the model objectives.

To contrast process-based and data-driven modelling approaches, HMs attempt to simplify and mimic natural processes using mathematical expressions relevant to the hydrological processes (i.e., explicit descriptions of the perceived catchment response to rainfall-runoff using boundary conditions, initial conditions, mass balance, energy balance, etc.) (Beven, 2012), while DDMs estimate statistical relationships between explanatory/input (e.g., rainfall) and response/target (e.g., streamflow) variables using historical measurements enabling simulations of the target variable (Bishop, 2006). Although research on these two approaches has advanced independently in the hydrological sciences, there is an increasing interest in combining process-based theory and data-driven approaches to increase our understanding of the natural system and improve predictive performance beyond what is achievable by the individual models (Adombi et al., 2021; Karpatne et al., 2017).

In the hydrological sciences, research in combining theoretical knowledge (Adombi et al., 2021) with data-driven approaches has gained interest as initially proposed in theory-guided data science (Karpatne et al., 2017). Since then, there have been several applications of the combined approach, although many of these approaches do not follow a standard naming convention. For example, coupling process-based with DDMs has been referred to as physics-informed DDMs (Liang et al., 2019), hybrid models (Kurian et al., 2020), or simply post-processing (Frame et al., 2021; Nearing et al., 2020). As another example, the term ‘physics-informed neural networks’ has been used to represent an approach where partial differential equations (e.g., the governing equation for subsurface flow) are incorporated within artificial neural networks (Raissi et al., 2019; Shen and Lawson, 2021). Furthermore, several recent studies have used process-based model outputs as input to DDMs, demonstrating that such an approach can be used to improve the predictive performance of the standalone HM (Frame et al., 2021; Ghaith et al., 2019; Konapala et al., 2020; Kumanlioglu and Fistikoglu, 2019; Lu et al., 2021; Nearing et al., 2020; Quilty et al., 2022). Different naming conventions may introduce linguistic uncertainty (Montanari, 2011). Therefore, an initial attempt is made here to classify the combined approaches to distinguish the different methods in the current literature (see Section 2.2). However, this classification is expected to evolve, given the rising popularity of the combined (HM-DDM) approaches.

Among various combined approaches, one method gaining popularity is ‘correcting’ process-based simulations by using DDMs to estimate the model residuals (Cho & Kim, 2022; Li et al., 2021a, 2021b; Papacharalampous et al., 2020a; Sharma et al., 2021; Shen et al., 2022, 2021; Sikorska-Senoner & Quilty, 2021). Incorporating model errors in hydrological simulations has traditionally been used for uncertainty estimation (see Section 2.3.2), for example, stochastic resampling of model errors for assessing predictive uncertainty in streamflow simulations (Montanari & Koutsoyiannis, 2012; Sikorska et al., 2014). As such, Sikorska-Senoner & Quilty (2021) (hereinafter referred to as SSQ21) developed the conceptual-data-driven approach (CDDA) to include an ensemble of HMs, each paired with a DDM that simulates the HM residuals. The CDDA only characterizes uncertainty in the HM parameters, while the DDMs simulate the expectation of the residuals associated with each HM parameter set. A key outcome of SSQ21 showed that all of the nonlinear DDMs considered in the study improved the streamflow simulations generated by the lumped conceptual HM. Although the CDDA adopted conceptual HMs, any deterministic model can be used in the CDDA.

As uncertainty in streamflow simulations plays a significant role in water-related decisions, Quilty et al. (2022) (hereinafter referred to as Q22) modified the CDDA by including other sources of uncertainty neglected in the CDDA (input data, input variable selection (IVS), DDM parameters, and model output). Referred to as the stochastic conceptual data-driven approach (SCDDA), this new framework improved streamflow simulations in all study catchments compared to the CDDA. However, unlike the CDDA, which uses DDMs to simulate the residuals of the ensemble HMs directly, the DDMs within the SCDDA directly simulate streamflow using the ensemble HM simulations as DDM inputs; in other words, the

SCDDA proposed in Q22 did not follow the same approach for using DDMs as in the CDDA proposed in SSQ21. In contrast to Q22, this research proposes a new SCDDA, an analogue of the CDDA in SSQ21, where DDMs are used to simulate HM residuals in a stochastic framework that, similar to Q22, can account for the uncertainty in input data, IVS, parameters, and model output. The new SCDDA takes advantage of the multiple parameter sets generated by the routine used to optimize the DDM's hyperparameters in the CDDA.

In SSQ21 and Q22, only one HM structure, based on the Hydrologiska Byråns Vattenbalansavdelning (HBV) model, specifically, HBV-light (Seibert & Vis, 2012), was used, making it challenging to identify whether the performance of the CDDA and SCDDA was specific to the HM's structure. Therefore, along with HBV-light, two additional HMs, the Technische Universität Wien model (TUWmodel, Parajka et al., 2007) and modèle du Génie Rural à 4 paramètres Journalier (GR4J, Perrin et al., 2003) are used in this work to explore the performance of the CDDA and SCDDA with respect to different HM structures. For the DDMs, eXtreme Gradient Boosting (XGB, Chen & Guestrin, 2016) and Random Forests (RF, Breiman, 2001) were chosen based on a recommendation in the work of SSQ21, which found the two DDMs to be most suitable to simulate HM residuals (out of the eight DDMs that were considered). Furthermore, the Long Short-Term Memory Networks (LSTM, Hochreiter & Schmidhuber, 1997) were included due to their prominent role in hydrological applications involving deep learning (Shen, 2018). LSTM is an extension of Recurrent Neural Networks with the capacity to learn time dependencies over long timescales, which is important when simulating streamflow, as persistence is commonly manifested in hydrological processes (Hurst, 1951; Iliopoulou et al., 2018; Koutsoyiannis, 2021; Pagano & Garen, 2005). The competitive predictive performance of LSTM has been well recognized in hydrology and water resources and is becoming the model of choice for deep learning applications in these domains (Feng et al., 2021, 2020; Kratzert et al., 2018; Liu et al., 2022; Shen, 2018).

1.1 Objectives and Thesis Organization

The main goal of this research is to develop and test the new SCDDA in three catchments (the same locations in SSQ21 and Q22). It is hypothesized that the new SCDDA can improve the ensemble and stochastic HM as well as the CDDA. Thus, the research objectives presented in this thesis are to:

1. Apply the CDDA and the new SCDDA for all (nine) combinations of HMs and DDMs for daily streamflow simulation in three Swiss catchments.
2. Compare the performance of the CDDA and SCDDA against one another, as well as the ensemble and stochastic HMs, considering the different combinations of HMs and DDMs.
3. Propose and explore the use of a diagnostic tool to predict if the SCDDA can improve upon the reliability obtained by the CDDA.

The third objective is not only useful for measuring improvements by converting the CDDA to its stochastic counterpart, but it also serves a practical purpose. Suppose the diagnostic tool predicts that the SCDDA will have lower reliability than the CDDA. In this case, the SCDDA may be neglected, saving implementation time and computational costs. The same three catchments from SSQ21 and Q22 were used to develop the different models and address the research objectives.

The main novelty of this research lies in the use of multiple parameter sets explored by the optimization routine in the CDDA. The CDDA requires a (computationally demanding) search for optimal DDM hyper-parameters (settings external to the model that affect the calibration of parameters, model structure, etc.). In the CDDA, each HM in the ensemble is associated with a single set of DDM hyper-parameters and a single set of DDM hyper-parameters is used to estimate DDM parameters (which are responsible for mapping the DDM inputs to the HM residuals). However, by retaining all parameter sets explored during optimization, the uncertainty of the CDDA (HM and DDM) parameters can be used within the stochastic framework (leading to the new SCDDA). In addition, adopting nine different HMs and DDMs within the CDDA and SCDDA for simulating streamflow in multiple catchments allows for the exploration of a diagnostic tool, the coverage probability plot (CPP), to determine if the predictive performance of the CDDA can be improved using the SCDDA. The CPP is also referred to as the predictive probability-probability plot and has proven to be a valuable tool for converting deterministic simulations to stochastic ones (Koutsoyiannis & Montanari, 2022). Finally, the new SCDDA can also be viewed as the second stage in a post-processing framework, where after the CDDA is used to correct the HM outputs (first stage), the stochastic framework is used to refine the CDDA and assess its uncertainty through stochastic resampling (transforming the CDDA into the new SCDDA).

This thesis is organized as follows. Chapter 2 presents background information on HMs, combining process-based theory with data-driven approaches and the uncertainty of hydrological simulations. Chapter 3 describes the methods related to the CDDA and (new and original) SCDDA as well as the adopted HMs and DDMs. Chapter 4 provides the experimental setup for the CDDA and SCDDA. Chapter 5 presents the main results and includes a discussion on their significance. Finally, Chapter 6 provides concluding remarks, discusses opportunities for future research, and ends with several recommendations for improving the proposed SCDDA.

Chapter 2

Background

This chapter provides an overview of hydrological modelling, including a brief history of its development, classification of hydrological models, combining process-based theory with data-driven models, and the uncertainty of hydrological predictions along with its estimation.

2.1 Hydrological Models

Sometimes referred to as the “science of water” (National Research Council, 1991), hydrological science (or hydrology) emerged to better understand the complexities of water movement on earth - since it serves as a necessary resource for all terrestrial life - and to avoid life-threatening hazards (National Research Council, 2012). Hydrology is frequently associated with other disciplines such as meteorology, climatology, geomorphology, hydrogeology, and ecology, playing a significant role in earth system science (Blöschl, 2005). Typical applications of hydrology to water resources problems include, but are not limited to: water resources planning and management (Brown et al., 2009), hydraulic designs (Chow et al., 1988), and flood and drought forecasting (Dawson & Wilby, 2001; Konapala and Mishra, 2020). Although many of these applications require measuring various hydrological variables (e.g., precipitation, evaporation, streamflow) in both space and time, the limitation of measurement techniques invoked HMs to estimate and extrapolate hydrological variables where measurements were/are not available (Beven, 2012). For rainfall-runoff modelling, the Rational Method (Mulvany, 1851), the first of its kind, was proposed over 150 years ago and is still used today to solve practical engineering problems (e.g., estimating peak discharges for drainage systems) (Singh, 2018). Since then, various HMs have been developed by incorporating water movement, energy transfer, relevant hydrological processes, and statistical relationships (Devi et al., 2015) through the advancement of computation in the 1960s (Singh, 2018). Since HMs can be developed using various approaches, researchers have attempted to classify these model types, which can be applied to different modelling objectives.

2.1.1 Classification of Hydrological Models

While there is no universal agreement on the classification of HMs, it is possible to group HMs based on their spatial discretization, prediction mode, model structure, and stochasticity. Spatial discretization determines if the model can incorporate the spatial variability of the hydrological process. By discretizing a catchment into multiple components (e.g., grids, subcatchments, and hydrologic response units) and solving for the variables of interest at each unit, distributed and semi-distributed models incorporate spatial variability of the hydrological process. The other approach is to lump the catchment into a single unit with the hydrologic variables representing the average over the area (Beven, 2012). Since incorporating spatial variability can result in high computational costs, spatial discretization should be selected according to the modelling objectives. For example, distributed models are likely to be considered if the model aims to estimate a particular state variable (e.g., groundwater level) for various locations within a catchment. However, lumped models provide an acceptable solution if the goal is to simulate the river discharge at the catchment outlet.

The prediction mode determines if the model's objective pertains to simulation or forecasting. Adopting the definitions provided by Beven and Young (2013): simulation refers to the "quantitative reproduction of the behaviour of a system, given some defined inputs but without reference to any observed outputs," while forecasting refers to the "quantitative reproduction of the behaviour of the system ahead of time, but given observations of the inputs, state variables (where applicable), and outputs up to the present time." In summary, simulation and forecasting differ by the variable of interest with respect to time (e.g., the future) and the input data required to retrieve the model outputs. As "prediction" is sometimes used synonymously with simulation and forecasting in the hydrology literature, to avoid ambiguity, this work uses prediction only when simulation and forecasting can be used in the same context.

The following classification is associated with the choice of model structure, categorized into process-based hydrological models, HMs, and data-driven models, DDMs. HMs attempt to simplify the hydrological process using the perceived response of the catchment from rainfall (i.e., through mass balance, energy balance, initial conditions, boundary conditions, etc.). Furthermore, HMs can be divided into physically-based and conceptual models. Although some researchers use physically-based and process-based synonymously, both physically-based and conceptual models are referred to as process-based in this work since they focus on developing mathematical descriptions of the perceived hydrological process, whether through known scientific principles of energy and water fluxes (physically-based) or by macroscale relationships, which may not have a direct physical interpretation (conceptual) (Beven, 2012). Although physically-based models are often regarded as more theoretically correct, conceptual models are often the choice for operational surface-runoff applications due to their computational speed (compared to physics-based models), scale-dependent parameters, and flexibility in choosing relevant hydrological processes (Bergström & Graham, 1998). The other class of model structures are DDMs (sometimes referred to as statistical models and empirical models), which estimate statistical relationships between explanatory/input (e.g., rainfall) and response/target (e.g., streamflow) variables using historical measurements (Bishop, 2006). DDMs have gained widespread popularity in recent years due to technological development (e.g., graphical processing units) and computer science (specifically, in the areas of machine learning (ML) and deep learning (DL)), improving modelling capabilities in various scientific fields (Ching et al., 2018; Jha et al., 2018; Ravuri et al., 2021; Reichstein et al., 2019; Senior et al., 2020; Suh et al., 2021). ML can be further classified into supervised and unsupervised learning. Supervised learning uses labelled datasets (i.e., target variables), allowing the model to learn the relationship between the target and explanatory variables. In contrast, unsupervised learning disregards labelled datasets and attempts to gain insights (e.g., patterns, groupings) about the data (Hastie et al., 2009). Since supervised learning can be considered a data-driven counterpart to HMs (i.e., focused on predicting a hydrological variable), supervised learning has been the most popular ML application in hydrology (Sit et al., 2020). However, it should be noted that DDMs have been regarded as an inductive approach compared to the deductive formulation of HMs (Beven, 2012). Although induction is often viewed as weaker than deduction (Koutsoyiannis, 2021), DDMs have recently been shown to have comparable or more accurate predictive capability than HMs (Nearing et al., 2021). The cause for this is not well understood, as the departure of model simulations from observed values can be caused by various limitations and uncertainties (see Sections 2.1.3 and 2.3).

The final classification relates to the stochasticity of the model. While deterministic models only produce a single outcome for a given set of input variables and model parameters, stochastic models consider uncertainty when estimating a variable of interest (Montanari & Koutsoyiannis, 2012). Despite deterministic models being the most popular method (as many process descriptions are developed using a deterministic approach), the single-valued approach results in inaccuracies compared to the observed values, requiring uncertainty estimates (Montanari &

Koutsoyiannis, 2012). Uncertainty assessment is essential for decision-makers and the public (see Section 2.3), as they may foretell the level of risk involved in an upcoming extreme event (Koutsoyiannis & Montanari, 2022). It should be noted that some HMs adopt a statistical model structure but use the expected value (i.e., the first moment or the mean) as the output, thus, becoming a deterministic model. The following sub-section reviews the hydrological modelling process.

2.1.2 Hydrological Modelling Process

The hydrological modelling process serves as a guide for model objectives and enhancing model results. An outline of the hydrological modelling process is provided in Figure 1 (Beven, 2012).

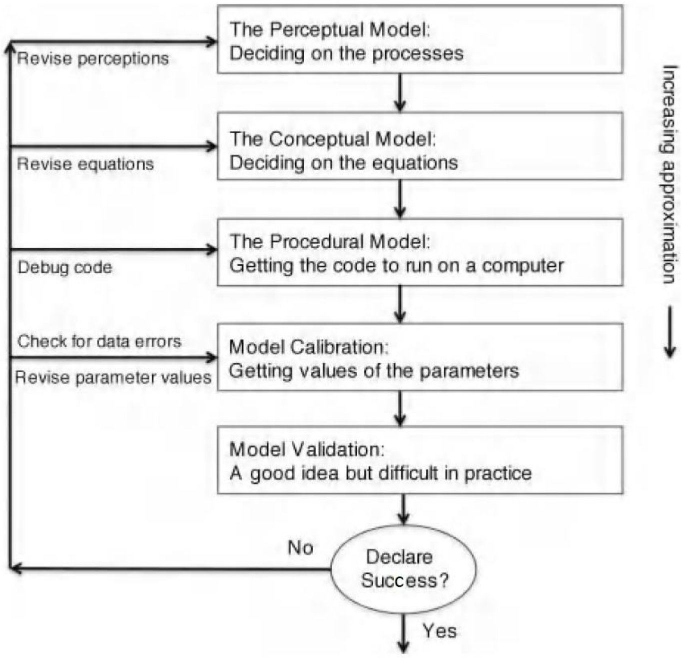

Figure 1. Outline of the hydrological modelling process (Beven, 2012)

Before selecting a model, it is essential to identify the modelling objectives as the purpose of the model guides the modelling process. The objectives can differ according to variable(s) of interest (e.g., streamflow, soil moisture, groundwater levels), the level of detail required from the model (e.g., meeting specific performance criteria), and project limitations (e.g., available data, funding, timelines). After specifying the objectives, the perceptual model of the rainfall-runoff process (for example) is determined (top of Figure 1). The perceptual model is defined as the perceptions of the catchment response to rainfall, which is limited to the modeller's knowledge of the hydrological process (Beven, 2012). The perceptual model does not consist of mathematical equations; instead, it is the awareness of the complexities of the processes occurring in the system that can be gained from prior experiences, such as from the field, in the lab, and in the literature. Hence, it is expected that experienced hydrologists will conceive a suitable perceptual model by identifying the critical hydrological processes for the given catchment. The conceptual model follows the next stage in the hydrological modelling process (not to be confused with conceptual HMs), where mathematical equations (including model structure) are determined using the perceptual model. Here, assumptions are made to simplify the complexities of the hydrological process to help describe them mathematically, for example, the initial and boundary conditions of the HMs.

Following the conceptual model, the next stage of the modelling process is the procedural model, where the conceptual model is translated into computer code for calibration and simulation. For calibration, the modeller chooses an optimization algorithm and the objective function(s) to determine suitable model parameters or structures that reflect the system behaviour. Furthermore, the dataset is traditionally split using a variant of split-sample testing (Klemeš, 1986). One or multiple segment(s) of data is used for calibration, while the remaining is used for verifying the model results. For HMs, in addition to a warm-up period to stabilize state variables (e.g., depth to the water table) (Kim et al., 2018), it is common to split the remaining dataset into two segments, one of which is used for calibration of model parameters or structures and the other for verifying the model. However, for DDMs (and ML algorithms, in particular), it is common to split the dataset into three segments: training, validation, and test sets (Xu & Goodacre, 2018). The training, validation, and test sets are used within DDMs to calibrate model parameters, tune the model hyper-parameters (i.e., settings external to the model that affect the calibration of parameters, model structure, etc.), and test the performance of the model out-of-sample, respectively. Since HMs are often assumed to not include hyper-parameters, the HM verification set is typically synonymous with the DDM's test set.

Suppose the model simulations in the calibration (HM) or training and validation (DDM) sets do not provide acceptable performance. In this case, the modelling process is reiterated by revising the perceptions, equations, code, and parameter values until achieving model objectives. However, it should be noted that model verification using the verification set (HM) or the test set (DDM) should only be used at the last stage of the modelling process. The model (HM or DDM) should not be updated based on the verification (HM) or test (DDM) set's performance as this introduces bias into the model simulations, providing the illusion of overly optimistic performance. Thus, the calibration (HM) or training and validation (DDM) sets should only be used to reiterate the modelling process.

2.1.3 Limitations of Hydrological and Data-driven Models

Although this sub-section is not intended to serve as an exhaustive list of problems related to HMs and DDMs, it is essential to review some common issues considered in the literature to understand the general limitations of both approaches.

First, HMs are based on perceptions of the hydrological system, making simplifications to describe them mathematically. However, the perceptions may be incorrect, and the simplifications may have a non-negligible effect on the simulation (Beven, 2001). The next issue with HMs is the problem of scale. The scale problem is the inability to consider the heterogeneity and non-linearity of the hydrological processes of interest for all scales due to the lack of measurement techniques (Beven, 2001). For example, soil parameters (e.g., hydraulic conductivity) may be estimated from point-scale geological surveys, which cannot be extrapolated due to their heterogeneous nature (Beven, 2009). As this problem is common in distributed physically-based modelling, conceptual models may be an alternative method through scale-dependent model structures aligned with available data (i.e., through hydrologic response units), often calibrated with historical records of observed streamflow. Although conceptual models can alleviate (to some extent) the problem of scale, the problem with developing scale-relevant theory persists. For example, inaccuracies of conceptual models can be caused by the inability to specify macroscale watershed behaviours due to heterogeneity (Nearing et al., 2021).

Furthermore, there are inherent problems with calibration for HMs related to searching for an optimal model (parameter set and/or model structure). The search space of parameters and model structures may lack a clear optimum (non-identifiability), and a single optimal model may be improper due to multiple optima or different goodness-of-fit criteria (non-uniqueness), both of which are common problems with model optimization (Beven, 2009). Although it is reasonable to assume an (acceptable) optimal model can be found within a search space, multiple combinations of structures and parameter set usually provide adequate performance. Hence defined as equifinality (Beven, 1993), this concept is used to reject the idea of a single optimal model; otherwise, deciding on a single set of model parameters or structure may be an arbitrary choice (Beven, 2012).

Another problem is related to the number of process descriptions in the model structure, where adding complexity to the model to increase predictive performance introduces additional tunable parameters. Suppose the perceived hydrological descriptions are correct for a given catchment but deteriorate model performance as more descriptions are added. In that case, this is often regarded as an overparameterized model (Beven, 2006).

Although the above issues have been attributed to HMs, DDMs also have significant limitations, some of which may have caused hydrologists to disregard DDMs (See et al., 2007). In the literature, DDMs have been more heavily criticized compared to their process-based counterpart mainly due to their high reliance on empirical estimates without any process knowledge. In addition, many hydrologists are skeptical of DDMs since they do not add any scientific knowledge or improved understanding of hydrology as the approach attempts to learn through inputs and outputs (See et al., 2007). As such, DDMs are often viewed as “black-box” models as it is difficult to relate the internal structure of the model to hydrological processes (Beven, 2012). However, there are numerous examples where DDMs have been useful when attempting to overcome the limitations of HMs. For example, although lumped conceptual HMs attempt to derive watershed-scale theories from data, DDMs have been tested on various catchments and shown to have, on average, better performance than HMs (Nearing et al., 2021). Hence, better watershed-scale theories could have been derived from data, but hydrologists were unable to find such relationships (Nearing et al., 2021). For predictions in ungauged basins, where observed streamflow is not available, one may assume that it is only possible to use theoretical knowledge (i.e., HMs) since DDMs require a training target (observed streamflow). However, training DDMs on neighbouring gauged catchments and applying the model to the ungauged catchments (transfer learning) shows promising performance in regions with no streamflow data (Kratzert et al., 2019). Similar to HMs, issues with calibration also exist for DDMs, including non-identifiability and non-uniqueness. However, the issue with over-parameterization is managed effectively in DDMs through regularization, which helps avoid overfitting (i.e., strong training performance but low testing performance) (Ng, 2004). For example, for an artificial neural network (ANN), the model may have thousands of tunable parameters (e.g., weights and biases) yet converge to a suitable model, unlike over-parameterized HMs. Mitigating over-parameterization is achievable since many DDMs incorporate regularization methods that assist in the bias-variance trade-off (Luxburg and Schölkopf, 2011). Such techniques are not widely used in HMs; therefore, poor-performing HMs can be caused by lack of regularization which may be considered as an over-parameterized model (Nearing et al., 2021). Despite these advantages, the “black-box” nature of DDMs remains a significant issue for some hydrologists.

The inability to understand hydrological processes of interest through DDMs may be a significant problem for modellers since they may be interested in several hydrological variables (perhaps, in an interdisciplinary approach) such as water quantity, quality, and particle tracking.

Unless these variables are explicitly modelled, this task seems unachievable purely through the data-driven approach. Furthermore, some hydrologists have compared DDMs to “curve-fitting exercises” if it is impossible to extract knowledge about the process (Zaherpour et al., 2019). To combat this issue, a subfield of data-driven modelling called interpretable ML, which seeks to make black-box models more explainable, has been growing in popularity (Du et al., 2019; Molnar, 2020). In a water resources case study, Lees et al. (2022) trained a Long-Short Term Memory network (LSTM) to simulate streamflow using meteorological time series and static catchment attributes as inputs, then linearly regressed the LSTM’s state vectors against satellite-derived soil moisture and snow depth. Their results showed that the LSTM learned from data alone how to represent soil moisture and snow processes internally when being trained to simulate streamflow. Lees et al. (2022) not only suggest that the LSTM is able to learn important hydrological processes internally (moving beyond curve-fitting exercises) but that it is also possible to utilize the internal states (of the LSTM) to simulate latent hydrological variables (e.g., soil moisture). Although further exploration of this interesting line of research is not considered herein, it is essential to recognize that DDMs may have the potential to learn hidden hydrological processes that hydrologists have not yet been capable of discovering (or measure). Nonetheless, if the model objectives prioritize prediction quality, hydrologists can take advantage of the predictive capability of DDMs (Beven, 2020). Hence, as this research is primarily concerned with improving the predictive performance of HMs, it is explored herein to what extent DDMs can improve upon HM simulations.

2.2 Combining Process-based Theory with Data-driven Approaches

In hydrology, research on combining theoretical knowledge (Adombi et al., 2021) with DDMs has been growing in popularity since the introduction of theory-guided data science (Karpatne et al., 2017). Various approaches exist (with different naming conventions) for combining theoretical knowledge with DDMs. However, strictly defining and classifying these "combined approaches" is outside the scope of this research. The combined approaches can be broadly classified into three categories to reduce linguistic uncertainty (see Section 2.3.1): informed, constrained, and error. As many of these approaches have been applied recently, more methods will likely be introduced in the near future, requiring the current classification to be updated. The following sub-sections review the literature on the three combined approaches.

2.2.1 Informed Approach

Among the three methods, the informed approach is the most popular likely due to its ease of application. The informed approach is a general method of infusing DDMs with relevant hydrological processes (Adombi et al., 2021) or learning physical relationships from DDMs for use in HMs. Since most DDMs are not limited to specific input variables (unlike HMs), a straightforward approach is to utilize HMs to produce relevant features to be used as input to a DDM. For example, actual evapotranspiration may be extracted from an HM calibrated for streamflow simulation and used in a DDM for streamflow forecasting. The ML literature refers to extracting relevant features from data as feature engineering (Heaton, 2016). Thus, the informed approach can be considered a feature engineering exercise for the DDM, using information from one or more HMs. Studies using the informed approach have consistently found improved predictive performance compared to a standalone HM (Ghaith et al., 2019;

Konapala et al., 2020; Kumanlioglu and Fistikoglu, 2019; Lu et al., 2021; Quilty et al., 2022). However, there are examples where the informed approach did not outperform a standalone DDM (Frame et al., 2021; Nearing et al., 2020). Recently, Liang et al. (2019) created a database using numerical simulations of different weather patterns, surface topography, vegetation, soil conditions, and contaminants. Features were extracted from this database and used to simulate surface water quantity and quality using DDMs, which showed that the combined approach had higher predictive performance than the standalone HM. In another study, Tongal and Booij (2018) demonstrated how baseflow separation coupled with DDMs could be used to improve streamflow simulation. A physics-guided architecture based on LSTM was developed by Daw et al. (2020), where temporal features of the input variables (e.g., depth of water and precipitation) were used to generate a latent variable (i.e., water density). The model was guided by the physical relationship between water depth and density when modelling the lake temperature. The physics-guided architecture outperformed a benchmark LSTM and LSTM based on the constrained approach (see Section 2.2.2).

Despite the popularity of the informed approach where hydrological processes or physical relationships are used as input to, or to guide the learning of, DDMs, it is possible to train DDMs to generate model parameters for HMs by using large-scale (multi-catchment) datasets. For example, Tsai et al. (2021) and Feng et al. (2022) used DDMs to simulate parameters for HMs trained across the conterminous United States. Their results demonstrate that the informed approach has an accuracy similar to a standalone DDM but with increased insight into the physical processes gained through the HM’s state variables. In the next sub-section, the constrained approach is explored and shown to be a promising combined approach for hydrological modelling.

2.2.2 Constrained Approach

The constrained approach alters the DDM’s objective/loss function with physical equations (Raissi et al., 2019). Although the constrained approach has sometimes been presented using different terminology, e.g., ‘physics-informed neural networks’ (Tartakovsky et al., 2020), this approach integrates theory into the loss function of the DDMs (Adombi et al., 2021). For example, it is possible to constrain the DDM with the governing equation for subsurface flow by supplementing the standard loss function (e.g., mean squared error) with the residuals of the governing equation as well as initial and boundary conditions to achieve physically realistic results (Adombi et al., 2021). Applications of the constrained approach are limited, likely due to the difficulty of its implementation (relative to the other combined approaches). However, the constrained approach has been used for subsurface flow problems governed by Darcy’s law (saturated conditions) (Tartakovsky et al., 2020), approximating solutions to the Richards equation and estimating parameters of the van Genuchten model (unsaturated conditions) (Depina et al., 2021) solving the coupled advection-dispersion and Darcy flow equations considering hydraulic conductivity that is space-dependent (He & Tartakovsky, 2021), water depth simulation (Mahesh et al., 2022), and data assimilation (He et al., 2020).

The constrained approach may benefit groundwater modellers mainly due to its computational speed and accuracy compared to physically-based distributed models. However, a significant limitation is that the model must be re-trained for new initial and boundary conditions (Shen & Lawson, 2021). Wang et al. (2020) extended the loss function constraints by including engineering control and expert knowledge for subsurface flow. The results demonstrated that the new approach achieved higher predictive capability, reliability, and generalizability than the standalone DDM. Another example of the constrained approach is provided by Jia et al. (2021), where flow and water temperature are simulated in river networks. The authors use a DDM informed by an HM and constrained by physical relationships between different river segments. The model resulted in superior performance compared to the standalone HM and DDM. Xie et al. (2021) used synthetic samples of extreme events to inform the DDM while constraining the loss function using heavy rainstorm events, rainless events, and monotonicity. The results showed that the combined approach increased the predictive performance of the DDM, including simulation of flood peaks, mitigating negative streamflow, and maintaining monotonicity. Another constrained approach directly introduces conservation laws (e.g., mass balance) into the DDM. Hoedt et al. (2021) constrained the internal structure of an LSTM to allow mass conservation within the DDM. Referred to as the Mass-Conserving LSTM (MC-LSTM), this approach enables users to apply the model for a wide range of problems, not just related to hydrology (such as traffic forecasting, where the number of cars in and out of the system is conserved). Although their results showed that the standalone LSTM outperformed MC-LSTM, the latter achieved more accurate high-volume flows, which the authors assumed was likely due to MC-LSTM ensuring mass conservation. In the next sub-section, the error approach is described, which is the combined approach adopted in this work.

2.2.3 Error Approach

The error approach focuses on the refinement of HM or DDM outputs and can be applied in two ways: post-processing the DDM outputs to conform with governing equations (e.g., Chen et al., 2021) or post-processing the HM outputs by correcting their residuals (errors) using DDMs. The former type of error approach is very similar to the constrained approach (Adombi et al., 2021); therefore, details of this approach are not discussed. However, the second type of error approach is desirable due to its simplicity and the amenability of estimating prediction uncertainty by converting deterministic predictions to probabilistic ones. In the error approach, an HM is simulated for a given catchment, and its residuals/errors (i.e., the difference between the observed and simulated results) are used as the target variable for a DDM. The simulation of residuals from the DDM is summed together with the HM simulations to provide an updated (or corrected) simulation (Sikorska-Senoner & Quilty, 2021). Although simple to implement, this approach has seldom been explored. Booker and Woods (2014) corrected HMs using flow duration curves and Random Forests (RF), showing significant performance gains compared to the standalone HM. Cho and Kim (2022) used LSTM to correct an HM applied to streamflow simulation, finding that the model bias significantly improved using the error approach compared to the standalone LSTM. Sharma et al. (2021) used LSTM to correct numerical weather forecasts, which improved bias for medium-range timescales. Sikorska-Senoner and Quilty (2021) used ensemble HMs where each ensemble member was corrected with a DDM. Their results indicate that any of their studied non-linear DDM improved the HM.

In contrast to the other combined approaches, the error approach has also been used to determine the prediction uncertainty by using the model error to convert deterministic predictions to probabilistic ones (see Section 2.3). Probabilistic predictions are useful for assessing the uncertainty of water resource systems and play a key role in water resources planning, management, and operations. Thus, probabilistic predictions generated by the error approach can supplement decision-making tasks in hydrology and water resources (e.g., flood forecasting, drought mitigation, reservoir operations). The new SCDDA developed and explored in this research is an error approach that can convert ensembles of deterministic HMs into stochastic coupled-data-driven models that produce probabilistic simulations. Aside from the new SCDDA presented here, to the best of the author's knowledge, only two other error approaches that produce probabilistic simulations exist in the hydrology and water resources literature. However, these two studies (described briefly below) only focused on a single HM (while this research considers three HMs).

In Papacharalampous et al. (2020a), an error approach was introduced where HM simulations based on a single model structure using different parameter sets (referred to as sister predictions) were used to generate a set of model errors (attached to each parameter set), which were simulated using quantile regression models at various quantiles. Probabilistic simulations were obtained by combining the HM simulations with the model error simulations (from the quantile regression models). Another error approach resulting in probabilistic simulations was proposed by Li et al. (2021b), where the error distribution of an optimized LSTM model (used to correct HM simulations) was estimated using a Markov Chain Monte Carlo algorithm. However, in both studies (Papacharalampous et al., 2020a; Li et al., 2021b), the authors do not evaluate the extent to which the probabilistic methods improve upon the deterministic HMs. In this research, the ensemble models (ensemble HM and CDDA) are extensively compared against their stochastic counterparts (stochastic HM and new SCDDA) to demonstrate the benefits of adopting the stochastic framework. The next sub-section provides an overview of the uncertainty of hydrological predictions, given its central importance to the new SCDDA and, more generally, the planning, management, and operation of water resources systems.

2.3 Uncertainty of Hydrological Predictions

Uncertainty estimation in hydrological simulation is one of the most critical subfields in hydrology, as it quantifies the reliability of the model reflecting real-life conditions (Montanari, 2011). Although most HMs are deterministic (see Section 2.1.1), uncertainty estimation is not a norm in practice or research, despite its importance in decision-making (Pappenberger & Beven, 2006). For example, hydraulic engineers use empirical techniques, such as safety factors and freeboard, to account for the uncertainty of the designers' knowledge about extreme rainfall/streamflow events (Montanari, 2011). Perhaps, uncertainty estimation is not a norm because it is challenging to implement or may be influenced by subjectivity (Pappenberger & Beven, 2006). However, through the improvement of technology, management, and risk assessment (dependent on uncertainty assessment), the number of victims of hydroclimatic disasters has significantly reduced since the 20th century, and there is potential for further improvement (Koutsoyiannis, 2020).

While it is common to encounter the term 'uncertainty,' its exact definition is also uncertain, as some have attributed the term to inexactness, imprecision, indeterminacy, vagueness, etc. (Beven, 2009). However, practical definitions of uncertainty are related to the model objectives or the focus of relevant questions, such as indeterminacy of hydrological simulations (Montanari, 2007). To provide context to the term 'uncertainty' adopted in this research, the rest of this section is dedicated to classifying (different types of) uncertainties and uncertainty assessment.

2.3.1 Classification of Uncertainties

Prediction uncertainty (sometimes described as global uncertainty) refers to the uncertainty related to the model output and actual/observed value of the variable of interest (Montanari, 2011). Researchers have classified prediction uncertainty into distinct types to assess the overall uncertainty and attempt to reduce the individual components of uncertainty. Uncertainty in hydrological predictions is generally grouped into aleatory and epistemic uncertainties. Aleatory uncertainty is interpreted as the inherent randomness from nature or natural variability, while epistemic uncertainty refers to non-random factors, mainly the uncertainty of our knowledge of the system (Beven, 2009). Since aleatory and epistemic uncertainties are often inseparable, they are typically evaluated in an integrated manner, lumping all uncertainties into a single source (Montanari, 2011). The other approach is to classify the individual uncertainties into separate, explicit sources (input, structure, parameter, etc.) and lump all remaining uncertainties (e.g., initial conditions) within the model error, which can be done without differentiating between aleatoric and epistemic uncertainty (Montanari & Koutsoyiannis, 2012). A description of uncertainty sources commonly considered in the hydrology and water resources literature is given below.

Input uncertainty refers to the uncertainty of data used as input to a model, which is important in hydrological forecasting applications since hydro-meteorological inputs (e.g., precipitation) are typically represented as ensembles, translating into the uncertainty of the forecasted variable (Han & Coulibaly, 2019). Structural uncertainty is defined as the inability of the model to reproduce the system dynamics given ideal inputs (i.e., inputs without uncertainty; for example, measurement errors). Since the model’s structure depends on (perhaps, biased) decisions of the modelling process, the structural uncertainty is typically classified as epistemic. In addition, extensive studies on hydrological modelling are focused on improving our understanding of the natural process. Therefore, many water resource practitioners attempt to reduce structural uncertainty (Renard et al., 2010). Another uncertainty source related to the model structure is parameter uncertainty. As many models require parameters to control the model behaviour, parameter uncertainty refers to the inability to estimate the true value of the parameters given incomplete and/or inconsistent data. Parameter uncertainty is influenced by the model structure, calibration method, and the consistency of the data (Montanari & Koutsoyiannis, 2012). Given the input data, model structure, and parameter(s), a model is simulated and compared with the observed data. The model error is the discrepancy between the simulated and observed values. The model error is especially useful since it includes all sources of uncertainty not explicitly accounted for in the model (uncertainty in initial conditions, boundary conditions, state variables, etc.). As a result, the model error can be used to estimate the model output uncertainty and assess the prediction uncertainty without directly accounting for all individual uncertainties.

Other uncertainties (which are rarely discussed) include linguistic and operation uncertainties (Montanari, 2011). As the name implies, linguistic uncertainty refers to the lack of clarity in terminology, as exemplified by the classification of the combined approaches in Section 2.2. Since coherent terminology and clarity in communication is essential for ease of understanding and mitigating errors, linguistic uncertainty plays a substantial role in all fields of science. Lastly, operation uncertainty is related to applying the model in real-life scenarios. Despite most HMs being calibrated with observed streamflow data, observations are prone to errors. Streamflow measurements are commonly based on stage-discharge relationships where discharge can significantly vary with a given stage measurement (Herschy, 2008) but are often delivered to the end-users in a deterministic way (Domeneghetti et al., 2012). Due to errors in measurement techniques, measured streamflow will often disobey the closure of mass balance, leading to uncertainty in the HMs. As the models in this work are applied in a research setting, errors due to measurement techniques are omitted, but interested readers are referred to Montanari (2011) for more information.

2.3.2 Uncertainty Assessment

A model's reliability and, therefore, confidence in the model's simulations is quantified through uncertainty assessment. Consequently, uncertainty assessment should be carried out carefully; otherwise, the credibility of the model and modeller may be diminished. It should be noted that uncertainty assessment, estimation, characterization, and quantification are used interchangeably in the literature on probabilistic methods (Montanari, 2011). Since uncertainty assessment in the literature typically uses probabilistic methods, non-probabilistic methods are not discussed here (although Beven (2009) and Montanari (2011) discuss these methods). Probabilistic methods use the probability theory, where uncertainty can be quantified by relying on one or more probability distributions. The popularity of probabilistic approaches for uncertainty assessment likely originates from the vast literature on probability, statistics, and stochastics, which can be used to alleviate the limitations of deterministic models (Montanari & Koutsoyiannis, 2012). One of the most commonly adopted uncertainty estimation methods uses a Bayesian approach, where the likelihood (the evidence, given model predictions) is used to update the prior probability of different models under consideration. The Bayesian approach is presumably the foundation of the Generalized Likelihood Uncertainty Estimation (GLUE, Beven & Binley, 1992) method, arguably the most popular uncertainty assessment method among hydrological modellers.

Despite its popularity, the GLUE method remains controversial, primarily due to the assumptions needed to compute the informal likelihood function (Montanari, 2005).

Limitations of the GLUE method led to the development of a blueprint for converting a deterministic model to a stochastic one, centred on using the model error distribution to estimate the (deterministic) model output uncertainty (Montanari & Koutsoyiannis, 2012). As discussed in Section 2.2.3, one method of combining HMs with DDMs is through the error approach. Using the error approach, SSQ21 proposed the CDDA, where each member of the ensemble HM (based on multiple parameter sets using a single structure) is paired with a DDM (one per HM ensemble member) to correct the HM simulations. A fundamental limitation of the CDDA is that the method only considered parameter uncertainty in the HM when generating the ensemble streamflow simulations. Therefore, a subsequent study by Q22 developed the stochastic CDDA (SCDDA) to include other sources of uncertainty neglected in the CDDA (input data, input variable selection (IVS), parameters, and model output). Since Q22 used HMs as input to the DDMs (informed approach, see Section 2.2.1), this approach inherently differs from the error approach adopted by the CDDA. Thus, a new SCDDA based on the error approach is introduced in this work, where uncertainty is estimated using the stochastic framework described in Quilty et al. (2019) (motivated by the blueprint introduced in Montanari & Koutsoyiannis, 2012) (see Chapter 3).

Before introducing the methods adopted in this research, it is essential to review the assumptions required for probabilistic approaches, which may determine the model limitations. First, the stochastic process (e.g., runoff) considered in this research is assumed ergodic since its statistical properties are deduced from an extended sample of the process (Montanari, 2011). In detail, as time tends to infinity, the sample statistical descriptions are assumed to be equivalent to the true statistical properties. Next, the stochastic process is assumed stationary, meaning the statistical properties are invariant to a shift in time origin (Koutsoyiannis, 2021). The assumption of stationarity adopted herein may seem impractical in light of environmental change. However, the stochastic framework introduced in Chapter 3 considers an approach for estimating the probability distribution of the model error that inherently accounts for heteroscedasticity (of the model error) and is amenable to real-time updates that may be useful for capturing ongoing environmental change. Although the methods considered herein require additional assumptions, they are described in Chapter 3.

Chapter 3

Methods

This chapter provides an overview of the main methods, including the ensemble-based conceptual-data-driven approach (CDDA), its stochastic version (SCDDA) (including a comparison between the original SCDDA from Quilty et al. (2022) and the new SCDDA, proposed in this research) as well as the different hydrological models, HMs, and data-driven models (DDMs).

3.1 Ensemble-based Conceptual-data-driven Approach (CDDA)

The CDDA utilizes an ensemble of HMs using a single model structure and multiple parameter sets to generate streamflow simulations, where a DDM is used to correct the residuals of each HM. The CDDA is given by the following equation (Quilty et al., 2022):

$$\begin{aligned} &Y_i\left(P_{t,\dots,t-D}, T_{t,\dots,t-D}, Q_{t-1,\dots,t-D} \middle| \Theta_{CDDA_i}\right) \\ &= y_i\left(P_t, T_t \middle| \Theta_{HM_i}\right) + \ r_i(P_{t,\dots,t-D}, T_{t,\dots,t-D}, Q_{t-1,\dots,t-D} | \Theta_{DDM_i}) \end{aligned} \qquad (1)$$

where Y_i, y_i, and r_i are the CDDA simulations, HM streamflow simulations, and DDM residual simulations for ensemble member i, respectively, $P_{t,\dots,t-D}, T_{t,\dots,t-D}$ and $Q_{t-1,\dots,t-D}$ are observed precipitation and air temperature at time lags $t, \dots, t-D$ and streamflow at time lags $t-1, \dots, t-D$ (where D is the maximum time lag), respectively, Θ_{CDDA_i}, Θ_{HM_i}, Θ_{DDM_i}, are the parameter sets associated with the i-th ensemble member of the CDDA, HM, and DDM, respectively, with $\Theta_{CDDA_i} = \{\Theta_{HM_i}, \Theta_{DDM_i}\}$.

While for a given HM (e.g., HBV-light, GR4J) the input variables remain fixed, the DDM can accept various input variables, not just limited to those in Equation 1, which may be useful for improving the CDDA simulations. The CDDA adopted in the research is of the general form:

$$Y_i\left(X_{HM}, X_{DDM} \middle| \Theta_{CDDA_i}\right) = y_i\left(X_{HM} \middle| \Theta_{HM_i}\right) + \; r_i(X_{DDM} | \Theta_{DDM_i}) \tag{2}$$

where X_{HM} and X_{DDM} represent the HM and DDM inputs, respectively. Since this work considers HMs with different input variable requirements, X_{HM} includes P_t as well as T_{mean_t} and/or PET_t (depending on the HM); where T_{mean} is the mean air temperature and PET is the potential evapotranspiration (Lindström & Bergström, 1992). However, for all HM-DDM combinations, X_{DDM} includes $P_{t,\ldots,t-D}$, $T_{\mathrm{min}_{t,\ldots,t-D}}$, $T_{\mathrm{max}_{t,\ldots,t-D}}$, $T_{\mathrm{mean}_{t,\ldots,t-D}}$, $PET_{t,\ldots,t-D}$, and $Q_{t-1,\ldots,t-D}$; where T_{min} and T_{max} are the minimum and maximum air temperatures, respectively. Since the new SCDDA applies the stochastic framework to the CDDA (see Section 3.3), the new SCDDA considers the exact same input variables as the CDDA.

The most significant limitation of the CDDA is that it only accounts for the HM parameter uncertainty (uncertainty associated with the true value of the estimated parameter(s)). Thus, the CDDA can only estimate confidence intervals (rather than prediction intervals) when quantifying uncertainty in the simulations (since prediction intervals account for the uncertainty associated with the prediction of the true value of a given hydrological variable) (Montanari & Koutsoyiannis, 2012). In other words, as a single DDM is used to correct the residuals of each HM ensemble member (one DDM for HM), the DDMs' output within the CDDA is related to the expected value of the HM residuals, not their distribution.

3.2 Stochastic Conceptual-data-driven Approach (SCDDA)

The SCDDA was introduced in Q22 to account for additional sources of uncertainty (i.e., input data, IVS, model output) not considered in the CDDA. The SCDDA is an extension of the earlier frameworks proposed by Montanari and Koutsoyiannis (2012) (which focused on HMs) and Quilty et al. (2019) (which focused on DDMs), where an HM is coupled with a DDM in a stochastic framework. The SCDDA makes use of the original equation proposed in Montanari and Koutsoyiannis (2012) for converting a deterministic HM into a stochastic one:

$$f_Q(Q) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{X}} f_e(Q - S(\Theta, X)|\Theta, X) f_\Theta(\Theta) f_X(X)\, d\Theta\, dX \tag{3}$$

where f_Q represents the probability density function (PDF) of the target variable (e.g., streamflow) to be simulated (or forecasted) and f_X, f_Θ, and f_e represent the PDF of the input data, the PDF of the parameter(s), and the PDF of the model error conditioned on the parameter(s) and input data, respectively, with Q, X, and Θ as target variables (i.e., streamflow), input data, and parameter(s), respectively, and S as the function for the deterministic model for streamflow. In Q22, the SCDDA is obtained by modifying Equation (3) to include (ensemble) HM simulations within X (along with other hydro-meteorological variables) and using a DDM for S (rather than an HM). Along with the original assumptions in Montanari and Koutsoyiannis (2012), it was further assumed in Q22 that the DDM parameters could be estimated independently of the HM parameters since the ensemble HM simulations were used as input to the DDMs (and represented input data uncertainty). The interested reader is referred to Quilty et al. (2022) for additional details on the original formulation of the SCDDA. The difference between the original SCDDA used in Q22 and the one proposed here is discussed in the next sub-section.

3.3 Comparing the New and Original SCDDAs

The SCDDA proposed in this work uses the ensemble HM simulations differently than in Q22. Notably, the SCDDA proposed in this research sums the HM simulations and the DDM residual simulations of HM (i.e., the error approach), while in Q22, the HM simulations were used as input to the DDMs (i.e., the informed approach). Figure 1 shows the difference between the original SCDDA presented in Q22 (A) and the new SCDDA proposed in this research (B).

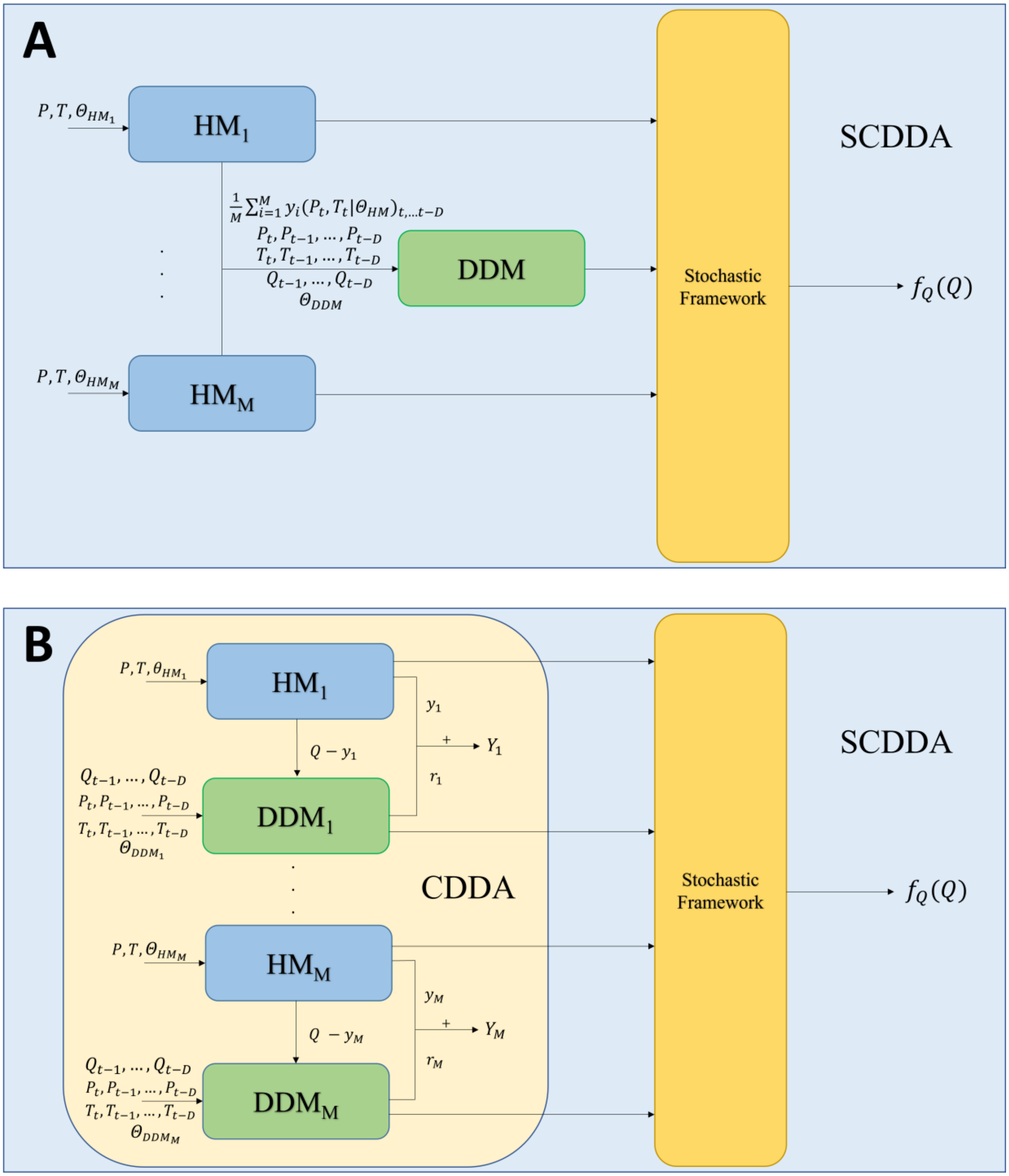
A
P, T, Θ_{HM_1}
HM_1
$\frac{1}{M}\sum_{i=1}^{M} y_i(P_t, T_t|\Theta_{HM})_{t,\ldots t-D}$
$P_t, P_{t-1}, \ldots, P_{t-D}$
$T_t, T_{t-1}, \ldots, T_{t-D}$
$Q_{t-1}, \ldots, Q_{t-D}$
Θ_{DDM}
DDM
P, T, Θ_{HM_M}
HM_M
Stochastic Framework
SCDDA
$f_Q(Q)$
B
P, T, θ_{HM_1}
HM_1
y_1
$+$
Y_1
$Q - y_1$
r_1
$Q_{t-1}, \ldots, Q_{t-D}$
$P_t, P_{t-1}, \ldots, P_{t-D}$
$T_t, T_{t-1}, \ldots, T_{t-D}$
Θ_{DDM_1}
DDM_1
CDDA
P, T, Θ_{HM_M}
HM_M
y_M
$+$
Y_M
$Q\ - y_M$
r_M
$Q_{t-1}, \ldots, Q_{t-D}$
$P_t, P_{t-1}, \ldots, P_{t-D}$
$T_t, T_{t-1}, \ldots, T_{t-D}$
Θ_{DDM_M}
DDM_M
Stochastic Framework
SCDDA
$f_Q(Q)$

Figure 2. The (original) SCDDA from Q22 (A) and its new version proposed in this research (B), using the notation from Q22

In Figure 2 (A), P and T are inputs to the HMs with parameter vector (Θ_{HM}), producing streamflow simulation (y). The DDM uses the ensemble mean of the HM simulations, P and T, and their time-lagged versions up to time $t - D$ (where D is the maximum time lag) along with DDM parameters (Θ_{DDM}) to produce the deterministic streamflow simulation. Optionally, the DDMs can also use time-lagged versions ($t - 1$, …, $t - D$) of observed streamflow (Q) as additional inputs. For the stochastic framework, given a new set of inputs, the distribution of the DDM model inputs (f_X, where the uncertainty is due solely to the ensemble HM simulations as the other inputs remain fixed at their observed values by randomly sampling HM parameter sets and using the HM simulations as inputs to the DDM), model parameters ($f_{\Theta_{DDM}}$), and errors ($f_{e|\Theta_{DDM},X}$) are stochastically sampled to estimate $f_Q(Q)$.

In Figure 2 (B), the HM residuals ($Q - y$) are used as the target variable for the DDM, and its associated simulation (r, the estimated HM residual) is summed with the HM simulation to produce a simulation from the CDDA for a given ensemble member (Y). For the new SCDDA proposed here, the ensemble of HMs and DDMs is simulated via stochastic resampling using the distribution of model parameters ($f_{\Theta_{HM}}$, $f_{\Theta_{DDM}|\Theta_{HM}}$) and errors ($f_{e|\Theta_{HM},\Theta_{DDM}}$) to estimate the streamflow distribution ($f_Q(Q)$). In this setup, the DDM parameters are conditioned by the HM parameters, while the error distribution ($f_{e|\Theta_{HM},\Theta_{DDM}}$) is related to the CDDA simulation (Y). It is also possible to formulate the SCDDA such that the HM and DDM parameters are jointly estimated (see Chapter 6).

By contrasting panels A and B in Figure 2, it can be seen that the new SCDDA directly incorporates the CDDA (B) while the SCDDA from Q22 (A) only adopts the ensemble HM simulations as input data and does not generate simulations by aggregating the outputs of the HMs and DDMs. An important benefit of having the CDDA 'built-in' to the SCDDA is that users can preserve both the HMs and DDMs generated by the CDDA and use the stochastic framework to estimate the uncertainty of the CDDA. As the new SCDDA uses stochastic resampling as the second stage in the post-processing framework, it is possible to revert to the CDDA if the SCDDA (as determined by the CPP) is not expected to improve model performance.

A significant difference between the SCDDA in Q22 and the new version proposed here is how HM parameter uncertainty is accounted for in the stochastic framework. The SCDDA in Q22 accounts for HM parameter uncertainty by using the ensemble HM simulations to represent input uncertainty in the DDM. Specifically, for a new set of model inputs, a single parameter vector (Θ_{HM}) from the ensemble of HM parameter vectors $\{\Theta_{HM_1}, \dots, \Theta_{HM_M}\}$ is randomly sampled, the HM associated with this parameter vector is used to generate a simulation and concatenated to the other hydro-meteorological variables and used as input to the DDM. In the new SCDDA proposed here, DDM parameter uncertainty is conditioned on the HM parameters. In detail, for a randomly selected HM parameter vector, a DDM parameter vector, conditioned on the randomly selected HM parameter vector, is chosen at random from $\{\Theta_{DDM_1}, \dots, \Theta_{DDM_N}\}$ (where N is the total number of DDM parameters for each HM ensemble member), since multiple DDMs are trained to simulate the residuals associated with each HM parameter vector. Thus, the function S in Equation 3 is a DDM in the original SCDDA in Q22, and the CDDA in the new SCDDA proposed here. A pseudo-code for the new SCDDA is provided in Section 4.4.

In the case studies explored herein, the input data uncertainty is not considered in the SCDDA since information on the input variable uncertainty was not available. IVS uncertainty was considered in two of the three DDMs, RF and XGB, since these methods inherently perform IVS when model parameters are calibrated. Thus, the model parameter uncertainty also includes IVS uncertainty in RF and XGB. However, DDMs that do not (inherently) perform IVS as part of the parameter calibration stage require IVS uncertainty to be estimated through explicit methods (e.g., via the bootstrap) if this source of uncertainty is to be considered (see, for example, Quilty & Adamowski (2020)). However, the LSTM implemented in this research did not account for IVS uncertainty. It was assumed that IVS would not significantly impact the model performance since the LSTM model structure inherently accounts for the relationship between the target variable and previous time lags of the explanatory variables, which was previously shown to be important to consider when simulating streamflow in the study catchments (Sikorska-Senoner & Quilty, 2021).

3.4 Hydrological Models

Previous work on the CDDA and SCDDA focused on a single HM structure, HBV-light (Seibert & Vis, 2012). To better assess the impact of the model structure on the performance of the CDDA and SCDDA, three different HM structures were considered in this research. The conceptual TUWmodel (Parajka et al., 2007; Viglione & Parajka, 2020) was adopted as the model is formulated based on the HBV model structure. Similarities and differences between the TUWmodel and HBV-light are discussed in Section 3.4.2. Furthermore, to assess the performance of CDDA and SCDDA when using a model with lower structural complexity and to benchmark the performance of such a model against higher complexity HMs (HBV-light and TUWmodel), the modèle du Génie Rural à 4 paramètres Journalier (GR4J, Perrin et al., 2003) was also adopted. Both TUWmodel and GR4J were calibrated using Bayesian Optimization with Gaussian Processes (BO, Snoek et al., 2012), a popular algorithm for tuning DDM hyper-parameters (Shahriari et al., 2016; Snoek et al., 2012b) that was recently used for calibrating process-based HMs (Ma et al., 2021; Ma et al., 2022). As the BO algorithm is known for finding suitable parameter sets at low computational costs, it is expected that the TUWmodel will not provide the same level of performance as achieved by HBV-light in SSQ21, which used a higher number of calibration iterations. However, utilizing BO to find suitable (but not necessarily optimal) model parameters may provide the opportunity for the CDDA and SCDDA to correct under-calibrated HMs leading to reliable simulations. The following three sub-sections briefly summarize the HMs used in this work.

3.4.1 HBV-light

An HBV variant, HBV-light (Seibert & Vis, 2012), was adopted as a (conceptual) lumped catchment rainfall-runoff model that simulates the catchment response to hydro-meteorologic input data through four routines (precipitation and snowmelt, soil moisture, groundwater, and routing). The model consists of 15 tunable parameters with T_{mean} and long-term averaged PET as inputs to HBV-light. The HBV-light simulations used in this research are from SSQ21, where the model parameters were calibrated using the Genetic Algorithm and Powell method (GAP, Seibert, 2000) and the Kling-Gupta Efficiency (KGE, Gupta et al., 2009) as the objective function. For more information on HBV-light and the calibration procedure, see Sikorska-Senoner et al. (2020).

3.4.2 TUWmodel

Another HBV-based model adopted here, TUWmodel, a lumped catchment rainfall-runoff model consisting of three major routines (precipitation and snowmelt, soil moisture, and routing) and 15 tunable parameters. While most of the processes are similar, the main differences between TUWmodel and HBV-light are summarized below:

1. For the snow routine, TUWmodel uses a threshold temperature interval to distinguish rain, snow, or a mixture of both, while HBV-light uses a single temperature threshold where meltwater and rainfall are contained in the snow until it exceeds a certain threshold with a refreezing component
2. PET is required as a user-defined input to TUWmodel and is calculated by HBV-light.
3. The triangular transfer function for routing.

Along with PET, P, and T_{mean} are also used as inputs to TUWmodel. The parameters in TUWmodel were calibrated using BO, with additional details described in Section 4.2. The mathematical background of the TUWmodel can be found in Parajka et al. (2007).

3.4.3 GR4J

GR4J is a lumped rainfall-runoff model with four tunable parameters, where P and PET are used as input to simulate streamflow (Perrin et al., 2003). Due to the model’s parsimony, robustness, computation speed, and simplicity, GR4J has become popular in the hydrology domain and has been shown to provide competitive performance when benchmarked against other HMs (Darbandsari & Coulibaly, 2020; Gaborit et al., 2017; Kunnath-Poovakka & Eldho, 2019; Oudin et al., 2008; Perrin et al., 2003; Wijayarathne & Coulibaly, 2020). The four parameters in GR4J that require calibration include the maximum capacity of the production store (mm), the catchment water exchange coefficient (mm/d), the maximum capacity of the routing store (mm), and the time base of the unit hydrograph (d). GR4J is also coupled with the Cema-Neige snow routine (GR4JCN, Valery, 2010) in Section 5.5 to explore the impact the snow routine has on model performance; in this case, GR4JCN has two additional parameters, ponderation coefficient (dimensionless) and degree-day factor (mm/°C/d) that require calibration. Since GR4JCN is only adopted in a single experiment and follows the same model development procedure as GR4J, GR4J is primarily referred to throughout the text. The GR4J parameters were calibrated using BO according to the details provided in Section 4.2. The mathematical formulation of GR4J can be found in Perrin et al. (2003).

3.5 Data-driven Models

Based on the recommendations in SSQ21, XGB and RF were evaluated in the CDDA and SCDDA. Furthermore, due to its increasing popularity in the hydrological modelling literature, LSTM was also adopted within the CDDA and SCDDA. Brief descriptions of the three DDMs are outlined in this section.

3.5.1 eXtreme Gradient Boosting (XGB)

A recent tree-based ensemble learning method, eXtreme Gradient Boosting (Chen and Guestrin, 2016), accurately simulated HM residuals (related to streamflow) in SSQ21. In contrast to RF, where predictions are made by bagging an ensemble of trees, the trees in the XGB are combined sequentially by scaling each tree according to a learning rate (also known as boosting), similar to the Gradient Boosting method (Hastie et al., 2009). In short, XGB is an efficient ensemble learning method that results in a parsimonious structure through regularization and inherently measures input variable importance while providing competitive performance compared to existing tree-based methods (Chen & Guestrin, 2016). Recent applications of XGB in the hydrology and water resources domains include flash flood risk assessment (Ma et al., 2021), prediction of dew point temperature (Dong et al., 2022), detecting leakage in urban water distribution networks (Wu et al., 2021), water quality prediction (Wang et al., 2022), and modelling lake bathymetry (Liu & Song, 2022).

3.5.2 Random Forests (RF)

Based on the family of decision tree models, RF was first introduced by Breiman (2001), and has gained widespread popularity due to its high performance, flexibility, amenability to perform quantile regression, and ability to measure each input variables' importance, among other useful qualities. RFs generate a bootstrapped dataset by randomly selecting samples from the training data with replacement, build multiple decision trees using a random subset of input variables for each root and node in each tree, and generate a final set of predictions by taking the mean of the outputs from all decision trees. By using the bootstrapped dataset and aggregating predictions across multiple trees, also known as bagging (Breiman, 1996), the diversity of decision trees created by RF assists in the bias-variance trade-off (Luxburg & Schölkopf, 2011).

In addition to streamflow simulation and forecasting (Papacharalampous & Tyralis, 2018; Schoppa et al., 2020), RF has been used for numerous hydro-meteorological applications, including the classification of severity of mid-winter ice breakups (de Coste et al., 2022), estimating regional groundwater fluoride concentrations (Rosecrans et al., 2022), downscaling spatial resolution of soil moisture satellite products (Triantakonstantis et al., 2022), prediction of the seasonal freeze-thaw cycle (Zhong et al., 2022), spatial interpolation of climate surfaces (precipitation and air temperature) (Tan et al., 2021), regional flood frequency analysis (Desai & Ouarda, 2021). For a detailed exploration of RF within water resources, see Tyralis et al. (2019b).

As noted above, RF intrinsically measures input variable importance. This useful feature of RF was formulated into a new IVS method, Guided Regularized Random Forests, by Deng and Runger (2013), and was shown to select a lower number of input variables that provide similar (or better) performance than the original input variable set when used in RF. In Q22, Guided Regularized Random Forests were used for IVS in the RF-based SCDDA; hence all RFs in this study use the Guided Regularized Random Forests.

For the input variable importance score, the residual sum of squares is calculated for each split to measure the total decrease in node impurities from splitting on the input variable, which is then averaged across all trees. The input variable importance score is then normalized. Afterwards, an importance weight and a regularization coefficient are used to calculate a penalty weight vector for all input variables. The penalty weight vector is used to guide the IVS procedure within RF. For more information on Guided Regularized Random Forests, see Deng and Runger (2013) and Quilty et al. (2022).

3.5.3 Long-Short Term Memory Network (LSTM)

Along with various deep learning neural networks applied to model time series, the Recurrent Neural Network (RNN) has been used for embedding sequential memory (time-based correlation) in the network architecture. The main limitation of RNN is that it is incapable of learning long-term dependencies due to vanishing or exploding gradients when training models with backpropagation (Hochreiter & Schmidhuber, 1997), which is undesirable for simulating streamflow that exhibits long-term dependence. Hence, a modified configuration of the RNN, the Long-Short Term Memory network (LSTM), overcomes this weakness of the RNN, with the capability of storing long-term information through cell states (Hochreiter & Schmidhuber, 1997). Compared to the RNN, the LSTM includes a cell state that stores long-term information and multiple gates (i.e., the forget gate, input gate, output gate), controlling the flow of information within the network. The forget gate controls the flow of information from the cell state to the forget gate. The input gate controls what new information can be updated in the cell state, and the output gate controls the information that passes from the cell state to the next hidden state. The output of the LSTM is connected through a single neuron dense layer that simulates the target variable. For a detailed description of the LSTM in the context of a large-scale rainfall-runoff modelling case study, see Kratzert et al. (2018).

The LSTM has become increasing popular relative to other DDMs in hydrology in the last three years due to its ability to process large datasets, inherently capture long-term dependencies, and accurately predict hydrological variables (Chen et al., 2020; Fan et al., 2020; Gauch et al., 2021; Kratzert et al., 2018; Rahimzad et al., 2021). In particular, several recent studies have focused on combining HMs with LSTMs. For example, Frame et al. (2021), Konapala et al. (2020), Lu et al. (2021), and Nearing et al. (2020) used HMs as inputs to the LSTM, showing instances where the

LSTM was able to improve the simulation quality of the standalone HMs. Furthermore, similar to the approach adopted here, LSTMs have also been used to simulate the residuals of the HM outputs. For instance, Cho and Kim (2022), Han (2021), and Sharma et al. (2021) used LSTM to correct the HM outputs, improving upon the streamflow simulations of the standalone HM. However, to date, no studies have used LSTM within a stochastic framework to correct the outputs of multiple HMs.

Chapter 4

Experiment Setup

This chapter describes the experimental setup, including information on the study area, HM setup, DDM setup, stochastic simulation, and performance assessment.

4.1 Study Area

The experiments use the same three Swiss mountainous catchments as presented in SSQ21.

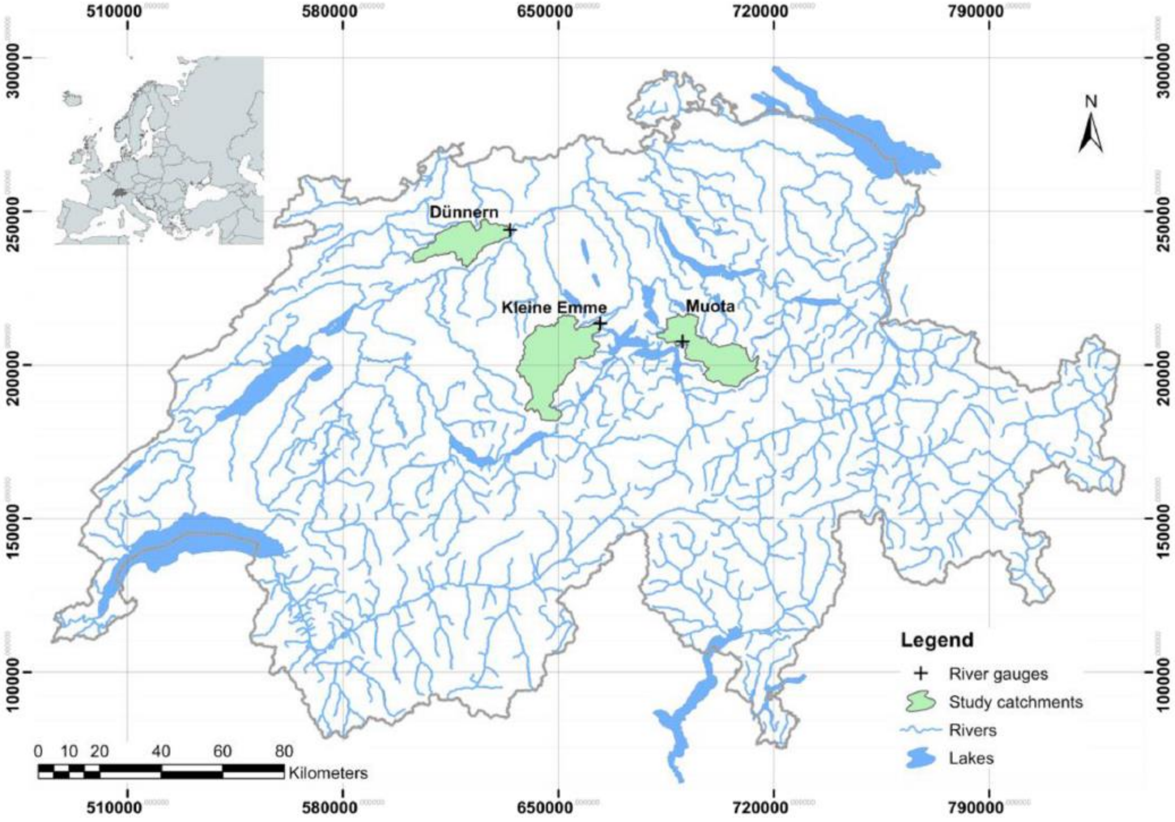

Figure 3. Locations of the three Swiss catchments in the Swiss coordinate system (Sikorska-Senoner, 2021)

All catchments in Figure 3 have an insignificant contribution from glaciers or human impacts during the time period used in this case study (1981-2014). Considering the dominant hydrological processes, the Dünnern catchment (234 km^2) is driven by rainfall, while both Kleine-Emme (478 km^2) and Muota (317 km^2) are driven by a mixture of rainfall and snowmelt processes. As the goal of this study is to simulate the streamflow at the outlet of the catchment, all models were built with basin-averaged variables using Thiessen’s polygon method. The following variables were considered as potential model inputs to the various HMs and DDMs, including precipitation (mm/d), minimum, maximum, and mean air temperature (˚C), and potential evapotranspiration (mm/d) at daily time steps. Measurements of all variables, including the streamflow at the catchment outlet (mm/d), were available during 1981-2014 and sourced from the Swiss Federal Office for the Environment (FOEN). Dataset partitioning followed the same calendar years used in Q22, where 1981-1984 was used as warm-up for the HMs, 1985-2004 as calibration, 2005-2009 as validation, and 2010-2014 for testing. Apart from the warm-up period, both HMs and DDMs followed the same dataset partitioning (i.e., the DDMs did not require a warm-up period). Summary statistics of the variables used to develop the HMs and DDMs are included in Appendix A.

4.2 HM Setup

Of the three HMs adopted in this research, HBV-light used the same model setup as presented in SSQ21, where the GAP method was used to calibrate the model with KGE as the objective function. The parameter ranges considered during the calibration of the individual HBV-light models are listed in Sikorska-Senoner et al. (2020). For GR4J and TUWmodel, BO is used to calibrate the model parameters. In detail, using BO, each ensemble member was optimized using

100 iterations (including 15 initial evaluations) with the KGE as the objective function. Although the CDDA in SSQ21 used 1000 ensemble members, their analysis showed that, for the studied catchments, approximately 100 members led to a performance that did not significantly differ from the 1000 member ensembles (as measured by the continuous ranked probability score, CRPS). Since the number of ensemble members has a significant impact on the computation time, 200 members were generated for each of the three HMs and catchments to ensure stable CRPS was achieved for each HM across all catchments. For HBV-light, the first 200 members of the original 1000-member ensemble were selected from SSQ21. The data were partitioned according to 4.1. Both TUWmodel and GR4J used the PET estimates from HBV-light (see Section 3.4.1, 3.4.2, and 3.4.3 for the input requirements for HBV-light, TUWmodel, and GR4J, respectively). TUWmodel and GR4J were implemented using TUWmodel (Viglione & Parajka, 2020) and airGR (Coron et al., 2022, 2017) R packages, respectively. The parameter ranges considered by BO for TUWmodel and GR4J models are listed in Appendix B.

4.3 DDM Setup

4.3.1 XGB and RF Setup

Since the DDMs do not consider explicit state variables as in HMs, the warm-up period (1981-1984) was not used as part of the DDM training. Instead, the same calibration (1985-2004) and validation periods (2005-2009) adopted for the HM were used to train and validate the DDMs. The DDMs used the residual of the HM simulations as the target variable and considered time-lagged versions of P, T_{min}, T_{max}, T_{mean}, PET, and Q as input variables, with the maximum time lag (D) for each input variable determined by the conditional mutual information (Brown et al., 2012) (see, SSQ21). To minimize the number of input variables while ensuring a sufficient number of previous time lags, the DDMs considered the current and previous nine time-lagged versions of P, T_{min}, T_{max}, T_{mean}, PET, as well as the previous nine prior days of Q, using the same maximum time lag ($D = 9$) as in SSQ21. Although the maximum time lags can be considered as a hyper-parameter to be optimized by BO, considering a single maximum time lag significantly reduces the computation time and is thus followed in this work. The same XGB and RF hyper-parameters and their ranges as described in SSQ21 and Q22, respectively, are also used here for BO hyper-parameter tuning. Thus, for brevity, the BO setup for XGB and RF is not discussed in detail. Due to the memory size requirements and computational cost, for each HM ensemble member, the BO configuration for RF and XGB used 16 iterations (in addition to four initial evaluations) and 25 iterations (in addition to five initial evaluations), respectively. The XGB and RF models were implemented using the xgboost (Chen et al., 2021) and RRF (Deng, 2013) R packages, respectively.

4.3.2 LSTM Setup

In addition to the maximum of time lag considered for each input variable (P, T_{min}, T_{max}, T_{mean}, and PET), the LSTM requires a time window (sequence length) hyper-parameter as it learns from a sequence of data. Preliminary experimentation showed that a 90-day sequence length provided suitable performance while keeping the computation time fast. The remaining hyper-parameters that required careful tuning were optimized using BO using the same objective function as other HM and DDM models and are listed below along with the ranges considered during optimization, which are values similar to those adopted in Alizadeh et al. (2021) (where integers are denoted by ‘L’):

- Learning rate (1e-4, 1e-1)

- Number of hidden units (16L, 128L)
- Dropout rate (0.2, 0.5)

The LSTM used the Adam optimizer (Kingma & Ba, 2015) for training via backpropagation and the same number of iterations (25) and initial evaluations (five) for BO as XGB. However, unlike XGB and RF, the target and explanatory/input were normalized using each variable's mean and standard deviation from the training set to enable faster convergence during model training. The CDDA requires training multiple DDMs for each HM ensemble member. Thus, to keep model training fast while enabling high-quality simulations, the batch size (used for training models in smaller batches of data for computational time) and number of training epochs (the number of times that the DDM will evaluate the training data) were set to 256 and 5, respectively. The batch size was set according to the upper limit used in Alizadeh et al. (2021), while the number of training epochs was determined based on trial-and-error. It is important to note that the goal of optimizing the individual LSTM models within the CDDA and SCDDA was not to achieve the best possible performance for every single model but to generate a set of models that provided high performance with diverse simulations; thus, the performance of individual LSTM models may be further improved using other settings. LSTM was implemented using Tensorflow (Abadi et al., 2015) and Keras (Chollet, 2015) with scikit-optimize for BO (Head et al., 2021)

4.4 Stochastic Simulation

Similar to Q22, the stochastic (simulation) framework used in the new SCDDA is formulated using two modes: an offline mode, where PDFs (e.g., input, parameter, model error) are estimated, and an online mode, where the distribution of streamflow is estimated using a new set of inputs and the previously estimated PDFs. As mentioned in Section 3.3, input data uncertainty was not considered in the SCDDA presented here since uncertainty in the input variables was not available; thus, all input variables remain fixed at their observed values during stochastic simulation. The parameter uncertainty is represented by the HM parameter sets obtained by the GAP (HBV-light) or BO (TUWmodel or GR4J) methods and the DDM parameter sets (obtained by BO) that are attached to each HM parameter set. During the online mode, parameter uncertainty is estimated by first sampling (uniformly at random) from the HM parameter sets and then by sampling (uniformly at random) from the DDM parameter set associated with the earlier sampled HM parameter set (i.e., each of the 200 HM parameter sets is considered at each step of the stochastic simulation and determine/condition the DDM parameter sets that are sampled from thereafter). Using the randomly sampled HM and DDM parameters, the streamflow is simulated on the validation set by combining HM and DDM (i.e., the CDDA) outputs. The combined (CDDA) simulation for the validation set is compared against the observed streamflow (also for the validation set) to generate a set of residuals, which are used to estimate the conditional PDF of the CDDA model error. The CDDA (represented by the randomly sampled HM and DDM parameter sets) is simulated for the new set of inputs, and the simulation is used to conditionally sample a model error (i.e., from the conditional PDF of the CDDA model error). Since the optimization of the CDDA considers multiple DDMs for each HM, the new SCDDA exploits the different DDM parameter sets associated with each HM parameter set to give a more realistic assessment of simulation uncertainty (i.e., the probability distribution of the true value of variable to be simulated), opposite of the CDDA that considers only a single DDM parameter set.

To implement the new SCDDA in an operational setting, the online mode of the stochastic simulation is carried out using the following steps:

1. For new input data at time t, (P_t, T_{mean_t}, and if applicable, PET_t) an HM parameter vector (Θ_{HM_i}) is sampled uniformly at random from $\{\Theta_{HM_1}, \ldots, \Theta_{HM_M}\}$ and used alongside the new input data to generate an HM simulation.
2. A DDM parameter vector, conditioned on the randomly selected HM parameter vector from Step 1, $\Theta_{DDM_j}|\Theta_{HM_i}$, is sampled uniformly at random from the parameter set $\{\Theta_{DDM_1}, \ \ldots, \Theta_{DDM_N}|\Theta_{HM_i}\}$ and is used alongside P_t, T_t (and if applicable, PET_t), previously observed flow (Q_{t-1}) and their lagged values (e.g., $P_{t-1}, \ldots, P_{t-D}$, $T_{\mathrm{max}_{t-1}}, \ldots, T_{\mathrm{max}_{t-D}}$, $T_{\mathrm{mean}_{t-1}}, \ldots, T_{\mathrm{mean}_{t-D}}$, $T_{\mathrm{min}_{t-1}}, \ldots, T_{\mathrm{min}_{t-D}}$, $Q_{t-2}, \ldots, Q_{t-D}$) to generate a DDM simulation (i.e., a simulation of the HM residual from Step 1).
3. HM and DDM simulations are summed together to retrieve a streamflow simulation from the CDDA for the new inputs.
4. An error is sampled from the conditional PDF of the CDDA model error ($f_{e|\Theta_{DDM_j},\Theta_{HM_i}}$) and added to the CDDA simulation for the new inputs.
5. Steps 1-4 are repeated K times (200 here) to generate an estimate of f_Q.

Similar to Q22, the conditional PDF of the CDDA model error is estimated using the K Nearest Neighbours (KNN) resampling approach (Sikorska et al., 2014), where the CDDA simulation for new input data (i.e., from the test set) is used to conditionally sample model errors from the same CDDA's validation set. The same method can be utilized to retrieve the stochastic HM by disregarding the DDM component and resampling the model error from the HM as described in the original blueprint paper (Montanari & Koutsoyiannis, 2012). While the KNN approach was adopted here, other approaches could be used to estimate the conditional PDF of the model error, such as those described in Papacharalampous et al. (2020a, 2020b), or Tyralis et al. (2019a). Furthermore, if the modeller prefers, the DDMs can use other available explanatory variables and/or disregard using previously observed streamflow time lags as input. A pseudo-code for the new SCDDA is provided in Algorithm 1.

Algorithm 1: The new SCDDA

```
for $m$ = 1:$L$ # length of the time series
    for $n$ = 1:$K$        # number of stochastic simulations
        $\Theta_{HM_i} \epsilon \{\Theta_{HM_1}, \dots, \Theta_{HM_M}\}$
        $\Theta_{DDM_j} \epsilon \{\Theta_{DDM_1}, \dots, \Theta_{DDM_N} | \Theta_{HM_i}\}$
        $y = HM(P_t, T_{\mathrm{mean}_t} | \Theta_{HM_i})$        # inputs may vary depending on HM
        $r = DDM(P_{t,\dots,t-D}, T_{\mathrm{mean}_{t,\dots,t-D}}, T_{\mathrm{min}_{t,\dots,t-D}}, T_{\mathrm{max}_{t,\dots,t-D}}, Q_{t-1,\dots,t-D} | \Theta_{DDM_j})$
        $CDDA = y + r$
        # obtain $CDDA_{validation}$ by running CDDA for entire validation set
        $e = Q_{validation} - CDDA_{validation}$
        # e contains the CDDA's validation set errors
        $SCDDA_{m,n} = CDDA + KNN(e, CDDA_{validation}, CDDA)$
```

Finally, since the stochastic models (stochastic HM and SCDDA) require training data to estimate the parameter PDF and the validation data to estimate the conditional PDF of the model error, the stochastic models are only evaluated on the test set (representative of out-of-sample performance).

4.5 Performance Assessment

Deterministic and probabilistic metrics are used to assess the performance of the different approaches (HM, CDDA, and SCDDA) on the test set (to measure out-of-sample performance). The deterministic metrics are calculated using the mean of the ensemble or stochastic simulations, while the probabilistic metrics consider all ensemble members or stochastic simulations. The performance of the ensemble and stochastic HMs (HBV-light, TUWmodel, and GR4J) serves as a benchmark for the CDDA and SCDDA, respectively, while the SCDDA is also compared against the CDDA to better understand the added value of adopting the stochastic framework.

An intercomparison of HMs is considered where HBV-light is used as the benchmark (as it was used in SSQ21 and Q22), TUWmodel is compared against HBV-light to explore the potential of BO to find suitable model parameters, given that both HMs have the same number of parameters and a similar model structure. GR4J is compared against HBV-light and TUWmodel to explore whether similar performance can be achieved with a simplified model structure. The intent is similar in Section 5.5, where GR4J is compared against GR4JCN, although the purpose of the comparison is to decipher whether CDDA and SCDDA can inherently account for snow processes absent from GR4J. The CDDA and SCDDA variants based on the different DDMs (XGB, RF, and LSTM) are compared against one another and the ensemble and stochastic HMs, to ascertain whether there are any HM-DDM combinations that consistently perform better than the others.

The performance assessment uses several deterministic metrics, including the mean absolute error (MAE), root mean squared error (RMSE), Nash Sutcliffe Efficiency (NSE), KGE, and percent bias (PBIAS) (Althoff and Rodrigues, 2021). In addition, the following probabilistic metrics are considered: average width (AW, Papacharalampous et al., 2020a), mean continuous ranked probability score (CRPS, Gneiting and Raftery, 2007), and the alpha index (α_R, Renard et al., 2010). Since these performance metrics are commonly used in the hydrology and water resources literature, the earlier sources should be referred to for further details.

The α_R, a measure of a probabilistic simulation's reliability, is estimated from the coverage probability plot (CPP, see Montanari & Koutsoyiannis, 2012) by measuring the area between the CPP and the bisector. The CPP has been given different names in the literature, for example, the predictive quantile-quantile plot (Eslamian, 2014) and the predictive probability-probability plot (Koutsoyiannis & Montanari, 2022). The CPP provides a visual assessment of the probabilistic simulations, characterizing the simulations' profile and helping diagnose issues with the simulations' spread and bias. For example, narrow/sharp simulations indicate that the observed values lie more frequently than expected on the tail ends of the simulated distributions, while large/over-dispersed simulations indicate that the observed values lie more frequently than expected on the middle quantiles of the simulated distribution. Hence, simulations generated by the ensemble HMs, stochastic HMs, the CDDA, and the SCDDA are evaluated via the CPP to visually assess the simulations' performance, complementing the other (deterministic and probabilistic) metrics.

Given that CPPs can characterize the profile of ensemble and probabilistic simulations, the CPPs for the three different HMs and the nine different CDDA and SCDDA variants are analyzed for the three study catchments (see Section 4.3). The validation set CPPs for the ensemble models (HM and CDDA) were compared against the test set CPPs of their stochastic counterparts (stochastic HM or SCDDA) to determine whether the CPPs could be used as a diagnostic tool to predict whether the stochastic framework can be used to improve upon the reliability of the ensemble models.

One of the main limitations of the SCDDA is the computational time needed for training the DDMs and running the online mode of the stochastic simulation. Computation time can be reduced by selecting models with lower complexity and/or selecting fewer ensemble members, although this may reduce model performance. Thus, the probabilistic metrics are computed for various ensemble sizes to depict the trade-off between model performance and ensemble size. In addition to the probabilistic metrics mentioned above, the decomposed CRPS (Hersbach, 2000) is also considered to further assess the impact of ensemble size on reliability and sharpness, as the CRPS jointly considers reliability and sharpness in a single metric. The deterministic metrics were computed using the hydroGOF R package (Zambrano-Bigiarini, 2020), while the CRPS was estimated using the verification R package (NCAR – Research Applications Laboratory, 2015); the remaining metrics were calculated using custom R scripts.

Chapter 5

Results and Discussion

This section investigates the new SCDDA, comparing it against the ensemble and stochastic HMs and the CDDA. Unless otherwise stated, the various metrics and plots in the section are for the test set (2010-2014) (see Appendix C for the training (1985-2004) and validation (2005-2009) set performance). It should be noted that all metrics are reported to two decimal places except PBIAS, which is rounded to the nearest decimal place (given as a percent), as returned by the hydroGOF package (Zambrano-Bigiarini, 2020). The MAE, RMSE, AW, and CRPS have units of mm/d while NSE, KGE, and α_R are unitless. The PBIAS is reported as a percentage (%).

5.1 Assessment of Ensemble HMs and CDDA Variants

Table 1 summarizes the deterministic performance of the ensemble HMs (using the mean of the ensemble streamflow simulations) for Dünnern, Kleine-Emme, and Muota catchments.

Table 1. Deterministic performance of HBV-light, TUWmodel, and GR4J using the mean of the ensemble streamflow simulations for the test set.

Criteria	HBV-light	TUWmodel	GR4J
		Dünnern	
MAE	0.54	0.73	0.62
RMSE	1.00	1.09	1.05
NSE	0.75	0.70	0.72
KGE	0.80	0.68	0.76
PBIAS	-1.4	22.1	2.4
		Kleine-Emme	
MAE	0.80	1.12	0.97
RMSE	1.57	1.69	1.65
NSE	0.69	0.63	0.65
KGE	0.84	0.76	0.81
PBIAS	-3.5	15.6	-3.4
		Muota	
MAE	1.17	2.66	2.75
RMSE	2.17	4.05	4.03
NSE	0.82	0.39	0.39
KGE	0.82	0.51	0.58
PBIAS	-8.5	-25.7	-4.6

In Table 1, HBV-light shows superior deterministic performance over GR4J and TUWmodel with lower MAE and RMSE and higher NSE and KGE scores. Although GR4J and TUWmodel have similar NSE as HBV-light in Dünnern and Kleine-Emme catchments, they show unsatisfactory performance in the Muota catchment, with NSE values below 0.5 (Moriasi et al., 2007) as well as MAE and RMSE that are 86-135 % higher. The discrepancy in the deterministic model performance between the two HBV variants is likely caused by the different calibration procedures since the models have a similar model structure. Regarding PBIAS, TUWmodel significantly underperforms compared to the other models, while GR4J and HBV-light show similar performance. Despite the poor model performance in the Muota catchment, GR4J in Dünnern and Kleine-Emme catchments shows deterministic performance that is competitive with HBV-light for most metrics.

The probabilistic metrics of the three ensemble HMs are given in Table 2.

Table 2. Probabilistic performance of HBV-light, TUWmodel, and GR4J for the test set.

Criteria	HBV-light	TUWmodel	GR4J
		Dünnern	
AW	0.52	1.98	1.30
α_R	0.59	0.47	0.74
CRPS	0.47	0.54	0.49
		Kleine-Emme	
AW	1.19	3.05	2.17
α_R	0.78	0.58	0.82
CRPS	0.68	0.83	0.77
		Muota	
AW	1.03	4.60	3.67
α_R	0.58	0.71	0.66
CRPS	1.05	2.20	2.35

Considering the sharpness of the three HMs (AW in Table 2), it is evident that HBV-light consistently results in the lowest AW for each catchment. Despite sharp simulations being desirable, they may not be useful if deemed unreliable. Since the goal is to have sharp and reliable simulations, sharper simulations may only be justified if they maintain an acceptable level of reliability. Here, the α_R for HBV-light indicates lower reliability than GR4J for all catchments. However, HBV-light should not be disregarded due to its lower reliability since the CRPS (which simultaneously accounts for sharpness and reliability) shows that HBV-light outperforms TUWmodel and GR4J across all catchments. Comparing the two ensemble HMs calibrated using BO, GR4J provides sharper simulations (lower AW) while having similar or higher reliability levels when compared to TUWmodel. Of note, all three HMs did not overfit the training data and, in some cases, had a better out-of-sample performance (training set results are not shown for brevity).

The ensemble HM results from Table 1 and Table 2 make it possible to identify some potential strengths and weaknesses of using BO to calibrate HMs. It is important to recognize that the goal of BO is to find reasonable parameters with few model evaluations since they are designed for computationally expensive problems, for example, deep learning models (Shahriari et al., 2016). Indeed, with only a small fraction of model evaluations (calibration runs) compared to the HBV-light calibration procedure, BO identified parameters that lead to high and moderate performance for GR4J and TUWmodel, respectively, for Dünnern and Kleine-Emme basins. However, BO did not lead to a satisfactory (deterministic or probabilistic) performance in the Muota basin.

Next, the deterministic performance of the CDDA variants is summarized in Table 3.

Table 3. Deterministic performance of the CDDA variants for all combinations of HMs (HBV-light, TUWmodel, GR4J) and DDMs (XGB, RF, and LSTM) using the mean of the ensemble streamflow simulations for the test set.

Criteria	HBV-light			TUWmodel			GR4J		
	XGB CDDA	RF CDDA	LSTM CDDA	XGB CDDA	RF CDDA	LSTM CDDA	XGB CDDA	RF CDDA	LSTM CDDA
					Dünnern				
MAE	0.39	0.40	0.30	0.35	0.36	0.33	0.34	0.38	0.31
RMSE	0.86	0.86	0.84	0.76	0.79	0.84	0.72	0.76	0.80
NSE	0.81	0.81	0.83	0.85	0.84	0.83	0.87	0.85	0.85
KGE	0.81	0.81	0.86	0.84	0.84	0.86	0.84	0.84	0.83
BIAS	-7.5	-7.5	-4.0	-5.2	-5.3	-3.1	-3.2	-3.1	-2.3
					Kleine-Emme				
MAE	0.61	0.62	0.64	0.59	0.62	0.66	0.59	0.64	0.62
RMSE	1.26	1.30	1.33	1.18	1.22	1.34	1.16	1.21	1.27
NSE	0.80	0.79	0.78	0.82	0.81	0.78	0.83	0.81	0.80
KGE	0.87	0.87	0.88	0.89	0.89	0.89	0.88	0.88	0.86
BIAS	-3.4	-3.3	-3.1	-3.2	-3.3	-1.9	-3.7	-3.6	-2.6
					Muota				
MAE	0.94	0.99	0.78	0.92	1.04	1.12	1.02	1.21	0.95
RMSE	1.94	2.02	1.81	1.75	1.92	2.18	1.79	2.00	1.95
NSE	0.86	0.85	0.88	0.89	0.86	0.83	0.88	0.85	0.86
KGE	0.83	0.82	0.89	0.91	0.91	0.90	0.91	0.91	0.87
BIAS	-8.3	-8.7	-2.9	-2.8	-3.0	-0.7	-3.8	-3.8	-0.7

From Table 3, the results show that the three CDDA variants (XGB, RF, and LSTM) provide similar deterministic performance for each HM except for PBIAS. The LSTM CDDA variant consistently resulted in better PBIAS in all cases, with the most significant difference shown in the Muota basin. Comparing the CDDA performance across the HMs, all variants provide strong deterministic performance in each basin. For example, the CDDA variants achieve KGE above 0.8 for the Dünnern (0.81-0.86), Kleine-Emme (0.86-0.89), and Muota (0.82-0.91) basins. It was anticipated that given the superior deterministic performance of HBV-light, as shown in Table 1, the corresponding CDDA variants would also provide the highest performance. However, the results demonstrate that poorly performing HMs can be substantially improved through the CDDA, leading to competitive models. Furthermore, in several cases, the PBIAS of the TUWmodel CDDA variants is much lower than that of the HBV-light CDDA variants, despite the TUWmodel having the highest PBIAS amongst the HMs (Table 1). To assess the generalization performance of the DDMs used within the CDDA, the deterministic metrics for the training, validation and test sets were evaluated. This analysis showed that XGB and RF were overfitting the training data, with training set KGE of ~0.99 being common, yet both XGB and RF provided a test set KGE of ~0.85. However, LSTM exhibited the most stable performance with a KGE of ~0.85 across the three sets (training, validation, and test).

The probabilistic performance of the CDDA variants is summarized in Table 4.

Table 4. Probabilistic performance of CDDA variants for all combinations of HMs (HBV-light, TUWmodel, and GR4J) and DDMs (XGB, RF, and LSTM) for the test set.

Criteria	HBV-light			TUWmodel			GR4J		
	XGB CDDA	RF CDDA	LSTM CDDA	XGB CDDA	RF CDDA	LSTM CDDA	XGB CDDA	RF CDDA	LSTM CDDA
				Dünnern					
AW	0.82	0.46	0.41	1.78	1.32	1.17	1.31	0.73	0.83
α_R	0.76	0.67	0.77	0.93	0.92	0.94	0.94	0.77	0.83
CRPS	0.31	0.34	0.26	0.25	0.27	0.25	0.25	0.30	0.24
				Kleine-Emme					
AW	1.72	1.03	0.95	3.01	2.10	1.92	2.28	1.17	1.67
α_R	0.88	0.78	0.71	0.92	0.89	0.82	0.90	0.74	0.85
CRPS	0.47	0.51	0.54	0.43	0.47	0.50	0.44	0.51	0.48
				Muota					
AW	1.88	1.15	0.76	4.72	3.08	2.77	4.04	1.93	2.04
α_R	0.70	0.63	0.67	0.93	0.83	0.84	0.90	0.68	0.82
CRPS	0.77	0.86	0.69	0.67	0.80	0.87	0.73	0.98	0.74

Considering the sharpness of the simulations (AW in Table 4), the RF CDDA and LSTM CDDA produce similar AW scores, while XGB CDDA provides simulations with a higher spread. Notably, the AW of XGB CDDA is nearly double that of RF CDDA and LSTM CDDA, considering GR4J in the Muota catchment. In terms of reliability, the XGB CDDA and LSTM CDDA result in the highest α_R across the three catchments, while RF CDDA frequently results in lower reliability than the former approaches. Despite the similar levels of sharpness for RF CDDA and LSTM CDDA, the higher reliability of LSTM CDDA indicates that LSTM should be preferred (instead of RF) when simulating streamflow residuals from the HMs in the study catchments. When considering the CRPS, the best performing CDDA variant for each HM provides similar CRPS across each basin. Thus, it can be seen that RF CDDA frequently underperforms compared to its XGB and LSTM counterparts, XGB CDDA tends to have higher reliability than the RF and LSTM CDDA variants, but this comes at the cost of a higher AW.

Summarizing results from Table 1, Table 2, Table 3, and Table 4, the CDDA improves the ensemble HM performance. For example, for the three catchments, the CDDA improved the RMSE by 7-20%, 20-57%, and 23-55% for the HBV-light, TUWmodel, and GR4J, respectively. However, it should be noted that there were several instances where the CDDA variants had higher PBIAS than the ensemble HMs, the exception being the TUWmodel and most LSTM variants. For instance, the best performing HBV-light CDDA for the Dünnern catchment increased PBIAS from 1.4% to 4%. For the probabilistic performance, the XGB CDDA maintains or increases the ensemble HMs' AW, while the RF and LSTM CDDA tend to decrease the AW, thus increasing the sharpness. In most cases, the reliability of the HMs was improved through the CDDA. However, a decrease in reliability was found when using RF CDDA with GR4J and HBV-light for the Kleine-Emme basin. Finally, the CDDA variants improved the ensemble HMs' CRPS by 18-45%, 40-70%, and 34-69% for HBV-light, TUWmodel, and GR4J, respectively.

5.2 Assessment of the Stochastic HMs and SCDDA Variants

The deterministic and probabilistic performance of the stochastic approaches (stochastic HM and SCDDA) are presented in and, respectively.

Table 5. Deterministic performance of the stochastic HM (SHM) and SCDDA variants using the mean of the stochastic streamflow simulations for the test set.

Criteria	HBV-light				TUWmodel				GR4J			
	SHM	XGB SCDDA	RF SCDDA	LSTM SCDDA	SHM	XGB SCDDA	RF SCDDA	LSTM SCDDA	SHM	XGB SCDDA	RF SCDDA	LSTM SCDDA
							Dünnern					
MAE	0.54	0.41	0.41	0.31	0.54	0.36	0.37	0.33	0.63	0.35	0.37	0.33
RMSE	0.99	0.86	0.86	0.85	1.03	0.78	0.81	0.84	1.08	0.73	0.77	0.83
NSE	0.76	0.81	0.81	0.83	0.73	0.85	0.84	0.83	0.71	0.87	0.85	0.83
KGE	0.81	0.85	0.85	0.88	0.71	0.80	0.81	0.81	0.69	0.84	0.84	0.82
PBIAS	0.4	0.7	0.2	-0.2	-3.2	-3.4	-3.8	-1.3	1.3	-1.9	-2.0	-0.5
							Kleine-Emme					
MAE	0.81	0.63	0.64	0.63	0.97	0.59	0.63	0.65	0.98	0.60	0.65	0.64
RMSE	1.55	1.26	1.3	1.37	1.61	1.18	1.23	1.32	1.59	1.17	1.23	1.29
NSE	0.69	0.80	0.78	0.77	0.67	0.82	0.81	0.79	0.68	0.83	0.81	0.79
KGE	0.83	0.87	0.88	0.86	0.74	0.84	0.85	0.82	0.76	0.88	0.89	0.83
PBIAS	1.7	1.4	1.7	0.5	0.9	-1.8	-1.1	-0.6	1.5	-0.9	-0.7	0.1
							Muota					
MAE	1.17	0.98	1.00	0.84	2.77	0.96	1.06	1.25	2.92	1.00	1.19	1.08
RMSE	2.24	2.04	2.08	1.97	3.94	1.89	2.07	2.42	4.12	1.84	2.08	2.13
NSE	0.81	0.84	0.84	0.86	0.42	0.87	0.84	0.79	0.37	0.87	0.84	0.83
KGE	0.77	0.78	0.78	0.83	0.40	0.78	0.79	0.74	0.34	0.83	0.82	0.78
PBIAS	-6.1	-5.0	-5.6	-3.5	-6.2	-5.7	-6.0	-4.6	-5.7	-5.1	-5.6	-2.9

Table 6. Probabilistic performance of the stochastic HM (SHM) and SCDDA variants for the test set.

Criteria	HBV-light				TUWmodel				GR4J			
	SHM	XGB SCDDA	RF SCDDA	LSTM SCDDA	SHM	XGB SCDDA	RF SCDDA	LSTM SCDDA	SHM	XGB SCDDA	RF SCDDA	LSTM SCDDA
						Dünnern						
AW	2.54	2.45	2.20	2.03	3.94	3.01	2.79	2.86	4.19	2.67	2.49	2.73
α_R	0.98	0.95	0.98	0.89	0.83	0.82	0.83	0.81	0.93	0.87	0.91	0.86
CRPS	0.38	0.29	0.29	0.24	0.40	0.28	0.28	0.26	0.44	0.26	0.28	0.24
						Kleine-Emme						
AW	4.68	4.26	3.71	4.08	6.63	5.06	4.71	5.11	6.52	4.61	4.06	4.94
α_R	0.91	0.87	0.92	0.85	0.86	0.79	0.83	0.82	0.88	0.83	0.86	0.85
CRPS	0.59	0.46	0.46	0.47	0.68	0.47	0.48	0.49	0.70	0.45	0.47	0.47
						Muota						
AW	5.12	5.12	4.79	4.43	12.41	7.44	6.76	8.38	12.97	6.86	6.22	8.21
α_R	0.89	0.93	0.91	0.94	0.95	0.83	0.88	0.89	0.94	0.86	0.88	0.89
CRPS	0.88	0.75	0.77	0.64	1.90	0.74	0.79	0.90	2.07	0.75	0.86	0.78

The results from Table 5 show that the SCDDA variants provide superior deterministic performance for most metrics compared to the stochastic HMs. Analyzing each SCDDA for a given HM reveals that the performance of the mean of the stochastic simulations has low variance across the models. Similar to the CDDA variants, the LSTM SCDDA has lower PBIAS compared to the other variants. The probabilistic scores (Table 6) reveal that the stochastic HMs have higher AW than the SCDDA variants but similar reliability. Interestingly, there seems to be no discernible pattern as to which DDM leads to the highest reliability and lowest CRPS amongst SCDDA variants. However, in all cases, the SCDDA variants result in superior CRPS scores compared to the stochastic HMs. Thus, given that the SCDDA variants have similar reliability but sharper simulations and lower CRPS than their stochastic HM counterparts, the SCDDA variants should be preferred to the stochastic HM for the study catchments.

Comparing the stochastic methods with the ensemble HM from Table 1 and Table 2, the SCDDA variants substantially improve most performance metrics. For example, for the three catchments, the SCDDA variants improved the RMSE 5-20%, 22-53%, and 21-54% considering HBV-light, TUWmodel, and GR4J, respectively. Furthermore, the SCDDA improved the CRPS by 27-49%, 41-66%, and 39-68% considering the same HMs. It is important to note that, unlike the CDDA, the LSTM SCDDA improved the PBIAS of the ensemble HMs for all cases. Thus, it can be said that LSTM SCDDA dominated the HM across all performance metrics.

Comparing the CDDA variants (Table 3 and Table 4) with the corresponding SCDDA variants (Table 5 and Table 6) interesting outcomes are found. First, not all SCDDA variants significantly improve the MAE, RMSE, NSE or KGE of the corresponding CDDA. Although most SCDDA variants tend to have similar deterministic scores as their CDDA counterparts, some SCDDA variants have significantly poorer performance. For example, considering TUWmodel for the Muota catchment, the KGE of XGB CDDA and XGB SCDDA was 0.91 and 0.78, respectively, representing a relative difference of 17 %. However, this could be due to deficiencies with the TUWmodel (rather than the stochastic framework) as there was a higher relative difference (28 %) in KGE for the ensemble HM (0.51, see Table 1) versus the stochastic HM (0.40, see Table 5) than the CDDA versus SCDDA (for the same HM and catchment). Considering the probabilistic performance, the AW of the SCDDA is higher compared to the CDDA, suggesting that the SCDDA is more conservative than the CDDA. In terms of reliability, the SCDDA had higher (lower) reliability than its CDDA counterparts that have α_R< 0.85 (α_R> 0.85). This interesting finding (explored further in Section 5.4) suggests that there may be a level of reliability beyond which CDDA cannot be further improved by the stochastic framework. Considering both sharpness and reliability, the SCDDA variants that improved the reliability of their CDDA counterparts tend to have similar or improved CRPS.

To visualize the streamflow simulations generated by the CDDA and SCDDA as well as the ensemble and stochastic HMs Figure 4-6 include time series plots of the different models' simulations. The mean simulation is included for each model along with the 95% confidence (ensemble HM and CDDA) or prediction (stochastic HM and SCDDA) intervals.

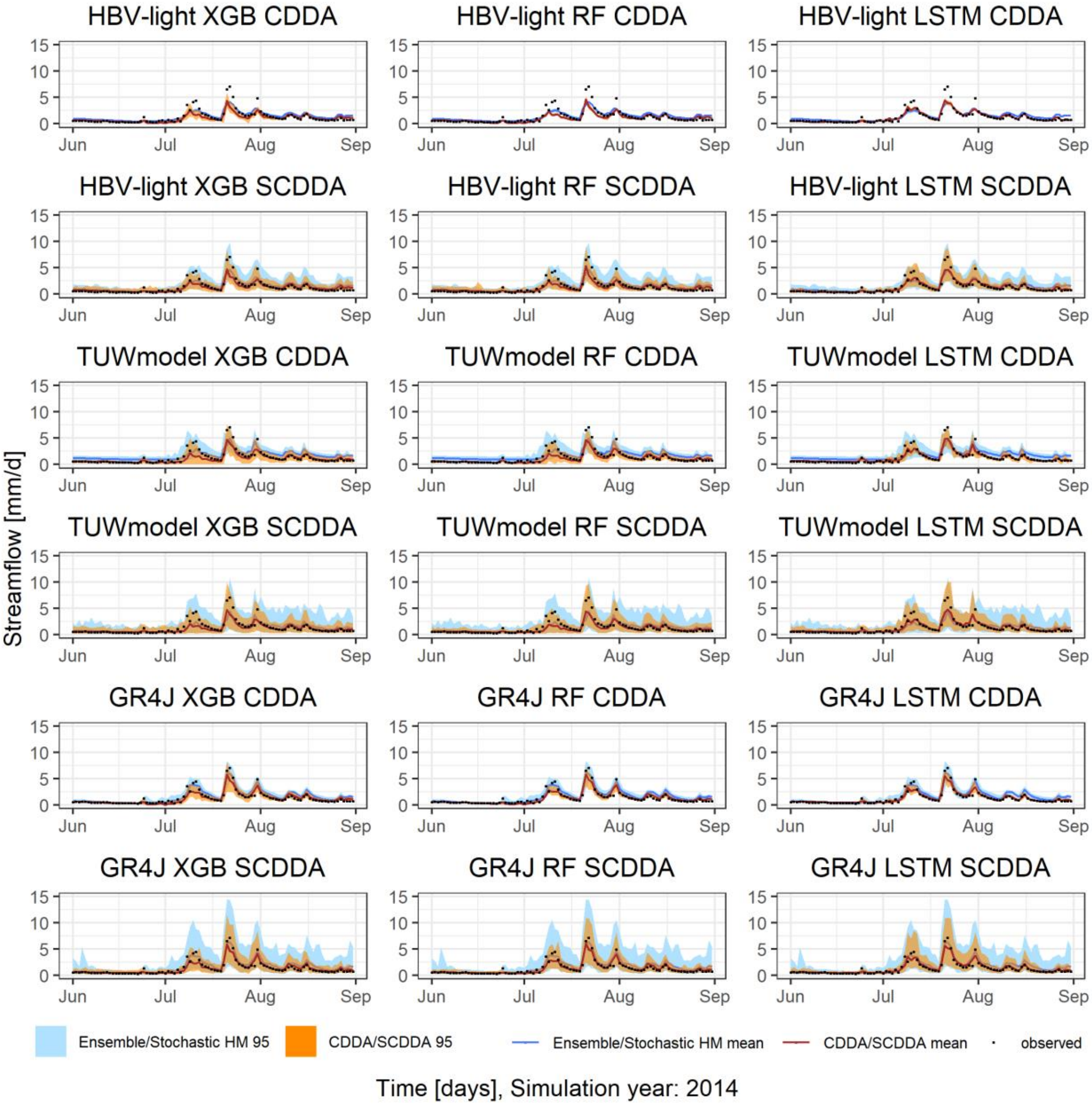

Figure 4. A comparison of the mean streamflow simulation and its 95% (confidence or) prediction intervals in 2014 using CDDA and SCDDA variants in the Dünnern catchment. The ensemble HM and stochastic HM are compared against their CDDA and SCDDA counterparts, respectively.

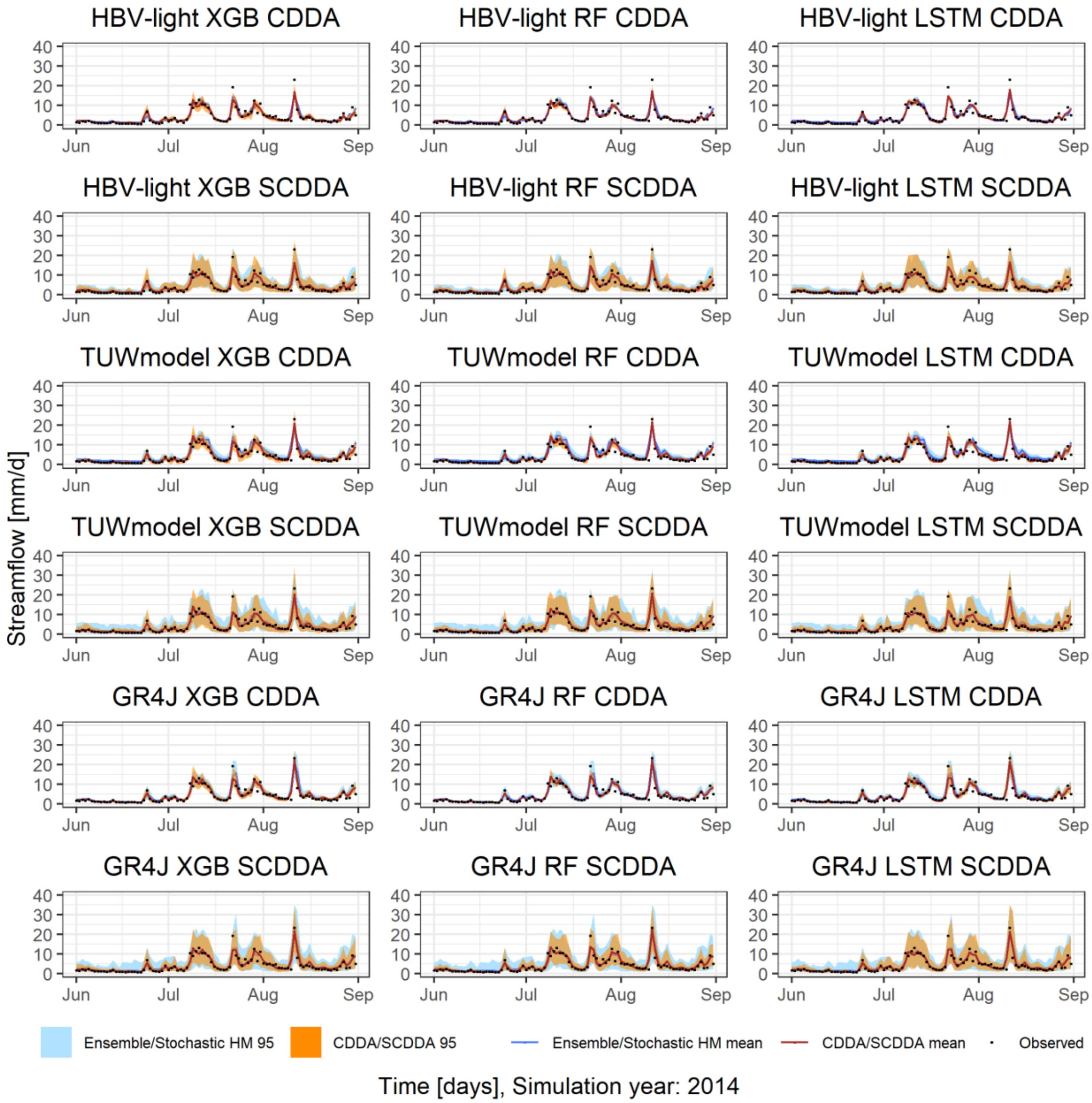

Figure 5. A comparison of the mean streamflow simulation and its 95% (confidence or) prediction interval in 2014 CDDA and SCDDA variants in the Kleine-Emme catchment. The ensemble HM and stochastic HM are compared against their CDDA and SCDDA counterparts, respectively.

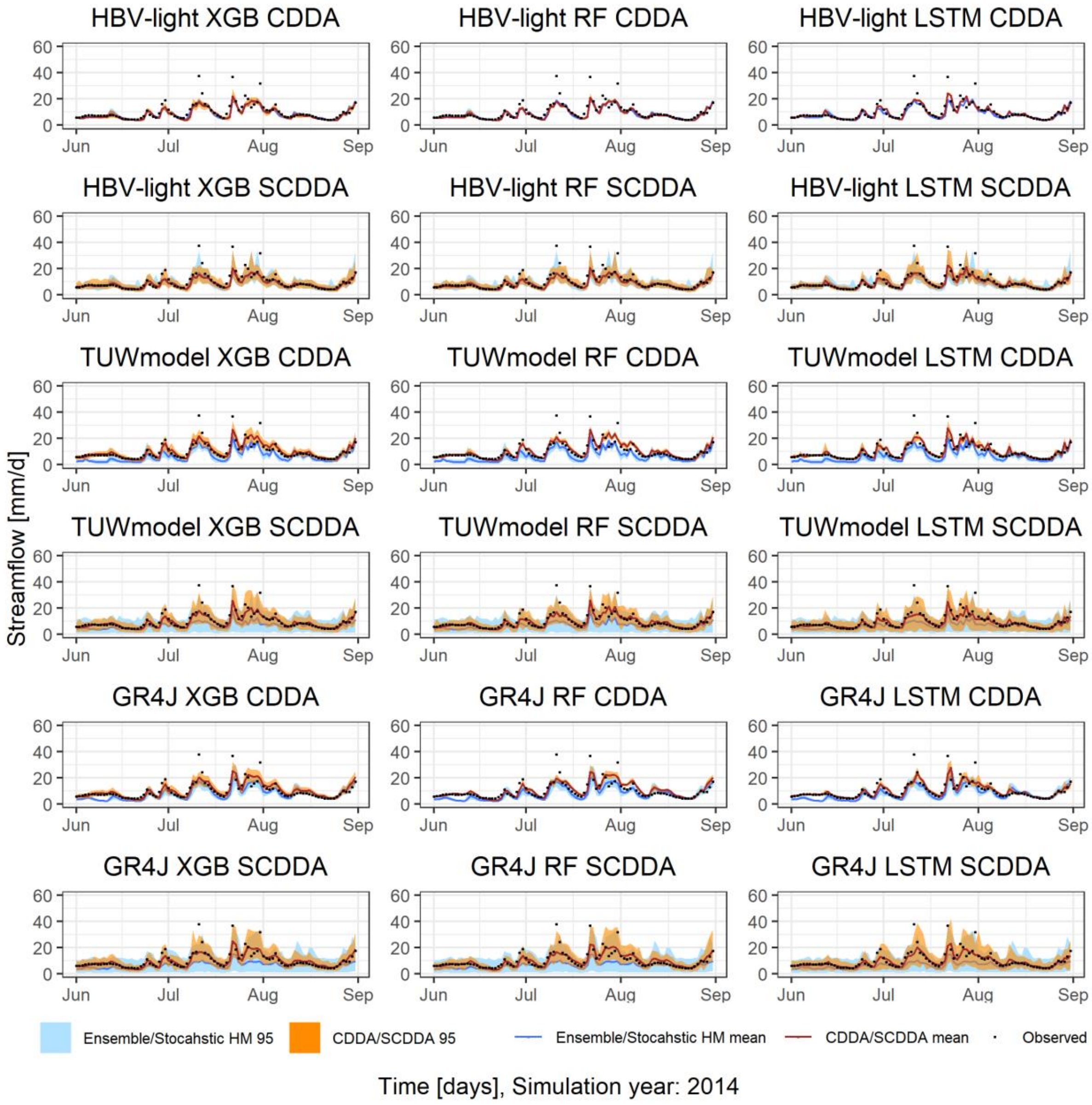

Figure 6. A comparison of the mean streamflow simulation and its 95% (confidence or) prediction interval in 2014 CDDA and SCDDA variants in Muota catchment. The ensemble HM and stochastic HM are compared against their CDDA and SCDDA counterparts, respectively.

The time series plots in Figure 4-6 illustrate several items worth mentioning. The AW scores reported above indicate that the CDDA variants result in sharp simulations, which is confirmed in Figure 4-6, although some CDDA variants have extremely sharp simulations. For example, the 95% confidence intervals of the HBV-light RF and LSTM CDDA variants are concentrated at the mean of the simulated streamflow for all catchments. Thus, the 95% confidence interval is unable to capture high streamflow events. However, the CDDA variants with sharp simulations appear to improve upon the low and mid-flow simulations of the corresponding ensemble HM. Considering the stochastic approaches, the SCDDA seems to compensate for the CDDA and provide conservative simulations by increasing their spread. For instance, many high flow events missed by the ensemble HM and CDDA are captured by the SCDDA, though not all observations are covered within the 95% prediction interval (as is expected). The SCDDA is also able to noticeably reduce the bias of the ensemble HM. In particular, the SCDDA seems to improve the bias of the ensemble HM simulations for GR4J and TUWmodel in the Muota catchment (Figure 6). Finally, Figure 4 and Figure 6 reveal that, in general, the SCDDA variants have sharp simulations for low flows and wide simulations for high flows. This result is likely due to the KNN algorithm used for estimating the conditional PDF of the model error, which inherently accounts for the heteroscedasticity of the model error (Sikorska et al., 2015)

Analyzing the distribution of the streamflow simulations (for specific events, such as floods) may extract additional characteristics, such as modality, to better understand the ensemble and probabilistic simulations. Here, raincloud plots (Allen et al., 2019) were used to enhance the visualization of the simulations’ distribution, which combines a boxplot, a jittered scatter plot, and a probability density plot. The boxplot provides the summary statistics (e.g., the median), the jittered scatter plot shows the raw data (which can be used to identify outliers), and the density plot shows the distribution to check the spread and the modes of the simulated streamflow. In Figure 7-9, an example high flow event was extracted from the test set to visualize the distribution of the streamflow simulations generated by the different models.

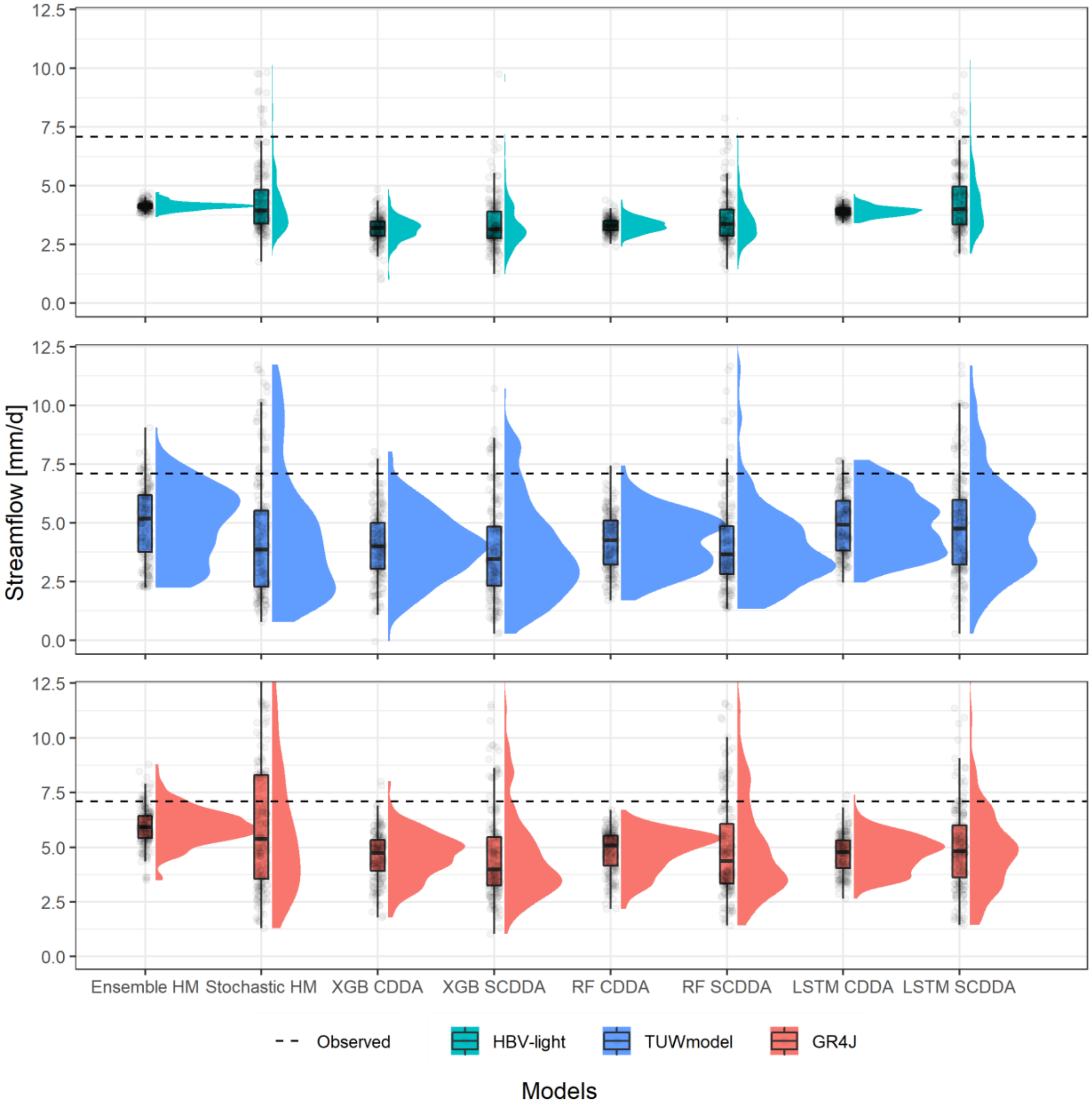

Figure 7. Raincloud plot of a high flow event (2014-07-22) for the ensemble and stochastic HM (HBV-light, TUWmodel, and GR4J) as well as the CDDA, and SCDDA variants in the Dünnern catchment. Note: the area under the curve of the density plot is scaled to the sharpest model.

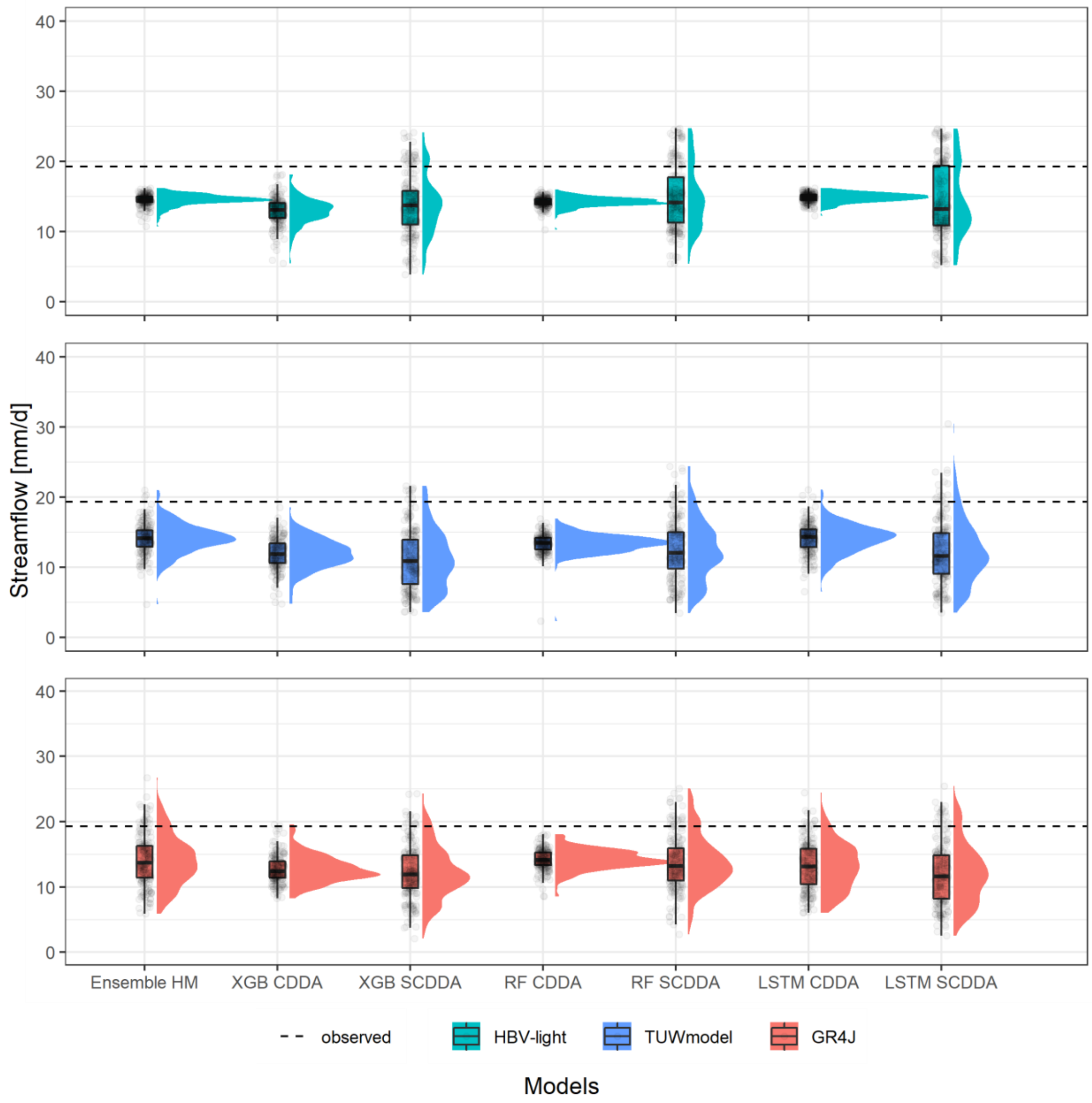

Figure 8. Raincloud plot of a high flow event (2014-07-22) for the ensemble and stochastic HM (HBV-light, TUWmodel, and GR4J) as well as the CDDA, and SCDDA variants in the Kleine-Emme catchment. Note: the area under the curve of the density plot is scaled to the sharpest model.

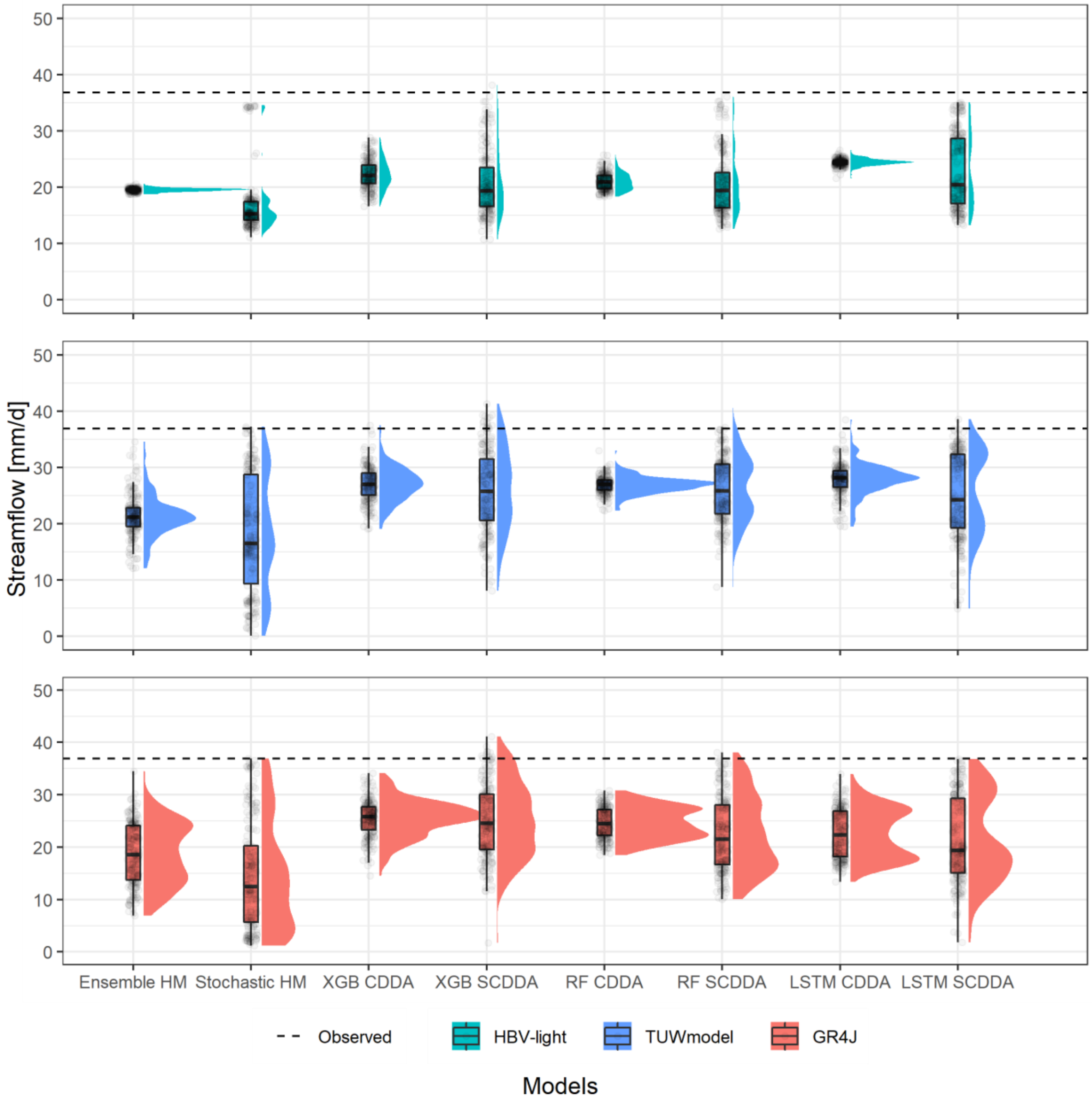

Figure 9. Raincloud plot of a high flow event (2014-07-22) for the ensemble and stochastic HM (HBV-light, TUWmodel, and GR4J) as well as the CDDA, and SCDDA variants in the Muota catchment. Note: the area under the curve of the density plot is scaled to the sharpest model.

Figure 7-9 show that HBV-light provides sharp simulations where no ensemble members capture the high streamflow event. Furthermore, HBV-light and its corresponding CDDA generally result in symmetric distributions. However, the modality of the streamflow simulations' distribution can change substantially by incorporating multiple uncertainty sources (e.g., parameters, model output) via the stochastic framework. For example, the ensemble simulations generated by HBV-light in Kleine-Emme, and TUWmodel in Muota, are approximately symmetric about their median, while their SCDDA counterparts produce simulations with multiple modes that often contain the observed value. In general, the SCDDA tends to capture the high streamflow events more so than its ensemble HM and CDDA counterparts. Although there are many cases where the SCDDA captures the observed streamflow (and, in many cases, has a simulation distribution that shifts upwards surrounding higher flows), the distribution widens, assigning a non-negligible probability to lower streamflow ranges. For example, the LSTM SCDDA using TUWmodel in the Kleine-Emme catchment has a wide simulation distribution, which better captures the observed value, although the mode of its distribution is much lower compared to its ensemble HM counterpart. In the following sub-section, the CPP is used to assess the reliability of the ensemble and stochastic models and is shown how the CPP can be used as a diagnostic tool to predict whether the stochastic framework (i.e., SCDDA) can further improve the CDDA.

5.3 CPPs as a Diagnostic Tool

The validation and test set CPPs for the ensemble and stochastic HMs as well as the CDDA, and SCDDA variants are given in Figure 10-12 for the three catchments. Utilizing the shape of the CPPs, it is possible to classify the streamflow simulation profile. First, when analyzing the ensemble HM results for all basins, GR4J and HBV-light provide sharp simulations, which confirms earlier results. For TUWmodel, the CPPs show that the ensemble HM has a large bias. For example, the Dünnern and Kleine-Emme basins for TUWmodel distinctly result in over-prediction. Furthermore, the validation set CPPs of the ensemble HMs and CDDA variants are very similar to the test set CPPs. Thus, the validation set CPPs can be used to predict the reliability of the test for the study catchments and the HMs and HM-DDM combinations explored herein.

When analyzing the CPPs, a pattern can be seen with respect to the ensemble HMs and the CDDA variants. An important observation is that all HBV-light and most GR4J simulations for the ensemble HM have sharp simulations, and their CDDA counterparts results in the same profile, suggesting that HMs with sharp simulations tend to result in CDDA with sharp simulations. However, it is difficult to find a common pattern with over-and under-predictions. Notably, TUWmodel has CPPs that show over-prediction; however, the corresponding CDDA variants produces CPPs with different shapes. Next, looking at the validation set CPPs for the CDDA (TUWmodel in Figure 10, GR4J in Figure 11, and TUWmodel in Figure 12), the CDDA variants that lie close to the bisector have corresponding SCDDA variants that are less reliable. This result is a visual representation of the previous section, where the CDDA variants with $\alpha_R > 0.85$ had SCDDA counterparts that were less reliable. One explanation is that, for a highly reliable CDDA, the error approaches white noise (is purely random); thus, the KNN-based stochastic resampling of model errors adds ‘noise’ to the streamflow simulation. Therefore, for a CDDA that already has high reliability, the stochastic resampling scheme results in SCDDA variants with lower reliability. However, the important discovery is that the CDDA’s CPPs from the validation set can be used to predict the reliability of the SCDDA (on the test set).

In general, the CDDA and SCDDA have similar or higher reliability than their ensemble or stochastic HM counterparts. Another important finding is that most of the SCDDA variants produce CPPs that are either close to the bisector line or indicate large simulations. Since reliable and conservative (large) simulations are critical in water resources applications (e.g., flood forecasting), the SCDDA can be a useful framework for hydrologists as well as water scientists and practitioners. Further considering the practicality of the above mentioned results, suppose simulation quality is only of interest. Then, using the CPPs as a diagnostic tool, modellers and/or users can decide to implement the stochastic resampling scheme depending on the CDDA’s validation set CPP. Thus, in a practical setting, if the CPP shows that the CDDA is sufficiently reliable, the user may benefit from lower computation time when generating simulations and/or forecasts in real-time (since the stochastic resampling scheme may be abandoned).

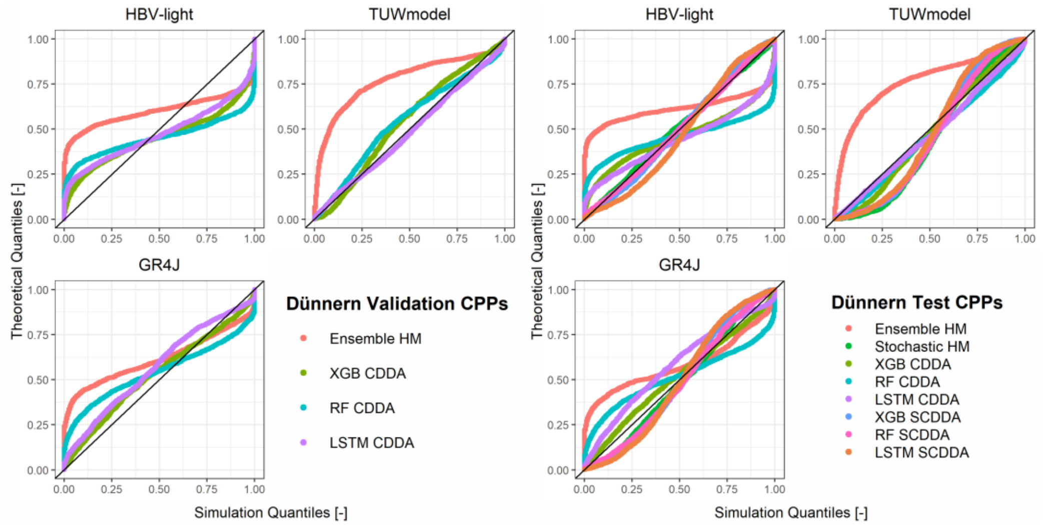

Figure 10. Validation (left) and test set (right) CPPs for the ensemble and stochastic HMs (HBV-light, TUWmodel, and GR4J) as well as the CDDA and SCDDA variants for the Dünnern catchment.

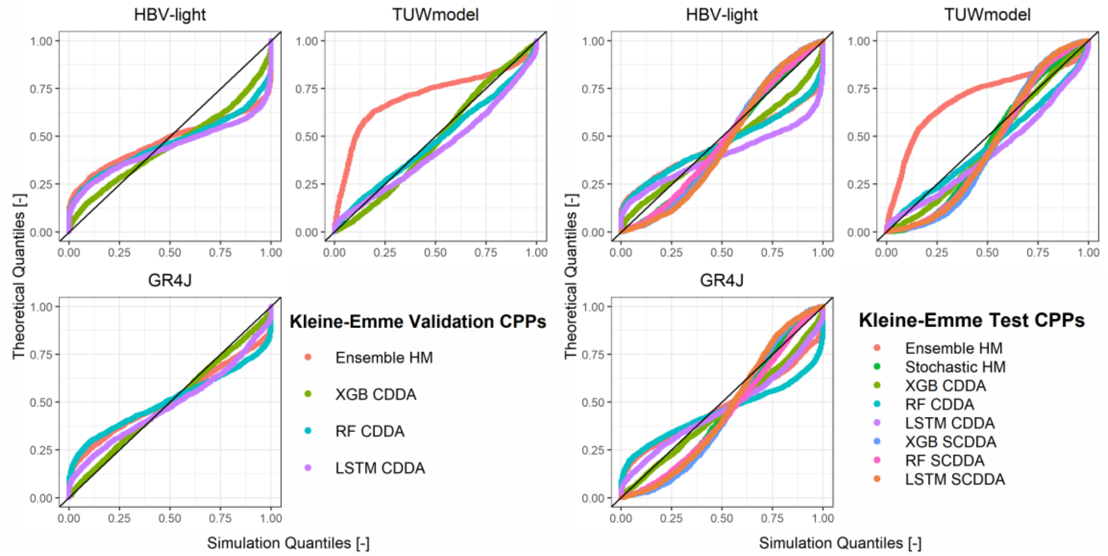

Figure 11. Validation (left) and test set (right) CPPs for the ensemble and stochastic HMs (HBV-light, TUWmodel, and GR4J) as well as CDDA and SCDDA variants for the Kleine-Emme catchment.

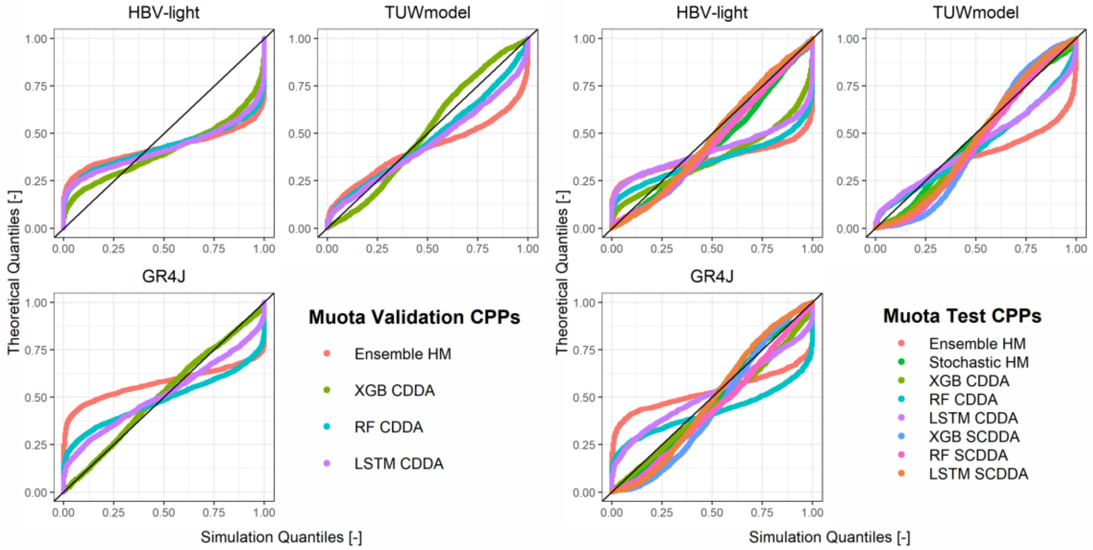

Figure 12. Validation (left) and test set (right) CPPs for the ensemble and stochastic HM (HBV-light, TUWmodel, and GR4J) as well as CDDA and SCDDA variants for the Muota catchment.

5.4 Effect of Ensemble Size on Model Performance

Since the computational demand of DDMs and/or stochastic resampling may deter users from using the CDDA and/or SCDDA, it is critical to assess the possibility of reducing the computational requirements while providing similar performance levels. One way to achieve this is to evaluate the effect of ensemble size (the number of ensemble members) on model performance to find a lower number of ensemble members that characterize the simulation uncertainty to a similar degree as ensembles with more members. The probabilistic performance metrics from Table 2, Table 4, and Table 6 were chosen to evaluate the effect of ensemble size on model performance. Although the AW and α_R can indicate changes in the sharpness and reliability, respectively, of an ensemble or probabilistic simulation for increasing (or decreasing) ensemble size, it is much more challenging to understand the overall improvement in model performance using these metrics compared to a single metric that simultaneously considers sharpness and reliability (such as the CRPS). Of practical value to the following analysis, it is possible to decompose the CRPS into reliability and sharpness components (Hersbach, 2000), known as the reliability CRPS and potential CRPS, where the latter is defined as the CRPS that would be achieved if the simulation was perfectly reliable and sensitive to the average spread of the ensemble simulation. For more information on CRPS decomposition, see Hersbach (2000). In what follows, the AW, α_R, CRPS as well as the reliability and potential components of the CRPS, are evaluated for various ensemble member sizes.

In Figure 13, the AW is plotted as a function of ensemble size for the ensemble and stochastic HM, CDDA, and SCDDA.

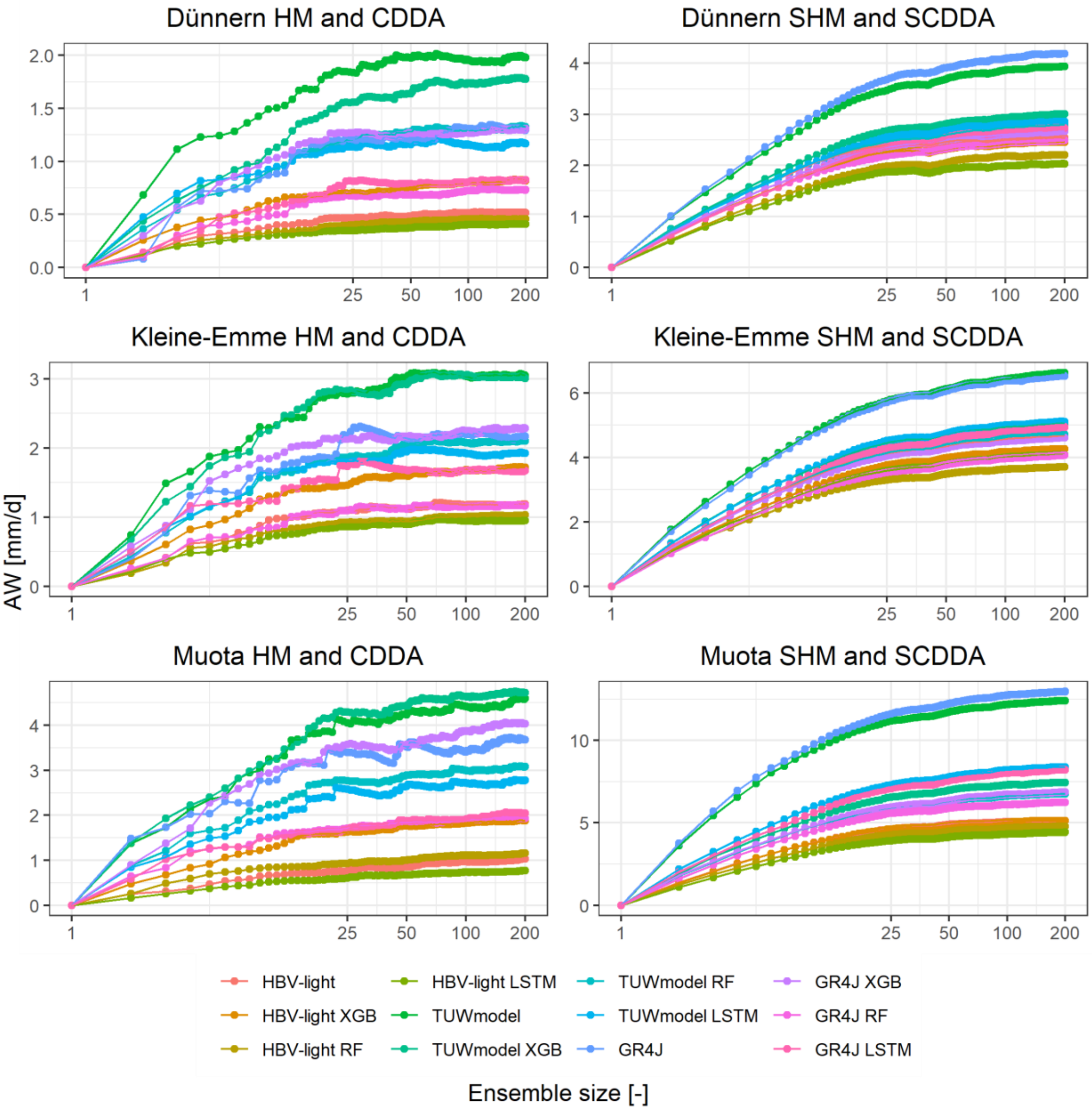

Figure 13. AW vs. ensemble size for the ensemble HM (HBV-light, TUWmodel, and GR4J), CDDA (XGB, RF, and LSTM) (left), and their stochastic counterparts (SHM and SCDDA) (right) for the test set.

From Figure 13, it appears that most ensemble HMs and CDDA variants have similar AW after 100 ensemble members. One noticeable difference between the HMs is that models calibrated by BO seem to be less stable, with fluctuating AW for increasing ensemble member sizes. This outcome is likely caused by BO searching diverse parameter spaces resulting in highly variable parameter sets (whereas the GAP method, due to the high number of calibration runs, converges to similar parameter sets with low variance). Similar levels of instability in the AW are also found for the CDDA variants adopting the HMs calibrated by Bo, regardless of the adopted DDM. For the stochastic HM and SCDDA, all models appear to have similar curves as the ensemble size increases. Notably, beyond 25 ensemble members, the AW seems to have stabilized although; however, an inflection point occurs around 40 ensemble members, causing a substantial change in AW that eventually stabilizes. Although the cause of this inflection point is unknown, the change in AW from 100-200 members is minor. Next, the effect of ensemble size on α_R is shown in Figure 14.

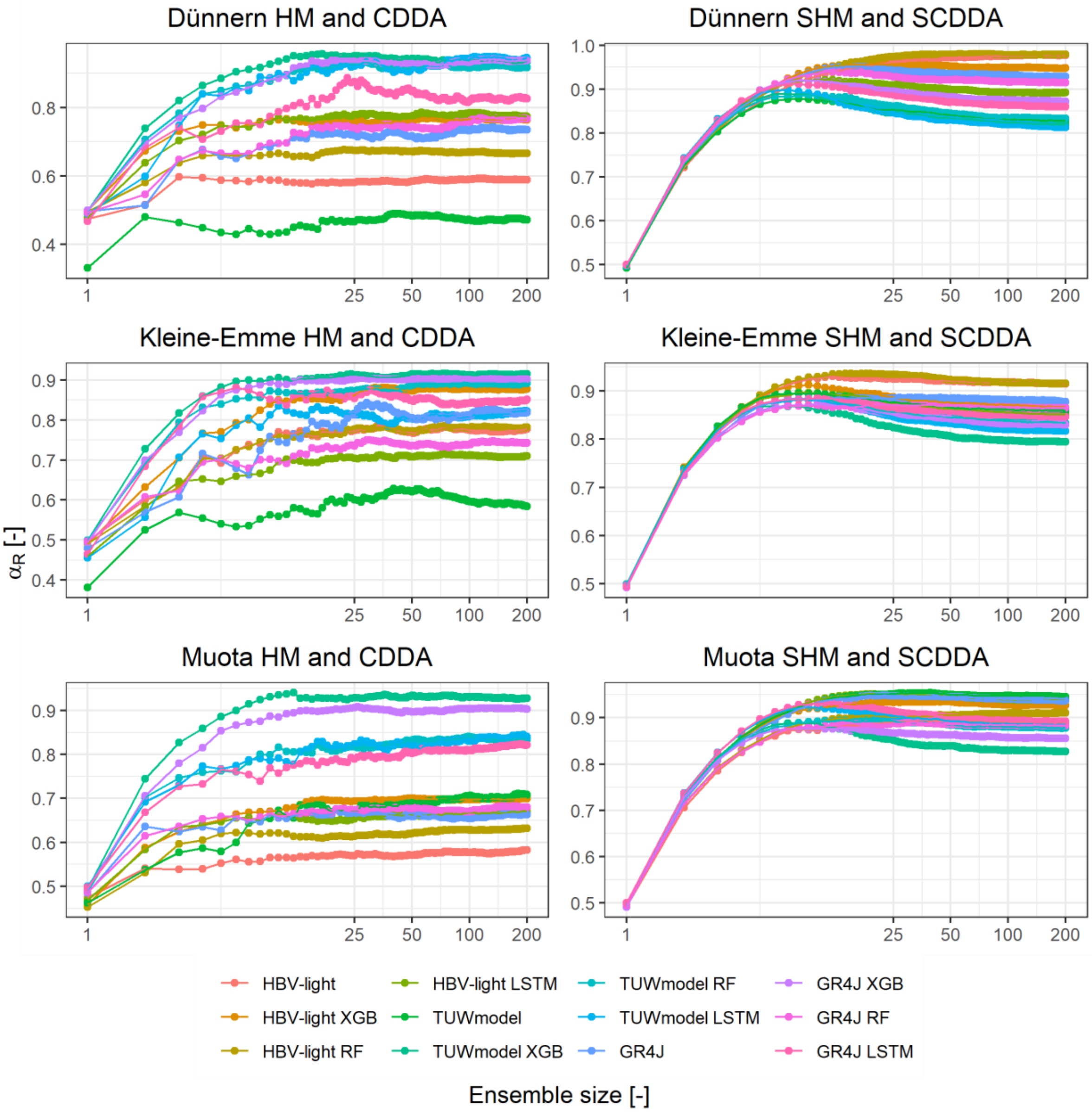

Figure 14. α_R index vs. ensemble size for the ensemble HM (HBV-light, TUWmodel, and GR4J), CDDA (XGB, RF, and LSTM) (left), and their stochastic counterparts (SHM and SCDDA) (right) for the test set.

Regarding the reliability of the ensemble HM and CDDA (Figure 14), it appears that the α_R for most models is relatively stable beyond 50 members. However, for the stochastic models, all models seem to rapidly increase in reliability from one to approximately 15 members, then slowly stabilize around 100 members. For some models (e.g., TUWmodel XGB SCDDA), it appears that the reliability of the model asymptotically decreases from 15 to 100 ensemble members. The instability of the α_R for a low number of ensemble members may be due to a poor estimation of the simulation uncertainty, which could be explained by the stabilization of the around 100 ensemble members. Finally, the CRPS and its decomposed metrics (potential CRPS and reliability CRPS) for varying ensemble sizes are shown in Figure 15-17 for the three catchments.

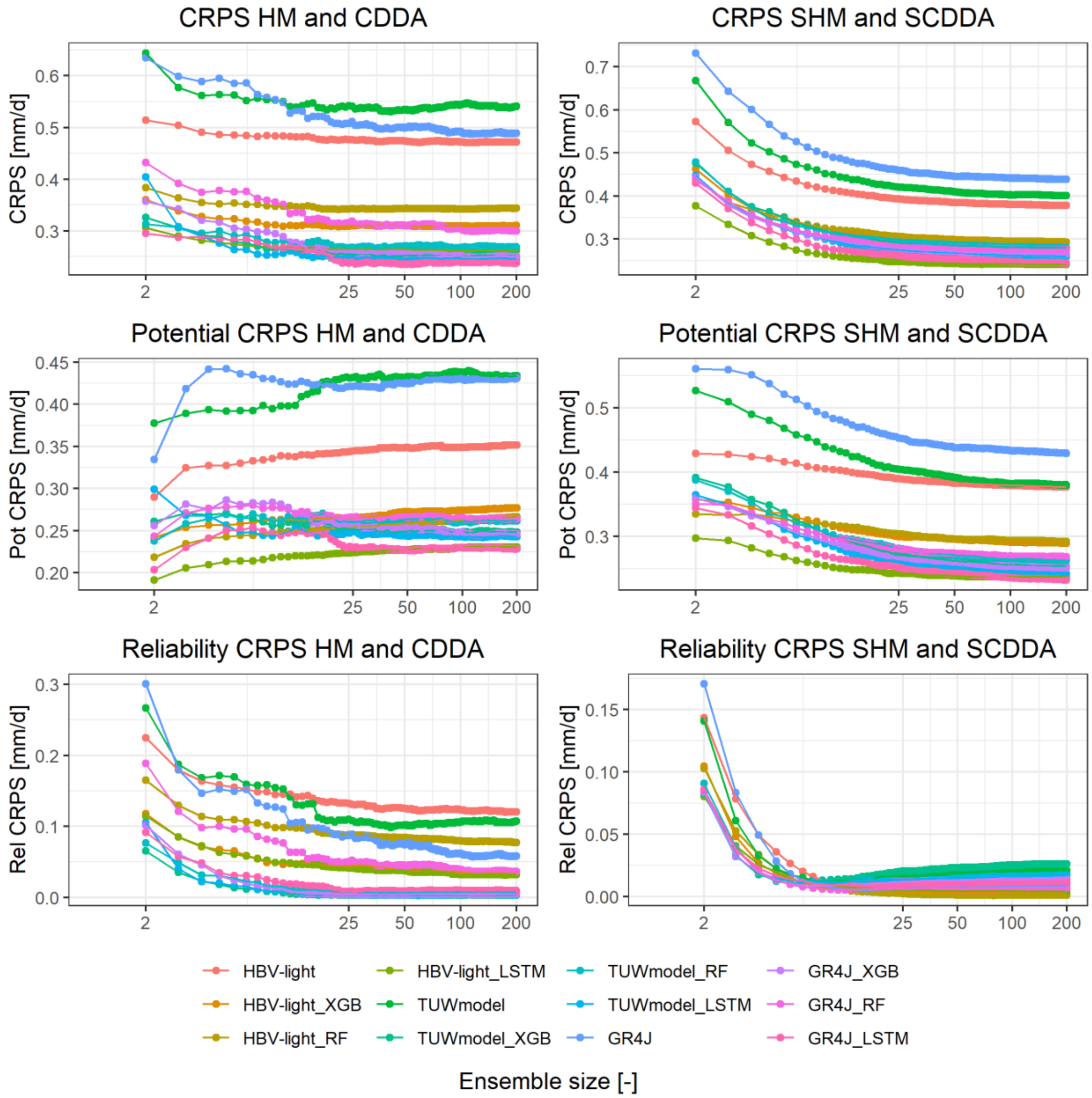

Figure 15. CRPS, potential CRPS (Pot CRPS), and reliability CRPS (Rel CRPS) vs. ensemble size for the ensemble HM (HBV-light, TUWmodel, and GR4J), CDDA (XGB, RF, and LSTM) (left), and their stochastic counterparts (SHM and SCDDA) (right) for the Dünnern catchment for the test set.

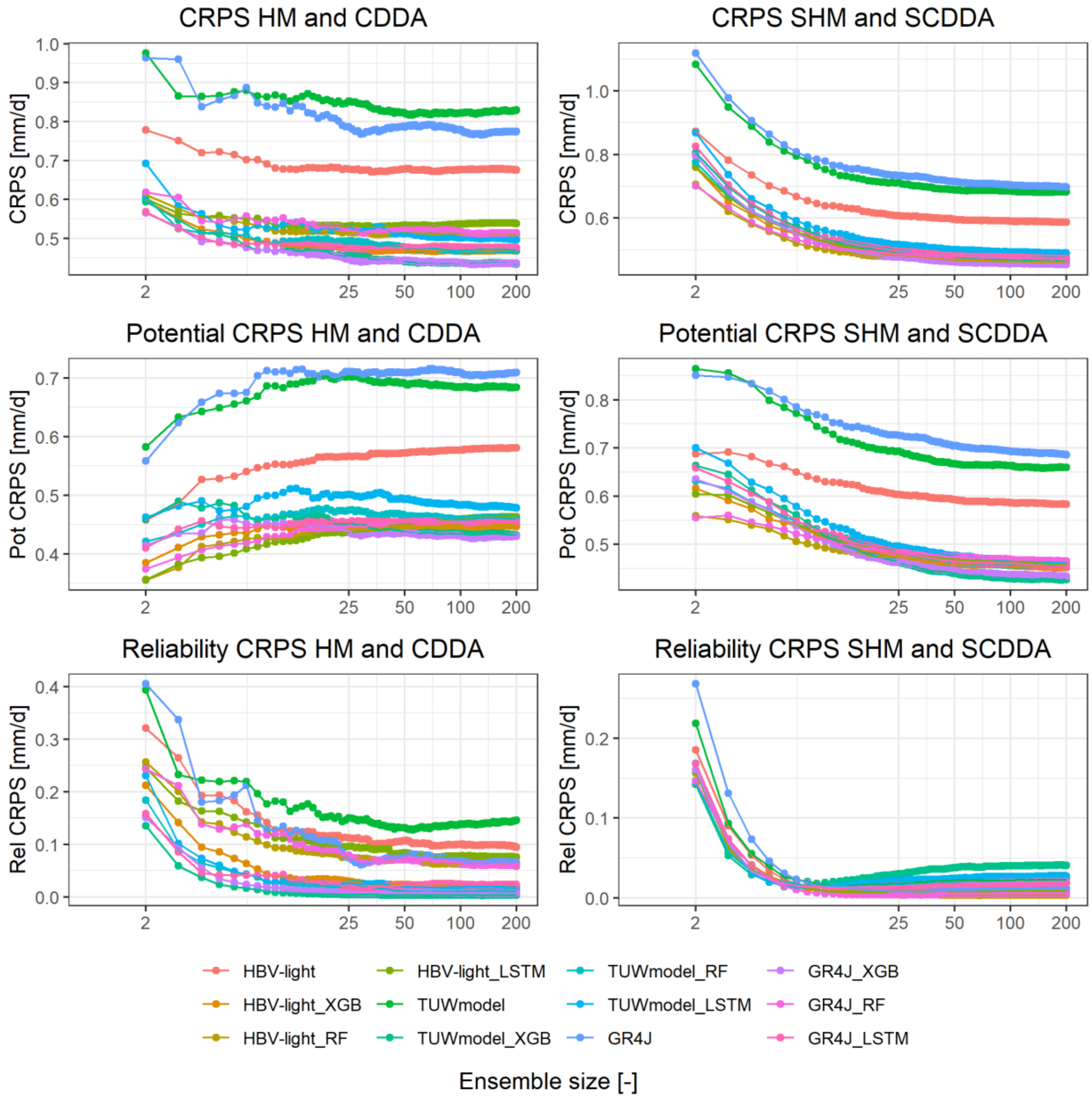

Figure 16. CRPS, potential CRPS (Pot CRPS), and reliability CRPS (Rel CRPS) vs. ensemble size for the HM (GR4J, HBV-light, TUWmodel), CDDA (XGB, RF, and LSTM) (left), and their stochastic counterparts (SHM and SCDDA) (right) for the Kleine-Emme catchment for the test set.

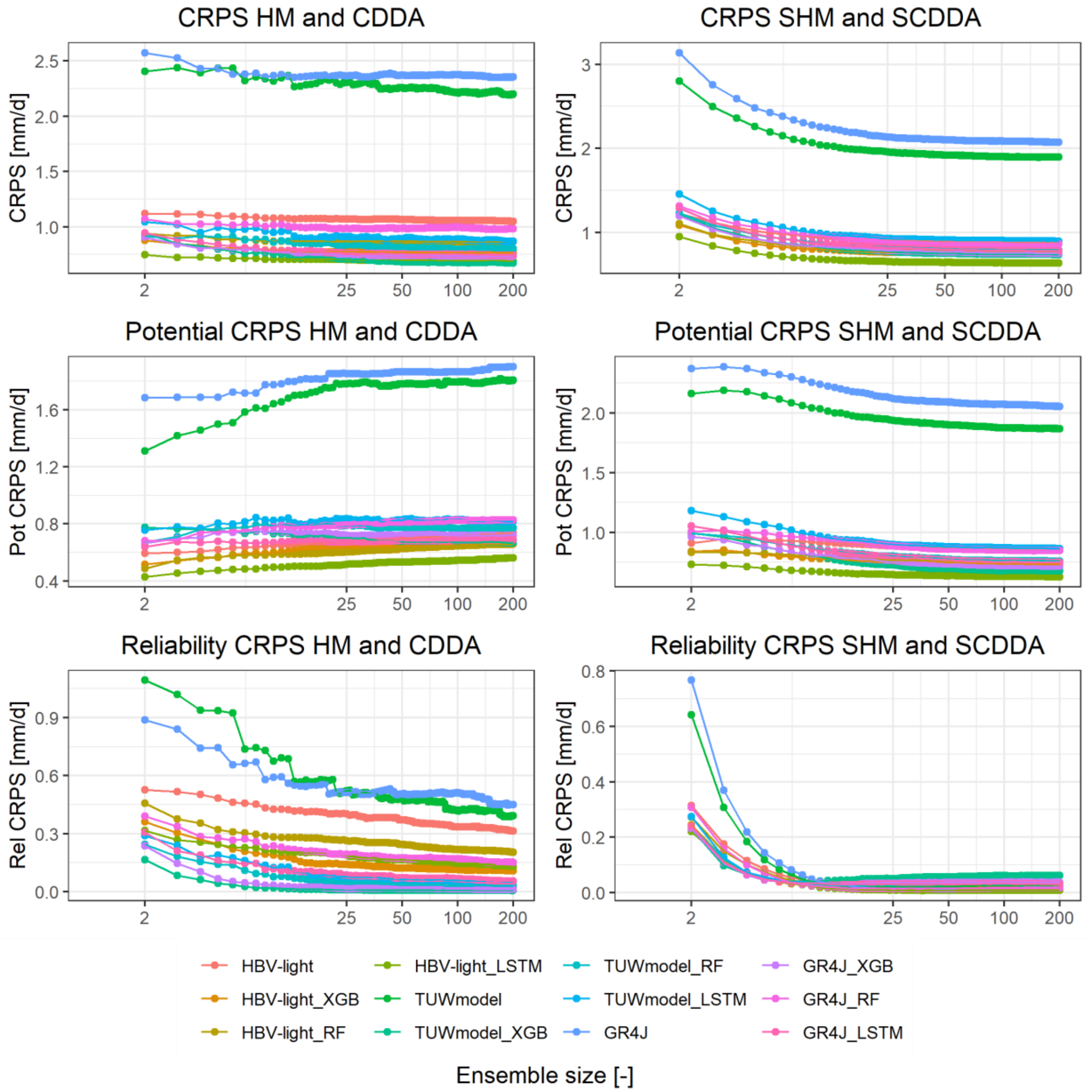

Figure 17. CRPS, potential CRPS (Pot CRPS), reliability CRPS (Rel CRPS) vs. ensemble size for the ensemble HM (HBV-light, TUWmodel, and GR4J), CDDA (XGB, RF, and LSTM) (left), and their stochastic counterparts (SHMs and SCDDA) (right) for the Muota catchment for the test set.

It appears that most models stabilize at approximately 100 ensemble members according to the potential CRPS of the ensemble HM, CDDA, and the stochastic models in all catchments. However, the potential CRPS of some models may continue to improve beyond 200 members, for example, the stochastic HM using GR4J in Dünnern and Kleine-Emme catchments, although the improvement may not be significant. Comparing the results of the potential CRPS with the AW from Figure 13, it appears that the increase in the AW with an increasing number of ensemble members does not meaningfully impact the potential CRPS beyond 100 members. Next, analyzing the reliability CRPS shows that, for most models, it stabilizes around 100 ensemble members, similar to the α_R. Furthermore, the shape of the α_R curves in Figure 14 closely match the reliability CRPS curves for the stochastic HM and SCDDA. Although the potential CRPS and reliability CRPS of some models in Figure 15-17 do not appear to stabilize by 200 members, the overall CRPS suggests that most models provide a stable estimation of simulation uncertainty by 100 members, showing little improvement in performance beyond this point. Therefore, the analysis of ensemble size versus model performance indicates that approximately 100 members are required for a stable estimation of simulation uncertainty in the study catchments.

It is important to recognize that the CRPS decomposition also estimates how much the sharpness and reliability of the simulations contribute to the overall CRPS. In detail, for most ensemble models (ensemble HM and CDDA), the reliability CRPS is a significant portion of the overall CRPS. In contrast, the potential CRPS contributes the most to the overall CRPS for the stochastic models (stochastic HM and SCDDA), with a minimal contribution coming from the reliability CRPS. Comparing the overall CRPS of CDDA and SCDDA variants, it was previously determined that the SCDDA could improve upon (or at least maintain) the

CRPS of its CDDA counterpart that is less reliable. Thus, it appears that the stochastic framework approach tends to convert the ensemble models (ensemble HM and CDDA) into more reliable models at the cost of simulation sharpness. This result enhances the previous conclusion that the SCDDA tends to generate simulations that are more conservative while improving the reliability of its (less reliable) CDDA counterpart.

5.5 Effect on Model Performance Using a Snow Module in GR4J

Among the three HMs, GR4J does not incorporate a snow routine, although snow processes play a significant role in the hydrology of two study catchments (Kleine-Emme and Muota). Not including snow processes in an HM can be viewed as an error of the perceptual model (see Section 2.1.2), where deciding on the important hydrological processes is the first step of modelling and iteratively refined until the model is deemed to be satisfactory (Beven, 2012). However, since the CDDA and SCDDA are used to correct the residuals of the ensemble HM, they may also implicitly account for processes not included in the HM (e.g., overcoming the need to couple a snow module with GR4J).

In what follows, GR4J is used to test whether a similar performance could be achieved by the CDDA and SCDDA when an important hydrological process is and is not explicitly accounted for in the model. Therefore, the CemaNeige (Valery, 2010) snow module is coupled with GR4J (GR4JCN) and compared against GR4J (without the snow module) using the CDDA and SCDDA variants. With the same calibration procedure as GR4J (see Section 4.2), the CRPS for GR4JCN in Dünnern, Kleine-Emme, and Muota catchments is estimated as 0.47, 0.68, and 2.22, representing approximately a 4%, 12%, and 6% improvement over

GR4J, respectively, on the test set. The minimal improvement in KGE with the snow routine may be caused by BO unable to find suitable parameter sets with the current setup. Since Kleine-Emme showed the most substantial improvement in performance using GR4JCN, this catchment was chosen to evaluate any differences in performance (KGE, α_R, and CRPS) between the GR4J and GR4JCN CDDA and SCDDA variants. In Table 7, the test set performance for the ensemble and stochastic HMs using GR4JCN as well as GR4JCN CDDA and SCDDA variants is summarized for the Kleine-Emme catchment

Table 7. Performance of ensemble and stochastic HMs using GR4JCN as well as the GR4JCN CDDA and SCDDA variants for the Kleine-Emme catchment.

Models	KGE	α_R	CRPS
Ensemble HM	0.81	0.88	0.68
Stochastic HM	0.77	0.89	0.64
XGB CDDA	0.88	0.90	0.46
XGB SCDDA	0.88	0.85	0.46
RF CDDA	0.88	0.77	0.52
RF SCDDA	0.89	0.88	0.48
LSTM CDDA	0.86	0.87	0.47
LSTM SCDDA	0.83	0.84	0.48

It is possible to identify that GR4J and GR4JCN CDDA and SCDDA variants provide similar results when comparing Table 7 with Table 5 and Table 6. Considering all three DDMs, the difference in CRPS for GR4J and GR4JCN CDDA and SCDDA variants is 0.02 or less. Similarly, the α_R differs by 0.03 or less, while the deterministic KGE is the same. Therefore, it appears that the CDDA and SCDDA have the potential to correct for hydrological processes absent from the HM. This result is aligned with Lees et al. (2022) where it has shown that the LSTM is able to reproduce hydrological processes from data (historical streamflow, meteorological variables, and catchment attributes). Thus, it appears that XGB and RF also share this ability. However, to generalize this finding, a much larger experiment with additional catchments, HM parameter optimization methods, and more flexible HMs where model structure complexity can be closely controlled (e.g., Raven (Craig et al., 2020)) should be considered.

Chapter 6

Conclusions, Future Work, and Recommendations

The conceptual data-driven approach (CDDA) can improve the simulations of ensemble hydrological models (HMs) by correcting their residuals/errors using data-driven models (DDMs). The research introduces a new stochastic CDDA (SCDDA) that can account for additional sources of uncertainty not considered in the CDDA (e.g., model output uncertainty). Here, the new SCDDA is tested using nine HM-DDM combinations (three HMs and three DDMs) and compared against the CDDA as well as ensemble and stochastic HMs for daily streamflow simulation in three Swiss catchments. The models are evaluated using several (deterministic and probabilistic) metrics and graphical aids (time series plots, raincloud plots, etc.). The coverage probability plot (CPP) is proposed as a diagnostic tool for predicting when the out-of-sample reliability of the ensemble models (ensemble HMs and CDDA) can be improved by the stochastic framework. Experiments showed that the new SCDDA could significantly improve the ensemble HM simulations across most performance metrics with improvements in the mean continuous ranked probability score (CRPS) of 27-68%. While the SCDDA improved upon the CRPS of the CDDA by up to 15%, it did not consistently outperform the CDDA. The CPPs showed that ensemble HMs with narrow simulations tended to result in CDDAs with narrow simulations, and unreliable ensemble models (CDDA and ensemble HMs) were improved using the stochastic framework. Meanwhile, CDDA variants with high reliability (α_R > 0.85) had SCDDA counterparts with lower reliability. Regardless, all SCDDA variants had reliable and/or conservative simulations, making them a valuable tool for decision-making. Studying probabilistic performance as a function of ensemble size (number of ensemble members) revealed that an ensemble size of 100 members led to stable performance. In one of the snow-dominated catchments (Kleine-Emme), the HM without a snow routine led to lower performance than when it was included. However, neglecting the snow module had no discernible impact on the deterministic performance and negligible impact on the probabilistic performance of the CDDA and SCDDA, indicating that both approaches have the potential to account for missing processes in HMs.

The new SCDDA, in conjunction with the CPP, can benefit hydrologists and water resource practitioners in several respects. First, hydrological modellers that already have access to ensemble HMs can use the new SCDDA to improve the predictive capabilities of their model. The new SCDDA can be especially useful where reliable and/or conservative simulations are required. If users are hesitant to apply the stochastic framework (e.g., for hydrological simulation along large river networks), the CPP can be checked for the CDDA to see if implementing the new SCDDA is worth the computational investment. However, if it is decided by the modeller or user that DDMs are too computationally demanding, an ensemble HM can easily be converted to a stochastic HM using the stochastic framework and (potentially) achieve levels of reliability similar to the new SCDDA. Finally, although not explored in this work, the new SCDDA is not limited to streamflow simulation and can be paired with other HMs to help address diverse problems related to geochemistry, land use effects, channel hydraulics, etc.

Given the flexibility of the SCDDA, it is possible to identify several potential improvements for the framework. As the DDMs are not restricted to the input variables used in this work, performance may be improved using other variables generated by the HM, such as actual evaporation and soil moisture. Furthermore, other HM structures and optimization algorithms may enhance the SCDDA simulations. For example, future work may involve using a blended model structure (Mai et al., 2020) with simultaneous calibration of structure and parameters (Chlumsky et al., 2021) within SCDDA and testing its efficacy in diverse catchments (Newman et al., 2014). In future studies, BO could also be used (alongside other optimization algorithms) for jointly estimating HM and DDM parameters by optimizing HM parameters and DDM hyper-parameters simultaneously. If successful, this approach may significantly reduce the computation time required for the proposed SCDDA. A set of recommendations are included below for improving upon the new SCDDA and generalizing the results obtained in this work:

1. *Large-scale experiments*: with the rising availability of big datasets in hydrology (Addor et al., 2017; Alvarez-Garreton et al., 2018; Arsenault et al., 2016; Chagas et al., 2020; Coxon et al., 2020; Fowler et al., 2021), it is possible to improve the performance of the new SCDDA by aggregating hydro-meteorological data from multiple catchments along with their catchment attributes. However, implementing the new SCDDA within large-scale experiments may significantly increase the computational requirements and restrict the DDM candidates to those that can be trained in batches (e.g., LSTM; see, for example, Klotz et al., 2022). However, once the HMs and DDMs have been calibrated/trained (offline mode), the new SCDDA is much faster when running in simulation/online mode. Using big datasets spanning specific regions or the globe, the new SCDDA could be designed as a regional or global model. In such cases, the HM and DDM parameters could be used to (generally) represent regional or global hydrological processes, while the model error

could be used to tailor the model simulations to specific locations (e.g., a catchment outlet of interest).

2. *Online estimation of model error*: the KNN method for stochastic resampling of model errors inherently account for heteroscedasticity in model error. However, by applying the KNN method to rolling or growing time-windows (i.e., shifting or increasing (the size of) the validation set with each new measurement of the input variables), environmental changes could be better accounted for when estimating the conditional probability density function of the model error. In this way, the assumption of stationarity of the model error could be relaxed.
3. *Online estimation of DDM parameter uncertainty:* instead of considering the DDM parameters to be static quantities, online machine learning methods, such as online recurrent extreme learning machines (Park & Kim, 2017), could be used to relax the assumption of stationarity of the model parameters.
4. *Forecasting with ensemble meteorological forcings*: by using meteorological forecasts (e.g., precipitation, potential evapotranspiration) as input to new SCDDA, it can easily be converted into a forecasting framework. Since many operational meteorological forecasting products output ensembles (e.g., Global Ensemble Forecast System (Zhou et al., 2022)), the new SCDDA can naturally include this information as a form of input data uncertainty.
5. *Augmenting the input variable set*: only a small number of potential input variables (see Section 3.1) were considered in the new SCDDA. However, given that DDMs can accept various input variables (e.g., numerical weather predictions from different products, time-based indices to reflect seasonality, state variables from HMs), it is

highly recommended that all potentially useful inputs be available to the modeller be considered when building models using the new SCDDA. If the number of input variables is exceedingly large, different feature extraction (e.g., variational autoenconders; Lopez-Alvis et al., 2022) or input variable selection methods (e.g., conditional mutual information; Quilty et al., 2016) could be used to reduce the number of potential input variables before the new SCDDA is implemented.

By implementing the above recommendations, the new SCDDA can be further improved and used as a hydrological modelling tool to help address water resources planning, management, and operational issues, at local, regional, and global scales.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Yangqing, J., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences 21, 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Adombi, A.V.D.P., Chesnaux, R., Boucher, M.A., 2021. Review: Theory-guided machine learning applied to hydrogeology—state of the art, opportunities and future challenges. Hydrogeology Journal 1–13. https://doi.org/10.1007/S10040-021-02403-2/TABLES/2

Alizadeh, B., Ghaderi Bafti, A., Kamangir, H., Zhang, Y., Wright, D.B., Franz, K.J., 2021. A novel attention-based LSTM cell post-processor coupled with bayesian optimization for streamflow prediction. Journal of Hydrology 601, 126526. https://doi.org/10.1016/J.JHYDROL.2021.126526

Allen, M., Poggiali, D., Whitaker, K., Marshall, T.R., Kievit, R.A., 2019. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Research 4. https://doi.org/10.12688/WELLCOMEOPENRES.15191.1

Althoff, D., Rodrigues, L.N., 2021. Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment. Journal of Hydrology 600, 126674. https://doi.org/10.1016/J.JHYDROL.2021.126674

Alvarez-Garreton, C., Mendoza, P.A., Pablo Boisier, J., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., Ayala, A., 2018. The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies-Chile dataset. Hydrology and Earth System Sciences 22, 5817–5846. https://doi.org/10.5194/HESS-22-5817-2018

Anand, J., Gosain, A.K., Khosa, R., Srinivasan, R., 2018. Regional scale hydrologic modeling for prediction of water balance, analysis of trends in streamflow and variations in streamflow: The case study of the Ganga River basin. Journal of Hydrology: Regional Studies 16, 32–53. https://doi.org/10.1016/J.EJRH.2018.02.007

Arsenault, R., Bazile, R., Ouellet Dallaire, C., Brissette, F., 2016. CANOPEX: A Canadian hydrometeorological watershed database. Hydrological Processes 30, 2734–2736. https://doi.org/10.1002/HYP.10880

Bergström, S., Graham, L.P., 1998. On the scale problem in hydrological modelling. Journal of Hydrology 211, 253–265. https://doi.org/10.1016/S0022-1694(98)00248-0

Beven, K., 2020. Deep learning, hydrological processes and the uniqueness of place. Hydrological Processes 34, 3608–3613. https://doi.org/10.1002/HYP.13805

Beven, K., 2012. Rainfall-Runoff Modelling: The Primer. John Wiley & Sons.

Beven, K., 2009. Environmental Modelling: An Uncertain Future? Routledge, New York.

Beven, K., 2006. A manifesto for the equifinality thesis. Journal of Hydrology 320, 18–36. https://doi.org/10.1016/J.JHYDROL.2005.07.007

Beven, K., 2001. How far can we go in distributed hydrological modelling? Hydrology and Earth System Sciences 5, 1–12. https://doi.org/10.5194/HESS-5-1-2001

Beven, K., 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Advances in Water Resources 16, 41–51. https://doi.org/10.1016/0309-1708(93)90028-E

Beven, K., Binley, A., 1992. The future of distributed models: Model calibration and uncertainty prediction. Hydrological Processes 6, 279–298. https://doi.org/10.1002/HYP.3360060305

Beven, K., Young, P., 2013. A guide to good practice in modeling semantics for authors and referees. Water Resources Research 49, 5092–5098. https://doi.org/10.1002/WRCR.20393

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer, New York. https://doi.org/10.1117/1.2819119

Blöschl, G., 2005. On the Fundamentals of Hydrological Sciences, in: Encyclopedia of Hydrological Sciences. John Wiley & Sons, Ltd, Chichester, UK. https://doi.org/10.1002/0470848944.hsa001a

Booker, D.J., Woods, R.A., 2014. Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. Journal of Hydrology 508, 227–239. https://doi.org/10.1016/J.JHYDROL.2013.11.007

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140. https://doi.org/10.1007/bf00058655

Brown, G., Pocock, A., Zhao, M.-J., Lujan, M., 2012. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection Ming-Jie Zhao Mikel Luján. Journal of Machine Learning Research 13, 27–66.

Brown, R.R., Keath, N., Wong, T.H.F., 2009. Urban water management in cities: historical, current and future regimes. Water Science and Technology 59, 847–855. https://doi.org/10.2166/WST.2009.029

Chagas, V.B.P., L. B. Chaffe, P., Addor, N., M. Fan, F., S. Fleischmann, A., C. D. Paiva, R., Siqueira, V.A., 2020. CAMELS-BR: Hydrometeorological time series and landscape attributes for 897 catchments in Brazil. Earth System Science Data 12, 2075–2096. https://doi.org/10.5194/ESSD-12-2075-2020

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., 2021. xgboost: Extreme Gradient Boosting.

Chen, X., Huang, J., Han, Z., Gao, H., Liu, M., Li, Z., Liu, X., Li, Q., Qi, H., Huang, Y., 2020. The importance of short lag-time in the runoff forecasting model based on long short-term memory. Journal of Hydrology 589, 125359. https://doi.org/10.1016/J.JHYDROL.2020.125359

Chen, Y., Huang, D., Zhang, D., Zeng, J., Wang, N., Zhang, H., Yan, J., 2021. Theory-guided hard constraint projection (HCP): A knowledge-based data-driven scientific machine learning method. Journal of Computational Physics 445, 110624. https://doi.org/10.1016/J.JCP.2021.110624

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M.M., Xie, W., Rosen, G.L., Lengerich, B.J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A.E., Shrikumar, A., Xu, J., Cofer, E.M., Lavender, C.A., Turaga, S.C., Alexandari, A.M., Lu, Z., Harris, D.J., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L.K., Segler, M.H.S., Boca, S.M., Swamidass, S.J., Huang, A., Gitter, A., Greene, C.S., 2018. Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface 15, 20170387. https://doi.org/10.1098/rsif.2017.0387

Chlumsky, R., Mai, J., Craig, J.R., Tolson, B.A., 2021. Simultaneous Calibration of Hydrologic Model Structure and Parameters Using a Blended Model. Water Resources Research 57, e2020WR029229. https://doi.org/10.1029/2020WR029229

Cho, K., Kim, Y., 2022. Improving streamflow prediction in the WRF-Hydro model with LSTM networks. Journal of Hydrology 605, 127297. https://doi.org/10.1016/J.JHYDROL.2021.127297

Chollet, F., 2015. Keras.

Chow, V. te, Maidment, D.R., Mays, L.W., 1988. Applied Hydrology. McGraw-Hill, New York.

Coron, L., Delaigue, O., Thirel, G., Perrin, D.D. and C., Michel, C., 2022. airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling. R News. https://doi.org/10.15454/EX11NA

Coron, L., Thirel, G., Delaigue, O., Perrin, C., Andréassian, V., 2017. The Suite of Lumped GR Hydrological Models in an R package. Environmental Modelling and Software 94, 166–171. https://doi.org/10.1016/j.envsoft.2017.05.002

Coxon, G., Addor, N., Bloomfield, J.P., Freer, J., Fry, M., Hannaford, J., Howden, N.J.K., Lane, R., Lewis, M., Robinson, E.L., Wagener, T., Woods, R., 2020. CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. Earth System Science Data 12, 2459–2483. https://doi.org/10.5194/ESSD-12-2459-2020

Craig, J.R., Brown, G., Chlumsky, R., Jenkinson, R.W., Jost, G., Lee, K., Mai, J., Serrer, M., Sgro, N., Shafii, M., Snowdon, A.P., Tolson, B.A., 2020. Flexible watershed simulation with the Raven hydrological modelling framework. Environmental Modelling & Software 129, 104728. https://doi.org/10.1016/J.ENVSOFT.2020.104728

Darbandsari, P., Coulibaly, P., 2020. Inter-comparison of lumped hydrological models in data-scarce watersheds using different precipitation forcing data sets: Case study of Northern Ontario, Canada. Journal of Hydrology: Regional Studies 31, 100730. https://doi.org/10.1016/J.EJRH.2020.100730

Daw, A., Thomas, R.Q., Carey, C.C., Read, J.S., Appling, A.P., Karpatne, A., 2020. Physics-guided architecture (pga) of neural networks for quantifying uncertainty in lake temperature modeling. Proceedings of the 2020 SIAM International Conference on Data Mining, SDM 2020 532–540.

Dawson, C.W., Wilby, R.L., 2001. Hydrological modelling using artificial neural networks. Progress in Physical Geography 25, 80–108. https://doi.org/10.1177/030913330102500104 de Coste, M., Li, Z., Dibike, Y., 2022. Assessing and predicting the severity of mid-winter breakups based on Canada-wide river ice data. Journal of Hydrology 607, 127550. https://doi.org/10.1016/J.JHYDROL.2022.127550

Deng, H., 2013. Guided Random Forest in the RRF Package.

Deng, H., Runger, G., 2013. Gene selection with guided regularized random forest. Pattern Recognition 46, 3483–3489. https://doi.org/10.1016/j.patcog.2013.05.018

Depina, I., Jain, S., Mar Valsson, S., Gotovac, H., 2021. Application of physics-informed neural networks to inverse problems in unsaturated groundwater flow. Georisk 16, 21–36. https://doi.org/10.1080/17499518.2021.1971251

Desai, S., Ouarda, T.B.M.J., 2021. Regional hydrological frequency analysis at ungauged sites with random forest regression. Journal of Hydrology 594, 125861. https://doi.org/10.1016/j.jhydrol.2020.125861

Devi, G.K., Ganasri, B.P., Dwarakish, G.S., 2015. A Review on Hydrological Models. Aquatic Procedia 4, 1001–1007. https://doi.org/10.1016/J.AQPRO.2015.02.126

Domeneghetti, A., Castellarin, A., Brath, A., 2012. Assessing rating-curve uncertainty and its effects on hydraulic model calibration. Hydrology and Earth System Sciences 16, 1191–1202. https://doi.org/10.5194/hess-16-1191-2012

Dong, J., Zeng, W., Lei, G., Wu, L., Chen, H., Wu, J., Huang, J., Gaiser, T., Srivastava, A.K., 2022. Simulation of dew point temperature in different time scales based on grasshopper algorithm optimized extreme gradient boosting. Journal of Hydrology 606, 127452. https://doi.org/10.1016/J.JHYDROL.2022.127452

Du, M., Liu, N., Hu, X., 2019. Techniques for interpretable machine learning. Commun ACM 63, 68–77. https://doi.org/10.1145/3359786

Eslamian, S., 2014. Handbook of Engineering Hydrology Modeling, Climate Change, and Variability. CRC Press.

Fan, H., Jiang, M., Xu, L., Zhu, H., Cheng, J., Jiang, J., 2020. Comparison of Long Short Term Memory Networks and the Hydrological Model in Runoff Simulation. Water 2020, Vol. 12, Page 175 12, 175. https://doi.org/10.3390/W12010175

Feng, D., Fang, K., Shen, C., 2020. Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales. Water Resources Research 56, e2019WR026793. https://doi.org/10.1029/2019WR026793

Feng, D., Lawson, K., Shen, C., 2021. Mitigating Prediction Error of Deep Learning Streamflow Models in Large Data-Sparse Regions With Ensemble Modeling and Soft Data. Geophysical Research Letters 48, e2021GL092999. https://doi.org/10.1029/2021GL092999

Feng, D., Liu, J., Lawson, K., Shen, C., 2022. Differentiable, learnable, regionalized process-based models with physical outputs can approach state-of-the-art hydrologic prediction accuracy. https://doi.org/10.48550/arxiv.2203.14827

Fowler, K.J.A., Acharya, S.C., Addor, N., Chou, C., Peel, M.C., 2021. CAMELS-AUS: Hydrometeorological time series and landscape attributes for 222 catchments in Australia. Earth System Science Data 13, 3847–3867. https://doi.org/10.5194/ESSD-13-3847-2021

Frame, J.M., Kratzert, F., Raney, A., Rahman, M., Salas, F.R., Nearing, G.S., 2021. Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics. JAWRA Journal of the American Water Resources Association, 57(6), 885-905. https://doi.org/10.1111/1752-1688.12964

Gaborit, É., Fortin, V., Tolson, B., Fry, L., Hunter, T., Gronewold, A.D., 2017. Great Lakes Runoff Inter-comparison Project, phase 2: Lake Ontario (GRIP-O). Journal of Great Lakes Research 43, 217–227. https://doi.org/10.1016/J.JGLR.2016.10.004

Gauch, M., Mai, J., Lin, J., 2021. The proper care and feeding of CAMELS: How limited training data affects streamflow prediction. Environmental Modelling & Software 135, 104926. https://doi.org/10.1016/J.ENVSOFT.2020.104926

Ghaith, M., Siam, A., Li, Z., El-Dakhakhni, W., 2019. Hybrid Hydrological Data-Driven Approach for Daily Streamflow Forecasting. Journal of Hydrologic Engineering 25, 04019063. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001866

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102, 359–378. https://doi.org/10.1198/016214506000001437

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

Han, H., 2021. Improving hydrologic modeling of runoff processes using data-driven models. Colorado State University. Libraries. https://doi.org/10.17616/R31NJMSY

Han, S., Coulibaly, P., 2019. Probabilistic Flood Forecasting Using Hydrologic Uncertainty Processor with Ensemble Weather Forecasts. Journal of Hydrometeorology 20, 1379–1398. https://doi.org/10.1175/JHM-D-18-0251.1

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-84858-7

He, Q.Z., Barajas-Solano, D., Tartakovsky, G., Tartakovsky, A.M., 2020. Physics-informed neural networks for multiphysics data assimilation with application to subsurface transport. Advances in Water Resources 141, 103610. https://doi.org/10.1016/J.ADVWATRES.2020.103610

He, Q.Z., Tartakovsky, A.M., 2021. Physics-Informed Neural Network Method for Forward and Backward Advection-Dispersion Equations. Water Resources Research 57, e2020WR029479. https://doi.org/10.1029/2020WR029479

Head, T., Kumar, M., Nahrstaedt, H., Louppe, G., Shcherbatyi, I., 2021. scikit-optimize. https://doi.org/10.5281/ZENODO.5565057

Heaton, J., 2016. An empirical analysis of feature engineering for predictive modeling. Conference Proceedings - IEEE SOUTHEASTCON 2016-July. https://doi.org/10.1109/SECON.2016.7506650

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting 15. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Herschy, R.W., 2008. Streamflow Measurement. CRC Press, London. https://doi.org/10.1201/9781482265880

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

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., Hochreiter, S., Klambauer, G., 2021. MC-LSTM: Mass-Conserving LSTM, in: Proceedings of the 38th International Conference on Machine Learning.

Hurst, H.E., 1951. Long-Term Storage Capacity of Reservoirs. Transactions of the American Society of Civil Engineers 116, 770–799. https://doi.org/10.1061/TACEAT.0006518

Iliopoulou, T., Papalexiou, S.M., Markonis, Y., Koutsoyiannis, D., 2018. Revisiting long-range dependence in annual precipitation. Journal of Hydrology 556, 891–900. https://doi.org/10.1016/J.JHYDROL.2016.04.015

Jain, Sharad Kumar, Mani, P., Jain, Sanjay K., Prakash, P., Singh, V.P., Tullos, D., Kumar, S., Agarwal, S.P., Dimri, A.P., 2018. A Brief review of flood forecasting techniques and their applications. https://doi.org/10.1080/15715124.2017.1411920 16, 329–344. https://doi.org/10.1080/15715124.2017.1411920

Jha, D., Ward, L., Paul, A., Liao, W., Choudhary, A., Wolverton, C., Agrawal, A., 2018. ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. Scientific Reports 8, 1–13. https://doi.org/10.1038/s41598-018-35934-y

Jia, X., Zwart, J., Sadler, J., Appling, A., Oliver, S., Markstrom, S., Willard, J., Xu, S., Steinbach, M., Read, J., Kumar, V., 2021. Physics-Guided Recurrent Graph Model for Predicting Flow and Temperature in River Networks, in: Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). Society for Industrial and Applied Mathematics, pp. 612–620. https://doi.org/10.1137/1.9781611976700.69

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Transactions on Knowledge and Data Engineering 29, 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

Kim, K.B., Kwon, H.H., Han, D., 2018. Exploration of warm-up period in conceptual hydrological modelling. Journal of Hydrology 556, 194–210. https://doi.org/10.1016/J.JHYDROL.2017.11.015

Kingma, D.P., Ba, J.L., 2015. Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. https://doi.org/10.48550/arxiv.1412.6980

Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrological Sciences Journal 31, 13–24. https://doi.org/10.1080/02626668609491024

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., Nearing, G., 2022. Uncertainty estimation with deep learning for rainfall–runoff modeling. Hydrology and Earth System Sciences 26, 1673–1693. https://doi.org/10.5194/hess-26-1673-2022

Konapala, G., Kao, S.C., Painter, S.L., Lu, D., 2020. Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. Environmental Research Letters 15, 104022. https://doi.org/10.1088/1748-9326/ABA927

Konapala, G., Mishra, A., 2020. Quantifying Climate and Catchment Control on Hydrological Drought in the Continental United States. Water Resources Research 56, e2018WR024620. https://doi.org/10.1029/2018WR024620

Koutsoyiannis, D., 2021. Stochastics of Hydroloclimatic Extremes A Cool Look at Risk. Kallipos, Athens.

Koutsoyiannis, D., 2020. Revisiting the global hydrological cycle: Is it intensifying? Hydrology and Earth System Sciences 24, 3899–3932. https://doi.org/10.5194/HESS-24-3899-2020

Koutsoyiannis, D., Montanari, A., 2022. Bluecat: A Local Uncertainty Estimator for Deterministic Simulations and Predictions. Water Resources Research e2021WR031215. https://doi.org/10.1029/2021WR031215

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrology and Earth System Sciences 22, 6005–6022. https://doi.org/10.5194/HESS-22-6005-2018

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019. Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. Water Resources Research 55, 11344–11354. https://doi.org/10.1029/2019WR026065

Kumanlioglu, A.A., Fistikoglu, O., 2019. Performance Enhancement of a Conceptual Hydrological Model by Integrating Artificial Intelligence. Journal of Hydrologic Engineering 24, 04019047. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001850

Kunnath-Poovakka, A., Eldho, T.I., 2019. A comparative study of conceptual rainfall-runoff models GR4J, AWBM and Sacramento at catchments in the upper Godavari river basin, India. Journal of Earth System Science 2019 128:2 128, 1–15. https://doi.org/10.1007/S12040-018-1055-8

Kurian, C., Sudheer, K.P., Vema, V.K., Sahoo, D., 2020. Effective flood forecasting at higher lead times through hybrid modelling framework. Journal of Hydrology 587, 124945. https://doi.org/10.1016/J.JHYDROL.2020.124945

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., de Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., Dadson, S.J., 2022. Hydrological concept formation inside long short-term memory (LSTM) networks. Hydrology and Earth System Sciences 26, 3079–3101. https://doi.org/10.5194/hess-26-3079-2022

Li, D., Marshall, L., Liang, Z., Sharma, A., Zhou, Y., 2021a. Bayesian LSTM With Stochastic Variational Inference for Estimating Model Uncertainty in Process-Based Hydrological Models. Water Resources Research 57, e2021WR029772. https://doi.org/10.1029/2021WR029772

Li, D., Marshall, L., Liang, Z., Sharma, A., Zhou, Y., 2021b. Characterizing distributed hydrological model residual errors using a probabilistic long short-term memory network. Journal of Hydrology 603, 126888. https://doi.org/10.1016/J.JHYDROL.2021.126888

Liang, J., Li, W., Bradford, S.A., Šimůnek, J., 2019. Physics-Informed Data-Driven Models to Predict Surface Runoff Water Quantity and Quality in Agricultural Fields. Water 2019, Vol. 11, Page 200 11, 200. https://doi.org/10.3390/W11020200

Lindström, G., Bergström, S., 1992. Improving the HBV and PULSE-models by use of temperature anomalies. Vannet i Norden.

Liu, J., Yuan, X., Zeng, J., Jiao, Y., Li, Y., Zhong, L., Yao, L., 2022. Ensemble streamflow forecasting over a cascade reservoir catchment with integrated hydrometeorological modeling and machine learning. Hydrology and Earth System Sciences 26, 265–278. https://doi.org/10.5194/HESS-26-265-2022

Liu, K., Song, C., 2022. Modeling lake bathymetry and water storage from DEM data constrained by limited underwater surveys. Journal of Hydrology 604, 127260. https://doi.org/10.1016/J.JHYDROL.2021.127260

Lopez-Alvis, J., Nguyen, F., Looms, M.C., Hermans, T., 2022. Geophysical Inversion Using a Variational Autoencoder to Model an Assembled Spatial Prior Uncertainty. Journal of Geophysical Research: Solid Earth 127. https://doi.org/10.1029/2021JB022581

Lu, D., Konapala, G., Painter, S.L., Kao, S.C., Gangrade, S., 2021. Streamflow Simulation in Data-Scarce Basins Using Bayesian and Physics-Informed Machine Learning Models. Journal of Hydrometeorology 22, 1421–1438. https://doi.org/10.1175/JHM-D-20-0082.1

Luxburg, U. von, Schölkopf, B., 2011. Statistical Learning Theory: Models, Concepts, and Results. Handbook of the History of Logic 10, 651–706. https://doi.org/10.1016/B978-0-444-52936-7.50016-1

Ma, J., Rao, K., Li, R., Yang, Y., Li, W., Zheng, H., 2022. Improved Hadoop-based cloud for complex model simulation optimization: Calibration of SWAT as an example. Environmental Modelling & Software 149, 105330. https://doi.org/10.1016/J.ENVSOFT.2022.105330

Ma, J., Zhang, J., Li, R., Zheng, H., Li, W., 2021. Using Bayesian optimization to automate the calibration of complex hydrological models: Framework and application. Environmental Modelling & Software 105235. https://doi.org/10.1016/J.ENVSOFT.2021.105235

Ma, M., Zhao, G., He, B., Li, Q., Dong, H., Wang, S., Wang, Z., 2021. XGBoost-based method for flash flood risk assessment. Journal of Hydrology 598, 126382. https://doi.org/10.1016/J.JHYDROL.2021.126382

Mahesh, R.B., Leandro, J., Lin, Q., 2022. Physics Informed Neural Network for Spatial-Temporal Flood Forecasting. Lecture Notes in Civil Engineering 178, 77–91. https://doi.org/10.1007/978-981-16-5501-2_7

Mai, J., R. Craig, J., A. Tolson, B., 2020. Simultaneously determining global sensitivities of model parameters and model structure. Hydrology and Earth System Sciences 24, 5835–5858. https://doi.org/10.5194/HESS-24-5835-2020

Molnar, C., 2020. Interpretable machine learning. Lulu. com.

Montanari, A., 2011. Uncertainty of Hydrological Predictions. Treatise on Water Science 2, 459–478. https://doi.org/10.1016/B978-0-444-53199-5.00045-2

Montanari, A., 2007. What do we mean by 'uncertainty'? The need for a consistent wording about uncertainty assessment in hydrology. Hydrological Processes 21, 841–845. https://doi.org/10.1002/HYP.6623

Montanari, A., 2005. Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations. Water Resources Research 41, 1–13. https://doi.org/10.1029/2004WR003826

Montanari, A., Koutsoyiannis, D., 2012. A blueprint for process-based modeling of uncertain hydrological systems. Water Resources Research 48. https://doi.org/10.1029/2011WR011412

Moriasi, D.N., Arnold, J.G., Liew, M.W. van, Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Trans ASABE 50, 885–900. https://doi.org/10.13031/2013.23153

Mulvany, T., 1851. On the use of self-registering rain and flood gauges in making observations of the relations of rain fall and flood discharges in a given catchment. Transactions of the Institution of Civil Engineers of Ireland 4, 18–33.

National Research Council, 2012. Challenges and Opportunities in the Hydrologic Sciences. National Academies Press, Washington, D.C. https://doi.org/10.17226/13293

National Research Council, 1991. Opportunities in the Hydrologic Sciences. National Academies Press, Washington, D.C. https://doi.org/10.17226/1543

NCAR - Research Applications Laboratory, 2015. verification: Weather Forecast Verification Utilities.

Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Prieto, C., Gupta, H. v., 2021. What Role Does Hydrological Science Play in the Age of Machine Learning? Water Resources Research. https://doi.org/10.1029/2020WR028091

Nearing, G.S., Sampson, A.K., Kratzert, F., Frame, J., 2020. Post-Processing a Conceptual Rainfall-Runoff Model with an LSTM. https://doi.org/10.31223/OSF.IO/53TE4

Newman, A., Sampson, K., Clark, M., Bock, A., Viger, R., Blodgett, D., 2014. A large-sample watershed-scale hydrometeorological dataset for the contiguous USA [WWW Document]. Boulder, CO: UCAR/NCAR. https://doi.org/https://dx.doi.org/10.5065/D6MW2F4D

Ng, A.Y., 2004. Feature selection, L1 vs. L2 regularization, and rotational invariance, in: Twenty-First International Conference on Machine Learning - ICML '04. ACM Press, New York, New York, USA, p. 78. https://doi.org/10.1145/1015330.1015435

Oudin, L., Andréassian, V., Perrin, C., Michel, C., le Moine, N., 2008. Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resources Research 44. https://doi.org/10.1029/2007WR006240

Pagano, T., Garen, D., 2005. A Recent Increase in Western U.S. Streamflow Variability and Persistence. Journal of Hydrometeorology 6, 173–179. https://doi.org/10.1175/JHM410.1

Papacharalampous, G., Koutsoyiannis, D., Montanari, A., 2020a. Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: Methodology development and investigation using toy models. Advances in Water Resources 136, 103471. https://doi.org/10.1016/J.ADVWATRES.2019.103471

Papacharalampous, G., Tyralis, H., Koutsoyiannis, D., Montanari, A., 2020b. Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale. Advances in Water Resources 136, 103470. https://doi.org/10.1016/J.ADVWATRES.2019.103470

Papacharalampous, G.A., Tyralis, H., 2018. Evaluation of random forests and Prophet for daily streamflow forecasting. Advances in Geosciences 45, 201–208. https://doi.org/10.5194/ADGEO-45-201-2018

Pappenberger, F., Beven, K.J., 2006. Ignorance is bliss: Or seven reasons not to use uncertainty analysis. Water Resources Research 42. https://doi.org/10.1029/2005WR004820

Parajka, J., Merz, R., Blöschl, G., 2007. Uncertainty and multiple objective calibration in regional water balance modelling: case study in 320 Austrian catchments. Hydrological Processes 21, 435–446. https://doi.org/10.1002/HYP.6253

Park, J.M., Kim, J.H., 2017. Online recurrent extreme learning machine and its application to time-series prediction. Proceedings of the International Joint Conference on Neural Networks 2017-May, 1983–1990. https://doi.org/10.1109/IJCNN.2017.7966094

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology 279, 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Quilty, J., Adamowski, J., 2020. A stochastic wavelet-based data-driven framework for forecasting uncertain multiscale hydrological and water resources processes. Environmental Modelling and Software 130, 104718. https://doi.org/10.1016/j.envsoft.2020.104718

Quilty, J., Adamowski, J., Boucher, M.A., 2019. A Stochastic Data-Driven Ensemble Forecasting Framework for Water Resources: A Case Study Using Ensemble Members Derived From a Database of Deterministic Wavelet-Based Models. Water Resources Research 55, 175–202. https://doi.org/10.1029/2018WR023205

Quilty, J., Adamowski, J., Khalil, B., Rathinasamy, M., 2016. Bootstrap rank-ordered conditional mutual information (broCMI): A nonlinear input variable selection method for water resources modeling. Water Resources Research 52, 2299–2326. https://doi.org/10.1002/2015WR016959

Quilty, J.M., Sikorska-Senoner, A.E., Hah, D., 2022. A stochastic conceptual-data-driven approach for improved hydrological simulations. Environmental Modelling & Software 105326. https://doi.org/10.1016/J.ENVSOFT.2022.105326

Rahimzad, M., Moghaddam Nia, A., Zolfonoon, H., Soltani, J., Danandeh Mehr, A., Kwon, H.H., 2021. Performance Comparison of an LSTM-based Deep Learning Model versus Conventional Machine Learning Algorithms for Streamflow Forecasting. Water Resources Management 35, 4167–4187. https://doi.org/10.1007/S11269-021-02937-W/TABLES/3

Raissi, M., Perdikaris, P., Karniadakis, G.E., 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics 378, 686–707. https://doi.org/10.1016/J.JCP.2018.10.045

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., Mohamed, S., 2021. Skilful precipitation nowcasting using deep generative models of radar. Nature 597, 672–677. https://doi.org/10.1038/s41586-021-03854-z

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. https://doi.org/10.1038/S41586-019-0912-1

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010a. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resources Research 46, 5521. https://doi.org/10.1029/2009WR008328

Rosecrans, C.Z., Belitz, K., Ransom, K.M., Stackelberg, P.E., McMahon, P.B., 2022. Predicting regional fluoride concentrations at public and domestic supply depths in basin-fill aquifers of the western United States using a random forest model. Science of The Total Environment 806, 150960. https://doi.org/10.1016/J.SCITOTENV.2021.150960

Schoppa, L., Disse, M., Bachmair, S., 2020. Evaluating the performance of random forest for large-scale flood discharge simulation. Journal of Hydrology 590, 125531. https://doi.org/10.1016/J.JHYDROL.2020.125531

See, L., Solomatine, D., Abrahart, R., Toth, E., 2007. Hydroinformatics: Computational intelligence and technological developments in water science applications - Editorial. Hydrological Sciences Journal 52, 391–396. https://doi.org/10.1623/HYSJ.52.3.391

Seibert, J., 2000. Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. Hydrology and Earth System Sciences 4, 215–224. https://doi.org/10.5194/HESS-4-215-2000

Seibert, J., Vis, M.J.P., 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. Hydrology and Earth System Sciences 16, 3315–3325. https://doi.org/10.5194/HESS-16-3315-2012

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D.T., Silver, D., Kavukcuoglu, K., Hassabis, D., 2020. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710. https://doi.org/10.1038/s41586-019-1923-7

Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., de Freitas, N., 2016. Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE 104, 148–175. https://doi.org/10.1109/JPROC.2015.2494218

Sharma, S., Ghimire, G.R., Siddique, R., 2021. Machine Learning for Postprocessing Ensemble Streamflow Forecasts.

Shen, C., 2018. A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. Water Resources Research 54, 8558–8593. https://doi.org/10.1029/2018WR022643

Shen, C., Lawson, K., 2021. Applications of Deep Learning in Hydrology. Deep Learning for the Earth Sciences 283–297. https://doi.org/10.1002/9781119646181.CH19

Shen, Y., Ruijsch, J., Lu, M., Sutanudjaja, E.H., Karssenberg, D., 2022. Random forests-based error-correction of streamflow from a large-scale hydrological model: Using model state variables to estimate error terms. Computers & Geosciences 159, 105019. https://doi.org/10.1016/J.CAGEO.2021.105019

Shen, Y., Ruijsch, J., Lu, M., Sutanudjaja, E.H., Karssenberg, D., 2021. Random forests-based error-correction of streamflow from a large-scale hydrological model: Using model state variables to estimate error terms. Computers & Geosciences 105019. https://doi.org/10.1016/J.CAGEO.2021.105019

Sikorska, A.E., Montanari, A., Koutsoyiannis, D., 2014. Estimating the Uncertainty of Hydrological Predictions through Data-Driven Resampling Techniques. Journal of

Hydrologic Engineering 20, A4014009. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000926

Sikorska-Senoner, A.E., Quilty, J.M., 2021. A novel ensemble-based conceptual-data-driven approach for improved streamflow simulations. Environmental Modelling & Software 143, 105094. https://doi.org/10.1016/J.ENVSOFT.2021.105094

Sikorska-Senoner, A.E., Schaefli, B., Seibert, J., 2020. Downsizing parameter ensembles for simulations of rare floods. Natural Hazards and Earth System Sciences 20, 3521–3549. https://doi.org/10.5194/NHESS-20-3521-2020

Singh, V.P., 2018. Hydrologic modeling: progress and future directions. Geoscience Letters 2018 5:1 5, 1–18. https://doi.org/10.1186/S40562-018-0113-Z

Sit, M., Demiray, B.Z., Xiang, Z., Ewing, G.J., Sermet, Y., Demir, I., 2020. A comprehensive review of deep learning applications in hydrology and water resources. Water Science and Technology 82, 2635–2670. https://doi.org/10.2166/WST.2020.369

Snoek, J., Larochelle, H., Adams, R.P., 2012a. Practical Bayesian Optimization of Machine Learning Algorithms. Advances in Neural Information Processing Systems 4, 2951–2959.

Snoek, J., Larochelle, H., Adams, R.P., 2012b. Practical Bayesian Optimization of Machine Learning Algorithms, in: Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.

Suh, Y., Bostanabad, R., Won, Y., 2021. Deep learning predicts boiling heat transfer. Scientific Reports 11, 1–10. https://doi.org/10.1038/s41598-021-85150-4

Tan, J., Xie, X., Zuo, J., Xing, X., Liu, B., Xia, Q., Zhang, Y., 2021. Coupling random forest and inverse distance weighting to generate climate surfaces of precipitation and temperature with Multiple-Covariates. Journal of Hydrology 598, 126270. https://doi.org/10.1016/J.JHYDROL.2021.126270

Tartakovsky, A.M., Marrero, C.O., Perdikaris, P., Tartakovsky, G.D., Barajas-Solano, D., 2020. Physics-Informed Deep Neural Networks for Learning Parameters and Constitutive Relationships in Subsurface Flow Problems. Water Resources Research 56, e2019WR026731. https://doi.org/10.1029/2019WR026731

Tongal, H., Booij, M.J., 2018. Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. Journal of Hydrology 564, 266–282. https://doi.org/10.1016/J.JHYDROL.2018.07.004

Triantakonstantis, D., Tziachris, P., Kainz, W., Zhang, H., Wang, S., Liu, K., Li, X., Li, Z., Zhang, X., Liu, B., 2022. Downscaling of AMSR-E Soil Moisture over North China Using Random Forest Regression. ISPRS International Journal of Geo-Information 2022, Vol. 11, Page 101 11, 101. https://doi.org/10.3390/IJGI11020101

Tsai, W.P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., Shen, C., 2021. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. Nature Communications 2021 12:1 12, 1–13. https://doi.org/10.1038/s41467-021-26107-z

Tyralis, H., Papacharalampous, G., Burnetas, A., Langousis, A., 2019a. Hydrological post-processing using stacked generalization of quantile regression algorithms: Large-scale application over CONUS. Journal of Hydrology 577, 123957. https://doi.org/10.1016/J.JHYDROL.2019.123957

Tyralis, H., Papacharalampous, G., Langousis, A., 2019b. A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. Water 2019, Vol. 11, Page 910 11, 910. https://doi.org/10.3390/W11050910

Valery, A., 2010. Modélisation précipitations – débit sous influence nivale Elaboration d'un module neige et évaluation sur 380 bassins versants. AgroParisTech.

Viglione, A., Parajka, J., 2020. TUWmodel: Lumped/Semi-Distributed Hydrological Model for Education Purposes.

Wang, S., Peng, H., Liang, S., 2022. Prediction of estuarine water quality using interpretable machine learning approach. Journal of Hydrology 605, 127320. https://doi.org/10.1016/J.JHYDROL.2021.127320

Wijayarathne, D.B., Coulibaly, P., 2020. Identification of hydrological models for operational flood forecasting in St. John's, Newfoundland, Canada. Journal of Hydrology: Regional Studies 27, 100646. https://doi.org/10.1016/J.EJRH.2019.100646

Wu, J., Ma, D., Wang, W., 2021. Leakage Identification in Water Distribution Networks Based on XGBoost Algorithm. Journal of Water Resources Planning and Management 148, 04021107. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001523

Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., Shen, C., 2021. Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. Journal of Hydrology 603, 127043. https://doi.org/10.1016/J.JHYDROL.2021.127043

Xu, Y., Goodacre, R., 2018. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. Journal of Analysis and Testing 2, 249. https://doi.org/10.1007/S41664-018-0068-2

Zaherpour, J., Mount, N., Gosling, S.N., Dankers, R., Eisner, S., Gerten, D., Liu, X., Masaki, Y., Müller Schmied, H., Tang, Q., Wada, Y., 2019. Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models. Environmental Modelling & Software 114, 112–128. https://doi.org/10.1016/J.ENVSOFT.2019.01.003

Zambrano-Bigiarini, M., 2020. hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series.

Zhong, W., Yuan, Q., Liu, T., Yue, L., 2022. Freeze/thaw onset detection combining SMAP and ASCAT data over Alaska: A machine learning approach. Journal of Hydrology 605, 127354. https://doi.org/10.1016/J.JHYDROL.2021.127354

Zhou, X., Zhu, Y., Hou, D., Fu, B., Li, W., Guan, H., Sinsky, E., Kolczynski, W., Xue, X., Luo, Y., Peng, J., Yang, B., Tallapragada, V., Pegion, P., 2022. The Development of the NCEP Global Ensemble Forecast System Version 12. Weather and Forecasting 37, 1069–1084. https://doi.org/10.1175/WAF-D-21-0112.1

Appendices

Appendix A

Input Variable Summary Statistics

Table A1. Input variable summary statistics (mean, minimum, maximum, standard deviation, inter-quantile range) for the warm-up, training, validation, and test sets. Note: Precipitation uses wet days P > 1.

Variables	Warm-up					Training					Validation					Test				
	mean	min	max	sd	IQR	mean	min	max	sd	IQR	mean	min	max	sd	IQR	mean	min	max	sd	IQR
											Dünnern									
P	4.79	1.00	60.12	7.18	6.34	4.86	1.00	66.62	7.23	6.47	5.16	1.00	69.04	8.01	6.26	4.64	1.00	64.56	6.95	6.20
T_{mean}	9.90	-6.96	27.33	7.09	11.50	10.45	-16.77	28.69	7.34	11.34	10.86	-11.08	27.11	7.54	12.07	10.89	-13.12	27.93	7.36	11.51
T_{min}	5.95	-14.15	24.65	7.13	11.08	6.89	-21.82	23.19	6.76	10.26	7.48	-15.19	22.54	6.85	10.90	7.39	-18.39	21.25	6.56	10.16
T_{max}	14.27	-4.59	35.62	7.54	12.09	14.74	-11.63	37.87	8.42	12.92	15.10	-5.16	34.63	8.66	13.80	15.25	-9.15	36.79	8.64	13.26
PET	1.02	0.01	5.92	1.34	1.24	1.03	0.01	7.68	1.32	1.44	1.05	0.01	5.84	1.28	1.48	1.07	0.01	5.79	1.31	1.53
Q	2.46	0.22	22.03	2.30	2.20	1.89	0.09	32.15	2.17	1.62	1.84	0.15	38.35	2.39	1.67	1.65	0.14	24.18	1.99	1.36
											Kleine-Emme									
P	6.34	1.00	103.76	8.90	8.67	6.50	1.00	89.01	8.77	9.11	6.64	1.00	102.81	10.05	8.57	6.40	1.00	60.13	8.68	8.99
T_{mean}	8.32	-9.13	25.29	7.30	12.16	8.89	-16.61	25.55	7.41	12.07	9.17	-10.45	24.32	7.60	12.76	9.21	-12.96	25.19	7.42	12.02
T_{min}	5.25	-14.12	23.79	7.29	11.83	6.15	-20.28	20.93	7.00	11.29	6.46	-14.51	20.04	7.07	11.68	6.56	-15.07	21.32	6.88	11.01
T_{max}	11.72	-7.00	30.96	7.64	12.76	12.23	-13.41	31.61	8.22	13.30	12.53	-7.26	29.91	8.55	14.23	12.52	-10.80	31.26	8.38	13.67
PET	1.30	0.01	10.24	1.68	1.29	1.27	-0.31	10.04	1.44	1.21	1.26	0.01	9.87	1.42	1.26	1.24	-0.05	10.15	1.41	1.23
Q	2.90	0.39	37.58	3.15	2.33	2.80	0.27	39.26	3.14	2.35	2.81	0.24	73.45	3.77	2.24	2.77	0.33	35.74	2.80	2.04
											Muota									
P	7.51	1.00	88.15	10.33	10.58	7.99	1.00	105.26	10.65	10.75	7.84	1.00	106.19	10.46	10.41	7.43	1.00	65.44	10.06	9.79
T_{mean}	8.28	-8.55	26.18	7.16	11.56	8.78	-17.28	26.12	7.25	11.71	9.06	-10.25	26.35	7.50	12.10	9.21	-13.01	27.53	7.28	11.88
T_{min}	4.71	-15.37	23.28	7.14	11.21	5.63	-22.28	24.06	6.75	10.61	5.95	-14.22	21.04	6.85	11.30	6.06	-16.07	23.49	6.68	10.90
T_{max}	12.17	-5.06	33.45	7.63	12.54	12.63	-13.54	34.40	8.20	13.01	12.93	-7.77	31.60	8.59	13.41	13.09	-10.77	33.03	8.33	13.02
PET	1.12	0.01	7.31	1.36	1.51	1.11	0.01	7.49	1.33	1.51	1.11	0.02	9.05	1.31	1.40	1.02	0.01	6.89	1.19	1.38
Q	5.82	0.53	39.78	5.29	6.38	5.50	0.44	53.48	4.98	5.72	5.19	0.47	71.77	4.76	5.77	5.53	0.33	49.40	5.18	5.10

Appendix B

HM Parameter Ranges

Table B1. Parameter ranges for TUWmodel used in BO (retrieved from the TUWmodel R package, (Viglione & Parajka, 2020)).

.Parameter	Explanation	Minimum	Maximum	Units
SCF	Snow correction factor	0.9	1.5	-
DDF	Degree-day factor	0	5	mm/°C/d
Tr	Threshold temperature for rain above	1	3	°C
Ts	Threshold temperature for snow below	-3	1	°C
Tm	Threshold temperature for melt above	-2	2	°C
LPrat	Parameter for potential evaporation	0	1	-
FC	Field capacity	0	600	mm
BETA	Parameter for runoff production	0	20	-
k0	Storage coefficient for very fast response	0	2	d
k1	Storage coefficient for fast response	2	30	d
k2	Storage coefficient for slow response	30	250	d
lsuz	Threshold storage state	1	100	mm
cperc	Constant percolation rate	0	8	mm/d
bmax	Maximum base at low flows	0	30	d
croute	Free scaling parameter	0	50	d^2/mm

Table B2. Parameter ranges for GR4JCN used in BO (retrieved from the airGR R package (Coron et al., 2022)). Note: GR4J used the same parameter ranges as GR4JCN but did not consider x5 and x6.

Parameter	Explanation	Minimum	Maximum	Units
x1	Production store maximal capacity	0.9	1.5	mm/°C/d
x2	Catchment water exchange coefficient	0	5	mm/d
x3	One-day maximal capacity of routing reservoir	1	3	mm
x4	Unit hydrograph time base	-3	1	d
x5	Ponderation coefficient	0	1	-
x6	Degree-day factor	2	6	mm/°C/d

Appendix C

Results for Calibration and Validation of Ensemble HMs and CDDA Variants

Table C1. Summary table for ensemble HM and CDDA variants for the training set. Note: performance metrics for the stochastic models only applies to the test set (Section 4.4).

Criteria	HBV-light				TUWmodel				GR4J			
	HM	CDDA XGB	CDDA RF	CDDA LSTM	HM	CDDA XGB	CDDA RF	CDDA LSTM	HM	CDDA XGB	CDDA RF	CDDA LSTM
Dünnern												
MAE	0.57	0.25	0.17	0.32	0.77	0.18	0.17	0.37	0.68	0.17	0.17	0.34
RMSE	0.95	0.36	0.33	0.72	1.17	0.25	0.35	0.85	1.16	0.24	0.36	0.85
NSE	0.81	0.97	0.98	0.89	0.71	0.99	0.97	0.85	0.71	0.99	0.97	0.85
KGE	0.89	0.98	0.97	0.94	0.73	0.98	0.96	0.91	0.81	0.98	0.95	0.87
PBIAS	5.5	0.0	0.2	-0.8	22.0	0.0	0.1	-1.2	1.8	0.0	0.1	-1.5
AW	0.59	0.67	0.25	0.46	2.15	1.04	0.59	1.23	1.50	0.81	0.37	0.93
α_R	0.61	0.78	0.69	0.74	0.52	0.97	0.93	0.90	0.74	0.96	0.77	0.91
CRPS	0.50	0.18	0.14	0.27	0.57	0.11	0.12	0.28	0.54	0.11	0.13	0.26
Kleine-Emme												
MAE	0.86	0.30	0.26	0.67	1.23	0.20	0.27	0.68	1.04	0.21	0.27	0.65
RMSE	1.62	0.48	0.56	1.33	1.90	0.29	0.55	1.36	1.85	0.30	0.55	1.35
NSE	0.73	0.98	0.97	0.82	0.63	0.99	0.97	0.81	0.66	0.99	0.97	0.82
KGE	0.87	0.96	0.95	0.91	0.74	0.98	0.95	0.90	0.81	0.97	0.94	0.86
PBIAS	1.0	0.0	0.1	-0.3	18.5	0.0	0.2	-0.3	-1.3	0.0	0.2	-1.4
AW	1.21	1.13	0.51	0.95	3.09	1.32	0.9	1.9	2.24	1.08	0.57	1.67
α_R	0.75	0.91	0.8	0.73	0.56	0.92	0.93	0.87	0.83	0.98	0.78	0.86
CRPS	0.73	0.21	0.21	0.56	0.93	0.13	0.2	0.52	0.83	0.13	0.21	0.51
Muota												
MAE	1.12	0.41	0.38	0.69	2.93	0.18	0.38	1.11	3.19	0.20	0.42	0.88
RMSE	1.85	0.58	0.70	1.40	4.51	0.25	0.68	2.04	4.68	0.27	0.72	1.59
NSE	0.86	0.99	0.98	0.92	0.18	1.00	0.98	0.83	0.12	1.00	0.98	0.90
KGE	0.93	0.99	0.98	0.96	0.48	0.99	0.98	0.91	0.49	0.99	0.98	0.93
PBIAS	0.3	0.0	0.1	0.6	-20.4	0.0	0.1	2.1	0.6	0.0	0.1	1.1
AW	1.15	1.24	0.62	0.77	4.89	1.57	1.32	2.86	3.91	1.41	0.90	2.08
α_R	0.65	0.82	0.72	0.70	0.77	0.85	0.88	0.84	0.62	0.91	0.75	0.82
CRPS	0.98	0.29	0.31	0.61	2.45	0.12	0.28	0.85	2.74	0.13	0.32	0.67

Table C2. Summary table for ensemble HM and CDDA variants for the validation set. Note: performance metrics for the stochastic models only applies to the test set (Section 4.4)

Criteria												
	HBV-light				TUWmodel				GR4J			
		CDDA	CDDA	CDDA		CDDA	CDDA	CDDA		CDDA	CDDA	CDDA
	HM	XGB	RF	LSTM	HM	XGB	RF	LSTM	HM	XGB	RF	LSTM
Dünnern												
MAE	0.55	0.44	0.45	0.34	0.88	0.41	0.44	0.38	0.78	0.42	0.48	0.40
RMSE	1.14	1.02	1.02	0.98	1.36	0.95	1.00	1.01	1.48	0.98	1.08	1.15
NSE	0.77	0.82	0.82	0.84	0.68	0.84	0.82	0.83	0.62	0.83	0.79	0.77
KGE	0.80	0.79	0.79	0.83	0.68	0.84	0.84	0.88	0.74	0.84	0.83	0.82
PBIAS	-2.4	-8.9	-8.5	-4.0	23.2	-1.3	-1.0	-1.5	2.8	-1.8	-1.4	-1.4
AW	0.56	0.92	0.52	0.45	2.17	1.92	1.45	1.27	1.58	1.45	0.88	1.02
α_R	0.61	0.75	0.66	0.77	0.45	0.90	0.89	0.96	0.70	0.91	0.76	0.85
CRPS	0.48	0.35	0.39	0.29	0.66	0.29	0.32	0.29	0.63	0.31	0.38	0.31
Kleine-Emme												
MAE	0.85	0.63	0.64	0.66	1.24	0.63	0.66	0.70	1.05	0.62	0.66	0.67
RMSE	1.58	1.24	1.27	1.31	1.86	1.23	1.27	1.40	1.87	1.21	1.26	1.36
NSE	0.82	0.89	0.89	0.88	0.76	0.89	0.89	0.87	0.75	0.90	0.89	0.87
KGE	0.88	0.89	0.88	0.91	0.78	0.91	0.91	0.92	0.86	0.92	0.91	0.90
PBIAS	-5.3	-5.0	-5.0	-3.3	14.7	-2.1	-2.4	-1.8	-4.7	-2.8	-2.9	-2.0
AW	1.15	1.75	1.05	0.95	3.10	3.12	2.17	1.96	2.28	2.36	1.26	1.75
α_R	0.73	0.86	0.76	0.72	0.56	0.94	0.92	0.85	0.83	0.96	0.78	0.86
CRPS	0.72	0.48	0.53	0.55	0.93	0.46	0.49	0.53	0.86	0.46	0.53	0.52
Muota												
MAE	1.04	0.88	0.90	0.66	2.72	0.83	0.92	1.03	3.03	0.84	1.03	0.80
RMSE	1.78	1.61	1.65	1.39	3.99	1.52	1.65	1.96	4.20	1.47	1.68	1.56
NSE	0.86	0.89	0.88	0.92	0.30	0.90	0.88	0.83	0.22	0.91	0.88	0.89
KGE	0.92	0.92	0.92	0.95	0.51	0.95	0.94	0.92	0.53	0.95	0.94	0.92
PBIAS	-3.3	-4.1	-3.7	-0.7	-21.4	1.9	1.9	1.6	-0.2	0.5	0.7	1.0
AW	1.04	1.82	1.13	0.76	4.99	4.90	3.33	2.86	3.72	3.92	1.94	2.00
α_R	0.64	0.73	0.68	0.68	0.75	0.92	0.89	0.85	0.63	0.97	0.74	0.84
CRPS	0.92	0.71	0.78	0.57	2.22	0.59	0.68	0.78	2.61	0.61	0.82	0.61