Evaluation of ensemble flood forecast performance by using a semi-distributed multi-model approach compared to single models

MASTER THESIS

Sander de Groot

Enschede, Jan 2025

CIVIL ENGINEERING - MASTER INTEGRATED WATER MANAGEMENT

Evaluation of ensemble flood forecast performance by using a semi-distributed multi-model approach compared to single models

Master Thesis

to obtain the degree of Master of Science at the University of Twente

To be presented on January 16, 2025

Author: Sander de Groot (s2172755)

Institutions: University of Twente & Deltares

Supervisors:
Dr. Ir. M.J. Booij (UTwente)
Dr. Ir. F. Diermanse (Deltares)
Prof. Ir. J. Kwadijk (Deltares)

Date: 09-01-2025

Preface

With this thesis, I finish my Master's degree in Civil Engineering and Management at the University of Twente. Hence, it also marks the end of my student time here in Enschede. My research involved flood forecasting in the Overijsselse Vecht river basin using two hydrological models. Hopefully, the conclusions from my research can add to the current state of the art in flood forecasting of regional river basins and help water authorities in flood preparedness and mitigation.

Taking on this challenge has taught me a great deal about hydrology and modeling discharge in a river basin. I have gained a deeper understanding of flood forecasting and its associated challenges. I have had the opportunity to get a glimpse into the real-time flood forecasting system FEWS-Vecht, currently in use by waterboards responsible for the Vecht. The possibility of having a look around in this system has offered me many insights which I could use for my own research. I also have learned to perform programming tasks with use of Python, for data analyses, modeling, and result visualization.

Luckily, I was not on my own during my research and, therefore, I want to express my thanks to some people who helped me during my thesis. This thesis was not possible without the valuable guidance of my supervisor at the university Martijn Booij, who I could always contact for questions and could always visit for a discussion. I also wish to thank Ferdinand Diermanse for this daily supervision at Deltares. His expertise in the world of hydrology was very valuable for my own understanding of the topic. He was able to bring some comfort and relaxation at times when things seemed to go a little less well. I would also like to thank Jaap Kwadijk for providing his view on the challenges at hand. I also wish to express a special thanks to Klaas-Jan van Heeringen, who has sat down with me for many hours, learning me about FEWS-Vecht and fixing any related problems. I would like to thank everyone from the Deltares flood risk management department, and specifically the colleagues working on the JCAR-ATRACE program. From the start of my project they invited me to meetings with the involved partners in the Vecht area, and they all have shown a great interest in my work. I would like to thank Jeroen van der Scheer from waterboard Vechtstromen for sharing crucial data from the Vecht and providing me with the necessary knowledge about the Vecht basin. Together with Ivan and William I started my thesis at Deltares and, while having individual projects, were able to do some work together. I would like to thank them for their support, our many discussions, but most of all the fun moments we had together the past months.

Finally, I would like to thank my family and friends for their support during my thesis, but also for the amazing time I had here during my entire studies. Thanks to all my friends who also spent many days working on their thesis in 'Het Hok' at the university. It has been an amazing place to work and find some distraction where needed. Thanks to Amy, Maarten and Laura for helping me with my programming struggles. Thanks also to ChatGPT, with whom I had many digital conversations regarding Python coding. I also would like to thank Lars, Cindy and Floor & Nick for their hospitality by providing me with a place to sleep close to Delft. You made my logistics much more manageable, but most of all, made me feel welcome in your homes and showed me around in Delft and Leiden. Without these sleep addresses, it would not have been manageable to be in the office twice a week.

Hopefully you enjoy reading my thesis!

Sander de Groot

Enschede, January 2025

Summary

The increasing frequency of extreme precipitation events presents significant challenges to flood preparedness and mitigation globally. Reliable and accurate Flood Forecast and Warning Systems (FFWS) are crucial for ensuring flood safety. A key component of a FFWS is a hydrological model that translates precipitation into runoff. Instead of relying on a single hydrological model, integrating multiple model structures offers a promising approach to enhancing FFWS performance by capturing a broader range of uncertainties. Despite its potential, this multi-model approach is not widely implemented by flood forecasting institutions. Furthermore, limited collaboration and inconsistent communication between regional and national institutions have led to the use of varying models and methodologies within shared river basins.

This study evaluates and compares the flood forecasting performance of two semi-distributed hydrological models HBV and GR4H versus a model combination consisting of both, for the Vecht River in Germany and the Netherlands. The models operate on an hourly time step and are driven by ensemble weather forecasts. Both HBV and GR4H were calibrated and validated using historical precipitation and discharge data. Discharge forecasts were generated using 20 precipitation ensembles from COSMO-LEPS. The model states were updated through direct state updating based on empirical storage-discharge relationships. Forecast performance was evaluated using metrics Relative Mean Absolute Error (RMAE), Continuous Ranked Probability Score (CRPS), and the skilled CRPS (CRPSS). These scores were applied to both individual and multi-model forecasts. The multi-model was constructed by combining forecasts from HBV and GR4H. Finally, the performance of the multi-model was compared to that of the individual models.

Both models demonstrated adequate performance in simulating high flows during calibration. Performance was assessed through a multi-objective function focusing on high flows, combining a weighted Nash-Sutcliffe efficiency (NS_w) and relative volume error (RVE). During calibration, HBV showed better objective function values than GR4H among most sub-catchments. However, during validation, performance declined for both models, with some sub-catchments yielding unsatisfactory results reflected in reduced NS_w and increased RVE. HBV maintained slightly better performance than GR4H during validation.

Both models showed good forecast performance for the upstream (independent) sub-catchments. Forecast accuracy for downstream (dependent) sub-catchments was poor at short lead times ($<$12 hours) for both HBV and GR4H. This was caused by inaccurate state updating, discrepancies between observed and simulated discharges, and uncertainty in observed discharge from catchments upstream. The multi-model forecasts showed marginal improvements based on the evaluation metrics, with respect to the individual model forecasts. In addition, the multi-model forecasts substantially outperformed the weaker individual model, therefore reducing the possibility of a poor forecast. Although slightly increased performance was observed for some lead times in RMAE and CRPS(S), the results varied significantly across sub-catchments and forecast periods.

To increase robustness of the results, the number of models in the multi-models framework could be expanded. To improve the forecast performance of the individual models, a more effective state updating method is recommended for the downstream catchments. This study emphasized the need for greater data transparency and collaboration between institutions in the Vecht basin, as current data sourcing practices are fragmented. However, this study has also has shown the potential of the multi-model approach to increase forecast performance with respect to using a single model.

Contents

List of Figures

List of Tables

Abbreviation list

Abbreviation	Description
ARMA	Autoregressive Moving Average
BMA	Bayesian Model Averaging
CDF	Cumulative Density Function
COSMO-LEPS	Consortium for Small Scale Modeling – Limited area Ensemble Prediction System
CRPS	Continuous Ranked Probability Score
CRPSS	Continuous Ranked Probability Skill Score
DWD	Deutscher Wetterdienst
ECMWF	European Centre for Medium-Ranged Forecasts
EFAS	European Flood Awareness System
FEWS	Flood Early Warning System
FFWS	Flood forecasting & Warning System
GE	Grand Ensembles
Ge	Germany
GMT	Greenwich Mean Time
GR4J/GR4H	Modèle Génie Rural à 4 paramètres Journalier/Horaire
HBV	Hydrologiska Byråns Vattenbalansavdelning
ICON-EU	ICOsahedral Nonhydrostatic
IRC	International Radar Composite
IQR	Inter Quartile Range
JCAR-ATRACE	Joint Cooperation for Applied Scientific Research to Accelerate Regional Adaptation to Climate Extremes
KNMI	Koninklijk Nederlands Meteorologisch Instituut
MAE	Mean Absolute Error
MFB	Mean-Field Bias
MFBS	Mean-Field Bias Spatially corrected
NL	Netherlands
NLWKN	Niedersächsischer Landesbetrieb für Wasserwirtschaft, Küsten-und Naturschutz
NSE	Nash-Sutcliffe Efficiency
NSw	Weighted NSE
NRW	Nordrhein-Westfalen
NWP	Numerical Weather Prediction
P	Precipitation
PET	Potential Evapotranspiration
PSO	Particle Swarm Optimization
Q	Discharge
RMAE	Relative Mean Absolute Error
RME	Relative Mean Error
RMSE	Relative Mean Squared Error
ROC	Relative Operating Characteristic
RVE	Relative Volume Error
T	Temperature
WMO	World Meteorological Organization

1 Introduction

1.1 Problem context

The intensity of global extreme precipitation events around the world has increased in recent decades and with future climate change projections it will increase even more (IPCC, 2023; Tradowsky et al., 2023). However, the direct translation between extreme precipitation and flooding can be complicated. Due to the influence of space- and time-dependent factors, an extreme rainfall event does not necessarily lead to a flood event (Seneviratne et al., 2023). In addition, successful water management policies and measures have increased resilience to floods and reduced the number of water-related disasters compared to the 20th century (Seneviratne et al., 2023; Perera et al., 2019). An example of a measure to reduce flood risk is a flood forecasting & warning system (FFWS). The decline in flood disasters and casualties may be partly attributed to a large increase in operational FFWS over the same period (Perera et al., 2019). However, with the apparent increase in very extreme precipitation events, the influence of these events on floods can increase and will remain a likely threat. Especially in North America, Asia, and Europe, there is strong evidence that the intensity of extreme precipitation has increased since the 1950s (IPCC, 2023).

In the summer of 2021, extreme rainfall in Northwestern Europe caused a catastrophic flood, leading to severe damage and more than 200 fatalities. Especially within Germany and Belgium the flash flood-like event hit hard. The unprecedented precipitation event resulted in one of the largest flood disasters in Northwestern Europe in decades (Kundzewicz and Pińskwar, 2022; Lehmkuhl et al., 2022). Immediately after the event, numerous studies were carried out to try to explain the cause and find the probability of recurrence of similar events in Europe. Research of Ludwig et al. (2023) showed that water levels in some locations in Germany had far exceeded the statistical 100-year return period, which is beyond historically observed gauge records. Despite the low probability of the event, the forecasts indicated heavy and extreme precipitation up to six days beforehand, though, with spatial and temporal uncertainty. Warnings and notifications were sent to responsible authorities, but not all people received the warning and were also not told how to act upon the issued warning (Tradowsky et al., 2023). Additionally, FFWS in the area underestimated water levels at most locations, based upon the precipitation forecasts. Clearly, while sophisticated weather and forecasting models exist, they cannot always guarantee complete predictability of flood disasters.

In response to the summer floods of 2021 and the summer droughts of 2018 and 2022, a Joint Cooperation program has been established for Applied scientific Research on flood and drought risk management in regional river basins (JCAR). The program's goal is to Accelerate Transboundary Regional Adaptation to Climate Extremes (ATRACE). The program lays emphasis on improving integrated planning, development and management of regional river basins in Belgium, Germany, Luxembourg and the Netherlands to reduce flood and drought risk and prepare for climate change extremes. Next to this, the program strives for development of an international expert community by fostering long-term partnerships between knowledge institutes. Several (research) institutes and universities are affiliated with the program: Deltares, RWRH Aachen, UFZ Potsdam, Uliège, KU Leuven, Luxembourg Insititute of Science and Technology, TU Delft, IVM VU Amsterdam, University of Twente, and research organizations from France, Austria and Switzerland (Slager and Kwadijk, 2023). This research will contribute to JCAR by focusing on the performance of a flood forecast system under high flow conditions in a regional, transboundary river basin. While the program also has a large interest in drought forecasting, the main focus of this research is on flood forecasting.

1.2 State of the art in flood forecasting

The catastrophic flood event in summer 2021 has raised attention to the need for better climate change preparedness against floods. Although extreme events like the one from the summer of 2021 are impossible to avoid given the current climate change projections, impacts can be reduced by appropriate flood preparedness and protection (Lehmkuhl et al., 2022). The level of flood protection can be increased by structural or non-structural measures (Perera et al., 2019). The use of non-structural measures has multiple advantages over the use of structural measures. Yazdi et al. (2014) mention the environmental and economic benefit of using non-structural strategies instead of, or next to, structural measures. One of the non-structural measures to reduce flood impact is flood forecasting and warning. A flood warning system that incorporates forecasting (FFWS) has been proven a viable non-structural measure in increasing flood preparedness and reducing flood damage (Jain et al., 2018; Arduino et al., 2005; Tradowsky et al., 2023).

1.2.1 Flood Forecasting and Warning System

Typically, a flood event is termed as an event where a certain threshold (water level or discharge) is exceeded. The sequence of steps that take place within an FFWS from the moment a precipitation event is recognized, to the moment the threshold is exceeded is called the flood timeline (Figure 1). This period is also called the maximum potential warning time (Carsell et al., 2004). The first task along this flood timeline is the collection of (measured or forecasted) hydrometeorological data. Next, this data is evaluated by personnel or by a FFWS. An FFWS often combines a hydrological rainfall-runoff model to translate (expected) precipitation to (expected) river discharge, and often a hydraulic model to simulate the flood propagation and (expected) water levels (Arduino et al., 2005). The moment from when a threat is recognized by the interpreter until exceedance of the flood threshold is called the lead time. The lead time depends upon various factors such as catchment lag time, basin size, basin characteristics and the rainfall event (Jain et al., 2018). After threat recognition, responsible authorities need to be notified of the potential event. Then, the authorities must decide whether to communicate to the public or not. The last step is the response, and could for example include evacuation of people and property or implementing measures such as placing sandbags or wooden barriers. This period is also referred to as the mitigation time (Carsell et al., 2004). The ultimate goal of an FFWS is to provide accurate forecasts of hydrological conditions from meteorological input (WMO, 2013), and alert authorities and the public on an imminent flood as reliably and early as possible (Jain et al., 2018).

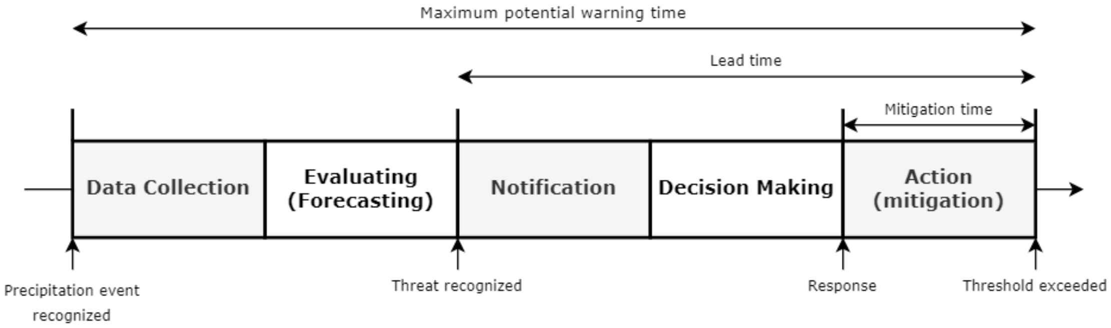


Figure 1: Flood timeline. The timeline shows the consecutive tasks from recognition of a precipitation event to the exceedance of a flood threshold (adapted from Carsell et al. (2004)).

1.2.2 Hydrological models

Hydrological models are a simplified representation of reality (Moradkhani and Sorooshian, 2008). The goal of a hydrological model is to better understand hydrological processes and improve decision-making in water resource planning, flood prediction, irrigation practices, etc. (Pechlivanidis et al., 2013). The focus of this study is on rainfall-runoff models, which are a type of hydrological models that translate precipitation to runoff in a river catchment (Sitterson et al., 2018). The models are forced by hydro-climatic time series and model parameters describe catchment characteristics. In the case of hydrological forecasting, hydro-climatic time series consists of expected weather variables like precipitation, evaporation and temperature. The output of the hydrological model is the simulated discharge (Verkade, 2008; Hapuarachchi et al., 2011).

When hydrological models are used for forecasting purposes, uncertainties from different sources can be distinguished: input data, model structure and model parameters (Wu et al., 2020; Kauffeldt et al., 2016; Gupta et al., 2005). Pechlivanidis et al. (2013) mention spatial and temporal variability of precipitation to be the main input data uncertainty in model predictions. Especially for short- to medium-range forecasts (2-15 days ahead), the largest uncertainty comes from meteorological inputs (Wu et al., 2020). To capture a part of the meteorological uncertainty, ensemble forecasts can be used (Cloke and Pappenberger, 2009; Thielen et al., 2013; Wu et al., 2020).

1.2.3 Ensemble forecasting

In general, flood forecasting can be divided into two parts: meteorological forecasting and hydrological forecasting. Meteorological forecasting models use the water and energy cycles in the atmosphere to predict future meteorological states, and hydrological forecasting models mostly the water cycle, to predict future discharge by direct surface runoff and runoff through soil layers (Das et al., 2022). Many FFWS depend on meteorological inputs from field observations or radar measurements. However, using this data provides forecasts for short (certain) lead times (1-2 days ahead). To provide early warning and sufficient response time, medium-ranged weather forecasts (2-15 days ahead) must be used (Das et al., 2022); Wu et al., 2020; Cloke and Pappenberger, 2009). To simulate medium-range forecasts, numerical weather prediction (NWP) techniques are used. NWP is a method of predicting future possible atmospheric states, by solving a set of differential equations (Teja et al., 2023). NWPs can produce a single deterministic prediction or a probabilistic prediction. The probabilistic, or ensemble forecast, considers multiple, equally probable predictions. By considering a range of likely scenarios, forecast uncertainty can be accounted for in the weather predictions (Wu et al., 2020; Kauffeldt et al., 2016; Jain et al., 2018). In Europe, most meteorological weather forecasts come from the European Centre for Medium-Range Weather Forecasts (ECMWF). This institute provides worldwide ensemble NWPs with lead times ranging from days to weeks ahead. Weather forecasts from ECMWF are used for example in the European Flood Awareness System (EFAS) and in Flood Early Warning System Delft (Delft-FEWS, Perera et al., 2019; Deltares, n.d.). The Rotal Meteorological Institute of the Netherlands (KNMI) also provides NWPs, called HARMONIE.

The outcomes of deterministic or probabilistic NWPs can be forced into hydrological models to obtain discharge forecasts. A deterministic forecast is subject to a larger uncertainty because it provides just one prediction, based on one set of initial conditions. This way, no indication of associated forecast uncertainty is taken into the prediction. For this reason, probabilistic forecasting has become increasingly popular (Wu et al., 2020; Thielen et al., 2013). One of the main advantages of ensemble forecasts is that it presents part of the forecast uncertainty by producing multiple outcomes for the same forecast period (Thielen et al., 2013). Hydrological forecasting is not only subject to meteorological uncertainties, but also to uncertainties associated with the hydrological model structure and parameters (Pechlivanidis et al., 2013; Kauffeldt et al., 2016; Wu et al., 2020). Using ensemble NWPs for future precipitation as driving input for hydrological models does capture a part of the meteorological uncertainty, however, the uncertainties related to model structure and parameters remain (Velázquez et al., 2011).

1.2.4 Multi-hydrological model forecasting

Relying on a single hydrological model may not be sufficient to account for all complex hydrological processes that take place in the environment. There may be many different model structures and parameter sets that are acceptable in reproducing the observed behavior of the system (Beven, 1993; Liu and Gupta, 2007). Due to limited skill in representing hydrological processes, the use of a single hydrological model is subject to statistical bias and underestimation of uncertainty (Liu and Gupta, 2007). To capture uncertainties related to model structure, a multi-model approach can be used. Besides quantification of model structure uncertainty, a multi-model approach can improve the accuracy and consistency of flood forecasts (Cloke and Pappenberger, 2009; Das et al., 2022; Teja et al., 2023). Multi-model approaches, driven by ensemble weather forecasts have been increasingly used in recent years and could be a promising approach to consider uncertainty in ensemble flood forecasting (Wu et al., 2020).

Wetterhall et al. (2013) asked flood forecasters to rank top priorities for improving the European Flood Awareness System, and they mentioned the use of multiple hydrological models as the top priority for increasing model robustness and better representation of model structure uncertainty. Use of a multi-model approach does however come with high implementation costs of new hydrological model systems, so the feasibility and performance of the models must be taken into account before selecting them (Kauffeldt et al., 2016).

Velázquez et al. (2011) researched the use of a multi-model approach for probabilistic flood forecasting. The study evaluated and compared the performance and reliability of different combinations of sixteen lumped hydrological models and meteorological forecasts for 29 catchments in France. The research showed that the multi-model with ensemble NWPs outperforms both the individual models with ensemble NWPs and the multi-model driven by deterministic weather predictions. There seems to be potential for improvement in hydrological forecasting, by using a multi-model approach in combination with ensemble NWPs. Velázquez et al. (2010) compared ensembles constructed from seventeen lumped hydrological models against their simple average counterparts. They found that for all 1061 considered catchments, the ensemble simulations were better than the mean average error of the aggregated simulation. This shows the added value of using more hydrological models in forecasting. Another recent study by Teja et al. (2023) tried to improve flood forecasting on a short- to medium-range timescale by finding suitable combinations of hydrological models and ensemble NWPs for a study area in India. Three lumped hydrological models and one distributed model were run singularly and combined, driven by deterministic and probabilistic NWPs from combined weather models (grand ensembles). The authors found the performance of the flood forecasts with use of the multi-model to be superior over the individual models. The use of grand ensembles (GE) in combination with the multi-model did not significantly increase performance, however, it did increase in the case of the individual models with GE.

The majority of the literature found on hydrological multi-model ensemble flood forecasting considers lumped hydrological models, run on a daily time step. Multi-model ensemble forecasting studies considering hydrological semi-distributed models, run on an hourly time step are far more scarce. Just recently, Thébault et al. (2024), combined a multi-model approach with semi-distributed models on an hourly time step. They used meteorological ensemble forecasts to predict discharge of 12 tributaries of the Rhône river. They found that explicitly considering uncertainty with this probabilistic super-ensemble improves the quality of the streamflow forecasts. Besides this paper, no other literature was found considering hydrological semi-distributed multi-model ensemble forecasting on an hourly time step.

1.3 Research gap

The literature review underscores the significant challenges posed by extreme precipitation events and emphasizes the necessity of a reliable and accurate FFWS. Notably, ensemble flood forecasting, through a multi-model approach, has been identified as a promising method for enhancing FFWS performance by accounting for a broader range of uncertainties. Despite this, a distinct gap exists in the literature regarding the impact of a multi-model approach with semi-distributed hydrological models operating on an hourly time step, driven by ensemble weather forecasts.

Given the increase in performance using a multi-model approach, it could be expected flood forecasting institutions to adopt the multi-model approach combined with ensemble weather forecasts, because of their better performance. However, current practice tends to rely on a single hydrological model, or one combination of a hydrological model and hydraulic model, to predict discharge and water levels (Velázquez et al., 2010). While evidence indicates that multi-model approaches combined with ensemble numerical weather predictions (NWPs) enhance forecast reliability and improve response accuracy, operational medium-range FFWS often still depend on a single-model framework.

Additionally, the flood event of 2021 highlighted inconsistencies in forecasting practices across regional authorities within the same river catchment. Different models and methodologies are frequently employed, with limited cooperation and communication across regional and national borders (Klein and van der Vat, 2024). The JCAR ATRACE program emphasizes the need for additional proof of the effectiveness of multi-model forecasting under extreme conditions, but also to foster greater collaboration among regional water authorities within shared river catchments.

1.4 Research aim

The aim of this research is: **to evaluate and compare the flood forecasting performance[3] of single semi-distributed hydrological models[1] versus a multi-model[2] approach, driven by ensemble weather forecasts on an hourly time step, for the Overijsselse Vecht River.**

[1] The hydrological models chosen for this study are HBV and GR4H. These models are both conceptual, lumped/semi-distributed rainfall-runoff models. In this study, the models are executed as semi-distributed for the Vecht and its sub-catchments. The models are widely used in hydrological modeling and forecasting studies and are also used in some operational FFWS around the world (Sun et al., 2020; Thielen et al., 2009; van Heeringen et al., 2013). The models have demonstrated satisfactory performance in simulating high-flow events in the majority of studies, indicating their suitability for this research (Lindström et al., 1997; Bouaziz et al., 2021; Perrin et al., 2003). For a detailed explanation of model selection, the reader is referred to De Groot (2024).

[2] A multi-model forecast is an approach where the forecasts of multiple hydrological models are combined to obtain more accurate forecast performance. Various studies have proven that this approach outperforms the use of a single hydrological model (Velázquez et al., 2011; Wu et al., 2020; Teja et al., 2023; Dion et al., 2021).

[3] The flood forecast performance of the models individually and combined as multi-model are compared and evaluated using forecast verification metrics.

1.5 Research questions

To achieve the research aim, three main research questions are formulated.

1. How do HBV and GR4H perform on high-flow simulations, based on historical data?
2. What is the flood forecast performance of the single hydrological models with the input of weather ensemble forecasts?
3. Does the use of multiple hydrological models provide better flood forecast performance compared to the use of a single hydrological model?

The first research question concerns the preparation of the hydrological models for simulating high flows accurately. This research question includes building HBV and GR4H and calibration and validation of the models. The second question will provide insight into how well the models can forecast a flood event when modeled singularly. The third question and main objective will provide insight if, and how the performance of the use of multiple hydrological models changes in comparison to the models used singularly.

1.6 Research scope

The focus of this study is on quantifying the performance of ensemble flood forecasts between single hydrological models and a multi-model approach. Further analysis of the forecasts regarding flood warning and mitigation are not considered in this study. This study is limited to providing insight and advice regarding the influence of hydrological models in a flood forecast system. Hence, it is also not the goal to develop a state of the art flood forecasting system for the Vecht River basin.

In addition, although increasing weather extremes are one of the main reasons for this research, climate change is not considered in this study. This research focuses only on historically observed events and does not apply climate change scenarios.

2 Study area, Models and Data

2.1 Study area

The Overijsselse Vecht catchment is chosen as study area. The Vecht basin is a regional, transboundary basin, flowing through Germany and the Netherlands (Figure 2). This catchment is within the JCAR-ATRACE program and there is a need for better understanding of hydrological processes during floods and droughts (Klein and van der Vat, 2024). Two of the main goals of JCAR-ATRACE are to improve integrated planning and development & management of regional river catchments to reduce transboundary flood risk (Slager and Kwadijk, 2023, Klein and van der Vat, 2024). Together with the research goal of this study, comparing flood forecast performance between single hydrological models and a model combination, the Vecht catchment is a suitable study area.

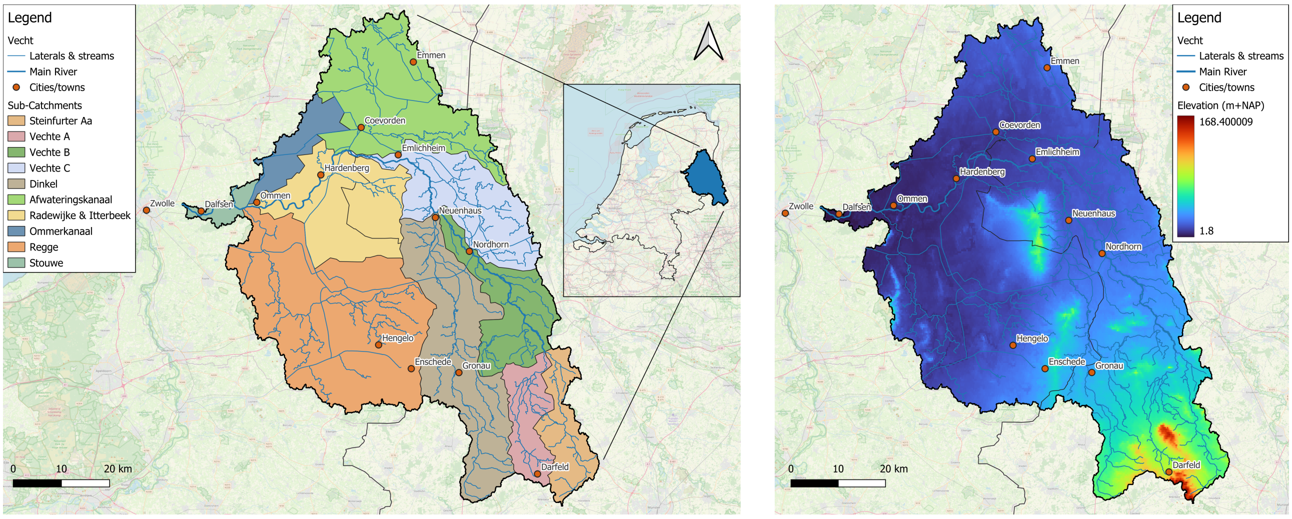


Figure 2: Left: Area of the Vecht river basin, located in the east of the Netherlands and partly in Germany. The map shows a division of the area into 10 sub-catchments. Right: Elevation map of the Vecht catchment

The Vecht originates in the small town of Darfeld in Germany. From here the stream heads northwest, confluencing with the Steinfurter Aa and Dinkel before entering the Netherlands. The stream continues southwest, confluencing with the Afwateringskanaal, Ommerkanaal and Regge, to finally end up in the Zwarte Water near the city of Zwolle (Figure 2). In total, the river stretches over a length of 167 km, of which 60 km is in the Netherlands. The elevation difference along the total trajectory is around 100 m from the origin to the outlet at the Zwarte Water (Figure 2). The total catchment area is around 4000 km^2, of which half is in the Netherlands. The discharge of the river is highly variable, from extreme low flows in summer (0-5 m^3/s) to extreme high flows in winter (250-500 m^3/s) (Verdonschot and Verdonschot, 2017). The duration of a flood wave along the Dutch part is approximately 14 hours (Spruyt and Fujisaki, 2021). The largest tributaries to the Vecht are the Steinfurter Aa, Dinkel, Afwateringskanaal, Regge and Ommerkanaal (Spruyt and Fujisaki, 2021). This study divides the complete Vecht catchment into 10 sub-catchments: Steinfurter Aa, Vechte A, Vechte B, Vechte C, Dinkel, Afwateringskanaal, Radewijke & Itterbeek, Ommerkanaal, Regge and Stouwe (Figure 2). This division is based on how sub-catchments of the Vecht are defined in FEWS-Vecht (the current operational FFWS for the Vecht). In this study, the independent lateral inflows (Steinfurter Aa, Vechte A, Dinkel, Afwateringskanaal, Ommerkanaal and Regge) are further referred to as the ‘upstream catchments’ and the dependent, main Vecht river catchments (Vechte B, Vechte C, Radewijke+Itterbeek and Stouwe), as the ‘downstream catchments’. Sub-catchment characteristics as surface area, mean discharge, maximum discharge and responsible authorities can be seen in Table 1. The mean and maximum discharges have been derived from available hourly discharge measurements (see Section 2.3.2).

Table 1: Characteristics of the sub-catchments considered in this study. Shown is the surface area, mean discharge (Oct-Apr), maximum measured discharge and date (among complete available data) and responsible authorities (Jungermann et al., 2012).

Sub-catchment	Area (km^2)	Contribution (%)	Mean discharge (Oct-Apr) (m^3/s)	Max. discharge (m^3/s)	Date max. discharge	Authority
Steinfurter Aa	204	5.0	2.8	71.6	2010-08-28	NRW
Vechte A	189	4.6	2.9	51.4	2010-08-27	NRW
Vechte B	320	7.8	10.5	92.0	2023-12-26	NLWKN
Vechte C	420	10.3	25.4	181.5	2023-12-27	NLWKN
Dinkel	671	16.4	9.1	82.3	2010-08-30	Vechtstromen & NLWKN
Afwateringskanaal	580	14.2	11.7	91.5	2008-01-22	Vechtstromen
Radewijkerbeek & Itterbeek	492	12.0	40.2	238.0	2023-12-27	Vechtstromen & NLWKN
Ommerkanaal	168	4.1	3.5	26.7	2014-05-28	Vechtstromen
Regge	1014	24.8	12.6	107.3	1998-10-29	Vechtstromen
Stouwe	34	0.8	51.6	300.0	2023-12-27	Drents-Overijsselse Delta
Total	**4092**	**100**				

The German authorities responsible for the Vecht are the institute Niedersächsischer Landesbetrieb für Wasserwirtschaft, Küsten- und Naturschutz (NLWKN) and local government Nordrhein-Westfalen (NRW). In the Netherlands, the regional authorities are the waterboards of Vechtstromen (upstream of Ommen) and Drents-Overijsselse Delta. The Dutch trajectory of the Vecht is highly modified by canalization and flow regulation structures. Over the years, the length reduced from 85 to 60 km due to canalization. Because of these measures, the water level dropped and weirs were built to regulate the discharge. At this moment there are a total of six weirs between Zwolle and the German border (Spruyt and Fujisaki, 2021). Figure 3 shows a top view of one of these weirs at De Haandrik, just downstream of the German border in the Netherlands.

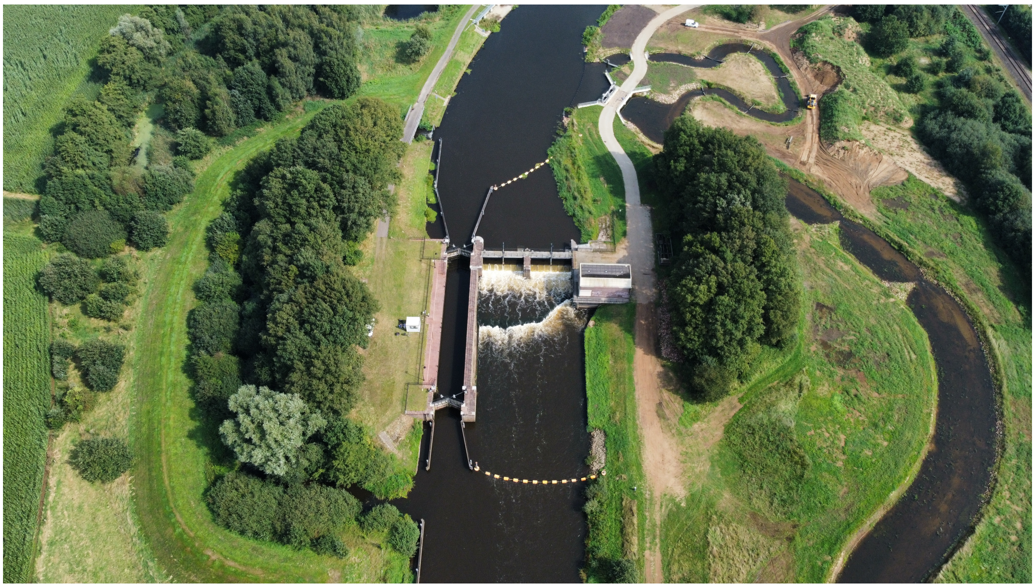


Figure 3: Weir De Haandrik in the Vecht and recently developed fish ladder (right). Photo: William Cazemier

2.2 Models

This section explains the models that are used in this study. First, Section 2.2.1 describes both hydrological models HBV and GR4H. Both model structures are shown, and underlying mathematical relations explained. Section 2.2.2 shows the existing FFWS (FEWS-Vecht) in use by the waterboards responsible for the Dutch part of the Vecht. This forecasting system is not used directly for forecasting, but used mainly as a tool to obtain and process data.

2.2.1 Hydrological models

This study evaluates the flood forecast performance of two rainfall-runoff models: HBV and GR4H. Both models are conceptual models, describing fluxes between water storages by means of several equations and parameters. The parameters require to be calibrated, since they often try to mimic hardly-, or immeasurable processes. Conceptual models are widely used in hydrological simulation and flood forecasting (Kan et al., 2017), making them a good fit for this research. The spatial resolution of both models is generally lumped, meaning spatial variability of inputs, parameters, boundary conditions and catchment characteristics are not considered (Pechlivanidis et al., 2013). Lumped models consider the catchment as one, homogeneous, area. However, this study has divided the Vecht into multiple sub-catchments, making the models semi-distributed. Part of this study is to construct both HBV and GR4H at a semi-distributed scale for the Vecht area and it sub-catchments. The models are often simulated on a daily time step. However, the response of the Vecht is rather in hours than in days. Figure 4 below shows the difference between hourly and daily measured discharge at station Wettringen (located at the outlet of Steinfurter Aa). Clearly, the response to a heavy precipitation event from summer 2010 leads to runoff almost instantly. Additionally, with the focus on predicting high-flow events, modeling with hourly time steps provides more temporal resolution in the forecasts. By using the daily discharge values, the true discharge peak is underestimated and information about the hydrograph shape is lost (Figure 4). To model on an hourly time step, both models require precipitation and potential evaporation time series on an hourly time step. Also, for calibration of the model parameters, historically observed discharge is required on an hourly time step.

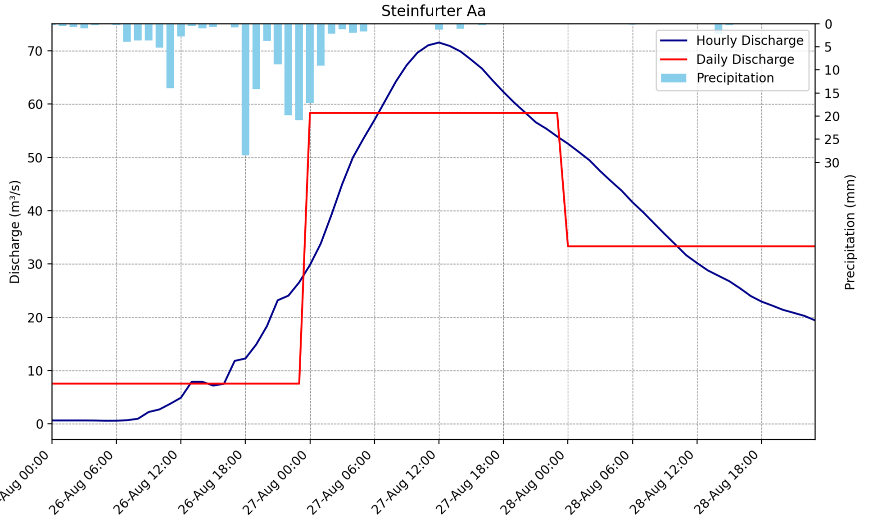


Figure 4: Observed hourly and daily discharge of station Wettringen (outlet of sub-catchment Steinfurter Aa) during a high-flow event in the summer of 2010. Precipitation is measured at rain gauge Steinfurt-Burgsteinfurt in mm/h (Figure 8).

The main difference between HBV and GR4H is the number of parameters they consider and how precipitation forcings are allocated to fast or slow runoff and the storages. This difference makes it interesting to investigate both models in their ability to simulate accurate runoff. The sections below describe the structure of HBV and GR4H and explain in detail how the final runoff is calculated.

HBV

The HBV model was first presented by the Swedish Meteorological and Hydrological Institute (Bergström and Forsman, 1973), and further developed and improved to the version of 1996 (Lindström et al., 1997). The model makes use of four subroutines: a precipitation/snow routine, a soil moisture/evaporation routine, and a quick and fast runoff routine (Lindström et al., 1997). Each component includes parameters that need to be calibrated. In this study the snow routine is not considered, assuming that all precipitation is rainfall. Demirel et al. (2013) mention eight parameters to be most important for a HBV model without snow routine: FC, LP, $BETA$, $CFLUX$, $ALFA$, KF, KS and $PERC$. These are also the parameters that this study will consider. An overview can be found in Table 2. The model considers three storage boxes: soil moisture (SM), upper response zone (UZ) and lower response zone (LZ). The upper response box describes quick runoff response, and the lower response box slow runoff response. Figure 5 shows a schematic of the HBV model. A detailed explanation of the fluxes that are shown in the figure is given below.

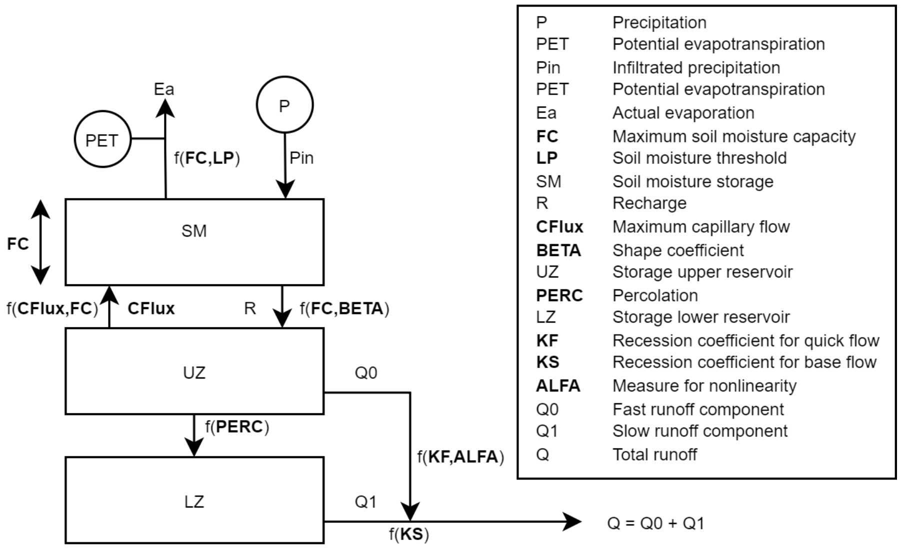

Ea
PET
f(**FC**,**LP**)
P
Pin
FC
SM
f(**CFlux**,**FC**)
CFlux
R
f(**FC**,**BETA**)
UZ
Q0
f(**PERC**)
f(**KF**,**ALFA**)
LZ
Q1
f(**KS**)
Q = Q0 + Q1
P Precipitation
PET Potential evapotranspiration
Pin Infiltrated precipitation
PET Potential evapotranspiration
Ea Actual evaporation
FC Maximum soil moisture capacity
LP Soil moisture threshold
SM Soil moisture storage
R Recharge
CFlux Maximum capillary flow
BETA Shape coefficient
UZ Storage upper reservoir
PERC Percolation
LZ Storage lower reservoir
KF Recession coefficient for quick flow
KS Recession coefficient for base flow
ALFA Measure for nonlinearity
Q0 Fast runoff component
Q1 Slow runoff component
Q Total runoff

Figure 5: Schematic of processes in the HBV model. The conceptualized parameters of the model are presented in bold in this figure. Figure adapted from Shrestha et al. (2009).

The soil moisture box is filled by infiltration of precipitation (P_{in}). The amount of infiltration depends on the precipitation (P), soil moisture level and field capacity (FC) and is described as

$$P_{in} = \begin{cases} P, & SM + P \leq FC \\ FC - SM, & SM + P > FC \end{cases} \tag{1}$$

All precipitation is infiltrated to the soil moisture box if the sum of the level in the box and the precipitation is smaller or equal to the value of FC. If not, only the amount to fill the soil moisture box to field capacity infiltrates. The excess water (P_e) is directed to the upper response box. Recharge (R) to the upper response box depends on P_{in}, the soil moisture deficit (SM/FC), and parameter β. The latter determines how much infiltrated precipitation becomes runoff. The relation is written as

$$R = P_{in} \left(\frac{SM}{FC} \right)^{\beta} \tag{2}$$

Actual evaporation of the model (E_a) depends on the level in the soil moisture box, potential evapotranspiration (PET) and the parameter LP. This parameter describes the limit for potential evapotranspiration. If the soil moisture storage exceeds $LP \cdot FC$, excess precipitation does not increase E_a, but contributes to runoff. The relations are written as

$$E_a = \begin{cases} PET\left(\frac{SM}{LP \cdot FC}\right), & SM < LP \cdot FC \\ PET, & SM \geq LP \cdot FC \end{cases} \tag{3}$$

Both storage boxes SM and UZ are updated each time step as

$$SM = SM + P_{in} - R - E_a \tag{4}$$

$$UZ = UZ + P_e + R \tag{5}$$

Outgoing flows from the upper zone box are the percolation ($PERC$), capillary flux (CF) and Q_0. $PERC$ requires calibration, and does not vary per time step. The percolation can never exceed the level in the upper zone box. If the percolation is higher than the available water in the box, all available water is percolated to the lower zone box. The capillary flux depends on the parameter $CFLUX$ and on the available water in the soil moisture and upper zone box. The fast runoff component Q_0 is calculated based on the level in the upper zone box and parameters KF and α. $PERC$, CF and Q_0 are described as

$$PERC = \begin{cases} PERC, & PERC \leq UZ \\ UZ, & PERC > UZ \end{cases} \tag{6}$$

$$CF = \max\left(CFLUX\left(1 - \frac{SM}{FC}\right); FC - SM; UZ\right) \tag{7}$$

$$Q_0 = \max\left(KF \cdot UZ^{1+\alpha}; UZ\right) \tag{8}$$

After each calculation the storages SM, UZ and LZ are updated with the in- or outgoing fluxes (similarly to Equation 4 & 5).

The lower zone box receives a percolation flux and returns a slow runoff, or base flow, component Q_1. This runoff component depends linearly on LZ, and cannot be higher than the level in the lower zone box. The relation is described by

$$Q_1 = \max(KS \cdot LZ; LZ) \tag{9}$$

Again, after calculation of Q_1, the lower zone box is updated.

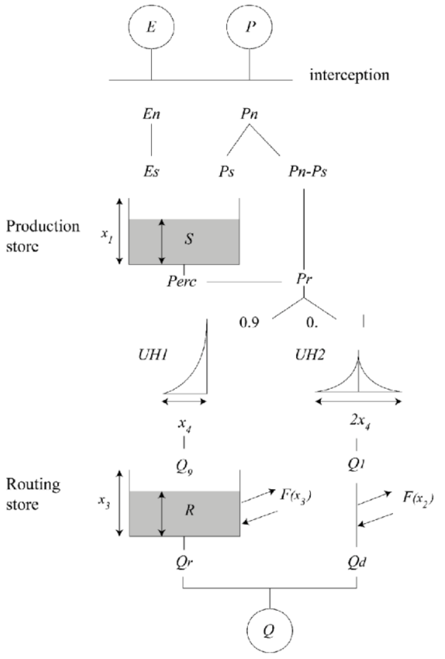
The last step is the determination of the total runoff Q. This can be done with a transformation routine to spread the discharge over multiple time steps, however this is not included in this study. This assumes that the sum of Q_0 and Q_1 is directly discharged per catchment within one time step, instead of spreading over successive time steps.

Table 2: Description of the HBV model parameters

Parameter	**Unit**	**Description**	**Parameter**	**Unit**	**Description**
FC	mm	Field capacity of the soil moisture box	ALFA	-	Quick runoff parameter
LP	-	Threshold for potential evapotranspiration	KF	h^{-1}	Recession coefficient for quick flow
BETA	-	Parameter that influences rate of recharge	KS	h^{-1}	Recession coefficient for base flow
CFLUX	mm/h	Maximum value of capillary flux	PERC	mm/h	Maximum percolation

GR4H

GR4H is an hourly version of the GR4J model, which is based on 4 free conceptualized parameters (Perrin et al., 2003). The model considers two storages: a production store (S) and a routing store (R) that divide rainfall into fast and slow streamflow responses. The fast response is accounted for as a fixed 10% and the slow response as a fixed 90% through two unit hydrographs ($UH1$ & $UH2$). In contrast with HBV, GR4H does take into account a time lag between a rainfall event and streamflow by means of these unit hydrographs. The four model parameters represent the maximum capacity of the production store (X_1), the groundwater exchange coefficient (X_2), the 1 day ahead capacity of the routing store (X_3), and the time base of the unit hydrograph (X_4). An overview of these parameters can be found in Table 3. A schematic of the processes in GR4H can be seen in Figure 6. A detailed explanation of the model is given below.



P	Precipitation
E	Potential evapotranspiration
Pn	Net precipitation
En	Net evapotranspiration capacity
Ps	Precipitation store
Es	Evaporation store
Pr	Total routing quantity
S	Production store
R	Routing store
PERC	Percolation
X1	Maximum capacity of production store
X2	Groundwater exchange coefficient
X3	1 day ahead capacity of routing store
X4	Time base of unit hydrograph
UH1	Unit hydrograph (90% routing)
UH2	Unit hydrograph (10% routing)
F	Groundwater exchange
Q1	Discharge from UH2
Q9	Discharge from UH1
Qr	Discharge reservoir
Qd	Discharge after exchange F
Q	Total runoff

Figure 6: Schematic of processes in the GR4H model. In bold the four conceptualized parameters of the model. Figure obtained from Perrin et al. (2003).

The first step in the model is to determine net rainfall P_n and net evapotranspiration capacity E_n. If the precipitation is equal to or larger than E (potential evapotranspiration), the net rainfall is equal to the difference between the rainfall P and E. In this case there will be no net evapotranspiration. If P is smaller than E, there is no net precipitation and the net evapotranspiration is equal to the difference of E and P. The relations are written as

$$\text{if } P \geq E,\ P_n = P - E,\ E_n = 0 \tag{10}$$

$$\text{if } P < E,\ P_n = 0,\ E_n = E - P \tag{11}$$

In case P_n is not zero, a part P_s fills the production store (S), which is determined by the actual level in the store and parameter X_1. It is described as

$$P_s = \frac{X_1 \left(1 - \left(\frac{S}{X_1}\right)^2\right) \tanh\left(\frac{P_n}{X_1}\right)}{1 + \frac{S}{X_1} \tanh\left(\frac{P_n}{X_1}\right)} \tag{12}$$

In the other case when E_n is not zero (Equation 11) the evaporation flux E_s calculates the amount of water that evaporates from the store. The evaporation store can never exceed the actual water in the production store. It is described as

$$E_s = \max\left(\frac{S\left(2-\frac{S}{X_1}\right)\tanh\left(\frac{E_n}{X_1}\right)}{1+\left(1-\frac{S}{X_1}\right)\tanh\left(\frac{E_n}{X_1}\right)}, S\right) \tag{13}$$

The storage is updated with

$$S = S - E_s + P_s \tag{14}$$

Next, percolation leakage $Perc$ from the production store is calculated as a function of the storage level as

$$Perc = S\left\{1-\left[1+\left(\frac{4S}{9X_1}\right)^4\right]^{-1/4}\right\} \tag{15}$$

The storage is updated, where $Perc$ can never exceed S.

$$S = S - Perc \tag{16}$$

the total amount of water that reaches the unit hydrographs (P_r) is given by

$$P_r = Perc + (P_n - P_s) \tag{17}$$

P_r is routed with the fixed 90% split by a unit hydrograph ($UH1$). The remaining 10% is routed by another unit hydrograph ($UH2$). The unit hydrographs are used to spread rainfall over successive time steps. The length base of the unit hydrographs, i.e. the amount of time steps, is described by X_4. $UH1$ has a time base of X_4 hours and $UH2$ of $2X_4$ hours. The values of X_4 cannot be lower than 12 hours. This is an assumption made from the minimal value of X_4 for GR4J, namely 0.5 days. X_4 can have every real value, but the model calculates on positive integers. This means $UH1$ has m ordinates, with m being equal to the smallest integer exceeding X_4. $UH2$ has ordinates n, where n is equal to the smallest integer exceeding $2X_4$. The distribution of the unit hydrographs is derived from S-curves (SH_1 & SH_2) and is written as

$$SH_1(t) = \begin{cases} 0, & t \leq 0 \\ \left(\frac{t}{X_4}\right)^{\frac{5}{2}}, & 0 < t < X_4 \\ 1, & t \geq X_4 \end{cases} \tag{18}$$

$$SH_2(t) = \begin{cases} 0, & t \leq 0 \\ \frac{1}{2}\left(\frac{t}{X_4}\right)^{\frac{5}{2}}, & 0 < t \leq X_4 \\ 1-\frac{1}{2}\left(\frac{t}{X_4}\right)^{\frac{5}{2}}, & X_4 < t < 2X_4 \\ 1, & t \geq 2X_4 \end{cases} \tag{19}$$

$UH1$ and $UH2$ ordinates are then calculated using

$$UH1(j) = SH1(j) - SH1(j-1) \tag{20}$$

$$UH2(j) = SH2(j) - SH2(j-1) \tag{21}$$

where j is an integer ranging up until the smallest integer exceeding X_4 and $2X_4$. $UH1$ and $UH2$ have a discharge as output, Q_9 and Q_1 which are calculated using

$$Q_9(t) = 0.9 \cdot \sum_{i=0}^{n} \mathrm{UH1}[i] \cdot P_r[t-i] \tag{22}$$

$$Q_1(t) = 0.1 \cdot \sum_{i=0}^{m} \mathrm{UH2}[i] \cdot P_r[t-i] \tag{23}$$

Next a groundwater exchange term F that acts on both the slow and fast runoff component is calculated as

$$F = X_2 \left(\frac{R}{X_3}\right)^{\frac{7}{2}} \tag{24}$$

X_2 is the water exchange coefficient, that can be either negative for water export, positive in case of water imports or zero when there is no exchange. X_3 is the maximum reference capacity of the routing store R. The higher the level in the routing store, the more exchange takes place. The level in the routing store is updated by using

$$R = \max(0, R + Q_9 + F) \tag{25}$$

Outflow Q_r from the slow runoff component is computed as

$$Q_r = R\left(1 - \left[1 + \left(\frac{R}{X_3}\right)^4\right]^{-\frac{1}{4}}\right) \tag{26}$$

The level in the routing store is updated with

$$R = R - Q_r \tag{27}$$

Outflow Q_d from the fast runoff component is subject to the same exchange F and is computed as

$$Q_d = \max(0, Q_1 + F) \tag{28}$$

The total discharge Q is obtained by summing Q_r and Q_d:

$$Q = Q_r + Q_d \tag{29}$$

Table 3: Description of the GR4H model parameters

Parameter	Unit	Description	Parameter	Unit	Description
X1	mm	Maximum capacity of the production store S	X3	mm	1 hour ahead capacity of routing store
X2	mm	Groundwater exchange coefficient	X4	h	Time base of the unit hydrograph

2.2.2 Forecast model system

For the Vecht catchment a version of the flood forecasting model system Delft-FEWS, called FEWS-Vecht is in use. FEWS-Vecht is a stand-alone Delft-FEWS application constructed in 2011 by Deltares and was commissioned by the former waterboard of Velt and Vecht (now part of waterboard Vechtstromen). This system enabled the waterboard to automate its flood forecasting and gain more insight into its operational flood forecasting method (van Heeringen, 2023).

Delft-FEWS was developed by Deltares and first introduced in 2002 (Werner et al., 2013). It is a real-time software infrastructure, designed for operational forecasting (Gijsbers et al., 2008). The system provides an open-shell structure, rather than previously more commonly used enclosed structures. In the enclosed structure approach, the forecasting system is built as a closed shell around the hydrological and hydraulic model. This so-called model-centric approach provides little flexibility in case of changes in the model or the data. This can cause a regular need for redesign or redevelopment of the system every time the conditions change (Werner et al., 2013). The open-shell structure used in Delft-FEWS does not contain hydrological or hydraulic models, but these can be coupled by the forecast operator. This allows for a much more modular and flexible use of the system (Gijsbers et al., 2008). The system has been used in many forecasting applications around the world and has been proven to be a state-of-the-art forecasting system (Werner et al., 2013; Gijsbers et al., 2008).

Within FEWS-Vecht, two hydrological models are incorporated: HBV and Walrus. Next to the hydrological models, a hydraulic model is coupled to FEWS-Vecht which calculates forecasted water levels. The hydraulic model implemented in FEWS-Vecht is Sobek-3. This model contains all relevant waterways and structures in the catchment. In this study, FEWS-Vecht is not used for forecasting. The first reason is because the main purpose of FEWS-Vecht is to provide real-time forecasts for administrators of the waterboards and other water authorities. Also, the Vecht system within FEWS-Vecht is described slightly different, being more complex than is intended in this study. The area is subdivided into more sub-catchments and reaches up until the IJsselmeer. Also, hydraulic processes of the Vecht are considered by the hydraulic model. Figure 7 shows a map from the interface of FEWS-Vecht of the different spatial resolutions of Walrus and HBV. Many of the sub-catchments upstream of Dalfsen are incorporated directly in this study from the Walrus model (as shown in Section 2.1). To keep focus on the hydrological aspect, keep the research feasible and maintain control over the models, FEWS-Vecht is used as tool mainly to acquire data from its vast archive and data processing functions. The last reason to exclude FEWS-Vecht as forecasting model is because GR4H is not connected and integrated in FEWS-Vecht.

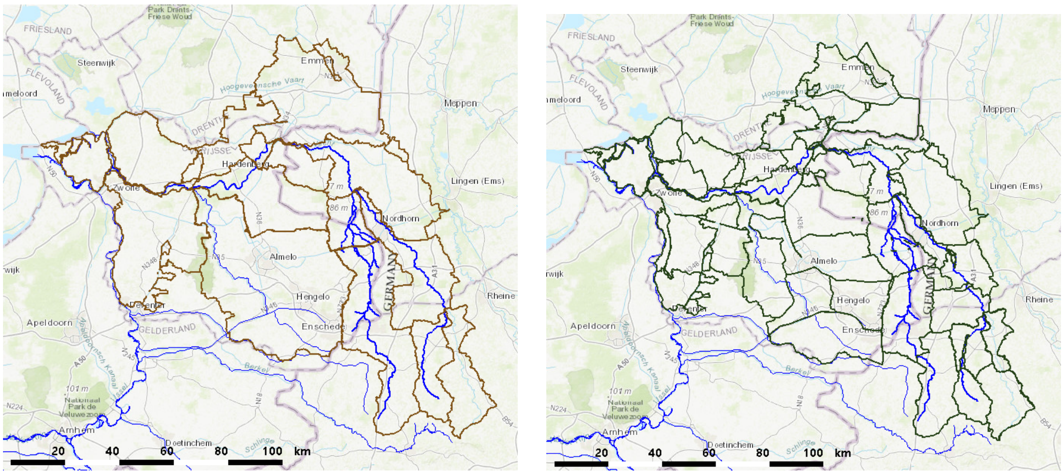


Figure 7: Division of sub-catchments of the Walrus model (left) and HBV model (right) in FEWS-Vecht. The Walrus model considers the Vecht as 14 sub-catchments and HBV as 36 sub-catchments (van Heeringen, 2023).

2.3 Data

In this section the available data used in the study is described. Section 2.3.1 shows available precipitation and evaporation data for the Vecht, and explains which data is used. Section 2.3.2 explains shortly the chosen discharge measuring stations along the Vecht and its tributaries. Lastly, Section 2.3.3 explains the precipitation ensemble forecasts that are available.

2.3.1 Precipitation and evaporation

The hydrological models are forced by historically observed precipitation and potential evaporation time series. Because the models must be run on an hourly time step, hourly precipitation values are required. For high flow prediction, hourly evaporation is less important because the relative contribution of evaporation to the water balance, compared to precipitation during high flow events is low. That is why daily measured values of evaporation are considered sufficient. Also, temperature is not considered because the amount of water storage in the form of snow in the study area is also insignificant for high flow prediction. If a model does consider an snow routine, it is not used and all precipitation is considered to be rainfall.

Precipitation is measured by rain gauges or can be estimated by radar and satellite. In the area there are various weather stations in use by the KNMI, DWD and NRW that measure precipitation. The KNMI has four stations in or close to the Vecht catchment that measure precipitation on an hourly time scale: Heino, Hupsel, Hoogeveen and Twente. The datasets from these stations are obtained from the KNMI website for the period 1995-2024. In Germany, the DWD has seven stations that measure hourly precipitation within or close to the Vecht catchment: Ahaus, Bad Bentheim, Coesfeld, Lingen, Lingen-Baccum, Ringe-Grossringen and Steinfurt-Burgsteinfurt. These datasets are obtained from the Opendata website from the DWD. This data is available from 2006 until 2024, except for Lingen and Lingen-Baccum. Data of gauge Lingen is available from 2006 until 2017, and Lingen-Baccum from 2022 to 2024. The NRW also has a handful of stations in the area, but after early analysis of the data it was observed that these stations deviated much from the DWD stations. Considering that the DWD is the official weather institution of Germany, the reliability and accuracy of their stations is considered to be the best of the two. The data from the NRW stations is therefore not used in this study. Satellite-estimated precipitation is not considered in this study, because the time interval is often on a daily scale and the spatial resolution is less than the local radar products. The locations of the weather stations can be seen in Figure 8.

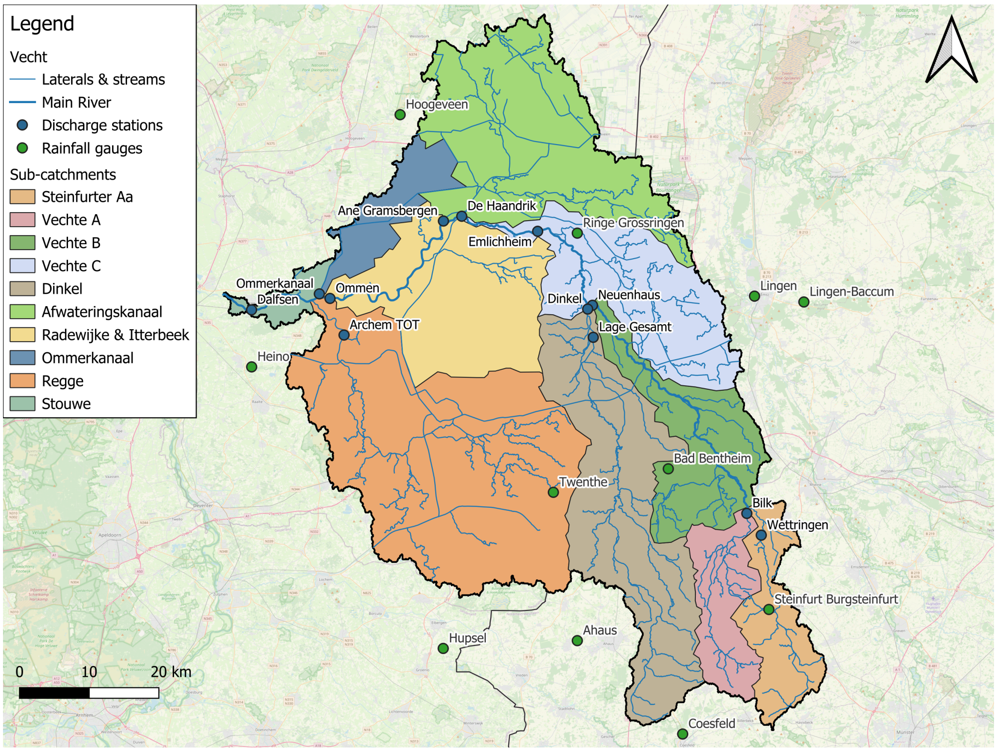


Figure 8: Locations of weather stations (green dots) and discharge stations (blue dots), from which data is used.

In addition to these weather stations, there is a variety of radar products available for the Vecht catchment area. The KNMI has precipitation radars located in Herwijnen and Den Helder, that provide precipitation maps real-time every 5 minutes. These estimates are prone to significant bias, and require post-processing (Imhoff et al., 2021). One technique that is used to correct the radar data is the Mean-Field Bias (MFB). This correction adjusts the radar estimates by comparing with measured rain gauge data, and applying a uniform constant factor across the entire radar area. This way, systematic over- or underestimation of the radar is corrected. Additionally, the radar is spatially corrected with a distance weighted interpolation, called MFBS, to account for local differences (KNMI, 2009). The data is available on a 2.4km x 2.4km grid (1998-2024) and on a 1km x 1km grid (2008-2024). Unfortunately, the time series are not fully continuous and there are some gaps in the data. Effectively, the MFBS 1km radar contains data from 2008 until 2024, missing the years 2009, 2013, 2014, 2016 and 2017. The MFBS 2.4km radar only contains additional data for the year 1998.

The KNMI also makes use of International Radar Composite (IRC). This radar product goes across the border and uses corrections of international weather stations, rather than just KNMI stations (Lempio et al., 2012). This radar data is on a 1km x 1km resolution and available from 2020 onward. Lastly, DWD makes use of Radolan on a 1km x 1km resolution, and is also available from 2020. In total, Radolan contains precipitation estimates from 17 radars combined with hourly measured values at the rain gauges (Deutscher Wetterdienst, 2024). Radolan overlaps with the east of the Netherlands and covers the entire area of the Vecht.

All radar products can be obtained from the archive of FEWS-Vecht. FEWS-Vecht can import the radar products and calculate per time step the catchment averaged precipitation, by using a spatial interpolation function. The result is a time series of precipitation per catchment. Because radar is an estimate of precipitation, the values can deviate from the weather stations, especially on locations far away from a station. Even by spatial correction using the measured rainfall of the weather stations, radar can under- or overestimate measured precipitation on locations far away from the weather stations. Figure 9 shows the available time series of the measured precipitation by the weather stations and radar products.

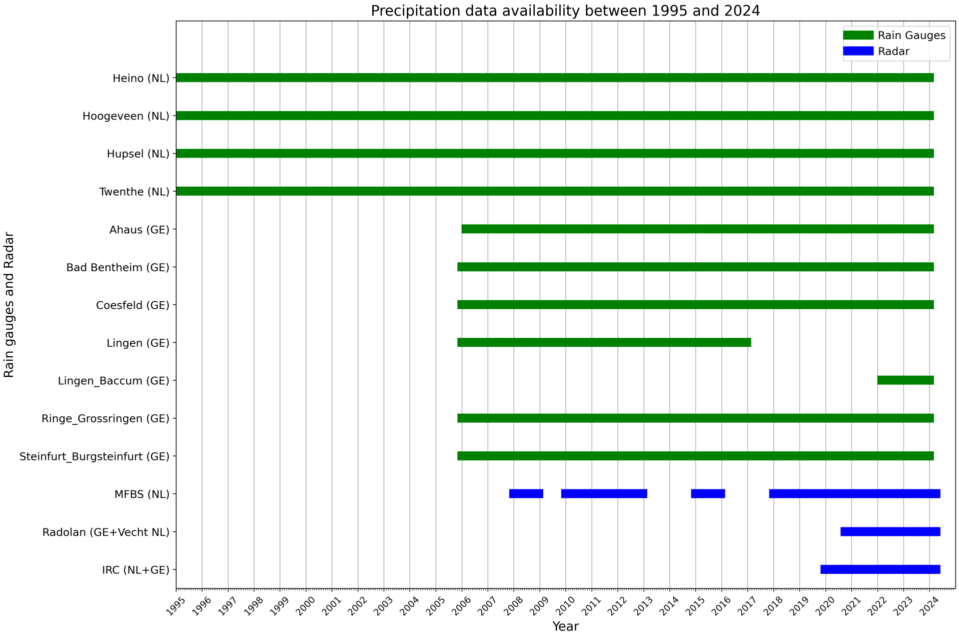


Figure 9: Precipitation data availability between 1995 and 2024. The green bars show the rain gauges in the Netherlands and Germany, and the blue bars the three radar products MFBS (for the Netherlands only), Radolan (Germany and Vecht Netherlands) and IRC (for both countries).

To determine which data to use, the radar products are compared to the weather stations. Table 4 shows for various periods of 6 months the difference between the summed averaged measured precipitation of all weather stations and the summed averaged precipitation measured by radar. Clearly, most radar products underestimate the measured precipitation from the weather stations. Radolan shows the largest deviations, both for the German and Dutch part of the Vecht. Precipitation is continuously underestimated, varying from -4.9% to -25.9%. MFBS shows the smallest deviations, between -2.1% and +3.2%. The IRC product shows precipitation deviations between -8.6% and 0.6% for all catchments. Especially for the German catchments, IRC shows large underestimations of precipitation. It must be noted that the IRC dataset contains most missing values (1.5%) compared to Radolan (0.6%) and MBFS (0.2%) over all periods. Based on this analysis it was chosen to use the MFBS (1km) radar for the Dutch catchments of the Vecht, because this radar product shows the largest similarity with the stations. All radars show a large underestimation in precipitation for the German catchments, so it is chosen to use the data from weather stations instead. For the German catchments, the weather station measured precipitation is interpolated. A simple interpolation, the nearest-neighbour method, is used to determine the allocation of precipitation to a catchment. This method simply assigns a precipitation value to a catchment of the nearest station. If there is no data, the second closest station is used, and so on. The result of this interpolation is used as input only for Steinfurter Aa, Vechte A, Vechte B and Vechte C. The corresponding weather stations are Steinfurt Burgsteinfurt, Steinfurt Burgsteinfurt, Bad Bentheim and Ringe Grossringen. This method is chosen because these weather stations are located mostly around the center of the corresponding sub-catchments, spatially covering the largest part of the catchment.

Table 4: Comparison of precipitation measurements from radar with weather stations for four half year periods The difference indicates an under- or overestimation of the catchment averaged radar precipitation compared to the total averaged measured precipitation of stations in the Vecht area (both Dutch and German stations)

Period		**Timesteps (h)**	**Stations (mm)**	**Radar (mm)**			**Radar difference (%)**				
				Radolan	**MFBS**	**IRC**	**Radolan (Tot)**	**Radolan (Ge)**	**MFBS (NL)**	**IRC (Tot)**	**IRC (Ge)**
1/7/2023 00:00	31/12/2023 23:00	4416	759.9	639.6	743.7	764.3	-15.8	-10.4	-2.1	0.6	-4.4
1/7/2022 00:00	31/12/2022 23:00	4416	308.7	290.1	308.2	288.0	-6.0	-11.2	-0.2	-6.7	-21.5
1/7/2021 00:00	31/12/2021 23:00	4416	371.9	320.4	383.7	353.6	-13.8	-19.6	3.2	-4.9	-23.3
1/10/2020 00:00	31/3/2021 23:00	4368	403.1	298.7	414.2	368.3	-25.9	-18.5	2.8	-8.6	-21.6
Total average			**460.9**	**387.2**	**462.5**	**443.6**	**-16.0**		**0.3**	**-3.8**	

2.3.2 Observed discharge

To assess the goodness of fit of the models, historically observed discharge is required of the Vecht and its tributaries. Like the precipitation time series, these are also required on an hourly time step. Within the Vecht catchment, there are many discharge measuring/estimation systems in use. However, not all prove to be useful because either measurement uncertainty is high or data is incomplete, and therefore untrustworthy or useless. For example, there are many weirs located in the Vecht that measure the water level, from which the discharge can be estimated. However, these structures affect the water flow and do not represent natural flow behavior. These type of measurements are therefore not preferred. Ideally, there is a discharge station at the outlet of each sub-catchment with a continuous data set in a free-flow environment that measures on hourly frequency. Unfortunately, there is not always a station located at the outlet, and the time series are not always consistent. Figure 10 shows the availability of the data from the selected discharge stations that best meet the requirements.

All stations are located at the outlet (or as closely as possible to the outlet) of each sub-catchment. This results in the following catchment-station combinations: Steinfurter Aa - Wettringen, Vechte A - Bilk, Vechte B - Neuenhaus, Vechte C - De Haandrik, Dinkel - Lage Gesamt (and Dinkel), Afwateringskanaal - Ane Gramsbergen, Radewijke+Itterbeek - Ommen, Ommerkanaal - Ommerkanaal, Regge - Archem TOT and Stouwe - Dalfsen. Figure 8 shows the location of the discharge stations.

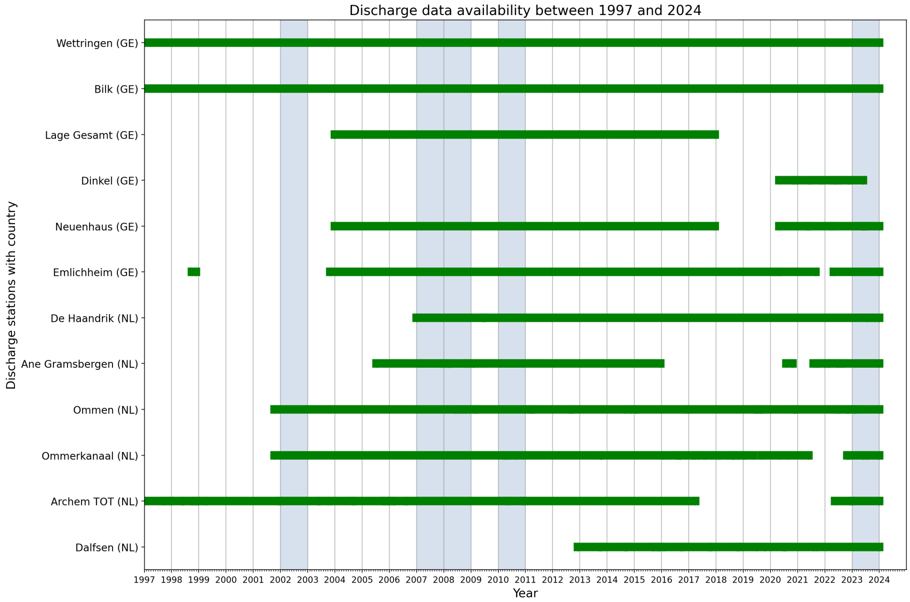


Figure 10: Discharge data availability between 1997 and 2024, for various discharge stations along the Vecht and tributaries. The years in light blue indicate the 5 years with the highest recorded discharges among all stations.

The discharge stations are selected based on their location and measurement reliability. The data set as seen in Figure 10 is composed of multiple sources. Most data originates from the open data websites of the NLWKN, NRW and Rijkswaterstaat. Data from these websites was provided in an internal document at Deltares until 2017. Additional data from that period onward has been obtained directly from the websites. Some data has also been obtained directly from the waterboards of Vechtstromen and Drents Overijsselse Delta. Lastly, some periods are filled with raw measured data from the FEWS-Vecht archive. An overview of all relevant discharge stations in the area, with explanation and source can be found in Appendix A.

Figure 10 also highlights five years in the data set which include a high-flow event (highlighted in blue). This is investigated to find years with high flows along the entire Vecht catchment. These high flows can be used for calibration of the models or for flood forecasting. For each year in the data set, the highest discharge measurement is selected from each discharge station. From this, the five highest recorded discharges per station are visualized in the heat map in Figure 11. The years with the highest flow values among all stations are 2002, 2007, 2008, 2010 and 2023 (respectively 5, 5, 8, 5 and 7 out of the 10 stations measured the highest discharge that year). It must be noted that not all discharge stations have continuous and complete data series (Figure 10). This means that for years with no or limited data, these years are not included in this analysis. An example is the winter of 1998, which is known for high flows in the Netherlands. However, this data set lacks data from that period for most Dutch discharge stations, hence it is not taken into account. By using this approach, overlapping periods with high discharge are found. The figure also shows the maximum value of the discharge that was measured during that year per station.

Some discharge stations measure inflow from canals into the main Vecht. The discharge time series of two canals, Afwateringskanaal and Ommerkanaal show unnatural flow behavior. This is likely the result of human interference in the canals or poor measuring equipment. The same odd flow patterns have been seen for Archem TOT, the measuring station of sub-catchment Regge. It is not clear where exactly this odd behavior originates from. To obtain smoother hydrographs a moving average is applied to the measuring stations of Afwateringskanaal, Ommerkanaal and Regge. The moving average window that is used is 12 hours. This window smooths the hydrographs enough so that small fluctuations were eliminated while keeping larger peaks. Figure 12 shows the raw discharge

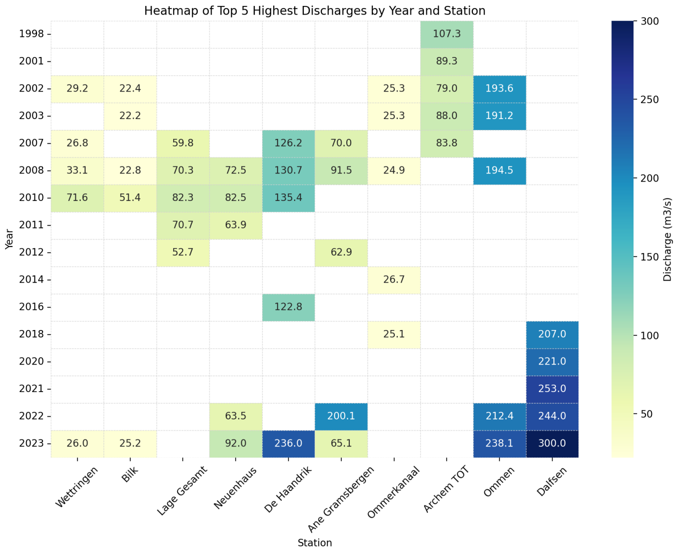


Figure 11: Heat map showing the top 5 highest measured discharges over the period 1997-2024, per station. Years that do not have any high measured discharge for any station are excluded from the figure. Years with the most discharge values are considered to be years that contain at least one high-flow event among most stations. These years are: 2002, 2007, 2008, 2010 and 2023.

observations versus the smoothed discharge over a period of 4 months.

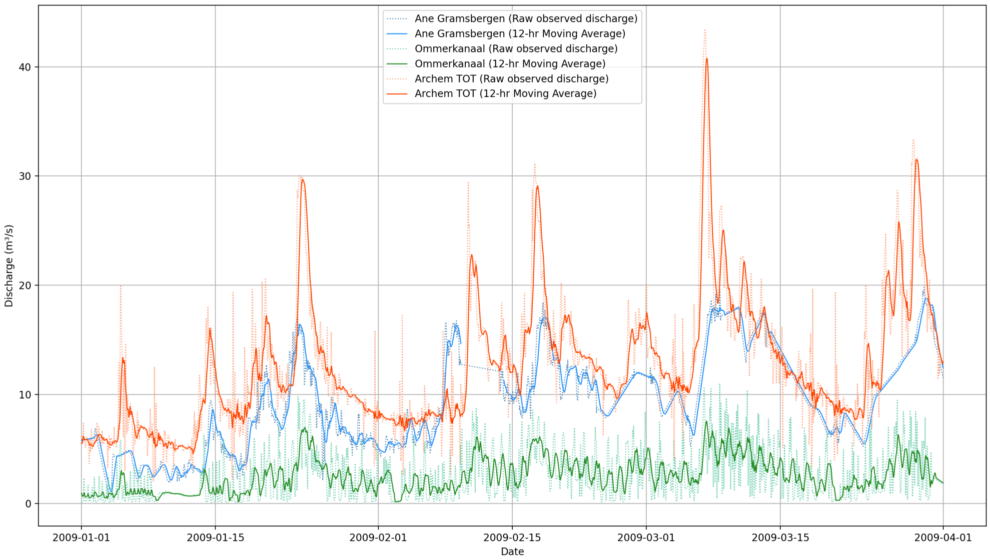


Figure 12: Smoothed discharge series of discharge stations Ane Gramsbergen (Afwateringskanaal), Ommerkanaal and Archem TOT (Regge). A moving average over 12 hours is applied to obtain the final smoothed discharge (solid lines).

2.3.3 Precipitation forecasts

There exist many NWP models that provide precipitation forecasts. Within FEWS-Vecht four models are incorporated: ECMWF, Harmonie, COSMO-LEPS and ICON-EU. Each model has its own forecast properties. Properties that distinguish the models are, for example, the issue of forecast, lead time, forecast resolution, spatial resolution and amount of forecast members. The forecast issue is the frequency at which a new forecast is made. Often, this occurs at an interval of 12 hours, but this can also be more frequent. The lead time is the time horizon of each forecast. Forecast resolution is the time step of the forecast. A forecast resolution of 3 hours means that every 3 hours in the forecast a precipitation value is forecasted. It is most beneficial to have this resolution resemble the hydrological model time step of one hour. Table 5 shows the properties of the NWP models, as applied in FEWS-Vecht. FEWS-Vecht has stored historical forecasts from all models in the archive which can be imported and downloaded. In this study the forecasted precipitation from COSMO-LEPS is used. COSMO-LEPS provides ensemble forecasts at relatively high forecast resolution (3-hourly time steps), medium spatial resolution (7km) and for sufficient lead times ahead (5.5 days) for the Vecht study area. Because the COSMO-LEPS forecasts accumulate precipitation over 3 hours, this must be disaggregated to hourly forecasts. A lead time of 5.5 days is longer than the residence time of the Vecht, which means that all precipitation that is translated to discharge will go through the complete system within the forecasting lead time. Also, in case of evacuation measures in the flood plains of the Vecht, Waterboard Vechtstromen indicated a minimal preparation time of 2-3 days. Forecasts with lead times of 5.5 days should therefore be sufficient. Forecasted potential evapotranspiration is not available, so measured daily PET from weather station Twenthe is used as input.

Table 5: Explanation of the available weather forecast models from the FEWS-Vecht archive (ECMWF, n.d., Marsigli et al., 2013, KNMI, 2017).

	ECMWF-ENS	ECMWF-HRES	COSMO-LEPS	ICON-EU	HARMONIE
Issue of forecast	00 and 12 UTC	00 and 12 UTC	00 and 12 UTC	00, 06, 12 and 18 UTC	00, 06, 12 and 18 UTC
Lead time	+15 days	+10 days	+5.5 days	+5 days	+2 days
Forecast resolution	6 hours	3 hours	3 hours	1 hour and 3 hours	1 hour
Spatial resolution	9 km	9 km	7 km	7 km	2.5 km
Forecast members	51	1	20	1	1
Administrator	ECMWF	ECMWF	COSMO	DWD	KNMI

Precipitation ensembles are not perfect and can provide differences in forecasted precipitation at larger lead times, compared to measured precipitation. Ideally, the ensemble mean, or median, of all ensembles should approximate the measured precipitation. However, it has been found that the COSMO-LEPS ensembles structurally overestimate precipitation compared to measured gauge precipitation and radar. Compared to IRC radar estimations this is in the order of +5% overestimation, but for sub-catchments Steinfurter Aa, Vechte A and Dinkel this is in the order of +50% to more than +100% overestimation of the ensemble mean compared to the IRC radar. The difference between the ensemble mean and measured rain gauge is much less, but still the ensembles structurally overestimate measured precipitation. Figure 13 shows the cumulative precipitation for a forecast issued at 2023-12-13 13:00 for sub-catchment Steinfurter Aa with the different precipitation data sources.

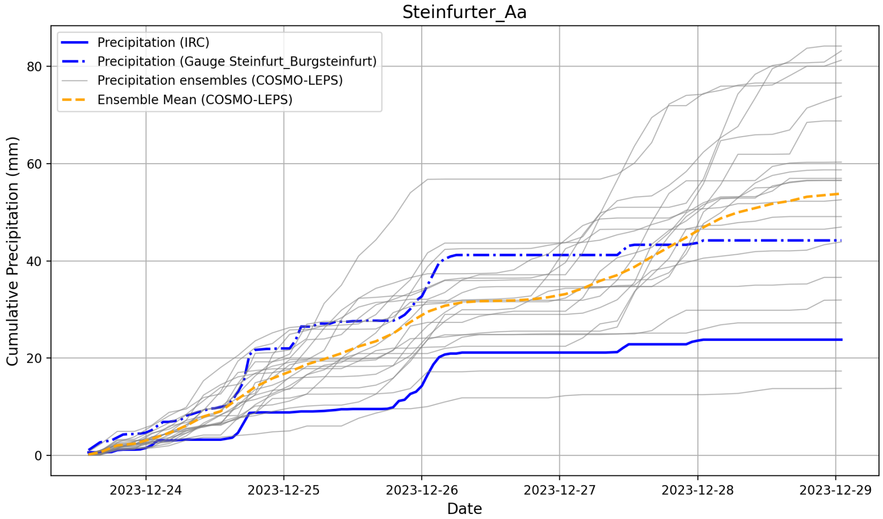


Figure 13: Cumulative precipitation ensembles versus measured rain gauge precipitation (Steinfurt-Burgsteinfurt) and IRC radar for sub-catchment Steinfurter Aa. The forecast is issued at 2023-12-23 13:00, with a lead time of 5.5 days.

3 Method

This chapter describes the steps taken to fulfill the objective of this study. The sections in this chapter are aligned with the research questions. Sections 3.1 and 3.2 describe the flow routing of the model and the calibration and validation of the hydrological models HBV and GR4H. Section 3.2 must give insight into how HBV and GR4H perform on high flow simulations, based on historical data. Section 3.3 explains the state updating procedure and shows the flood forecast model. Section 3.4 explains which evaluation metrics are chosen to evaluate the performance of the model forecasts. Finally, section 3.5 describes how the multi-model forecasts are constructed and evaluated. This section must lead to finding out if the use of a multi-model approach provides better flood forecast performance compared to single models. Figure 14 shows an overview of the applied method.

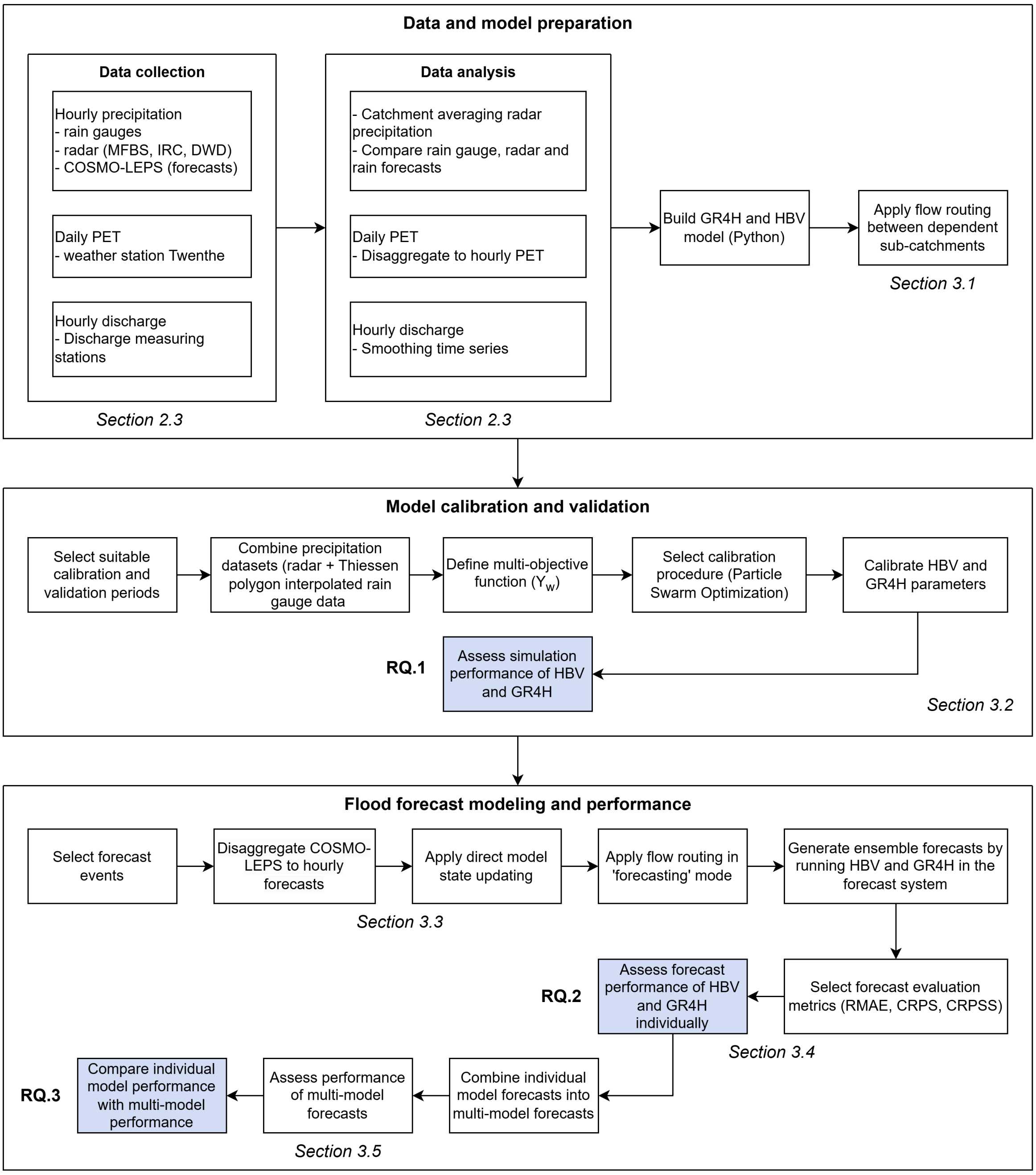

Data and model preparation
Data collection
Hourly precipitation
- rain gauges
- radar (MFBS, IRC, DWD)
- COSMO-LEPS (forecasts)
Daily PET
- weather station Twenthe
Hourly discharge
- Discharge measuring stations
Section 2.3
Data analysis
- Catchment averaging radar precipitation
- Compare rain gauge, radar and rain forecasts
Daily PET
- Disaggregate to hourly PET
Hourly discharge
- Smoothing time series
Section 2.3
Build GR4H and HBV model (Python)
Apply flow routing between dependent sub-catchments
Section 3.1
Model calibration and validation
Select suitable calibration and validation periods
Combine precipitation datasets (radar + Thiessen polygon interpolated rain gauge data
Define multi-objective function (Y_w)
Select calibration procedure (Particle Swarm Optimization)
Calibrate HBV and GR4H parameters
RQ.1
Assess simulation performance of HBV and GR4H
Section 3.2
Flood forecast modeling and performance
Select forecast events
Disaggregate COSMO-LEPS to hourly forecasts
Apply direct model state updating
Apply flow routing in 'forecasting' mode
Generate ensemble forecasts by running HBV and GR4H in the forecast system
Section 3.3
RQ.2
Assess forecast performance of HBV and GR4H individually
Select forecast evaluation metrics (RMAE, CRPS, CRPSS)
Section 3.4
RQ.3
Compare individual model performance with multi-model performance
Assess performance of multi-model forecasts
Combine individual model forecasts into multi-model forecasts
Section 3.5

Figure 14: Overview of the methodology. The method is divided in three main blocks: Data and model preparation, Model calibration and validation and Flood forecast modeling and performance. The blue boxes indicate the final step for each of the three research questions.

3.1 Flow routing

Both models must be able to simulate discharge at the outlet of each sub-catchment. This means the flow dependencies between the sub-catchments should be modeled. This includes discharge routing between an upstream catchment, towards the downstream catchment. This could be done by, for example, a hydraulic routing model (Refsgaard, 1997). This type of routing is based on dynamic equations that take river channel dimensions and hydraulic control structures into account. In this study a simple routing is assumed, keeping the focus of the study on the hydrological processes. This routing involves estimation of flow delay between the discharge at the outlet of an upstream sub-catchment, and the outlet of the subsequent sub-catchment downstream. Based on Spruyt and Fujisaki (2021), the duration of the discharge wave in the Dutch part of the Vecht is 14 hours over a distance of 60 km. This means an average propagation speed of 1.2 m/s is assumed. In this study, it is assumed that the same speed also applies to the German catchments. Because the elevation difference in Germany is larger, this speed may be an underestimate. However, the Vecht in Germany is characterized by a more meandering nature, leading to a possible reduction of propagation speed. In addition to assuming homogeneous propagation speed, the routing also does not take into account channel dimensions. Especially for the four downstream catchments, this could lead to errors in flood peak timing and magnitude.

The delay between the discharge measurement stations is calculated by dividing the distance between the stations by the assumed homogeneous propagation speed. The distance between the stations is estimated by Google Earth and MarinePlan. The latter is a navigation tool for recreational shipping and can calculate the distance of waterways. Table 6 shows the distance between the stations and calculated delay of the discharge wave, assuming 1.2 m/s flow speed. The lateral inflows Dinkel, Afwateringskanaal, Regge and Ommerkanaal join the main Vecht river just after the discharge measurement stations in the Vecht. That is why it is assumed that the water from these inflows takes the same amount of time to reach the next discharge station as the water in the main river channel. As an example, the simulated discharge of Vechte C at a certain time (t) is added to the simulated discharge of Vechte B and Dinkel from 7 hours earlier (t-7). This logic applies only to the models in ‘simulation’ mode. For discharge forecasting, available observed discharge is also used to forecast discharge of the downstream catchments. The delay, however, remains the same between ‘simulation’ and ‘forecasting’ mode. Section 3.3.4 describes the routing of the models in ‘forecasting’ mode.

Table 6: Flow routing between the catchments. The distance between the discharge measurement stations at the outlet of the catchments and the estimated flood wave delay (assuming a flow speed of 1.2 m/s) is shown.

Downstream Catchment	**Inflow Catchment(s)**	**Discharge Stations**	**Distance (km)**	**Delay (h)**
Vechte B	Steinfurter Aa	Wettringen → Neuenhaus	51	12
	Vechte A	Bilk → Neuenhaus	48	11
Vechte C	Vechte B	Neuenhaus → De Haandrik	29	7
	Dinkel	Lage Gesamt → De Haandrik	29	7
Radewijke + Itterbeek	Vechte C	De Haandrik → Ommen	32	7
	Afwateringskanaal	Ane Gramsbergen → Ommen	32	7
Stouwe	Radewijke + Itterbeek	Ommen → Dalfsen	12	3
	Ommerkanaal	Ommerkanaal → Dalfsen	12	3
	Regge	Archem TOT → Dalfsen	12	3

3.2 Model calibration and validation

3.2.1 Objective functions

The goal of the hydrological models is to accurately simulate hydrological behavior of the catchment. To achieve this, the model parameters must be chosen in such a way that the simulated discharge resembles the measured discharge as closely as possible. For conceptual models such as HBV and GR4H, most parameters do not represent a directly measurable catchment characteristic, so calibration is required (Madsen, 2000). This can be done either manually or automatically according to a search scheme and measures of the goodness-of-fit. To find the optimal parameters, a search algorithm called Particle Swarm Optimization (PSO) is used. More about this algorithm can be read in Section 3.2.2. There exist many performance measures to assess the goodness-of-fit of the model, depending on the objective of the model simulation. Madsen (2000) mentions 4 general objectives that measure different aspects of the hydrograph: a correct water balance, a good agreement of the hydrograph shape, a good agreement of high flows and a good agreement of low flows. Combining multiple of these objectives generally leads to an improved fit of the hydrograph during low- and high-flow phases (Pfannerstill et al., 2014). In this study a multi-objective function is used, combining multiple single objective functions. Since the main purpose of this study is to accurately simulate high-flow conditions, the objective function must put emphasis on parts of the hydrograph with high flows. Next to that, it is important that the water balance is correct, to ensure the same amount of water is simulated as is measured.

The first performance metric used to assess the goodness-of-fit of high flows is a weighted variant of the Nash-Sutcliffe efficiency (NSE). The weighted variant ($\mathrm{NS_w}$) was introduced by Hundecha and Bárdossy (2004) and places additional emphasis on high flows. A weight $\mathrm{W_i}$, equal to the observed discharge on day *i*, is multiplied by the difference between the observed and simulated discharge. In case of high observed discharge, the penalty for a large difference in observed and simulated discharge increases proportionally with the observed discharge. According to multiple studies, this weighted version shows better simulation results for high flow conditions, compared to the traditional NSE (Vormoor et al., 2018, Ten Berge, 2024). Like the normal NSE, the value ranges between -∞ and 1, with 1 being a complete resemblance between the observed and simulated discharge. $\mathrm{NS_w}$ is defined as

$$NS_w = 1 - \frac{\sum_{i=1}^{t}\left[w_i\big(\mathrm{Q}_{sim}^{i} - \mathrm{Q}_{obs}^{i}\big)^2\right]}{\sum_{i=1}^{t}\left[w_i\big(\mathrm{Q}_{obs}^{i} - \overline{Q_{obs}}\big)^2\right]} \tag{30}$$

where i is the time step, t the total number of time steps, Q_{sim} the simulated discharge, Q_{obs} the observed discharge, w_i an additional weight equal to Q_{obs} and $\overline{Q_{obs}}$ is the mean of Q_{obs}.

Another single objective function that describes the error in the water balance, the Relative Volume Error (RVE), is also used. This objective function calculates the summed relative difference in simulated and observed discharge over the simulation period. The RVE varies between -∞ and ∞, with a completely closing water balance resulting in a RVE of 0 (e.g., no water is lost or added). The RVE is formulated as

$$RVE = \frac{\sum_{i=1}^{t}(\mathrm{Q}_{sim}^{i} - \mathrm{Q}_{obs}^{i})}{\sum_{i=1}^{t}(\mathrm{Q}_{obs}^{i})} \tag{31}$$

By combining both objective functions, the multi-objective function $\mathrm{Y_w}$ can be formulated (**tenBerge2024Robust** This function combines the weighted NSE function, and also penalizes an incorrect water balance by dividing $\mathrm{NS_w}$ by 1 plus the absolute value of RVE. The perfect score of $\mathrm{Y_w}$ would be obtained with a $\mathrm{NS_w}$ of 1 and a RVE of 0, hence $\mathrm{Y_w}$ of 1. A similar multi-objective function Y was used by Akhtar et al. (2009). $\mathrm{Y_w}$ is formulated as

$$Y_w = \frac{NS_w}{1 + |RVE|} \tag{32}$$

3.2.2 Calibration technique

The calibration routine that will be used to find the model parameters is Particle Swarm Optimization (PSO). This search algorithm is used instead of, for example, Monte Carlo simulations, because PSO converges to increasingly better parameter sets, rather than taking random combinations each iteration (as Monte Carlo does). PSO is an automatic stochastic population-based optimization algorithm. It was first introduced by Eberhart and Kennedy (1995) as an inspiration from the behavior of schools of fish or birds as they collectively search for resources. The goal of such an optimization method is to find a global optimum by minimizing (or maximizing) an objective function (Gill et al., 2006). In this study, PSO tries to maximize the objective function, since the best and maximum value of the objective function $\mathrm{Y_w}$ is 1.

The PSO algorithm consists of a number of particles (swarm) that move within a given search space, coordinating their next position according to the particles' own local best and the swarm's global best. Initially, the particles are placed randomly within the search space with the purpose to compute the objective function. Every iteration, the local and global best of the particles are computed and stored. The position and velocity (direction) of each particle is updated according to best location of the particles until the optimal solution is found (Jahandideh-Tehrani et al., 2020).

In this study, the optimization problem has as many dimensions as it has parameters to calibrate. For GR4H, the particles would be placed within a 4 dimensional search space, bounded by minimum and maximum values of X1, X2, X3 and X4. The position of a particle, i.e. the parameter set, is represented as dimensional vector

$$X_i = (x_{(i,1)}, x_{(i,2)}, ..., x_{(i,D)}) \tag{33}$$

where i is the number of particles and D the number of dimensions. Similarly, the position change, or velocity (Equation 34), previous best particle position (Equation 35) and best swarm position (Equation 36) are given as

$$V_i = (v_{(i,1)}, v_{(i,2)}, ..., v_{(i,D)}) \tag{34}$$

$$P_i = (p_{(i,1)}, p_{(i,2)}, ..., p_{(i,D)}) \tag{35}$$

$$P_g = (p_{(g,1)}, p_{(g,2)}, ..., p_{(g,D)}) \tag{36}$$

where g is the index of the best particle in the swarm. To update each particle's velocity (Equation 37) and position (Equation 38) the following equations are used:

$$\mathrm{v}^{t+1}_{(i,d)} = \mathrm{w}\mathrm{v}^{t}_{(i,d)} + c_1\mathrm{r}^t_1(\mathrm{p}^t_{(i,d)} - \mathrm{x}^t_{(i,d)}) + c_2\mathrm{r}^t_2(\mathrm{p}^t_{(g,d)} - \mathrm{x}^t_{(i,d)}) \tag{37}$$

$$\mathrm{x}^{t+1}_{(i,d)} = \mathrm{x}^{t}_{(i,d)} + \mathrm{v}^{t+1}_{(i,d)}) \tag{38}$$

where d = 1, ..., D is the dimension of the search space, i = 1, 2, ..., N the number of particles and t the iteration number. w is the inertia weight and is used to control the search for a local or global optimum. At first, the inertia weight is high, which favors global search of the search space, and with increasing number of iterations, w decreases which is favorable for local exploration (Gill et al., 2006). c_1 and c_2 are constants that influence the degree of the particle's movement toward its own local best position and towards the best global position. r_1 and r_2 are random numbers between [0,1], therefore introducing stochastic behavior of the particles during their search.

PSO can be used as a function in Python and can be installed and downloaded from the pyswarm library. All the constants and parameters mentioned above can be changed according to the user's preference. However, in this study, only two parameters are varied: particle size and number of iterations. By default, 100 particles and 100 iterations are used; however, since the search space dimensions are large (4 and 8 dimensions), this number is reduced to maintain a feasible model runtime. The other parameters are kept at default values (Table 7). Additionally, minstep and minfunc can be specified, respectively, indicating the minimum step size of particle movement between iterations and minimum change in the objective function value. When this value is reached, the search algorithm stops. The algorithm has thus three ways to stop: reaching (1) the number of iterations, (2) the minimum step size, or (3) the minimum objective function change.

Table 7: Default parameter values of the PSO algorithm from pyswarm in Python

PSO parameter	**Swarm size (N)**	**Number of iterations (t)**	**Inertia weight (w)**	**c1**	**c2**	**minstep**	**minfunc**
Default value	100	100	0.5	0.5	0.5	1e-8	1e-8

The algorithm requires boundaries of the dimensions in the search space. These ranges must represent parameter values that are logically possible and describe hydrological processes in the model realistically. The ranges of the model parameters are obtained from literature. Demirel et al. (2013) used the ranges of model parameters as seen in Table 8. For this study, the same ranges are used. The time dependent parameters CFLUX, KF, KS PERC and X4 have been modified so their resolution agrees with the hourly time step of the model.

Table 8: Model parameter ranges (Demirel et al., 2013)

	Parameter	Unit	Range		Parameter	Unit	Range
HBV				*GR4H*			
	FC	mm	100 - 800		X1	mm	10 - 2000
	LP	-	0.1 - 1		X2	mm	-8 to +6
	BETA	-	1 - 6		X3	mm	10 - 500
	CFLUX	mm/h	0.1 - 1		X4	hours	0 - 96
	ALFA	-	0.1 - 2				
	KF	h^{-1}	0.00021 - 0.021				
	KS	h^{-1}	0.000021 - 0.0083				
	PERC	mm/h	0.00042 - 0.25				

3.2.3 Data processing

The MFBS radar has gaps in the time series (Figure 9). These years must be filled by other precipitation data. To do this, Thiessen polygons are defined around the weather stations in the area (Appendix B). Because MFBS only covers the Dutch sub-catchments, the Thiessen polygons are only drawn for the Dutch weather stations. It is chosen to use this interpolation method instead of the neirest neighbour method, because there are multiple stations covering the sub-catchments. The coverage of each weather station on the area of each catchment is multiplied by the precipitation as weight factor. The result can be seen in Table 9. Note that this applies only to the Dutch sub-catchments.

Table 9: Precipitation weights derived from Thiessen polygons, for the Dutch catchments.

Catchment	**Rain gauge**							
	Hoogeveen	**Hupsel**	**Heino**	**Twenthe**	**Ahaus**	**Coesfeld**	**Bad Bentheim**	**Ringe Grossringen**
Dinkel	-	-	-	0.25	0.095	0.20	0.36	0.095
Afwateringskanaal	0.36	-	-	-	-	-	-	0.62
Radewijke+Itterbeek	0.38	-	0.082	0.20	-	-	-	0.34
Regge	-	0.26	0.26	0.46	-	-	-	-
Ommerkanaal	1	-	-	-	-	-	-	-
Stouwe	-	-	1	-	-	-	-	-

The radar products are imported as grid files from the FEWS-Vecht archive, but these are not yet interpolated to catchment-averaged values. FEWS-Vecht can perform a spatial interpolation function which calculates precipitation from the radar grid files. The precipitation is averaged per sub-catchment as defined by the HBV model in FEWS-Vecht. The first processing step is to average these into the 10 sub-catchments considered in this study. As an example, sub-catchment Dinkel consists of 3 smaller areas in HBV in FEWS-Vecht. These 3 areas are simply averaged to one value to obtain the final catchment-averaged precipitation for the Dinkel.

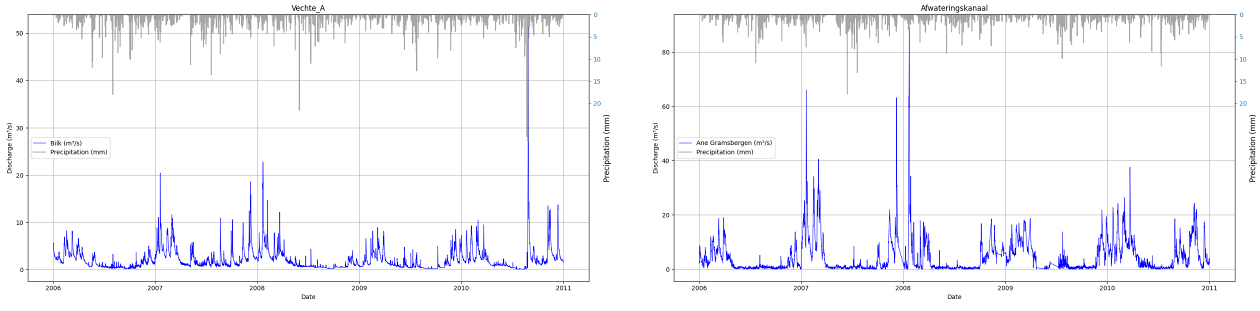
3.2.4 Calibration and validation periods

From the literature, no clear conclusions can be drawn on the question how long the calibration and validation period should be. A generally accepted method is to split the available data into two more or less equal sections for calibration and validation. However, many studies have followed different approaches in which calibration periods are longer than validation and vice versa. Generally, it is agreed that the use of a large part of the available data for calibration increases model performance (Arsenault et al., 2018).

Unfortunately, in this study, the available data is limited. There are not many periods available where there is a complete precipitation and discharge dataset available for all sub-catchments, which is crucial for calibration. As can be seen in Figure 10, there is a continuous period from 2013 to 2015, including 2015, where data is available for each discharge station. This period can be extended from 2007 up to and including 2015 if station Dalfsen is not considered. In theory, it is possible to use the observed discharge sums of Ommen, Ommerkanaal and Archem TOT (Regge) as proxy for the observed discharge at Dalfsen.

The period that is chosen for calibration is between the years 2006 until and including 2010, where 2006 will be a warm-up period, 2007, 2008 and 2010 the calibration period, and 2009 a year without objective function calculation. 2009 is left out, because there is no high flow observation measured by any of the stations (Figure 11). 2006 will be used as a warm-up period to fill initial storages and reach realistic simulated discharge. For this period, only precipitation and potential evaporation input is necessary. During the calibration period, some high-flow events take place, with a specific high-flow event in the winter of 2008 (mainly in the Netherlands) and the summer of 2010 (mainly upstream in Germany). Both events can be seen in Figure 15. Training the model for this period will therefore help in finding the parameter sets that describe high flows best. Unfortunately, the MFBS radar data during this period is discontinuous. Only for the years 2008 and 2010 there is data from the MFBS radar. For the years 2006 (warm-up), 2007 and 2009, rain gauge data is used for the Dutch catchments.

Hydrological model validation is the process of demonstrating the ability of the model to perform outside the calibration period (Refsgaard et al., 2005). When the model performance in the validation period is within acceptable limits or errors, the model is considered validated. To qualify if the models are validated, the obtained Y_w values are compared to the validation results of literature that have used the same objective function.



(a)

(b)

Figure 15: Hydrographs of Vechte A (a) and Afwateringskanaal (b) for the calibration period 2006-2010. In the summer of 2010 a high-flow event took place in the upstream German catchments, and in the winter of 2008 in most Dutch catchments.

Two validation periods are selected: (1) March 2023 - March 2024 and (2) January 2013 - January 2015. The first period includes a high-flow event during Christmas 2023, with the highest measured discharge from 1998 until now at some discharge stations. The second period is not characterized by any remarkable high-flow event and can be considered a robustness test for the models to see if they also perform under 'normal' hydrological conditions. This period is mainly selected according to the availability of data. The precipitation input used for the first validation period consists of MFBS for the Dutch catchments and IRC radar data for the German catchments. According to earlier analysis of the precipitation data (see Table 4), the precipitation sum for the German catchments deviates not much from the stations (-4.4%) during that period. It must be noted that deviations vary per sub-catchment, as Figure 13 has shown for Steinfurter Aa during the flood-peak of the Christmas event. Because a large deviation is only seen between the IRC radar and rain gauge Steinfurt-Burgsteinfurt it is still chosen to use IRC data instead of the interpolated German rain gauges for this period. For the second period, MFBS data are used for the year 2015, and Thiessen polygon weighted rain gauge data for 2013 and 2014. For the German catchments, the rain gauge data from DWD is used.

3.2.5 Experimental setup

Table 10 shows an overview of the calibration and validation setup. For the calibration years, no observed discharge series of Dalfsen is available. Hence, the goodness-of-fit for sub-catchment Stouwe cannot be measured. Because this sub-catchment is relatively small (0.8%), and depends mostly on discharge from the upstream catchments, it is decided to exclude Stouwe from calibration. The parameter set that is assigned to Stouwe will be the best performing parameter set from either Ommerkanaal or Radewijke+Itterbeek. These catchments are closest to Stouwe and are likely to resemble the characteristics of Stouwe best. The other discharge stations used are in line with the catchment-station combinations as defined in Section 2.3.2.

3.3 Flood forecast modeling

The following sections describe how the models are adapted to be able to perform forecast simulations. In this report, the term forecasting refers to the model runs that are made with input of historic meteorological ensemble predictions, in combination with historic observed discharge. In practice, this is also commonly referred to as a hindcast. First, the choice of events to apply the forecast simulations is explained. The meteorological ensemble forecasts are then described. Then, the data assimilation/updating process is explained, and an updating approach is chosen. Next, the ensemble forecast system is shown. Lastly, an explanation of the experimental setup for the forecasts is given.

Table 10: Calibration and validation periods for HBV and GR4H. The available precipitation data varies per period, and is composed of the best available time series.

	Period	Warm-up	Precipitation input		PET input
			NL	**GE**	**NL+GE**
Calibration	Jan 2006 - Jan 2011	Jan 2006 - Jan 2007	MFBS (2008, 2010) Rain gauges (2006, 2007, 2009)	Rain gauges (2006-2011)	Daily measured PET (Twenthe)
Validation	Mar 2023 - Mar 2024	Mar 2023 - Oct 2023	MFBS (2023-2024)	IRC (2023-2024)	Daily measured PET (Twenthe)
	Jan 2013 - Jan 2016	Jan 2013 - Jul 2013	MFBS (2015) Rain gauges (2013, 2014)	Rain gauges (2013-2015)	Daily measured PET (Twenthe)

3.3.1 Forecast events

Precipitation forecast ensembles are imported from the FEWS-Vecht archive. For the COSMO-LEPS ensembles, only forecasts are stored in the archive from 2021 onward. This significantly limits the choice for interesting high-flow periods. It is crucial to have observed discharge available for each sub-catchment for the most accurate discharge forecasts and forecast evaluation. Looking at the available observed discharge in Figure 10 for the period 2021-2024, there is one high flow event. This is the same high flow event that has been used for validation of the models (around the Christmas days of 2023). Since this event has recorded some of the highest discharges at several measuring stations, this is an interesting period to test the model performance singular, and combined. Unfortunately, no discharge is observed from either station Lage Gesamt or Dinkel. Discharge routed from the Dinkel sub-catchment towards the downstream catchments can therefore not be based on observed time series.

To test the models on different kinds of flow events, it is preferred to perform a second forecast on another period. Based on the observed discharge availability there remains just a period from roughly June 2022 to June 2023 for which there is data available for each sub-catchment. During this period, there was no significant high-flow event. Since it is the only period where a complete data set is available, it is still used to test the forecast ability of the models. The chosen period is in March 2023, where between 10 and 13 March a small flood wave was observed throughout the Vecht system. In contrast to the event of Christmas 2023, observed discharges prior to this increased flow of March are low. The period prior to the Christmas event was already wet, and discharge throughout the Vecht already increased. It is useful to test the March event because of the different hydrological conditions.

3.3.2 COSMO-LEPS processing

The COSMO-LEPS forecasts covering the periods mentioned above are imported from the FEWS-Vecht archive. These are forecasts issued every 12 hours at 01:00 and 13:00 (GMT+1) showing the 3 hour precipitation accumulations for 5.5 days (132 hours ahead). Similarly to the processing of the radar precipitation data, a spatial interpolation function is used to calculate the catchment averaged precipitation. Figure 16 below shows an example of this spatial interpolation for a precipitation forecast made on 10 March 2023 at 13:00 (GMT+1), for 19:00 that same day (6 hours ahead). Lastly, the 3 hour accumulations of precipitation are disaggregated to hourly forecasts. In the original data set, a forecast issued at 13:00 contains the first forecasted value at 16:00. This means that the accumulated precipitation forecasts of hour 13:00-14:00, 14:00-15:00, and 15:00-16:00 are given for hour 16:00. To obtain hourly values, the value for 16:00 has been spread evenly over the three previous hours. In the new hourly dataset, a forecast issued at 13:00 contains the first forecasted precipitation value at 14:00. This means that up until and including the hour of each forecast issue, observations are available. The hour containing the first forecasted value is from here on referred to as ‘forecast time’ and always starts at 02:00 or 14:00 (GMT+1).

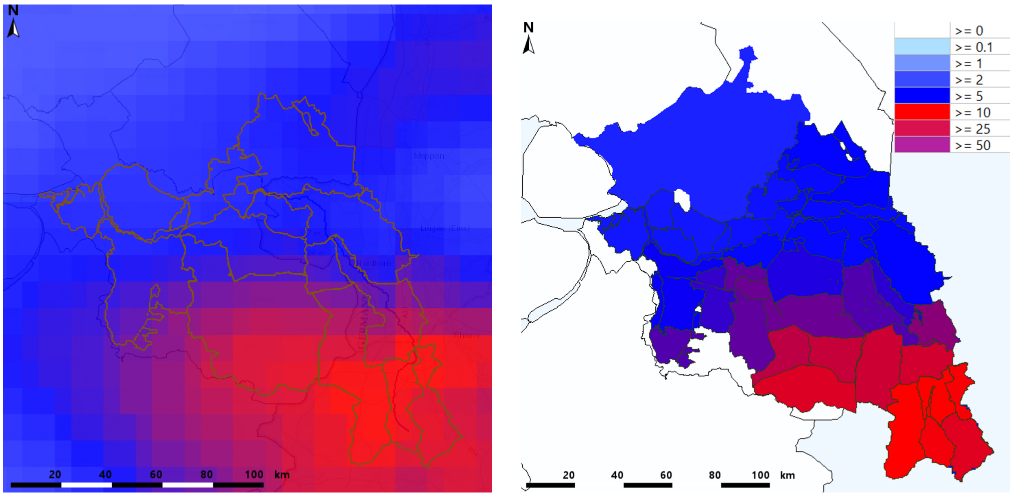


Figure 16: Forecasted precipitation by NWP COSMO-LEPS for 10 March 2023 19:00 (3 hour accumulations). Left: raw precipitation forecast. Right: catchment averaged precipitation forecast after spatial interpolation (image obtained from FEWS-Vecht)

3.3.3 Updating procedure

This section describes the updating procedure of the model during forecasting. In 'simulation' mode, the models calculate discharge based on fixed parameter sets and varying state variables, under input of precipitation and potential evapotranspiration. Models in simulation mode cannot be expected to perform well in a forecast environment (Werner et al., 2005), because they do not take into account the observed real-time discharge. When the models are operated in real-time, the observed discharge at the time step prior to the forecast time could be taken into consideration (in this report referred to as 'forecast issue time'). Up to this time, information about the observed variables is most up-to-date. The feedback process in which real-time data is combined with simulated data in each new (forecast) step to minimize model error is called data assimilation or updating (Refsgaard et al., 2005, Werner et al., 2005). To obtain accurate forecasts, a combination of observed and simulated discharges must be used (Werner et al., 2005).

Data assimilation approaches

Refsgaard (1997) defined four approaches to data assimilation. Figure 17 shows a schematic representation of the structure of a real-time forecasting model with the four approaches to data assimilation. The first approach is the updating of the input variables. Input variables are often considered the largest source of uncertainty in hydrological modeling and forecasting. An advantage of this approach is that the state variables are also automatically updated when the input variables are adapted. However, the problem is that the model itself is in the optimization loop, leading to ambiguity (Werner et al., 2005). Multiple combinations of input could lead to the same result, hence it is not completely clear which adjustments are truly reflective of the real hydrological processes and which simply lead to the smallest model error. The second approach considers updating of the state variables. This can be done by direct insertion, up to complex statistical filters (such as Kalman filtering approaches). The third approach is the update of the of model parameters. Refsgaard (1997) mentions that recalibration of the model at every time step has no real advantages, except for simple (black-box) forecasting systems. Especially for complex conceptual models with a large number of parameters, it is difficult to perform adaptive calibration. In addition to this, it is debated whether it is reasonable to apply changes to the model parameters over short intervals, because the parameters try to represent unique catchment characteristics, which are not likely to change much in short time. The last approach is the updating of the output variables. An output error of the model is statistically defined and is used to adjust the forecast. A commonly used statistical model is autoregressive moving average (ARMA). This method is also used in FEWS-Vecht to update the forecast models (van Heeringen, 2023).

From the four different approaches of data assimilation, the second approach, updating state variables, is used. State updating is often used in hydrological modeling and forecasting and is crucial for medium-range forecasts, since the initial states determine the model output (Werner et al., 2005; Demirel et al., 2013). There are multiple methods to update the initial states of the models. The updating procedure that will be used is the same as that used in Demirel et al. (2013). This method uses empirical relations between the simulated discharge and initial storage to determine the states at a given discharge at the forecast issue time. The same approach is also used in Benninga et al. (2017) to update the storages of a hydrological model. Other methods such as Ensemble Kalman Filtering are widely applied in data assimilation (Werner et al., 2005; Ridler et al., 2014). However, this method is difficult to implement, and multiple studies have shown simpler data assimilation methods to have performed nearly as well as the Ensemble Kalman filter (Refsgaard, 1997; Xiong and O'Conner, 2002).

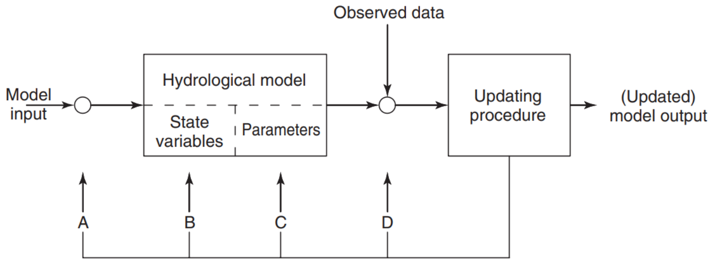


Figure 17: Structure of a forecasting model with data assimilation approaches (after Refsgaard (1997)). A: updating input variables, B: updating state variables, C: updating model parameters, D: updating output variables

Model storage updating

The model storage update procedure of Demirel et al. (2013) uses the empirical relations between simulated discharge and fast runoff of the HBV and GR4J model to divide the observed discharge between the fast and slow runoff components. These components are used to update the routing store (R) in the GR4J model and the upper- and lower response storages (UZ & LZ) in the HBV model. The production store (S) in GR4J and soil moisture storage (SM) in HBV are not directly related to discharge, so these are initiated by a calibrated model run until the forecast issue time.

A big difference between the studies of Demirel et al. (2013) and Benninga et al. (2017) and this study is that the models used in both articles are applied on a lumped spatial resolution. This study considers the Vecht catchment as a semi-distributed area with 10 sub-catchments, of which 4 are dependent on inflow from other sub-catchments. This poses a problem in the estimation of the storages for the downstream catchments, because the relation between simulated discharge and the storages is influenced by the discharge from the upstream catchment(s). Especially the sub-catchments at the end of the flow chain will continuously lose this relationship. To the authors' knowledge, this model storage updating procedure has not been applied to a semi-distributed model. The problem is visualized in Figure 18. The figures show the relationship between simulated discharge and the storages UZ (HBV) and R (GR4H) over the calibration period (2006-2011) for the upstream sub-catchment Dinkel and downstream sub-catchment Radewijke+Itterbeek. In the figure of Radewijke+Itterbeek, the simulated discharge is a cumulative of discharges from the sub-catchments upstream of Radewijke+Itterbeek, rather than the simulated discharge in this sub-catchment only. As an example, the observed discharge at 11-03-2023 13:00 is added to the plots as a vertical dotted line. Say that the model storages should be updated according to the observed discharge at this time, storage values should be used that correspond to the intersection of the observed discharge with the point cloud (blue dots in Figure 18. For upstream sub-catchment the Dinkel, this is a clear relation, however, for downstream sub-catchment Radewijke+Itterbeek this relation is lost. Clearly, the relation between the storages and the simulated discharge dissipates to the point where it is nearly impossible to find a value that corresponds to the observed discharge. Because of this large uncertainty, it is decided to try and find a logic in deciding which storage value to choose. This logic should be applied equally for the HBV and GR4H models, in order to vary as little as possible in updating the model storages. This logic applies only to the four downstream sub-catchments Vechte B, Vechte C, Radewijke+Itterbeek and Stouwe, because the relationship for the upstream sub-catchments is rather clear.

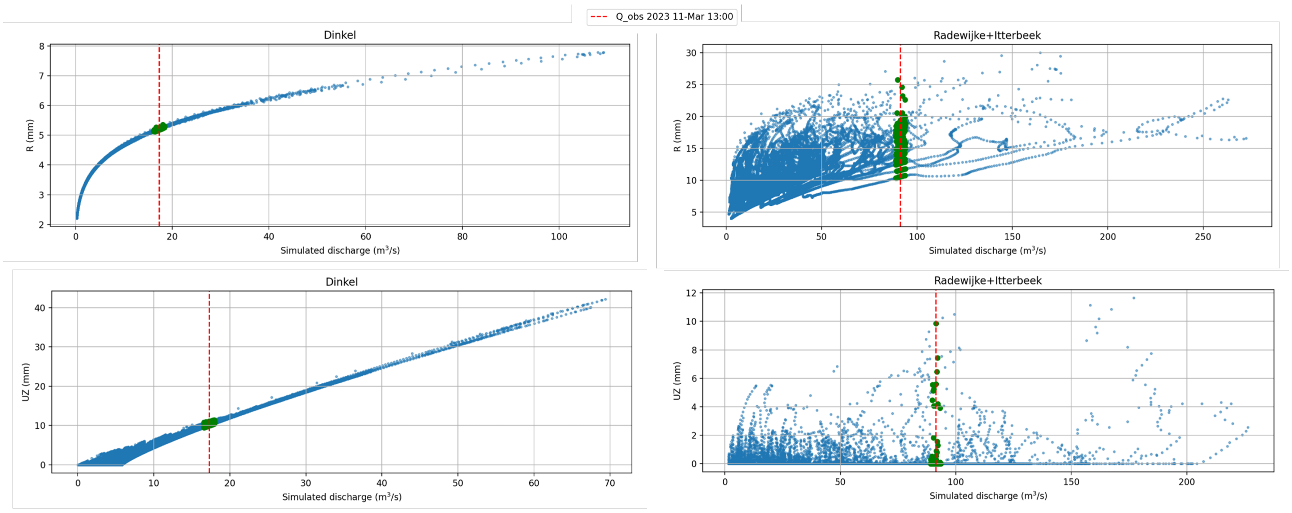


Figure 18: Empirical relationship between storages R (GR4H) and UZ (HBV) and simulated discharge over the calibration period 2006-2011. As an example, upstream sub-catchment Dinkel (left) shows a clear relationship, but this relationship diminishes further downstream at Radewijke+Itterbeek (right). The dashed red line shows the observed discharge at forecast issue time 11-03-2023 13:00. The green dots show storage values at the intersection of the observed discharge and the simulated discharge, with a 1% margin of the maximum simulated discharge as bufferzone.

Two different logics are applied to determine the initial storages of the downstream sub-catchments. A distinction has been made between low-flow (baseflow) conditions and increased flow. In case the observed discharge at the forecast issue time consists of baseflow only, all storages of the downstream sub-catchments are put at 0. The discharge generated at the forecast time in these catchments will only consist of the observed or simulated discharge from the sub-catchment(s) upstream. This proved to provide the best estimate of the initial storages. For all other flow regimes, another logic is applied that considers estimation of the storages based on the values obtained from the upstream sub-catchments, in combination with boundary conditions. For example, the value of the routing store (R) in GR4H cannot exceed the parameter X3. Because of this, it is catchment-specific what values of R are computationally possible. Next to this boundary condition, the maximum values for the routing store are determined by looking at the highest occurring value from the long-term calibrated simulation run. For Vechte B the chosen value of UZ, LZ, and R is the average obtained from Steinfurter Aa and Vechte A. If this value exceeds the boundary conditions, the boundary condition is chosen. For Vechte C the chosen UZ and LZ becomes the same values as Dinkel, and R the same as Vechte B. This different approach was necessary because the values of the routing store for the Dinkel are much lower than of Vechte C due to the calibration of X3. Using the same routing store value as the Dinkel for Vechte C would in this case lead to underestimation of the routing store of Vechte C. For this reason the value from Vechte B is adopted, because the shape and magnitude of the relationship graphs of Vechte B and Vechte C are mostly similar. The storage values for Radewijke+Itterbeek and Stouwe are estimated by taking the values of Afwateringskanaal and Regge. For the upper storage of HBV, the boundary value of Radewijke+Itterbeek and Stouwe is set to 0, because this visually gave the best results. Lastly, for forecasting period 1 (Christmas 2023), no observed discharge is available for sub-catchment Dinkel. This means that it is unclear which storage value to choose. Only for this period, the values of UZ and LZ are determined by the average of Steinfurter Aa and Vechte A, and the value of R is chosen by taking the value of Regge. An overview of the logic can be seen in Table 11.

Table 11: Storage update logic for sub-catchment Dinkel (missing observed discharge during Christmas 2023) and the downstream sub-catchments. The values of the storages of HBV (UZ, LZ) and GR4H (R) are estimated using (averaged) values from the upstream sub-catchments. When this results in impossible or inaccurate values (according to relationship between simulated discharge and storages), boundary conditions are used. A distinction is made between normal and baseflow conditions in the determination of these boundary values.

Catchment	**Estimation of UZ (HBV)**	**Estimation of LZ (HBV)**	**Estimation of R (GR4H)**	**UZ bound (mm)**		**LZ bound (mm)**		**R bound (mm)**	
				Normal flow	**Baseflow**	**Normal flow**	**Baseflow**	**Normal flow**	**Baseflow**
Dinkel	Average Steinfurter Aa + Vechte A	Average Steinfurter Aa + Vechte A	Average Regge	-	-	-	-	10	10
Vechte B	Average Steinfurter Aa + Vechte A	Average Steinfurter Aa + Vechte A	Average Steinfurter Aa + Vechte A	30	0	22	0	20	0
Vechte C	Dinkel	Dinkel	Vechte B	100	0	22	0	20	0
Radewijke+ Itterbeek	Afwateringskanaal	Afwateringskanaal	Afwateringskanaal	0	0	300	0	20	0
Stouwe	Regge	Regge	Regge	0	0	300	0	70	0

3.3.4 Ensemble forecast system

The last step to obtain a forecasting model system is the application of the flow routing through the model. In ‘simulation’ mode this consists of the simulated discharge delay from catchment to catchment. In ‘forecasting’ mode a combination of observed and simulated discharge is used to determine the forecasted discharge. Especially for short forecast lead times, the observed discharge at an upstream catchment is the best estimation of the discharge later in time at a catchment located downstream. The delay between the catchments does not change and remains the same (Table 6). However, as long as there is observed discharge available, i.e. the current lead time is smaller than the flow routing delay, observed discharge from the catchment upstream is used to determine the forecasted discharge in a catchment downstream. As an example, a forecast is issued at 23 December 13:00. At this time step it is assumed the last observed discharge is measured. At the first forecast time at 14:00 (t=0) the forecasted discharge at the most downstream catchment Stouwe consists of the observed discharge of Radewijke+Itterbeek at 11:00 (t-3), observed discharge of Regge at 11:00 (t-3), observed discharge of Ommerkanaal at 11:00 (t-3) and the simulated discharge in sub-catchment Stouwe based on the forecasted precipitation in the area. At a lead time of 3 hours at 17:00 (t=3) there is no observed discharge available of the catchments upstream of Stouwe, so the simulated discharge is used. This creates an uncertainty, because there will be differences in observed discharge and simulated discharge, even though the storages have been updated to match the observed discharge at the forecast issue time. Because sub-catchment Stouwe and Radewijke+Itterbeek are connected in series, the forecasted discharge of Radewijke+Itterbeek does also consist of observed discharge from Vechte C and Afwateringskanaal. According to the flow delay between these sub-catchments the observed discharge is used until 21:00 (t=7), and from 21:00 the simulated discharge. This process continues upstream up to Vechte A and Steinfurter Aa. Note that the forecasted discharge from the six upstream catchments never consists of observed discharge, only simulated discharge. Figure 19 shows a flow diagram of the ensemble forecast system. The colors group the 10 sub-catchments, starting at the bottom right with Vechte A and Steinfurter Aa. The diamond shaped boxes indicate input into the system, being either the forecasted ensemble precipitation $P(t)$ or observed discharge $Q_{obs}(t)$. The squared boxes represent the processes that take place in the calibrated hydrological model. The output from these boxes is the simulated discharge $Q_{sim}(t)$. The circular boxes represent forecasted discharge $Q_{forecast}(t)$. The flow routing delay is indicated by the arrows between the boxes. For each new forecast that is issued (every 12 hours) the storages are updated and the forecasted discharge is calculated with the new precipitation ensemble predictions and observed discharge. In a real-time prediction system it would make sense to apply the updating of storages also in between the issued forecasts (so at every time step). However, since it is not the goal to provide real-time forecasts in this study this is not applied here.

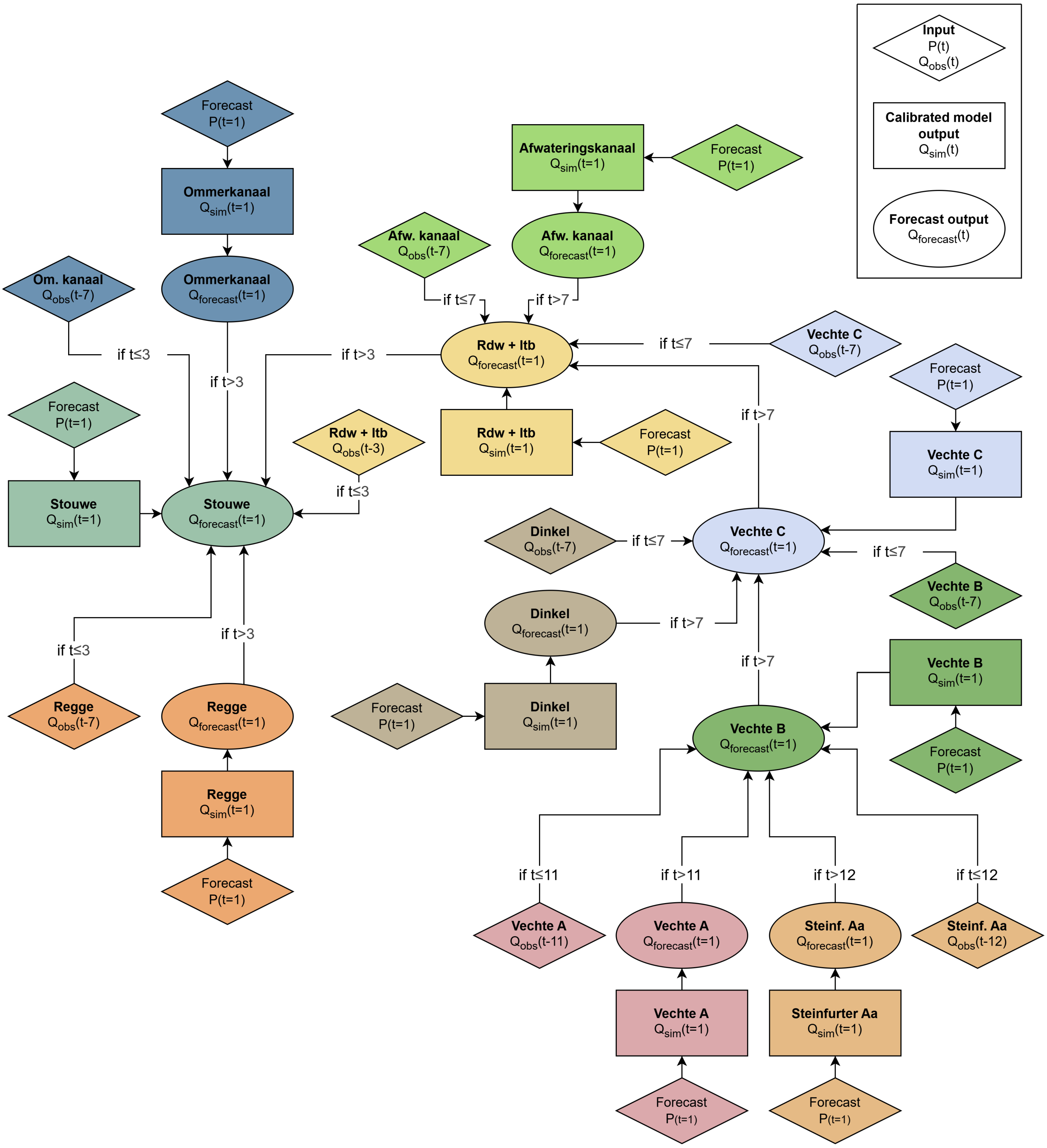


Figure 19: Flow diagram of the ensemble forecast system. The chart represents the Vecht system as described by the 10 sub-catchments. The diagram shows the input of the model ($P(t)$ and $Q_{obs}(t)$) in the diamond shaped boxes, the calibrated model output ($Q_{sim}(t)$) in the squared boxes and the forecast output ($Q_{forecast}(t)$) in the circular boxes. The arrows describe the flow routing between the sub-catchments, based on the time step (t) in hours ahead from the forecast time.

3.3.5 Experimental setup

From the two selected historical events, a short period is selected for which forecasts are produced. This is done so that each event consists of ten forecasts. The first forecast is selected at multiple days before the highest observed discharge was observed. The ten forecasts span a period of (10x12=) 120 hours, plus the lead time of 132 hours of the last forecast, resulting in a total of 252 hours (10.5 days) that cover the events. For both events, this is a sufficient amount of time to capture the highest flood wave peak. Forecast 1 covers the Christmas 2023 event and the first forecast is issued at 23-12-2023 13:00 and the last at 28-12-2023 01:00. During this period the observed discharge for sub-catchment Dinkel is not available. This means the storages of Dinkel can not be updated based on observed discharge. It is chosen to leave this discharge out of the analysis because using the simulated discharge of the Dinkel would lead to large jumps in the forecasted discharge of the downstream catchments at the time step where the model switches from observed discharge to simulated discharge of catchments that include the discharge from the Dinkel (so downstream of Vechte C). For evaluation of the forecasts, these gaps are filled by means of linear interpolation between the first known points. The inputs of the system are the 20 ensemble member COSMO-LEPS precipitation forecasts at hourly resolution and the daily measured PET from weather station Twenthe.

The second forecast period is issued at 08-03-2023 01:00 and the last forecast is issued at 12-03-2023 13:00. For this forecast all sub-catchments have an observed data series available. Unlike past simulations, discharge station Dinkel is used instead of Lage Gesamt. Station Lage Gesamt has no data available during this period, hence raw discharge data from the FEWS-Vecht archive are imported from station Dinkel. This discharge series, together with the series from De Haandrik and Neuenhaus are smoothed using a moving average over 6 hours. An overview of the setup of the forecasts can be seen in Table 12.

Table 12: Setup for ensemble forecasting. Forecast 1 consists of 10 forecasts during the period 23-12-2023 13:00 to 28-12-2023 01:00. Forecast 2 consists of 10 forecasts during the period 08-03-2023 01:00 to 12-03-2023 13:00.

	First Forecast	Last forecast	No. forecasts	Precipitation Forecasts	Resolution	PET input	Remarks
Forecast 1	23-12-2023 13:00	28-12-2023 01:00	10	20 ensemble COSMO-LEPS	Hour	Daily measured (Twenthe)	No discharge time series available for sub-catchment Dinkel
Forecast 2	08-03-2023 01:00	12-03-2023 13:00	10	20 ensemble COSMO-LEPS	Hour	Daily measured (Twenthe)	Observed discharge series from discharge station Dinkel is used instead of Lage Gesamt

3.4 Flood forecast performance evaluation

This section explains how the ensemble forecast system is evaluated. Forecast evaluation (also called forecast verification) is the process of assessing the quality of a forecast (Anctil and Ramos, 2019). This can be done visually by inspecting the discharge forecasts in a hydrograph alongside the observed discharge. However, this only allows for evaluation of one single forecast at a time. To evaluate the quality of the forecast system as a whole, the results of many more forecasts should be used. Forecast quality is often described by forecast attributes (e.g., bias, accuracy, reliability, sharpness, and resolution) and numerically quantified by evaluation metrics. Section 3.4.1 shows for each forecast attribute a handful of forecast evaluation metrics that are commonly used to quantify the quality of a forecast system. From these metrics, several are chosen to be used to evaluate the forecast system in this study. Sections 3.4.2, 3.4.3 and 3.4.4 elaborate further on the chosen evaluation metrics.

3.4.1 Forecast attributes and evaluation metrics

Forecast quality can be assessed using a wide range of evaluation metrics (or scores) (Anctil and Ramos, 2019). Since forecast quality is described by multiple forecast attributes, it is also necessary to use a variety of evaluation metrics to obtain a good understanding of overall forecast performance (Demargne et al., 2010; Anctil and Ramos, 2019). Depending on the goal of the forecast system, multiple metrics should be selected to avoid mis-evaluation (drawing wrong conclusions), over-evaluating (excessive or redundant use of metrics that evaluate the same qualities) or under-evaluation (leave out evaluation of a forecast attribute) (Anctil and Ramos, 2019).

Metrics can evaluate a single-valued (deterministic) forecast, or ensemble forecasts (probabilistic). Deterministic metrics are, for example bias, mean absolute error (MAE), relative mean error (RME), relative mean squared error (RMSE) and Pearson correlation coefficient (Verkade et al., 2013; Anctil and Ramos, 2019). These metrics evaluate forecast attributes bias and accuracy. Bias refers to the difference between the average forecast and the average observation and accuracy to the average difference between individual forecasts and observations (Anctil and Ramos, 2019). Normally, the ensemble mean or median is used to create a deterministic forecast from an ensemble set (Anctil and Ramos, 2019; Brown et al., 2010). Probabilistic metrics that take the uncertainty from all ensemble members into account are, for example, the Brier score, continuous ranked probability score (CRPS) and a variant that scores the skill of the forecast with respect to a reference into account (CRPSS). These metrics are widely used in forecasting studies to assess the overall forecast accuracy (Velázquez et al., 2011; Benninga et al., 2017; Anctil and Ramos, 2019). Reliability is the statistical consistency between the measured and the simulated discharge. Metrics that evaluate reliability are, for example, the reliability diagram and rank histogram (Benninga et al., 2017; Anctil and Ramos, 2019). Sharpness is the tendency to forecast high and low flow extremes, instead of forecasts close to mean or climatological probabilities. To evaluate sharpness forecast frequency, histograms can be used. The last evaluation attribute, resolution, is the ability to correctly forecast (non)occurrence of events. Relative operating characteristic (ROC) curves can be used to evaluate resolution (Velázquez et al., 2011; Benninga et al., 2017).

The forecast evaluation metrics that are used to evaluate the quality of the forecasts are the relative mean absolute error (RMAE) and the continuous ranked probability score (CRPS) and its skilled variant (CRPSS). By using these metrics, the attributes bias and accuracy are assessed. Other metrics to evaluate the forecast reliability, sharpness and resolution, such as ROC and rank histograms are not used because they require a large sample size (many forecasts) for evaluation. Since this study forecasts only for two short flood events consisting of 10 forecasts, the number of forecasts is insufficient. The RMAE is the same as the relative mean error, however, the absolute value of the errors is calculated such that negative and positive errors will not cancel each other out. All three metrics are explained in more detail in Section 4.2.2, 4.2.3 and 4.2.4.

3.4.2 Relative Mean Absolute Error

RMAE measures the average absolute difference between the forecast and the observed discharge, relative to the observed discharge. The RMAE is calculated using:

$$RMAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Q_f(i) - Q_o|}{Q_o} \tag{39}$$

where n is the number of forecasts, Q_f the discharge forecast, and Q_o the observed discharge. The RMAE can take any value between 0 and infinity, where a value of 0 indicates perfect resemblance to the observed discharge. Similarly to CRPS, the RMAE can be averaged among the number of forecasts to obtain a single-valued estimation of the error (Equation 39). It is common practice to use the ensemble mean (Verkade et al., 2013) or median, as value of Q_f instead of averaging the errors of the individual ensemble members. In this study, the ensemble median will be used to evaluate the RMAE of the single model and MHM forecasts.

3.4.3 Continuous Ranked Probability Score

The CRPS is a commonly used forecast evaluation score in the forecasting literature to measure the general accuracy of probabilistic forecasts (Pappenberger et al., 2015; Verkade et al., 2013; Teja et al., 2023; Velázquez et al., 2010). CRPS compares the distance between the cumulative distribution function (CDF) of the forecast with the CDF of the observed discharge (Trinh et al., 2013). The area differences between CDFs are equal to the integral of the squared difference between all forecast-observation pairs (Teja et al., 2023). The CRPS can be calculated with the following equation (Velázquez et al., 2010):

$$CRPS(F, y) = \int_{-\infty}^{\infty} (F(x) - H(x \geq y))^2 dx \tag{40}$$

where F is the CDF derived from the ensemble forecast, x the predicted variable (discharge) and y the corresponding observed value. $H(x \geq y)$ describes the Heavyside function, which is equal to 1 for forecasted values larger than the observed value and 0 for forecasted values lower than the observation (Velázquez et al., 2010). Figure 20 shows the principle of CRPS. The blue line shows the CDF of the observed discharge, and the red line the CDF of the ensemble forecasts. The CRPS has the unit of the variable to be evaluated, in this case m^3/s. Equation 40 defines the CRPS for one single forecast. In practice, the CRPS is averaged among multiple forecasts, obtaining the mean CRPS for multiple events (Verkade et al., 2013; Teja et al., 2023). The CRPS is always positive, and a value of 0 indicates a perfect (deterministic) simulation. For deterministic forecasts, the CRPS is reduced to the mean absolute error (Pappenberger et al., 2015). Because the observed discharge varies by catchment, CRPS cannot be directly compared between different areas or seasons (Benninga et al., 2017). In order to compare between the catchments the CRPS can be normalized against the standard deviation or against a reference CRPS. Because it is useful to compare the forecast performance between the sub-catchments the CRPS will compared against a reference system. This procedure is explained in the next section.

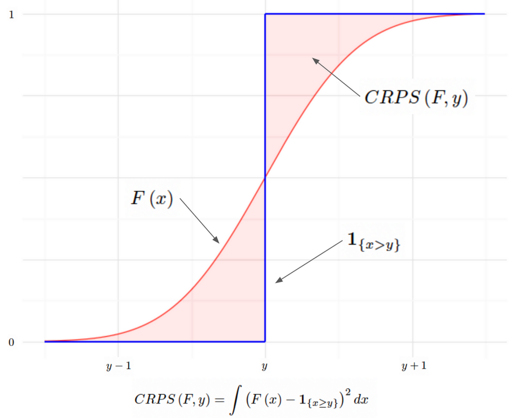


Figure 20: Principle of CRPS. The red line describes the CDF of the forecasted discharge. The area in red is the calculated CRPS, based on the CDF of the observed discharge (blue line). Figure obtained from Faran (2023).

3.4.4 Continuous Ranked Probability Skill Score

The CRPS can be normalized with a reference CRPS to obtain the continuous rank probability skill score (CRPSS). The CRPSS is defined as:

$$CRPSS = 1 - \frac{CRPS_{for}}{CRPS_{ref}} \tag{41}$$

The reference forecast should be a robust benchmark to the forecast to be evaluated (Bennett et al., 2014). The reference forecast can be seen as an alternative to the actual modeled forecasts. CRPSS is unit-less and a value of 1 indicates that the CRPS of the modeled forecast is 0, and thus a perfect forecast simulation. CRPSS below 0 indicates that the reference forecast has a better CRPS score than the modeled forecast. As a reference system, persistency and hydrological climatology are often used (Pappenberger et al., 2015). Hydrological climatology reference forecast consists of the average past observed discharge on the same calendar day (or hour) as the modeled forecast. A persistency reference forecast consists of the last observed discharge before the forecasting time, as forecast for the entire lead time. Benninga et al. (2017) and Bennett et al. (2014) investigated the CRPS of both reference forecasts against lead times up to 10 and 9 days, respectively. They found that reference forecasts based on climatology perform the worst at all lead times, and reference forecasts based on persistency only show low error at short lead time (up to 2 days). Bennett et al. (2014) recommends to use a reference set that consists of an ensemble of past observed precipitation series over a long historical period. By running the hydrological model with these ensembles, a reference forecast is obtained. This reference set shows smaller errors at all lead times compared to the persistency and climatology forecast.

The persistence reference forecast is chosen as the reference for assessing the model forecast, although Benninga et al. (2017) and Bennett et al. (2014) showed smaller errors at all lead times using the historical precipitation ensemble reference. For this approach, a large data set of historical precipitation is required. Due to limited and irregular hourly radar and German rain gauge time series (Figure 9) this reference set is not used. In the evaluation of the CRPSS results, it should be taken into account that the reference forecast based on persistency is easy to beat at longer lead times.

3.5 Comparison of single model forecast to multi-model forecast

This section explains how the ensemble forecasts from the individual models are combined and assessed. In total there are three ensemble forecast combinations: (1) NWP + HBV, (2) NWP + GR4H and (3) NWP + MHM. NWP refers to the weather prediction model, which is the COSMO-LEPS model for all combinations. MHM refers to the combined ensemble forecasts of HBV and GR4H. The number of ensemble members in the mentioned combinations are 20, 20 and 40. In order to assess the multi-model combination, the results from the individual hydrological model forecasts should be combined. Duan et al. (2007) used the Ensemble Bayesian Model Averaging (BMA) approach for multi-model combination. BMA is a statistical procedure that applies higher weights to better performing predictions and lower weights to worse performing ones. Another approach is to use an algorithm to assign larger weights to the model(s) that lead to better predictability under similar prediction conditions (Devineni et al., 2008). Other studies applied a simpler method where the ensemble members of the individual models are grouped together (Dion et al., 2021; Teja et al., 2023; Velázquez et al., 2011). In this study, the same grouping approach is applied. Applying weights to ensemble members or to better forecasts is not chosen because of the limited number of models (one NWP model and two hydrological models) and the limited number of unique ensemble members. By combining all the ensemble forecasts, a larger spread in possible outcomes is obtained, likely to better capture the true observed event. The ensemble forecasts from the multi-model combination are evaluated by using the same criteria as the single model ensembles forecasts. This way, the multi-model ensembles can be directly compared to the performance of the single models. The goal of this method is to find an answer to the last research question: does the use of the multi-model forecast provide better flood forecast performance compared to use of single hydrological models?

4 Results

4.1 Calibration and Validation

This section shows the results of calibration and validation of HBV and GR4H. In Section 4.1.1 and 4.1.2 the optimal parameter set and corresponding objective function values that are found for each sub-catchment are presented. Section 4.1.3 shows the validation results for both validation periods. The last section, 4.1.4, summarizes all results and answers the first research question about the performance of HBV and GR4H on high-flow simulations.

4.1.1 HBV Calibration

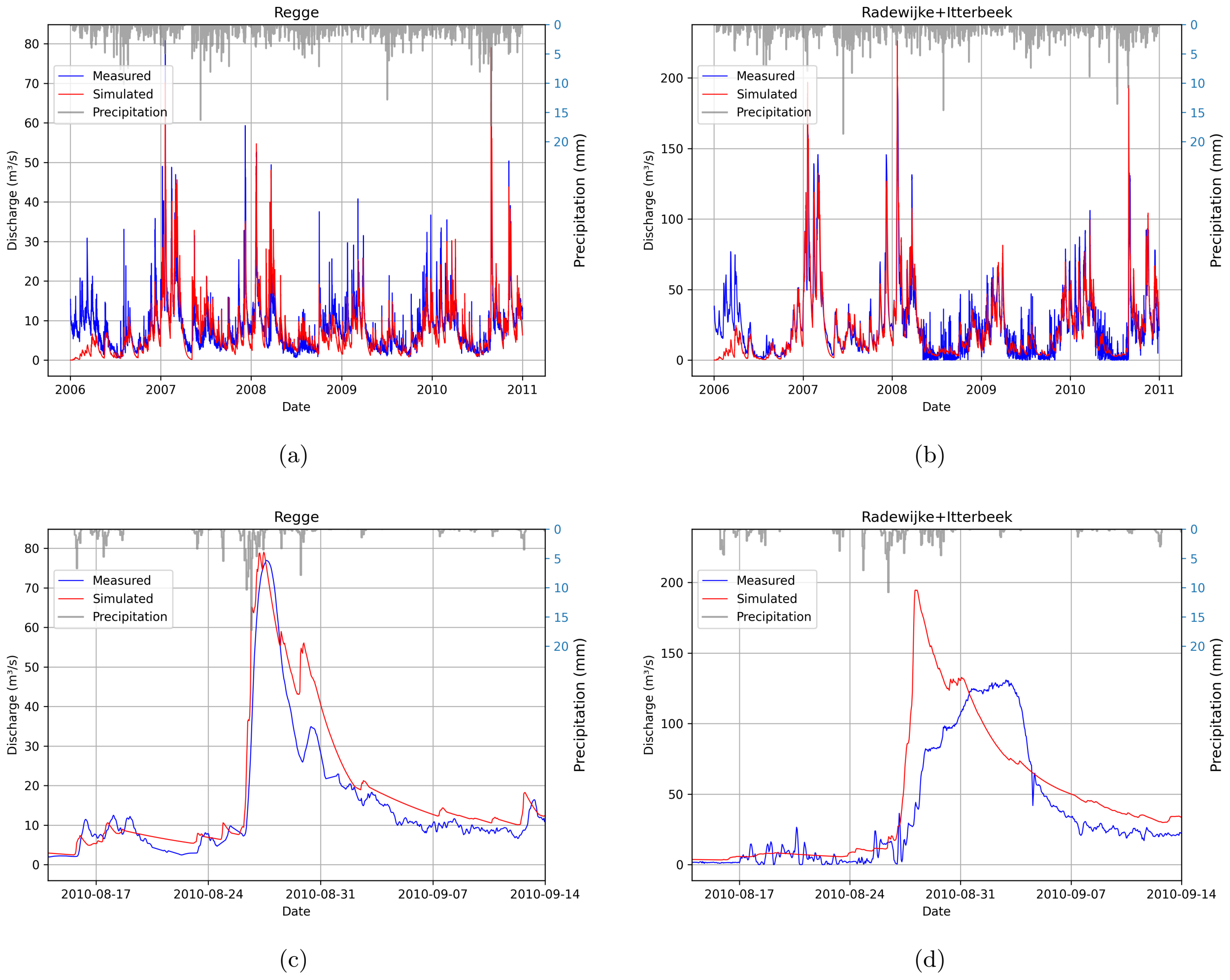
HBV has been calibrated three times by the PSO optimization algorithm. Twice with 50 particles and 25 iterations per catchment, and once with 200 particles and 25 iterations per catchment. In total 11.700, 11.700 and 46.800 calibration runs have been performed. Because the number of parameters to be calibrated for HBV is larger than GR4H, it was decided to perform more runs, with more particles, to increase the likeliness of finding the global optimum. In Appendix C.1, figures showing the evolution of the particles per iteration against Y_w of the third calibration run can be seen. All particles reach the final objective function value relatively quickly (within 10-15 iterations).

Neither of the three calibration runs yielded the exact same parameter set and objective function values. Probably, the particles converge too quickly to a local optimum that yield approximately the same Y_w. The final obtained highest values of Y_w were comparable between the three runs for all sub-catchments, however, with different parameter combinations. The optimal parameter set and objective function values of the three calibration runs can be found in Appendix C.2. The chosen parameter set and the corresponding values of the objective function are composed of the best results from the three runs. Run 3, with the most particles, produces most of the highest Y_w values. The final parameter set of Radewijke+Itterbeek is chosen from calibration run 1, however, this does not yield the highest Y_w. This parameter set consisted of less values at the boundary levels of the parameters, and the change in Y_w was minimal (-0.006). The finally chosen parameter set (Table 13) has been run in the model once to test if the produced objective function values did not change much. The objective function values did not change after this run and all sub-catchments kept the same values. Since Radewijke+Itterbeek performs better than Ommerkanaal, the parameter set of Radewijke+Itterbeek is assigned to Stouwe. The obtained Y_w values are between 0.84 and 0.91 for all sub-catchments. This is slightly higher than what was found in the literature. Ten Berge (2024) used the same objective function to calibrate the HBV model for high flows in the Lesse. She found Y_w values on average of 0.82 among multiple climate testing schemes using the SCEM-UA calibration algorithm. Benninga et al. (2017) found values of Y_w around 0.8 using an almost identical objective function. Akhtar et al. (2009) found function values between 0.61 and 0.92 for different types of HBV models, for multiple river basins.

Figure 21 below shows the observed and simulated hydrogaphs of Regge and Radewijke+Itterbeek, respectively the worst and best performing sub-catchments. The model seems to accurately simulate the highest flow peaks, however the model also underestimates smaller peaks during winter and base-flow in summer. Since focus is mainly on simulating high-flow events, the HBV model performs well based on the hydrograph figures and values of the multi-objective function. Figure 35 in Appendix C.3 shows the hydrographs of all sub-catchments.

Table 13: Best performing HBV parameter set according to the objective function values. The table is the result of the combination of the best performing sets from a total of three calibration runs. The resulting set yields the best combined function values of Y_w.

	FC	**LP**	**BETA**	**CFLUX**	**ALFA**	**KF (10^{-3})**	**KS (10^{-3})**	**PERC**	**Y_w**	**NS_w**	**RVE**
	[mm]	**[-]**	**[-]**	**[mm/h]**	**[-]**	**[h^{-1}]**	**[h^{-1}]**	**[mm/h]**			**(10^{-5} %)**
Steinfurter Aa	247.08	0.47	4.24	0.040	0.11	12.57	6.87	0.06	0.90	0.90	0.70
Vechte A	457.33	0.98	5.31	0.014	0.10	9.05	6.14	0.11	0.87	0.87	1.99
Dinkel	213.39	0.50	2.32	0.025	0.21	3.78	5.77	0.03	0.87	0.87	1.96
Afwateringskanaal	215.94	0.27	4.21	0.042	0.57	1.99	8.33	0.16	0.84	0.84	2.43
Ommerkanaal	110.47	0.49	5.15	0.015	0.11	7.07	1.25	0.07	0.85	0.85	-2.82
Regge	447.95	0.53	2.09	0.025	0.10	8.79	4.56	0.10	0.84	0.84	2.09
Vechte B	323.44	1.00	4.39	0.029	0.10	0.21	8.33	0.25	0.86	0.86	-0.02
Vechte C	305.86	0.99	6.00	0.021	0.44	0.21	3.38	0.07	0.86	0.86	0.00
Radewijke+Itterbeek	520.16	0.34	1.19	0.034	1.97	0.40	0.02	0.25	0.91	0.91	-10.26
Stouwe	520.16	0.34	1.19	0.034	1.97	0.40	0.02	0.25	-	-	-



(a)

(b)

(c)

(d)

Figure 21: Observed and simulated (HBV) hydrographs of Regge (a) and Radewijke+Itterbeek (b) for the calibration period 2006-2010. Figures c and d show a short period during summer 2010 for the same sub-catchments (2010-08-14 till 2010-09-14).

4.1.2 GR4H Calibration

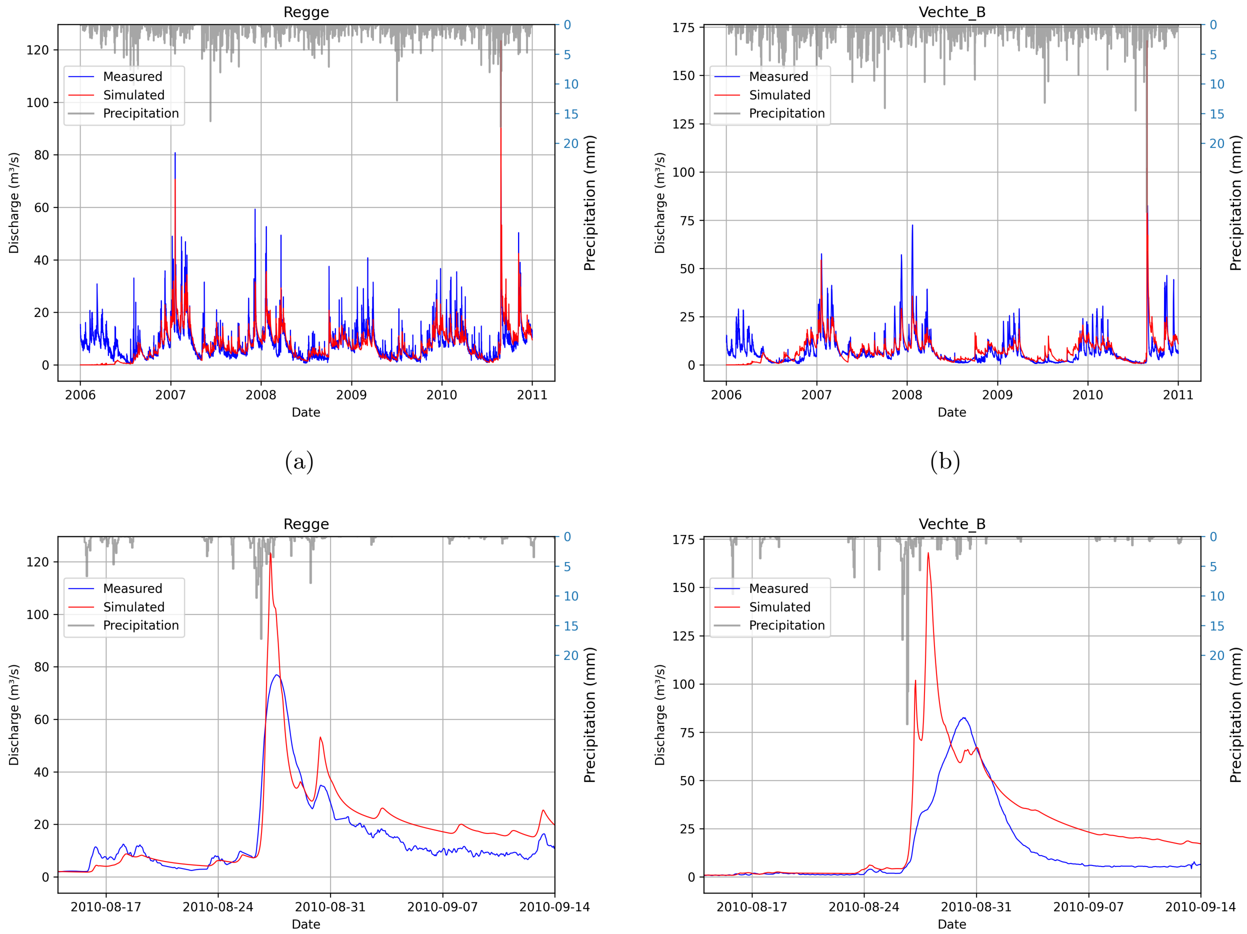
GR4H has been calibrated twice by the PSO optimization algorithm. Once with 25 particles and 25 iterations per catchment and once with 50 particles and 25 iterations per catchment. This corresponds to a total number of 5.850 and 11.700 model runs. All particles were able to converge to the final objective function value within the given iterations. Figures showing the evolution of the particles per iteration against Y_w of calibration run 2 can be found in Appendix C.1. Both calibration runs did not yield the exact same result, raising the suspicion that the global optimum was not found, just local optima. However, the corresponding objective function values of Y_w were comparable between both runs. Only sub-catchment Regge showed a large difference between calibration run 1 (0.822) and run 2 (0.561). The optimal parameter sets and corresponding objective function values NS_w, RVE and Y_w of both calibration runs can be found in Appendix C.2.

The final set of optimal parameters has been composed by combining the best performing set of both runs, according to Y_w. This parameter set has been run in the model once, to see if the dependent downstream catchments of Vechte B, Vechte C and Radewijke+Itterbeek did not decrease in model performance. Their performance increased slightly, so the combined parameter set is used. This set and the corresponding objective function values can be seen in Table 14. Because Ommerkanaal performed better than Radewijke+Itterbeek, Stouwe was assigned the parameter set of Ommerkanaal. In general, the Y_w of GR4H in the calibration is between 0.63 and 0.82. No literature has been found that used the GR4H model in combination with the same objective function. However, Ten Berge (2024) found function values of Y_w between 0.89 and 0.92 for GR6J, a different version of the GR4J model.

Table 14: Best performing parameter set for GR4H according to the objective function values. This table is a combination of the best performing parameter sets from two calibration runs.

	X1 [mm]	X2 [mm]	X3 [mm]	X4 [h]	Y_w	NS_w	RVE (10^{-5} %)
Steinfurter Aa	517.87	-1.36	84.34	12.00	0.75	0.75	1.35
Vechte A	447.32	-1.07	153.63	17.63	0.76	0.76	3.08
Dinkel	455.19	-0.23	10.00	75.25	0.75	0.75	3.71
Afwateringskanaal	108.53	-1.26	48.66	22.40	0.79	0.79	-20.48
Ommerkanaal	46.81	-1.60	184.90	13.34	0.76	0.76	-4.47
Regge	730.50	-0.80	27.69	14.35	0.82	0.82	7.79
Vechte B	523.93	-0.27	38.13	12.00	0.63	0.63	25.76
Vechte C	248.14	-0.38	47.12	47.87	0.75	0.75	5.56
Radewijke+Itterbeek	10.00	-6.73	46.74	12.48	0.76	0.76	5.20
Stouwe	46.81	-1.60	184.90	13.34	-	-	-

Figure 22 shows the hydrographs of Regge and Vechte B for the calibration period (2006-2010), respectively, the worst and best performing sub-catchments. The model seems to have difficulty simulating medium-high discharge, e.g. discharge that occurs during wintertime. The large peak in 2010 is overestimated in both catchments by nearly a factor of two. For the other sub-catchments the same trend is observed that the highest discharge peak is overestimated, and the regular winter discharge underestimated (Figure 36 in Appendix C.3). In general, the timing and shape of the simulated discharge seems to be comparable to the observed discharge. It seems like there is a trade-off between simulating the small discharge peaks during winter and simulating the high-flow peaks, where tuning the parameters in favor of one is disadvantageous for the other. Despite this, the GR4H model performs decently based on the multi-objective function.



(a)

(b)

(c)

(d)

Figure 22: Observed and simulated (GR4H) hydrographs of Regge (a) and Vechte B (b) for the calibration period 2006-2010. Figures c and d show a short period during summer 2010 for the same sub-catchments (2010-08-14 till 2010-09-14).

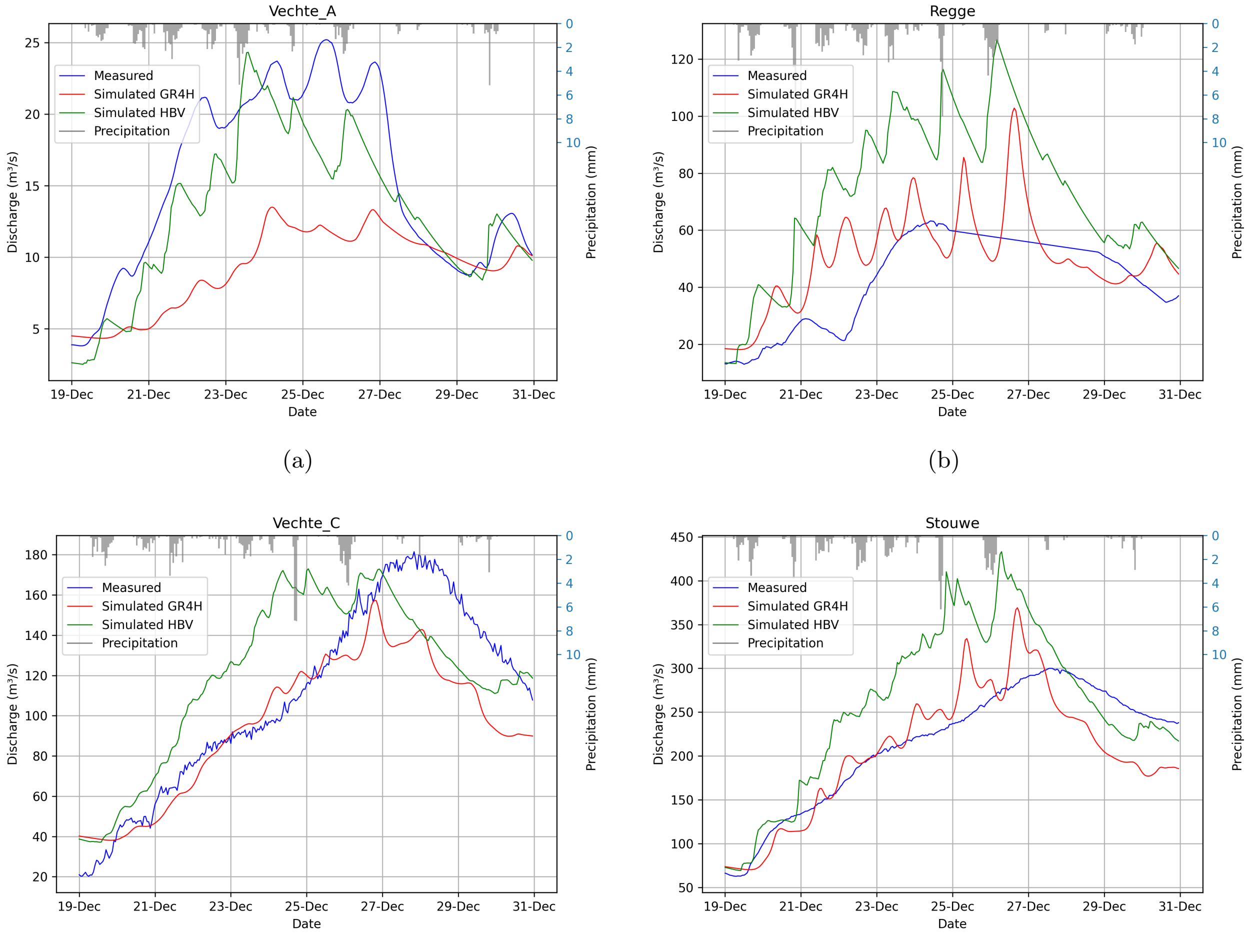
4.1.3 HBV & GR4H Validation

Both models are validated on the two independent periods: (1) March 2023 - March 2024 and (2) 2013 up to and including 2015. The result on the objective function value Y_w can be seen in Table 15. The complete results, including NS_w and RVE values, can be found in Appendix D.1. Both models performed less for both validation periods. For some sub-catchments the models were not able to perform with Y_w above 0.5. For HBV validation 1, Regge even scored a negative function value, indicating that the mean of the observed discharge would be a better approximation than the simulated one. This outlier raises the question whether the observed discharge at Regge for this period is accurate. For the other sub-catchments HBV performs satisfactory for both validation periods, however, the difference with the calibration is large for some sub-catchments. Ten Berge (2024) found only a slight reduction of Y_w (up to -0.1) between calibration and validation of the HBV model. Benninga et al. (2017) also found reduced validation results in the order of -0.1 Y_w. GR4H shows a notable poor performance for Steinfurter Aa and Vechte A during validation 1. This is caused by both poor NS_w and large underestimation of RVE. Both models show significant deviations from the observed water balance for both validation periods. There seems to be no clear relation between over- or underestimation of the RVE between the periods or between the models.

Figure 23 shows the simulated discharge from HBV and GR4H for Vechte A, Regge, Vechte C, and Stouwe for the period of 19 December 2023 to 31 December 2023 (Validation 1). Vechte A and Regge are, respectively, the worst performing catchments during validation 1 for GR4H and HBV, and Vechte

Table 15: Objective function values Y_w for validation 1 and 2. Shown is also the absolute difference of Y_w compared to the calibration.

	Validation 1 ($\mathbf{Y_w}$)		Diff. Calibration		Validation 2 ($\mathbf{Y_w}$)		Diff. Calibration	
	HBV	**GR4H**	**HBV**	**GR4H**	**HBV**	**GR4H**	**HBV**	**GR4H**
Steinfurter Aa	0.77	0.36	-0.14	-0.39	0.61	0.45	-0.29	-0.30
Vechte A	0.52	0.24	-0.35	-0.52	0.65	0.40	-0.22	-0.36
Dinkel	-	-	-	-	0.83	0.66	-0.04	-0.09
Afwateringskanaal	0.51	0.49	-0.33	-0.31	0.53	0.42	-0.31	-0.38
Ommerkanaal	0.70	0.55	-0.15	-0.21	0.78	0.86	-0.07	+0.10
Regge	-0.26	0.49	-1.1	-0.33	0.52	0.56	-0.32	-0.26
Vechte B	0.66	0.37	-0.21	-0.26	0.85	0.64	-0.01	+0.01
Vechte C	0.72	0.83	-0.13	+0.08	0.74	0.67	-0.12	-0.08
Radewijke+Itterbeek	0.79	0.61	-0.11	-0.14	0.74	0.82	-0.17	+0.06
Stouwe	0.64	0.80	-	-	0.87	0.76	-	-



(a)

(b)

(c)

(d)

Figure 23: Simulated and observed hydrographs of HBV and GR4H for the high-flow event during Christmas 2023 for Vechte A (a), Regge (b), Vechte C (c) and Stouwe (d).

C and Stouwe two of the better performing ones. The complete validation period 1 is cropped to only the few days that cover the high-flow event. Generally, it can be noticed that both models show a spiked behavior, e.g. the response on precipitation input is fast. For catchment Vechte A, the observed discharge shows similar behavior, however for the downstream catchments, response is slower and the observed discharge wave is smoother and spread out over time. This would make sense because wave attenuation is not taken into account in the model, just a simple flood wave delay. Despite that delay, the model seems to model the high-flow peak too early for the downstream catchments Vechte C and Stouwe. This could indicate that the flow speed of 1.2 m/s is an overestimation of reality. Hydrographs of all catchments for both models for both validation periods can be found in Appendix D.

4.1.4 Conclusion RQ.1

The first research question is: 'How do HBV and GR4H perform on high-flow simulations, based on historical data?' To answer this question both models have been calibrated and validated. The calibration of both models on the years 2007, 2008 and 2010 yields good objective function values Y_w for the catchments modelled by HBV. On average, Y_w is 0.87, indicating good resemblance of simulated and observed flow. GR4H performed slightly worse on the calibration period, yielding an average Y_w of 0.75. Based on the objective function values, the optimal parameter sets have been chosen and used for validation.

Two validation periods are chosen to test the models. Validation 1 includes one of the highest peak flow events of recent history. On this validation period, the models performed worst, with remarkably large differences between the catchments and between both models. The average Y_w of HBV for validation 1 and 2 is 0.56 and 0.71. GR4H performed on average 0.53 and 0.62. Both models have difficulty with correctly simulating the extreme high-flow event during Christmas 2023. For the relatively 'normal' validation period 2, both models show better simulation performance, but still substantially below the calibration period. The conclusion from these results is that both models decrease sharply in performance for extreme periods of high-flow. Additionally, it could be that the model parameters are overfitted on the calibration data, resulting in decreased model performance outside of this period. However, the difference between individual sub-catchments is large, and downstream catchments seem to perform slightly better than upstream catchments. However, the timing of peak discharge of the downstream catchments seems to be too early.

4.2 Flood forecast performance

This section shows the results of the performance of the forecasting system using HBV, GR4H, and the multi-model (HBV+GR4H). Section 4.2.1 shows a general visual evaluation of the forecast system for the individual model forecasts and for the multi-model forecasts. Section 4.2.2, 4.2.3 and 4.2.4 show the performance of the system by means of evaluation metrics RMAE, CRPS and CRPSS. The RMAE and CRPS scores are first evaluated for the individual model forecasts and then the multi-model is assessed. The CRPSS is then assessed for the multi-model forecasts only.

4.2.1 General forecast evaluation

Ensemble forecasts are generated using HBV, GR4H, and the multi-model for both forecast periods December 2023 and March 2023. The result of one of these forecasts (issued on 23-12-2023 13:00 and 08-03-2023 13:00) for each period can be seen in Figure 24. Spaghetti plots of HBV, GR4H and the multi-model for the most downstream catchment Stouwe are shown. The ensemble median shows the tendency of the ensembles and the inter quartile range (IQR) provides a measure of spread and variability of the ensembles (Teja et al., 2023). As an indicator of discharge magnitude, official warning levels as used by waterboard WDOD are added as dashed horizontal lines. These discharge levels have been empirically derived from the Qh-relation at discharge station Dalfsen (data obtained from WDOD). This figure can be found in Appendix E.

Individual model evaluation

Generally it can be seen that the forecast system has difficulty forecasting accurate discharges for short lead times (<24 hours). Both HBV and GR4H forecasts deviate slightly from the observed discharge at the start of the forecast for both forecast periods. This is likely caused by three factors: (1) uncertainty in observed discharge sums, (2) difference between observed and simulated downstream discharge, and (3) inaccurate data assimilation. All three factors are briefly explained in the following paragraphs.

At the first forecast time, the forecasted discharge of Stouwe consists for the largest part of the observed discharge of Radewijke+Itterbeek, Regge and Ommerkanaal from 3 hours earlier. This sum can lead to a different forecasted discharge as is actually measured. For Stouwe this effect seems to be minimal for this specific forecast, however, for Vechte C this effect is more prominent (see Appendix F.1).

Also, clearly the effect of changing from observed discharge to simulated discharge can be seen. The forecasts of March 2023 show a sudden decrease in discharge of all ensembles after just a few hours. Up to a lead time of 3 hours, the observed discharge from the upstream catchments Radewijke+Itterbeek, Regge and Ommerkanaal is used for the forecasted discharge at Stouwe. After this time, the discharge forecasts from these catchments are used. For this specific forecast, the forecasted discharge at these catchments is smaller than the actual measured discharge the hour before, leading to a sudden decrease in discharge in that time step. The same is observed for the Christmas event, however, more hidden in the forecasted discharge of the downstream catchments. As explained in Section 3.3.5, the missing observed discharge from sub-catchment Dinkel has been left out of the analysis. This results in gaps in the hydrographs of sub-catchments downstream of Dinkel. These gaps have been linearly interpolated. For Stouwe this can be seen in the hydrograph between lead time 10-17 hours. After this time, the forecasted discharge from Dinkel (through downstream catchments Vechte C and Radewijke+Itterbeek) is added to the forecasted discharge of Stouwe. The forecasted discharge from Dinkel has not been initialized on observed discharge, such that it is highly uncertain if the forecasted discharge from Dinkel is accurate. Because the linear interpolation shows a sharp increase above the measured discharge, it is likely that the forecasted discharge from Dinkel is an overestimation of the true discharge.

The third reason why the forecasted discharge is off at short lead times is due to inaccurate data assimilation. For the downstream catchments, no clear storage-discharge relation could be determined, so the initialization is an estimation, which could lead to an over- or underestimation of the discharge forecasts.

After approximately 12-24 hours, the main uncertainty changes from the error in initialization to uncertainties from the precipitation forecasts. Both models show a large ensemble spread, both overestimating and underestimating the observed discharge. The IQR often captures the flood event for most lead times by using the HBV model, however, it is mostly underdispersed for GR4H. The ensemble median of the HBV model is a decent deterministic estimator of the observed discharge for both periods. The ensemble median of GR4H is overestimating the observed discharge for shorter lead times during the Christmas event and slightly underestimating at longer lead times. During March 2023 the ensemble median of GR4H is underestimating the observed discharge heavily. There is a large variability in forecast results between the sub-catchments, but also between each issued forecast.

Multi-model evaluation

Spaghetti plots of the combined model show both the strengths and weaknesses of the individual

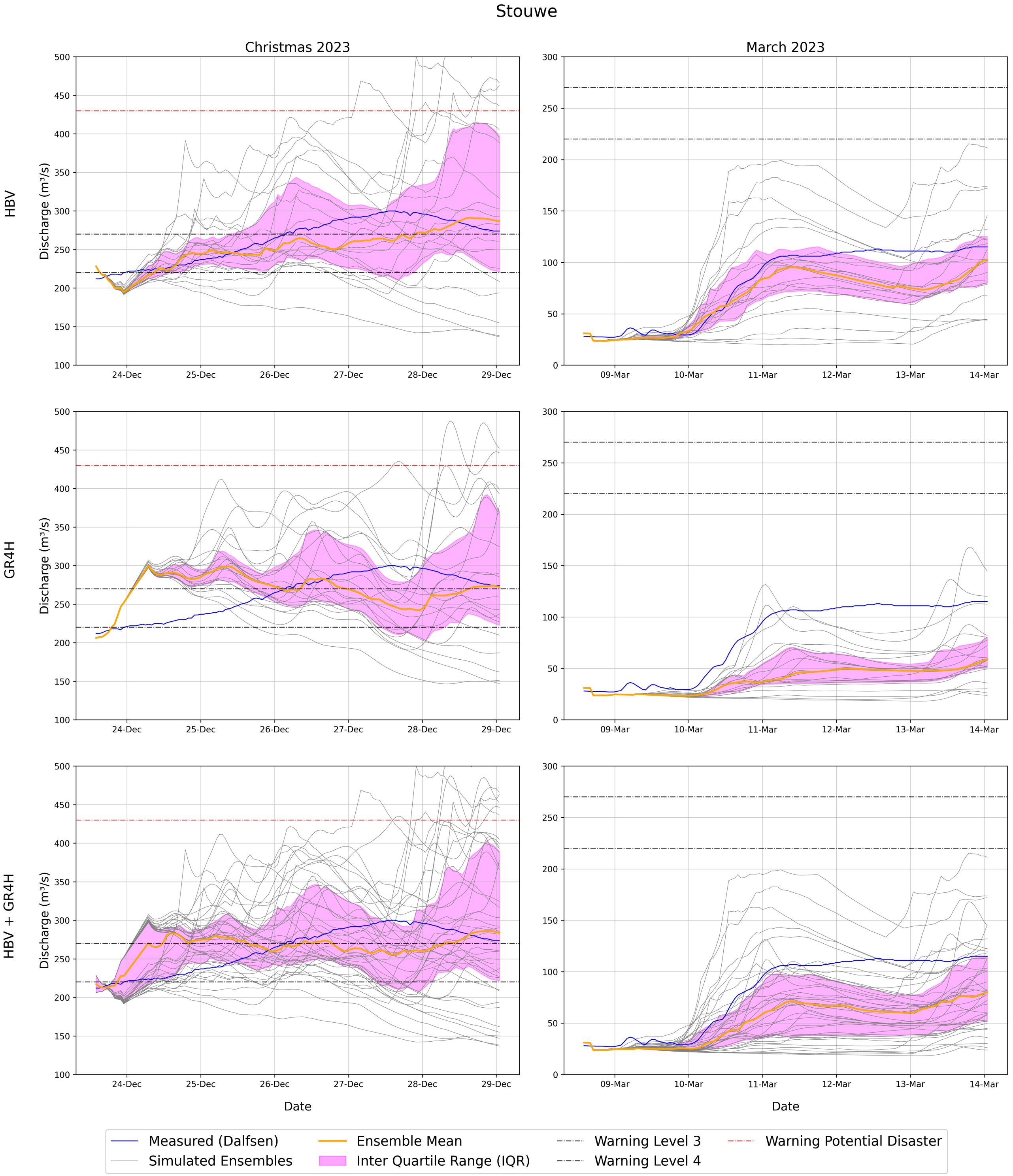


Figure 24: Spaghetti plots of the ensemble forecasts (grey), measured discharge (blue), ensemble median (orange) and inter quartile range (pink) for sub-catchment Stouwe for both forecast periods for HBV, GR4H and the multi-model. Both HBV and GR4H contain 20 ensembles and the multi-model 40 ensembles. The dashed horizontal lines represent the warning levels at which the water authorities increase the state of alertness. Note that the y-axes are not similar between the events, due to the large difference in peak discharge.

model forecasts. Because HBV and GR4H show opposite behavior during the first few hours during the Christmas event, the IQR of the combined models does capture the observed discharge better, however, with an increased uncertainty range. At longer lead times, the multi-model forecast is similar to the individual model forecasts for the Christmas event. For the March 2023 event, the weakness of the GR4H model (under-dispersion) is also visible in the multi-model forecast. In every case, the multi-model forecast is better than the worst performing model, but it takes the weak parts of the models equally into account as well as the better performing parts. Hence, if one of the models shows accurate performance among all ensembles, and the other a poor performance among all ensembles at a similar part of the hydrograph, the multi-model forecasts will automatically become an average of the two. Based on only one forecast and only one sub-catchment, it is difficult to determine the general performance of the forecast system and say with certainty if the multi-model forecasts are an improvement to the single models. In Appendix F.1 a similar assessment of the spaghetti plots of sub-catchments Ommerkanaal and Vechte C can be found for the same forecast. To quantify the forecast performance of the complete system, the forecast evaluation criteria RMAE, CRPS and CRPSS are used.

4.2.2 RMAE

Individual model evaluation

To assess the overall bias in the ensemble forecasts, the RMAE is calculated. This is done using the ensemble median for all forecasts. The resulting figures show the RMAE against the lead time for all 10 forecasts. Figure 25 shows the result of the individual models. The colors indicate the spread of the 10 forecasts for the two forecast periods. A similar plot for the multi-model forecasts can be found in Appendix F.2. The upstream sub-catchments show for both models and periods low RMAE values for short lead times (<12 hours). With increasing lead times, RMAE values increase, as well as the spread between the forecasts. The opposite can be seen for most downstream sub-catchments. At short lead times (<12 hours) the RMAE values range between 0 and 0.6 for both periods and for both models. This indicates poor initialization at the forecast time for the downstream sub-catchments. For the Christmas event, most RMAE values are lower at a lead time of 132 hours, than at a lead time of 0 hours. Based on these figures, no model seems to structurally outperform the other. Both models show a sharp increase in RMAE values with increasing lead time for the upstream sub-catchments, however, lower maximum RMAE values for the downstream sub-catchments. This makes sense because the discharge at the downstream sub-catchments is much higher than the upstream sub-catchments. Large deviations in the ensemble median from the observed discharge are less likely to occur. This can be clearly seen for the Christmas event, where the maximum RMAE of the downstream sub-catchments for both models is around 0.6, while the maximum RMAE values of the same period for the upstream sub-catchments exceed 1.0.

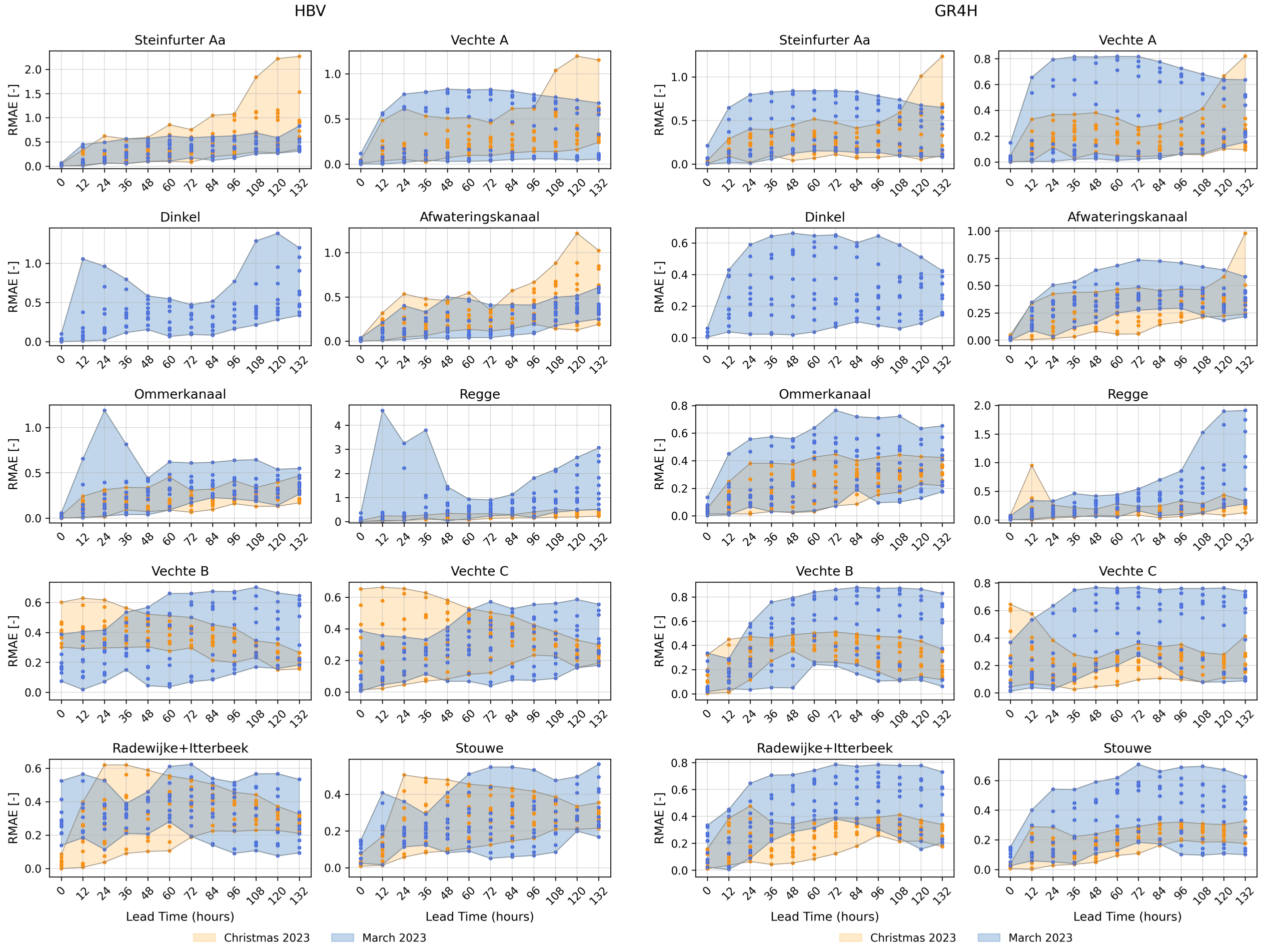


Figure 25: RMAE of the ensemble median across the 10 forecasts (10 values per lead time) for HBV (left) and GR4H (right) for all sub-catchments. In blue the spread of the March event and in orange the spread of the Christmas event. Note that the y-axes vary per sub-catchment, to increase visibility of the figures.

Multi-model evaluation

To clearly visualize the difference in RMAE between the models individually and combined, the median is taken over the 10 forecasts per lead time. In this way, all model combinations and forecast periods can be plotted in one figure. In addition, outliers such as the RMAE values for Regge are not taken into account (Figure 25, HBV, March 2023). Figure 26 shows for the 10 sub-catchments the individual and multi-model RMAE results. Like in Figure 25, the periods can not be directly compared due to different discharge magnitudes, however, the model results per period can be directly compared. Generally, the multi-model RMAE values are somewhere in between the individual model values for all lead times. The multi-model always shows lower RMAE than the worst individual model (solid line always below one of the individual model lines). The tendency of the multi-model RMAE seems to be towards the best performing individual model. For some catchments, the multi-model shows lower RMAE values than both individual models. This is the case for the March event of sub-catchment Dinkel for lead times between 24-54 hours and between 78-132 hours. For all upstream catchments there are lead times for which the multi-model RMAE is below both individual model RMAE values, indicating lower errors and hence (slightly) better performance. For the downstream sub-catchments the multi-model RMAE seems to be more of an average between the two individual models. Based on this bias evaluation, the use of the multi-model can slightly decrease the RMAE for various lead times for the upstream sub-catchments, but show no clear decrease for the downstream catchments.

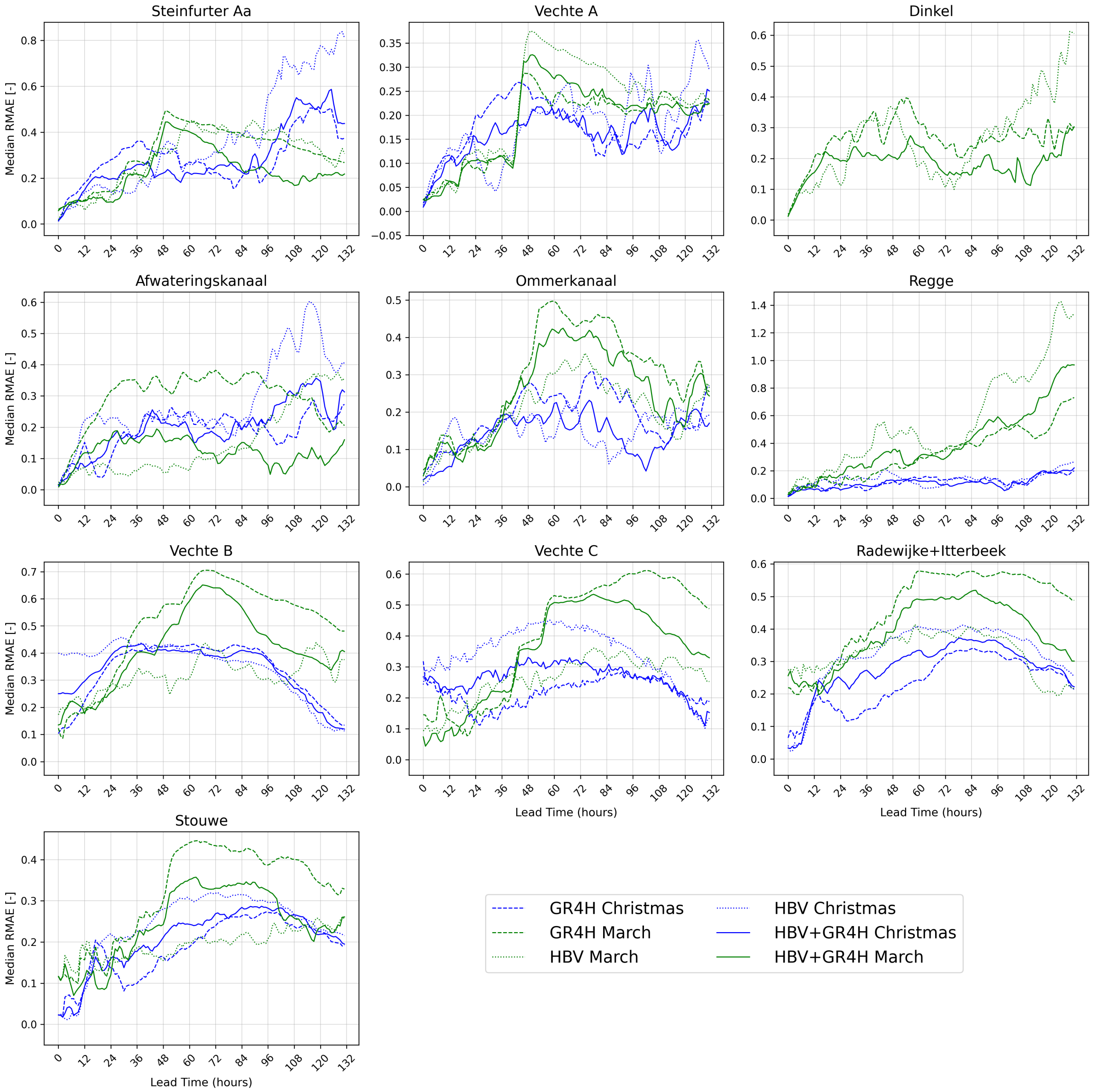


Figure 26: RMAE against lead time. RMAE has been calculated with the ensemble median, aggregated by taking the median RMAE over all forecasts. The colors indicate the forecast period (blue: Christmas, green: March) and the line types the HBV, GR4H and multi-model. Note that the y-axes vary per sub-catchment to increase visibility of the figures

4.2.3 CRPS

The mean CRPS has been calculated for HBV and GR4H over the 10 forecasts for both forecast periods for all sub-catchments. The value of CRPS has the unit of the discharge (m^3/s) and a low CRPS indicates good performance with respect to the observed discharge. As reference, the mean CRPS of the persistency forecast is calculated. The results can be seen in Figure 27. The persistency forecasts show high CRPS values for long lead times, which is expected because observed discharge at lead time 0 cannot be an accurate forecast for a long time ahead. Although the periods are plotted side by side, it is not possible to directly compare the periods with each other. As an example, a maximum CRPS of 4 m^3/s during the Christmas event 2023 is a relatively much better score than a maximum CRPS of 4 m^3/s during March 2023, because absolute discharges during the Christmas event are much higher. This analysis will therefore focus mostly on the differences and trends between the individual models and the multi-model, rather than the magnitude of the CRPS.

Individual model evaluation

The HBV model does show in general lower CRPS values for most sub-catchments for both periods

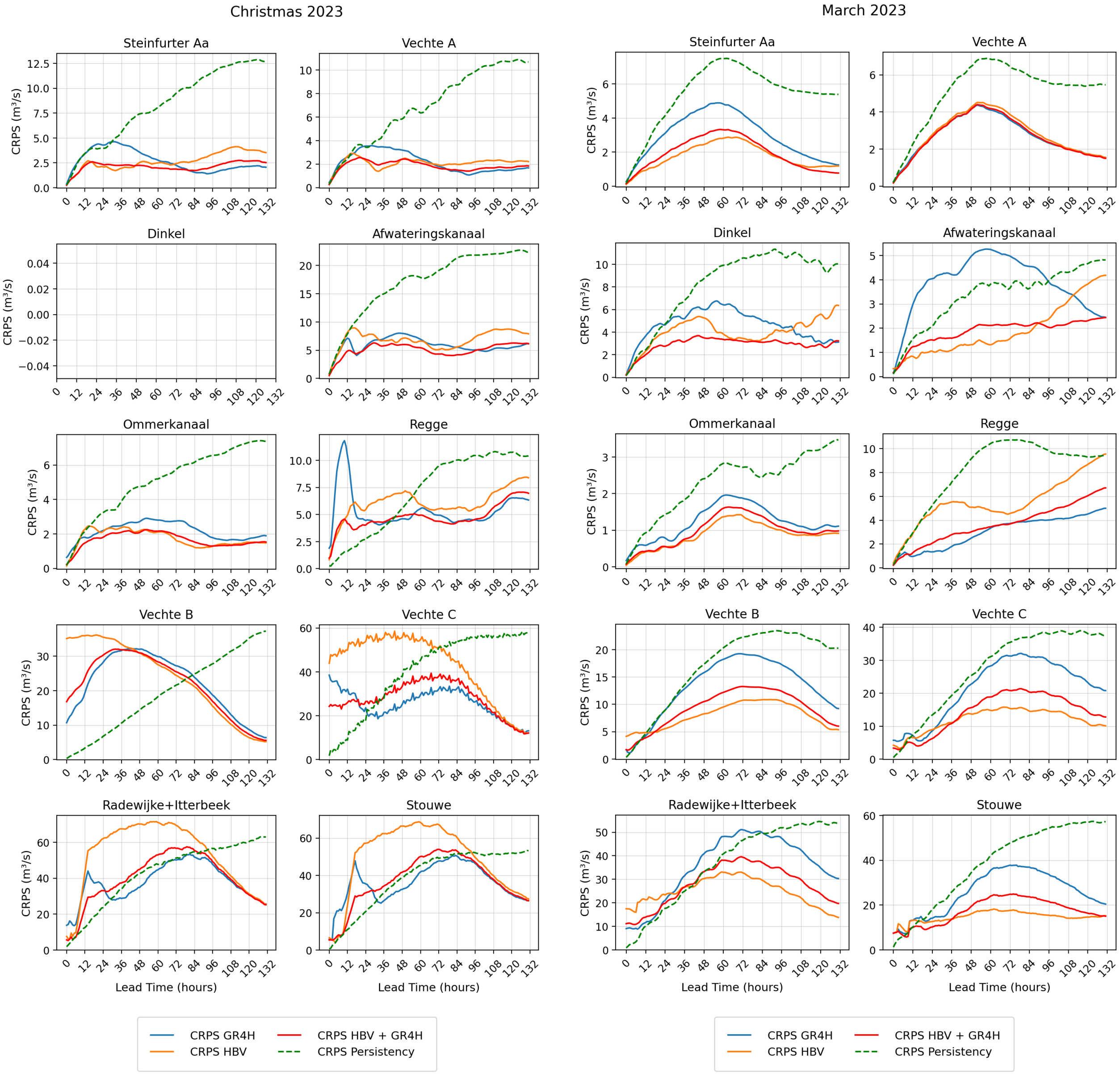


Figure 27: CRPS of the individual models (HBV in orange, GR4H in blue), multi-model (red) and reference forecasts (green dashed) of both periods for all sub-catchments. The figures to the left show CRPS for the Christmas event, and the figure to the right the CRPS of the March event. As there is no observed discharge available during the Christmas event for sub-catchment Dinkel, an empty plot is shown.

compared to GR4H. Only for the downstream sub-catchments during the Christmas event does the GR4H model perform slightly better than the HBV model. The upstream sub-catchments show a small CRPS for short lead times (<12 hours) and are generally lower than the reference forecast (green dashed line). For longer lead times (>12 hours) the upstream sub-catchments also show lower CRPS than the reference forecast. This indicates that the forecasts for the upstream sub-catchments show better skill than the reference forecast. The downstream catchments show poor skill during the Christmas 2023 event, but beat the reference forecast at most lead times during March 2023. A clear difference is observed between the upstream catchments and the downstream catchments. The upstream catchments challenge the persistency forecast and generally show better skill. The downstream catchments show at a lead time of 0 hours already an increased CRPS (due to poor state updating), and are only able to outperform the persistency forecast at long lead times. At sub-catchment Stouwe all models outperform the reference forecast only after 96 hours during the

Christmas event.

Multi-model evaluation

The multi-model forecasts show no clear consistent increase in performance, with respect to the individual models based on the CRPS. However, the multi-model CRPS tends to stick more to the better performing model. This trend can be seen for the upstream and downstream sub-catchments for both forecast periods. This would indicate that the multi-model is able to reduce the negative performance of the worser model. This is clearly visible for Regge during the Christmas period. A large CRPS peak is visible for GR4H, however, the multi-model CRPS shows this peak only slightly. Also for all other lead times the multi-model CRPS is located closely to the best performing model, and occasionally also below. This is also observed for most other sub-catchments. The most remarkable result is for sub-catchment Dinkel during March 2023. The multi-model shows lower CRPS, compared to the individual models for all lead times. A similar result can be seen for sub-catchment Afwateringskanaal during the Christmas event. Overall, the multi-model CRPS mostly shows a slight improvement for at least a few lead times among the total forecast horizon.

4.2.4 CRPSS

Multi-model evaluation

The CRPS is skilled (CRPSS) by benchmarking it with the persistency reference forecast. By doing so, the catchments and periods can be directly compared. Figure 28 shows for all sub-catchments the CRPSS against lead time for the Christmas and March event. As long as the CRPSS is above 0, the CRPS of the model forecast shows higher skill than the CRPS of the reference forecast. For most upstream sub-catchments this is the case for both periods. The skill of these sub-catchments starts around 0 or above and increases steadily with lead time. However, downstream sub-catchments show very poor initial skill at short lead times (<24 hours) compared to the reference forecast. Even at lead times of several days some downstream catchments have not been able to beat the reference forecast. Especially Vechte B (brown line in Figure 28) and Vechte C (pink line in Figure 28) show very poor skill for the Christmas event. The skill for Vechte B at a lead time of 0 hours is -20, which means that the CRPS at the start of each forecast is more than 20 times higher than the persistency CRPS. This is caused by poor storage updating at the forecast time in combination with a very low CRPS for the persistency forecast at the forecast time (lead time 0 hours). Because both models individually show a large difference between the discharge at lead time 0 and the measured discharge, the multi-model forecasts are also far off. The March event shows better skill for most sub-catchments. Except for sub-catchment Radewijke+Itterbeek, the reference forecast is outperformed within 12 hours, and four out of the six upstream catchments outperform the reference forecast at all lead times.

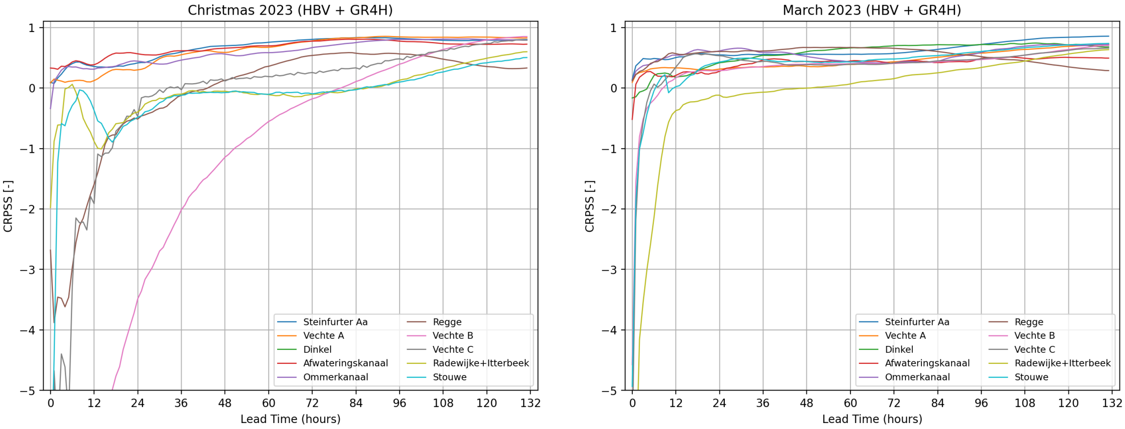
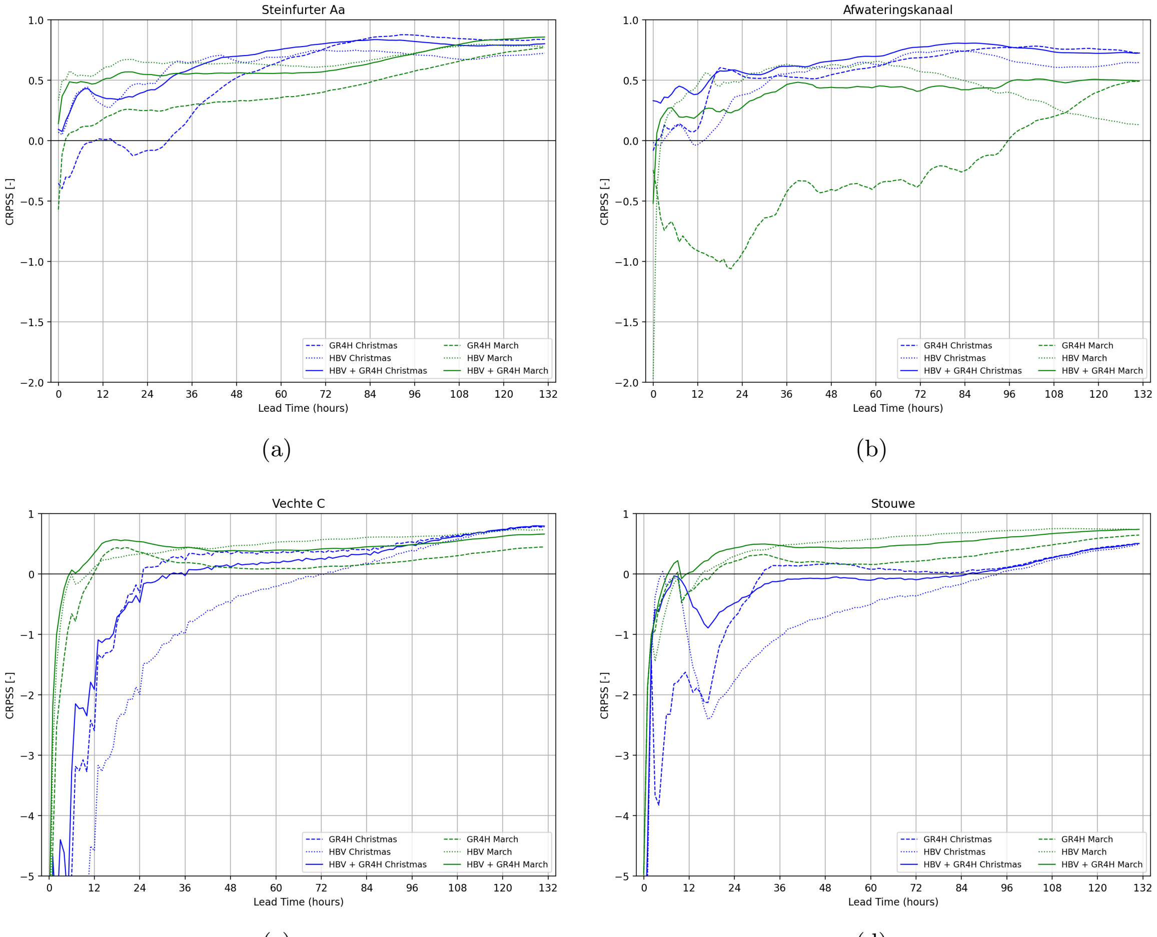


Figure 28: Skill of the multi-model forecasts for each sub-catchment, expressed by CRPS compared to CRPS of the persistency reference forecasts. The left figure shows the skill for the Christmas event and the right figure the skill of the March event.

To be able to compare the individual model forecast skill with the multi-model forecast skill, the CRPSS of HBV, GR4H and the multi-model are plotted for sub-catchment Steinfurter Aa, Afwateringskanaal, Vechte C and Stouwe in Figure 29. For each of these sub-catchments the skill of the multi-model forecast is close to the best skilled individual model, or a slight improvement. Especially for catchment Afwateringskanaal for the Christmas event the skill at a lead time of 0 hours is much better than any of the individual model forecast skill. Also at longer lead times, the multi-model forecasts show an increased skill. For lead times around 12-36 hours the multi-model also shows increased skill for both forecast periods for sub-catchment Stouwe. Based on these figures it seems that the multi-model forecasts show an increased skill and are less sensitive to low skill of one of the individual models. This can be seen in the figures of Afwateringskanaal and Stouwe. Individual model GR4H shows poor skill for all lead times for sub-catchment Afwateringskanaal, however the multi-model forecast skill is above zero for all lead times. For the Christmas period of sub-catchment Stouwe, the multi-model skill is less influenced by the dip in skill of both the HBV and GR4H model (between lead times 12-24 hours).



(a)

(b)

(c)

(d)

Figure 29: Skill of all model combination forecasts for four sub-catchments: Steinfurter Aa (a), Afwateringskanaal (b), Vechte C (c) and Stouwe (d). The colors indicate the forecast period (blue: Christmas, green: March) and the line types the model combinations.

4.2.5 Conclusion RQ.2 and RQ.3

The second research question is: 'What is the flood forecast performance of the single hydrological models with input of weather ensemble forecasts?' Both models show inaccurate discharge forecasting at short lead times (<24 hours) for the downstream sub-catchments. The upstream sub-catchments are generally modeled more accurately at short lead times. Based on visual evaluation of the spaghetti plots, the ensemble forecasts of GR4H seem to be underdispersed, while HBV shows better ensemble spread around the observed discharge. Based on the RMAE results, no model seems to clearly outperform the other. The same trends between both models have been found. These trends are mainly high initial error combined with a large spread at short lead times for the downstream sub-catchments, low error at short lead times that increase fast with longer lead time for the upstream catchments, and lower maximum errors for the downstream sub-catchments. For all upstream sub-catchments, both model forecasts showed lower CRPS values than the persistency forecast, indicating good forecast skill. However, both models showed poor skill for most lead times for the downstream sub-catchments.

The third and last research question is: 'Does the use of multiple hydrological models provide better ensemble flood forecast performance compared to the use of a single hydrological model?' Based on the RMAE, the multi-model forecasts show a performance that is in between the performance of the individual models for most lead times. However, for the upstream sub-catchments the multi-model performance also shows smaller error than both individual models for certain lead times. For the downstream sub-catchments, there seems to be no clear improvement of model performance based on the RMAE, compared to the individual model forecasts. Based on the CRPSS the multi-model forecasts showed skill close to the best-skilled individual model, or a slight improvement. Based on these findings, there is no clear proof that the use of a multi-model approach (consisting of two hydrological models) consistently improves flood forecasts. However, it can be said that the performance of the multi-model forecasts often clearly outperformed the worst model performance for the (lumped) upstream sub-catchments.

5 Discussion

This study has evaluated flood forecast performance using the hydrological models HBV and GR4H, and a multi-model consisting of the forecasts of both models combined. This chapter discusses the limitations of the study, which could have led to different results as well as interpretation of the results according to the literature.

5.1 Limitations in data, models and method

5.1.1 Data

This study has used historic measured discharge and precipitation time series in hourly time steps. Precipitation series were used from rain gauges in the catchment area and available radar measurements. In addition to these measurements, precipitation ensemble predictions from the COSMO-LEPS NWP model were used as input forcing for discharge forecasting.

Observed discharge
Discharge data was used from discharge measuring stations located in the main Vecht river and its tributaries. Given the fast-responding nature of the Vecht system, hourly data were essential for accurate flow routing and forecasting. However, such data were significantly less available than daily series, and few sub-catchments had long, continuous time series. After a 5-year calibration and two validation periods (a half-year and a 3-year period), the continuous data were nearly depleted. In addition, discharge has been measured using different methods, ranging from (relative) accurate acoustic methods to Q-h relationships. This could have led to inaccurate discharge measurements at some locations. Lastly, it was found that there are differences in discharge between stations located in close proximity (De Haandrik and Emlichheim). Overall, the hourly discharge data have been lacking in consistency and quality among the ten considered sub-catchments, leading to difficulties during calibration, validation, and forecasting.

Observed precipitation
To estimate precipitation, a combination of rain gauges and radar have been used. Rain gauges give accurate measurements, but only at the precise location of the measurement. Radar provides a higher spatial resolution, but with higher uncertainties. In this study, it was found that radar often underestimates cumulative precipitation compared to gauge measurements. Since input forcings are one of the main sources of uncertainty, the use of other precipitation data could have led to different results. Although no consistent deviation was found, data preprocessing could have been used to improve the precipitation time series used. The hourly precipitation time series were also found to be inconsistent and incomplete. For the German catchments, data was available from 2006. Radar in the Netherlands was available from 2008, but with multiple missing years. Radolan and IRC are available from 2020 onward. Similarly to the discharge data, there are many more, and longer, time series available on a daily time step.

Precipitation forecasts
To generate discharge forecasts, precipitation forecasts from the COSMO-LEPS numerical weather model were used. Unfortunately, only forecasts were stored in the FEWS-Vecht archive from 2021 onward. Due to reduced hourly observed discharge data, only 2 periods were found useful for assessment. This included the same high flow event used in validation, Christmas 2023. The other period did not include a high-flow event, but for all sub-catchments discharge time series were available. The use of more NWP models could have led to different results. COSMO-LEPS was selected mainly for its high forecast resolution (3 hour accumulations). However, since the models were made on an hourly time step, even higher forecast resolution could have been used. Unfortunately, these type of NWP models

either have no ensemble members (ICON-EU) or provide forecasts at short lead times (HARMONIE). In this study, the raw precipitation forecasts have not been processed. Multiple studies have shown that applying processing methods such as Quantile Regression Forest (Teja et al., 2023) or Quantile mapping (Benninga et al., 2017) to raw precipitation forecasts can improve model forecast performance. Using one of these processing techniques could have improved the forecast performance of the individual models, but also the multi-model performance.

5.1.2 Model

In order to model on an hourly time step, the GR4J model has been adapted to execute calculations in hours rather than in days. This has been done by changing the units of parameters from days to hours and feeding the model with hourly inputs. The minimum boundary value of X4 has been directly adopted from 0.5 days (as proposed in Perrin et al. (2003)) to 12 hours. Two sub-catchments obtained an X4 of exactly 12 hours during calibration. Probably, the 0.5 day minimum is purely a computational boundary value for modeling at daily time step. Reducing this boundary would maybe have led to other parameter values for the sub-catchments and also different model performance. In addition, Bennett et al. (2014) proposed additional changes to GR4J to provide a better discharge simulation on hourly time steps. This was done by applying factors to percolation, groundwater exchange (F), outflow of the routing store (Q_r) and S-curves of the unit hydrographs. These changes have been tested in this study for a single calibration run of the model, but this did not yield much different results to the multi-objective function. Since no other sources were found that explicitly describe the changes from GR4J to GR4H, it is unclear whether changes were necessary to gain improved discharge simulation of GR4H.

To make the hydrological models semi-distributed, discharge from the upstream sub-catchments was routed to sub-catchments downstream. This was done simply by applying a routing delay according to the distance of the sub-catchments outflow points and a flood wave propagation speed of 1.2 m/s. Different velocities have been tested, but did not show a consistent improvement of the multi-objective function. Including the propagation velocity as calibration parameter, the velocity can be modeled per sub-catchment. This could improve the timing of the modeled discharge peaks, and therefore increase model performance. In addition, more sophisticated methods could improve the flow routing between sub-catchments. For example, a hydraulic model could be used to describe the flow within the river. Since this would complicate the research, it was chosen to not include this. Also, Muskingum routing could have been used to describe the flow from one sub-catchment to the other. The use of either a hydraulic model or Muskingum routing will likely improve the simulated discharge in the downstream catchments.

In addition, only the natural river flow of the river was modeled. Influences from, for example, the complex canal network in the area have not been considered. The main reason for this is because during high flow conditions, the influence of the canal network is limited. However, during dry periods, the canals are used as inflow for the Vecht. By including this inflow into the models, the forecast results could have been different, especially for a heavy rain event after a prolonged drought. This type of high-flow event may be underestimated by the models because the inflow from the canals was not considered.

5.1.3 Method

This study has not performed a sensitivity analysis of the model parameters. For both models, the number of parameters to be calibrated was limited (only four for GR4H and eight for HBV). The most important parameters to be calibrated for HBV have been selected based on what has been done in Demirel et al. (2013). A sensitivity analysis could have led to another number of parameters to be calibrated, and hence different results.

The models were initialized at each issued forecast using a direct state variable updating method to simulate the last observed discharge at the start of each forecast. Model storages (UZ, LZ, and R) were determined through empirical relationships between simulated storage and simulated discharge. While this method proved accurate for upstream (lumped) catchments, the relationship deteriorated downstream due to flow routing from upstream catchments, complicating storage initialization to match the last measured discharge. For all downstream catchments, the error in initialization propagated through the system. Poor initialization of Vechte B can be seen in the forecasts for Stouwe as a sudden drop or an increase in discharge (depending on the sign of the discharge error). In order to select storage values for the downstream catchments, a distinction was made between baseflow and regular flow. This has helped slightly with determining the storage values, but for most downstream sub-catchments it was best to initialize the storages on zero. This meant that no discharge was generated from the storages. This proved to be the best choice for the downstream catchments under baseflow conditions, and also for Stouwe and Radewijke+Itterbeek during all flow conditions. Further challenges arose from the high sensitivity of UZ (HBV) and R (GR4H) to changes, contributing to inaccuracies and uncertainties in short lead-time forecasts for downstream catchments. Alternative procedures, such as Ensemble Kalman Filtering or statistical methods like auto-regressive moving average (ARMA), could improve forecast accuracy. Implementing these methods, particularly for downstream catchments, may significantly enhance model performance.

The forecasts produced by the individual models and the multi-model were evaluated using three metrics: RMAE, CRPS, and CRPSS. According to the literature, it is recommended to assess the forecasts using at least one metric for each key attribute: bias, accuracy, reliability, sharpness, and resolution. However, the chosen metrics do not cover all these attributes. Most metrics that quantify reliability and resolution require many forecasts in the analysis. This study had limited event-based forecasts (10 forecasts per period). This made it impossible to evaluate the forecasts using metrics such as rank histograms or ROC. The forecasting model could be improved by automating the forecast simulations, so that more forecasts can be generated faster. Increasing the number of forecasts would likely improve the performance of both individual models and the multi-model while also enhancing the statistical significance of the metrics used for evaluation.

The last limitation of the method used in this study is the number of models that have been used. For a multi-model study an absolute minimal number is two, which have been used in this study. Preferably, (many) more models are used to obtain a larger variability in model structures. In this study only two models have been used in the assessment due to time limitations in combination with programming limitations. Borman et al. (2007) suggest that a number of at least 6 models are required for accurate multi-model ensemble forecasting. It is therefore expected that if some additional models were added, the forecast performance of the multi-model would have been significantly higher than those of the individual models.

5.2 Interpretation of results

5.2.1 Model performance

The performance of HBV and GR4H was evaluated with the multi-objective function, combining NS_w and RVE. Performance during calibration was more than satisfactory for both models, however, HBV clearly outperformed GR4H (Table 13 & 14). For most sub-catchments, NS_w and NSE values for HBV exceeded those reported in the literature. In contrast, literature values for GR4H variants (GR4J and GR6J) were higher than those found in this study. Ten Berge (2024) found NS_w values around 0.8 for a lumped HBV model, compared to 0.84-0.91 in this study. Ten Berge (2024) found NS_w values around 0.9 for the GR6J model, compared to 0.63-0.82 in this study. Akhtar et al. (2009) and Benninga et al. (2017) used the normal variant of the NSE and found similar results. Akhtar et al. (2009) found NSE values ranging from 0.58 to 0.92 for six versions of a lumped HBV model. Benninga et al. (2017) found NSE values of 0.78 and 0.81 for a lumped HBV model. Zhang et al. (2015) found NSE values ranging from 0.84 to 0.95 for two versions of the GR4J model. None of the literature found used NS_w in combination with a semi-distributed hydrological model. This makes it difficult to find an explanation for the poorer performance of GR4H compared to variants of the model and the HBV model. Perhaps the GR4H model as used in this study is not correctly applied on an hourly time step, or the model is not well applicable on semi-distributed scale. Although NS_w was high for HBV and satisfactory for GR4H over the entire calibration period, both models clearly overestimated the highest peak event, while moderately high peaks were underestimated. Based on the evaluation of the hydrographs, during validation, GR4H also underestimated most high-flow peaks, showing little response to rainfall events.

5.2.2 Multi-model forecasts

This study has shown that combining flood forecasts from multiple hydrological models into one, creating multi-model forecasts, can slightly improve forecast performance. This has also been found in other studies. The results of Dion et al. (2021), Teja et al. (2023) and Velázquez et al. (2011) showed an increased model performance by grouping the forecasts of multiple hydrological models. The improvements found in the literature are larger and more significant than those found in this study. Perhaps, this could be due to the difference in the number of combined hydrological models in the literature. Dion et al. (2021) used eight models, Teja et al. (2023) used four models, and Velázquez et al. (2011) used 16 models. No other study was found that grouped two models into a multi-model forecast. Another reason could be the difference in spatial application of the models. All the literature found on multi-modeling has used lumped hydrological models. Using semi-distributed models increases the complexity of flood forecasting and could have resulted in the smaller forecast improvements found in this study.

Next to a small increase in forecast performance using the multi-model, this study has shown that using a combined forecast, reduces the possibility of having a very poor forecast. All three evaluation metrics showed that the multi-model forecast tends to be closer to the better of the two models. In case one of the models showed an extremely poor score, the multi-model score was affected much less by this poor score and often produced a decent score. This could have practical advantages for forecast systems relying on only one hydrological model. By adding at least one other to the system, the possibility of producing a better forecast than with only one model increases.

5.3 Generalizations

This study provides valuable insights into the impact of using multiple hydrological models in high-flow forecasting systems. However, the findings are specific to the Vecht basin and should be applied to other basins with caution. The models were developed to reflect the unique characteristics of the Vecht and its sub-catchments. In addition, not only the catchment characteristics and structure of the hydrological models determine the performance of a flood forecast. There are more sources of uncertainty, for example, the input, that determines the flood forecast performance. Also, the way the flood forecast system is designed, is of large influence on the performance. The choice to use a hydraulic model in the forecasting system has large influence on the hydrograph shape. Also, the choice for data assimilation has a large impact on the quality of the forecasts. By using another approach in the design of the forecast system it becomes more difficult to assign any possible differences in performance to the just the use of the multi-model.

6 Conclusion & Recommendations

6.1 Conclusion

RQ.1: How do HBV and GR4H perform on high-flow simulations, based on historical data?

Both models performed well in simulating high flows during the calibration period, yielding more than satisfactory multi-objective function values for all sub-catchments. HBV performed better on all parts of the hydrograph, compared to GR4H. GR4H overestimated extreme summer peaks, while underestimating enlarged flow during winter. The models were validated on two independent periods. Both models showed a decrease in performance for all sub-catchments for both periods. Especially for the extreme high flow event in validation period 1 the models showed poor performance for most sub-catchments. For both models, the downstream catchments showed better performance than the upstream catchments. However, timing of the flood peaks was often too early and too sharp for the downstream catchments. On validation period 2, with no extreme high flow event, both models performed better, but worse than the calibration period. To conclude, both models are able to simulate high-flows to satisfaction, however both models show large decrease in performance during extreme conditions.

RQ.2: What is the flood forecast performance of the single hydrological models with input of weather ensemble forecasts?

Both individual models HBV and GR4H have been used in a forecasting system for the Vecht. Forced by weather ensemble forecasts from COSMO-LEPS for two forecast periods, a total of 10 discharge forecasts per period per model were generated. These forecasts haven been evaluated visually with use of spaghetti plots, and quantitatively by evaluation metrics RMAE, CRPS and CRPSS. Both models showed difficulty in forecasting accurate discharge at short lead times (<24 hours) for the downstream catchments, compared to the upstream catchments. Contradictory, forecast performance of the downstream catchments at longer lead times were found to be slightly better than the upstream catchments. Forecast ensembles generated by GR4H tend to be under-dispersed for both periods for most sub-catchments. This is in accordance with the underestimation of flows found in validation. For HBV, no clear under- or over-dispersion of the ensembles was found.

RQ.3: Does the use of multiple hydrological models provide better ensemble flood forecast performance compared to the use of a single hydrological model?

The multi-model forecasts, consisting of the combined forecasts of HBV and GR4H, showed marginal improvement with respect to both individual model forecasts. For the upstream catchments, the RMAE and CRPS(S) metrics showed smaller errors when using the multi-model for several lead times. The downstream catchments showed smaller improvements in performance as the multi-model performance was mostly in between the individual model performance. However, the multi-model tends to reduce the effect of a poor forecast. It was found that the multi-model is less sensitive to extremely poor performance by one of the individual models, if the other model performs decently. The multi-model shows performance more closely to the better performing model than to the poor model. Generally, there is potential that a multi-model approach can improve forecast performance and is able to reduce the possibility of a poor forecast.

Research aim: to evaluate and compare the flood forecasting performance of single semi-distributed hydrological models versus a multi-model approach, driven by ensemble weather forecasts on an hourly time step, for the Overijsselse Vecht River.

The models HBV and GR4H were both successfully modeled on semi-distributed scale, and applied in a simple flood forecasting system. Despite many limitations of the applied method, a comparison between the model forecasts individually, and the multi-model forecasts was made. It was found that a multi-model consisting of only two hydrological models is able to improve forecast performance slightly for the upstream catchments. However, consistent improvements among all lead times and all subcatchments was not found. For the downstream catchments (with inflow from other catchments) the applied method proved to be unsuitable. This was, however, not the case for the upstream catchments. Despite the limitations, there is evidence that flood forecast systems using at least two hydrological models can improve both individual model performance, but also eliminate a poor forecast performance of one model.

6.2 Recommendations

This section provides suggestions to improve flood forecasts in multi-hydrological ensemble forecasting. First, general recommendations and further research regarding the methodology and results are given, followed by practical recommendations for the JCAR-ATRACE program, Deltares and water authorities in the Vecht basin.

6.2.1 General recommendations and further research

The results of this study show ensemble flood forecasting is subject to many uncertainties, coming from the input, models, but also the methods used. The applied data-assimilation technique proved to be unsuitable for semi-distributed models. The storages of the downstream catchments of the model were particularly hard to initialize. This has led to inaccurate forecasts at the start of each forecasts, of which the effects cascade through time to the other catchments downstream. This poor initialization has likely overshadowed possible improvements by the multi-model at the downstream catchments for short lead times. Therefore, it is recommended to only apply the direct storage updating method for lumped models. For further research that will make use of semi-distributed models, another updating technique is advised.

This study has combined the forecasts of two individual models to obtain the multi-models forecasts. However, most multi-modeling studies found in the literature use many more hydrological models. It is preferred to use more individual models and assess their combined performance. It is therefore recommended to increase the number of models in a multi-model study to at least three, but preferable more. It would also be of interest to combine different type of models (conceptual and physics-based). Also, by including more NWP models, uncertainties from the weather forecasts can be included and many more multi-model combinations can be made.

This study has focused on high flows only, while low flows pose an equally large future threat to the Vecht basin. Typically, hydrological models have difficulty modeling discharge in very long dry periods. It could be interesting to study drought forecasting in the Vecht basin.

6.2.2 Practical recommendations

In this study a transboundary river basin was evaluated. For accurate modeling, data from German and Dutch authorities was used. All data was scattered and obtained from multiple sources. Deltares provided a discharge data set, FEWS-Vecht has archived discharge and (forecasted) precipitation data, Vechtstromen and WDOD have their own databases and KNMI, DWD, NLWKN and NRW have data portals which can be found on the web. Every data source requested another approach in which data could be accessed. The metadata was often different, such that processing of the data to the preferred format was necessary. In this study all used data is stored and documented such it is clear where it originates from. However, this can be improved if the Dutch and German authorities would construct a central location where they combine their data. This would provide a continuous and automatically updating dataset. This data could include hourly and daily discharge and precipitation series. Practically, this is what FEWS-Vecht does in its archive, but this is not accessible for everyone. Such an initiative could be taken by JCAR-ATRACE (together with Deltares) in good consult with at least Vechtstromen, WDOD, NLWKN, NRW. In addition Rijkswaterstaat, KNMI and DWD could be included. It would be a large operation to achieve this, however it would increase the transparency of data sharing between all parties. Other studies that require data of the Vecht basin would also be able to take off faster when all data is stored somewhere central.

This study has found that there is much less hourly (discharge and precipitation) data available than daily data. The results of this study show response of the Vecht system (especially further upstream) is fast and hourly data and models are necessary to describe the discharge in the system. Also, there are many discontinuous time series in both precipitation and discharge measurements. Therefore it is recommended to increase the measurement network in the Vecht basin that measure on hourly time step. Currently there are four KNMI precipitation stations close to or in the Vecht basin. For an area as large at the Vecht, relying on only four precipitation stations is minimal. In the Vecht there are a handful of accurate acoustic discharge stations. It is recommended to place at least two more at the outlets of Dinkel and Regge (or repair/replace old ones). Both streams cover a large part of total Vecht basin, and have been found to lack (accurate) discharge data. To achieve this, Vechtstromen and NLWKN could invest in these measurement stations. Although discharge measurements in the German part of the Vecht where found to be reliable, an accurate measuring station could also be added to the network at for example Neuenhaus.

References

Akhtar, M., Ahmad, N., & Booij, M. J. (2009). Use of regional climate model simulations as input for hydrological models for the Hindukush-Karakorum-Himalaya region. *Hydrology and Earth System Sciences*, *13*(7), 1075–1089. https://doi.org/10.5194/hess-13-1075-2009

Anctil, F., & Ramos, M.-H. (2019). Verification Metrics for Hydrological Ensemble Forecasts. In *Handbook of hydrometeorological ensemble forecasting* (pp. 893–922). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39925-1{_}3

Arduino, G., Reggiani, P., & Todini, E. (2005). Recent advances in flood forecasting and flood risk assessment. *Hydrology and Earth System Sciences*, *9*(4), 280–284. https://doi.org/10.5194/hess-9-280-2005

Arsenault, R., Brissette, F., & Martel, J.-L. (2018). The hazards of split-sample validation in hydrological model calibration. *Journal of Hydrology*, *566*, 346–362. https://doi.org/10.1016/j.jhydrol.2018.09.027

Bennett, J. C., Robertson, D. E., Shrestha, D. L., Wang, Q., Enever, D., Hapuarachchi, P., & Tuteja, N. K. (2014). A System for Continuous Hydrological Ensemble Forecasting (SCHEF) to lead times of 9 days. *Journal of Hydrology*, *519*, 2832–2846. https://doi.org/10.1016/j.jhydrol.2014.08.010

Benninga, H.-J. F., Booij, M. J., Romanowicz, R. J., & Rientjes, T. H. M. (2017). Performance of ensemble streamflow forecasts under varied hydrometeorological conditions. *Hydrology and Earth System Sciences*, *21*(10), 5273–5291. https://doi.org/10.5194/hess-21-5273-2017

Bergström, S., & Forsman, A. (1973). Development of a conceptual deterministic rainfall-runoff model. *Nord. Hydrol*, *4*, 240–253.

Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources*, *16*(1), 41–51. https://doi.org/10.1016/0309-1708(93)90028-E

Borman, H., Breuer, L., Croke, B., Graeff, T., Hubrechts, L., Huisman, J. A., Kite, G., Lanini, J., Leavesley, G. H., Lindström, G., Seibert, J., & Willems, P. (2007). REDUCTION OF PREDICTIVE UNCERTAINTY BY ENSEMBLE HYDROLOGICAL MODELLING OF DISCHARGE AND LAND USE CHANGE EFFECTS. In L. Pfister & L. Hoffman (Eds.), *Uncertainties in the 'monitoring-conceptualisation-modelling' sequence of catchment research.*

Bouaziz, L. J. E., Fenicia, F., Thirel, G., de Boer-Euser, T., Buitink, J., Brauer, C. C., De Niel, J., Dewals, B. J., Drogue, G., Grelier, B., Melsen, L. A., Moustakas, S., Nossent, J., Pereira, F., Sprokkereef, E., Stam, J., Weerts, A. H., Willems, P., Savenije, H. H. G., & Hrachowitz, M. (2021). Behind the scenes of streamflow model performance. *Hydrology and Earth System Sciences*, *25*(2), 1069–1095. https://doi.org/10.5194/hess-25-1069-2021

Brown, J. D., Demargne, J., Seo, D.-J., & Liu, Y. (2010). The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environmental Modelling & Software*, *25*(7), 854–872. https://doi.org/10.1016/j.envsoft.2010.01.009

Carsell, K. M., Pingel, N. D., & Ford, D. T. (2004). Quantifying the Benefit of a Flood Warning System. *Natural Hazards Review*, *5*(3), 131–140. https://doi.org/10.1061/(ASCE)1527-6988(2004)5:3(131)

Cloke, H., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, *375*(3-4), 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005

Das, J., Manikanta, V., Nikhil Teja, K., & Umamahesh, N. V. (2022). Two decades of ensemble flood forecasting: a state-of-the-art on past developments, present applications and future opportunities. *Hydrological Sciences Journal*, *67*(3), 477–493. https://doi.org/10.1080/02626667.2021.2023157

De Groot, S. (2024, April). Master Thesis Literature Review.

Deltares. (n.d.). About Delft-FEWS. https://oss.deltares.nl/web/delft-fews/about-delft-fews

Demargne, J., Brown, J., Liu, Y., Seo, D.-J., Wu, L., Toth, Z., & Zhu, Y. (2010). Diagnostic verification of hydrometeorological and hydrologic ensembles. *Atmospheric Science Letters*, *11*(2), 114–122. https://doi.org/10.1002/asl.261

Demirel, M. C., Booij, M. J., & Hoekstra, A. Y. (2013). Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models. *Water Resources Research*, *49*(7), 4035–4053. https://doi.org/10.1002/wrcr.20294

Deutscher Wetterdienst. (2024). Analysen radarbasierter stündlicher (RW) und täglicher (SF) Niederschlagshöhen. https://www.dwd.de/DE/leistungen/radolan/radolan.html

Devineni, N., Sankarasubramanian, A., & Ghosh, S. (2008). Multimodel ensembles of streamflow forecasts: Role of predictor state in developing optimal combinations. *Water Resources Research*, *44*(9). https://doi.org/10.1029/2006WR005855

Dion, P., Martel, J.-L., & Arsenault, R. (2021). Hydrological ensemble forecasting using a multi-model framework. *Journal of Hydrology*, *600*, 126537. https://doi.org/10.1016/j.jhydrol.2021.126537

Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, *30*(5), 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Eberhart, R., & Kennedy, J. (1995). A new optimizer using particle swarm theory. *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, 39–43. https://doi.org/10.1109/MHS.1995.494215

ECMWF. (n.d.). Medium-range forecasts. https://www.ecmwf.int/en/forecasts/documentation-and-support/medium-range-forecasts

Faran, I. (2023). CRPS — A Scoring Function for Bayesian Machine Learning Models. *Towards Data Science*. https://towardsdatascience.com/crps-a-scoring-function-for-bayesian-machine-learning-models-dd55a7a337a8

Gijsbers, P., Werner, M., & Schellekens, J. (2008). *Delft FEWS: A proven infrastructure to bring data, sensors and models together* (tech. rep.). https://scholarsarchive.byu.edu/iemssconference/2008/all/89

Gill, M. K., Kaheil, Y. H., Khalil, A., McKee, M., & Bastidas, L. (2006). Multiobjective particle swarm optimization for parameter estimation in hydrology. *Water Resources Research*, *42*(7). https://doi.org/10.1029/2005WR004528

Gupta, H. V., Beven, K. J., & Wagener, T. (2005, October). Model Calibration and Uncertainty Estimation. In *Encyclopedia of hydrological sciences*. Wiley. https://doi.org/10.1002/0470848944.hsa138

Hapuarachchi, H. A. P., Wang, Q. J., & Pagano, T. C. (2011). A review of advances in flash flood forecasting. *Hydrological Processes*, *25*(18), 2771–2784. https://doi.org/10.1002/hyp.8040

Hundecha, Y., & Bárdossy, A. (2004). Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model. *Journal of Hydrology*, *292*(1-4), 281–295. https://doi.org/10.1016/j.jhydrol.2004.01.002

Imhoff, R., Brauer, C., van Heeringen, K.-J., Leijnse, H., Overeem, A., Weerts, A., & Uijlenhoet, R. (2021). A climatological benchmark for operational radar rainfall bias reduction. *Hydrology and Earth System Sciences*, *25*(7), 4061–4080. https://doi.org/10.5194/hess-25-4061-2021

IPCC. (2023, July). *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (P. Arias, M. Bustamante, I. Elgizouli, G. Flato, M. Howden, C. Méndez-Vallejo, J. J. Pereira, R. Pichs-Madruga, S. K. Rose, Y. Saheb, R. Sánchez Rodríguez, D. Ürge-Vorsatz, C. Xiao, N. Yassaa, J. Romero, J. Kim, E. F. Haites, Y. Jung, R. Stavins, ... C. Péan, Eds.; tech. rep.). Intergovernmental Panel on Climate Change. https://doi.org/10.59327/IPCC/AR6-9789291691647

Jahandideh-Tehrani, M., Bozorg-Haddad, O., & Loáiciga, H. A. (2020). Application of particle swarm optimization to water management: an introduction and overview. *Environmental Monitoring and Assessment*, *192*(5), 281. https://doi.org/10.1007/s10661-020-8228-z

Jain, S. K., Mani, P., Jain, S. 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. *International Journal of River Basin Management*, *16*(3), 329–344. https://doi.org/10.1080/15715124.2017.1411920

Jungermann, N., Hakvoort, H., & Versteeg, R. (2012, July). *Neerslag-afvoermodellen voor de Overijsselse Vecht* (tech. rep.). HKV.

Kan, G., He, X., Ding, L., Li, J., Liang, K., & Hong, Y. (2017). Study on Applicability of Conceptual Hydrological Models for Flood Forecasting in Humid, Semi-Humid Semi-Arid and Arid Basins in China. *Water*, *9*(10), 719. https://doi.org/10.3390/w9100719

Kauffeldt, A., Wetterhall, F., Pappenberger, F., Salamon, P., & Thielen, J. (2016). Technical review of large-scale hydrological models for implementation in operational flood forecasting schemes on continental level. *Environmental Modelling & Software*, *75*, 68–76. https://doi.org/10.1016/j.envsoft.2015.09.009

Klein, A., & van der Vat, M. (2024, April). *Scoping Study of the Vecht, Berkel and Oude IJssel river basins* (tech. rep.). JCAR ATRACE.

KNMI. (2009). *Neerslagklimatologie uit weerradar* (tech. rep.). KNMI. https://edepot.wur.nl/191836

KNMI. (2017). Weer- en klimaatmodellen. https://cdn.knmi.nl/system/readmore_links/files/000/000/394/original/NL_KNMI_Weermodel_A4_hires_100322.pdf?1661958779

Kundzewicz, Z. W., & Pińskwar, I. (2022). Are Pluvial and Fluvial Floods on the Rise? *Water*, *14*(17), 2612. https://doi.org/10.3390/w14172612

Lehmkuhl, F., Schüttrumpf, H., Schwarzbauer, J., Brüll, C., Dietze, M., Letmathe, P., Völker, C., & Hollert, H. (2022). Assessment of the 2021 summer flood in Central Europe. *Environmental Sciences Europe*, *34*(1), 107. https://doi.org/10.1186/s12302-022-00685-1

Lempio, G., Einfalt, T., Lobbrecht, A., & Lempio, G. (2012). *Considerations for compositing radar data from three countries* (tech. rep.).

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, *201*(1-4), 272–288. https://doi.org/10.1016/S0022-1694(97)00041-3

Liu & Gupta, H. V. (2007). Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, *43*(7). https://doi.org/10.1029/2006WR005756

Ludwig, P., Ehmele, F., Franca, M. J., Mohr, S., Caldas-Alvarez, A., Daniell, J. E., Ehret, U., Feldmann, H., Hundhausen, M., Knippertz, P., Küpfer, K., Kunz, M., Mühr, B., Pinto, J. G., Quinting, J., Schäfer, A. M., Seidel, F., & Wisotzky, C. (2023). A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe – Part 2: Historical context and relation to climate change. *Natural Hazards and Earth System Sciences*, *23*(4), 1287–1311. https://doi.org/10.5194/nhess-23-1287-2023

Madsen, H. (2000). Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. *Journal of Hydrology*, *235*(3-4), 276–288. https://doi.org/10.1016/S0022-1694(00)00279-1

Marsigli, C., Diomede, T., Montani, A., & Paccagnella, T. (2013, July). *COSMO Technical Report - The CONSENS (CONsolidation of COSMO ENSmble) Priority Project* (tech. rep.). Consortium for small scale modelling. https://doi.org/10.13140/RG.2.1.4233.6801

Moradkhani, H., & Sorooshian, S. (2008). General Review of Rainfall-Runoff Modeling: Model Calibration, Data Assimilation, and Uncertainty Analysis. In *Hydrological modelling and the water cycle* (pp. 1–24). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-77843-1{_}1

Pappenberger, F., Ramos, M., Cloke, H., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., & Salamon, P. (2015). How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology*, *522*, 697–713. https://doi.org/10.1016/j.jhydrol.2015.01.024

Pechlivanidis, I., Jackson, B., McIntyre, N., & Wheater, H. (2013). Catchment scale hydrological modelling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications. *Global NEST Journal*, *13*(3), 193–214. https://doi.org/10.30955/gnj.000778

Perera, D., Seidou, O., Agnihotri, J., Wahid, A., & Rasmy, M. (2019). Flood Early Warning Systems: A Review Of Benefits, Challenges And Prospects. *UNU-INWEH Report Series*, (8). https://doi.org/10.13140/RG.2.2.28339.78880

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*(1-4), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Pfannerstill, M., Guse, B., & Fohrer, N. (2014). Smart low flow signature metrics for an improved overall performance evaluation of hydrological models. *Journal of Hydrology*, *510*, 447–458. https://doi.org/10.1016/j.jhydrol.2013.12.044

Refsgaard, J. C. (1997). Validation and Intercomparison of Different Updating Procedures for Real-Time Forecasting. *Hydrology Research*, *28*(2), 65–84. https://doi.org/10.2166/nh.1997.0005

Refsgaard, J. C., Henriksen, H. J., Harrar, W. G., Scholten, H., & Kassahun, A. (2005). Quality assurance in model based water management – review of existing practice and outline of new approaches. *Environmental Modelling & Software*, *20*(10), 1201–1215. https://doi.org/10.1016/j.envsoft.2004.07.006

Ridler, M. E., van Velzen, N., Hummel, S., Sandholt, I., Falk, A. K., Heemink, A., & Madsen, H. (2014). Data assimilation framework: Linking an open data assimilation library (OpenDA) to a widely adopted model interface (OpenMI). *Environmental Modelling & Software*, *57*, 76–89. https://doi.org/10.1016/j.envsoft.2014.02.008

Seneviratne, S., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Ghosh, S., Iskandar, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S., Wehner, M., & Zhou, B. (2023, July). Weather and Climate Extreme Events in a Changing Climate. In *Climate change 2021 – the physical science basis* (pp. 1513–1766). Cambridge University Press. https://doi.org/10.1017/9781009157896.013

Shrestha, D. L., Kayastha, N., & Solomatine, D. P. (2009). A novel approach to parameter uncertainty analysis of hydrological models using neural networks. *Hydrology and Earth System Sciences*, *13*(7), 1235–1248. https://doi.org/10.5194/hess-13-1235-2009

Sitterson, J., Knightes, C., Parmar, R., Wolfe, K., Avant, B., Overview, A., & Muche, M. (2018). *An Overview of Rainfall-Runoff Model Types An Overview of Rainfall-Runoff Model Types An Overview of Rainfall-Runoff Model Types* (tech. rep.). https://scholarsarchive.byu.edu/iemssconferencehttps://scholarsarchive.byu.edu/iemssconference/2018/Stream-C/41Thisoralpresentation

Slager, K., & Kwadijk, J. (2023, July). *Accelerate Transboundary Regional Adaptation to Climate Extremes* (tech. rep.). JCAR ATRACE.

Spruyt, A., & Fujisaki, A. (2021, December). *Ontwikkeling zesde-generatie model Overijsselse Vechtdelta* (tech. rep.). Deltares. https://publications.deltares.nl/11205258_007_0007.pdf

Sun, Y., Bao, W., Valk, K., Brauer, C. C., Sumihar, J., & Weerts, A. H. (2020). Improving Forecast Skill of Lowland Hydrological Models Using Ensemble Kalman Filter and Unscented Kalman Filter. *Water Resources Research*, *56*(8). https://doi.org/10.1029/2020WR027468

Teja, K. N., Manikanta, V., Das, J., & Umamahesh, N. (2023). Enhancing the predictability of flood forecasts by combining Numerical Weather Prediction ensembles with multiple hydrological models. *Journal of Hydrology*, *625*, 130176. https://doi.org/10.1016/j.jhydrol.2023.130176

Ten Berge, A. (2024, April). *Robustness of hydrological models for simulating impact of climate change on high and low streamflow in the Lesse* [Doctoral dissertation, UTwente].

Thébault, C., Perrin, C., Legrand, S., Andréassian, V., Thirel, G., & Delaigue, O. (2024). What Can Be Expected from a Semi-Distributed Multi-Model Approach for Streamflow Forecasting? Tailoring the Structure and Size of a Super-Ensemble on the Rhône Basin. https://doi.org/10.2139/ssrn.5017925

Thielen, J., Bartholmes, J., Ramos, M.-H., & de Roo, A. (2009). The European Flood Alert System – Part 1: Concept and development. *Hydrology and Earth System Sciences*, *13*(2), 125–140. https://doi.org/10.5194/hess-13-125-2009

Thielen, J., Schaake, J., Martin, E., Pappenberger, F., & Pailleux, J. (2013). Hydrological ensemble prediction systems. *Hydrological Processes*, *27*(1), 1–4. https://doi.org/10.1002/hyp.9679

Tradowsky, J. S., Philip, S. Y., Kreienkamp, F., Kew, S. F., Lorenz, P., Arrighi, J., Bettmann, T., Caluwaerts, S., Chan, S. C., De Cruz, L., de Vries, H., Demuth, N., Ferrone, A., Fischer, E. M., Fowler, H. J., Goergen, K., Heinrich, D., Henrichs, Y., Kaspar, F., . . . Wanders, N. (2023). Attribution of the heavy rainfall events leading to severe flooding in Western Europe during July 2021. *Climatic Change*, *176*(7), 90. https://doi.org/10.1007/s10584-023-03502-7

Trinh, B. N., Thielen-del Pozo, J., & Thirel, G. (2013). The reduction continuous rank probability score for evaluating discharge forecasts from hydrological ensemble prediction systems. *Atmospheric Science Letters*, *14*(2), 61–65. https://doi.org/10.1002/asl2.417

van Heeringen, K.-J. (2023). *Handleiding FEWS Vecht 2023* (tech. rep.).

van Heeringen, K.-J., Filius, P., Tromp, G., & Renner, T. (2013). FEWS Vecht, a crossing boundaries flood forecasting system. *EGU General Assembly Conference Abstracts*, 13808.

Velázquez, J. A., Anctil, F., & Perrin, C. (2010). Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments. *Hydrology and Earth System Sciences*, *14*(11), 2303–2317. https://doi.org/10.5194/hess-14-2303-2010

Velázquez, J. A., Anctil, F., Ramos, M. H., & Perrin, C. (2011). Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. *Advances in Geosciences*, *29*, 33–42. https://doi.org/10.5194/adgeo-29-33-2011

Verdonschot, P. F., & Verdonschot, R. C. (2017). *Meetprogramma Overijsselse Vecht : nulsituatie 2017 en effecten maatregelen* (tech. rep.). https://doi.org/10.18174/440223

Verkade, J. (2008). *The value of flood warning systems* (tech. rep.).

Verkade, J., Brown, J., Reggiani, P., & Weerts, A. (2013). Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, *501*, 73–91. https://doi.org/10.1016/j.jhydrol.2013.07.039

Vormoor, K., Heistermann, M., Bronstert, A., & Lawrence, D. (2018). Hydrological model parameter (in)stability – "crash testing" the HBV model under contrasting flood seasonality conditions. *Hydrological Sciences Journal*, *63*(7), 991–1007. https://doi.org/10.1080/02626667.2018.1466056

Werner, M., Schellekens, J., Gijsbers, P., van Dijk, M., van den Akker, O., & Heynert, K. (2013). The Delft-FEWS flow forecasting system. *Environmental Modelling & Software*, *40*, 65–77. https://doi.org/10.1016/J.ENVSOFT.2012.07.010

Werner, M., Schellekens, J., & Kwadijk, J. C. J. (2005, October). Flood Early Warning Systems for Hydrological (Sub) Catchments. In *Encyclopedia of hydrological sciences*. Wiley. https://doi.org/10.1002/0470848944.hsa022

Wetterhall, F., Pappenberger, F., Alfieri, L., Cloke, H. L., Thielen-del Pozo, J., Balabanova, S., Daňhelka, J., Vogelbacher, A., Salamon, P., Carrasco, I., Cabrera-Tordera, A. J., Corzo-Toscano, M., Garcia-Padilla, M., Garcia-Sanchez, R. J., Ardilouze, C., Jurela, S., Terek, B., Csik, A., Casey, J., . . . Holubecka, M. (2013). HESS Opinions &amp;quot;Forecaster priorities for improving probabilistic flood forecasts&amp;quot; *Hydrology and Earth System Sciences*, *17*(11), 4389–4399. https://doi.org/10.5194/hess-17-4389-2013

WMO. (2013, May). *Integrated Flood Management Tools and Series. Flood Forecasting and Early Warning* (tech. rep.). World Meteorological Organization. www.wmo.int

Wu, W., Emerton, R., Duan, Q., Wood, A. W., Wetterhall, F., & Robertson, D. E. (2020). Ensemble flood forecasting: Current status and future opportunities. *WIREs Water*, *7*(3). https://doi.org/10.1002/wat2.1432

Xiong, L., & O'Conner, K. M. (2002). Comparison of four updating models for real-time river flow forecasting. *Hydrological Sciences Journal*, *47*(4), 621–639. https://doi.org/10.1080/02626660209492964

Yazdi, J., Salehi Neyshabouri, S. A. A., & Golian, S. (2014). A stochastic framework to assess the performance of flood warning systems based on rainfall-runoff modeling. *Hydrological Processes*, *28*(17), 4718–4731. https://doi.org/10.1002/hyp.9969

Zhang, X., Booij, M. J., & Xu, Y.-P. (2015). Improved Simulation of Peak Flows under Climate Change: Postprocessing or Composite Objective Calibration? *Journal of Hydrometeorology*, *16*(5), 2187–2208. https://doi.org/10.1175/JHM-D-14-0218.1

A Discharge data

Table 16: Explanation of available hourly measuring discharge stations. The stations in bold are used in the final dataset. The table also shows which sources are used in the final dataset.

Name	Country	From (d/m/Y)	Until (d/m/Y)	Source(s)	Comments
Bilk	GE	1/11/1975	31/5/2024	Deltares (1975-2017), NRW (1996-2024)	NRW data most reliable
Wettringen	GE	1/11/1975	31/5/2024	Deltares (1975-2017), NRW (1996-2024)	NRW data most reliable
Ohne	GE	1/11/2003	30/4/2024	NLWKN (2003-2017), FEWS (2018-2024)	Real-time FEWS data, with missing years 2018 and 2020
Gronau	GE	1/11/1996	31/5/2024	NRW	NRW data most reliable
Lage Gesamt	GE	1/1/2004	31/12/2017	NLWKN	No data after 2017 due to software issue
Dinkel	GE	1/5/2020	11/6/2023	FEWS	Real-time from FEWS database
Neuenhaus	GE	1/1/2001	30/4/2024	NLWKN (2001-2018), FEWS (2021-2024)	Stuwmeting. Data van Sebas tot 2017, vanaf 2020 real-time data uit FEWS (gat tussen 2017 en 2020)
Emlichheim	GE	1/10/1998	30/4/2024	NLWKN (2004-2018), FEWS (2018-2024), Vechtstromen (1998, Christmas '23)	NLWKN data complemented with real-time FEWS and Vechtstromen data
De Haandrik	NL	1/1/2007	28/5/2024	Vechtstromen	Christmas '23 contains corrected discharge
Ane Gramsbergen	NL	15/7/2005	30/4/2024	Deltares (2005-2016), FEWS (2020-2024), Vechtstromen (Christmas '23)	Real-time FEWS data contains many gaps
Ommen	NL	14/10/2001	1/5/2024	RWS (2001-2024), Vechtstromen (Christmas '23)	
Ommerkanaal	NL	14/10/2001	1/5/2024	RWS (2001-2024), Vechtstromen (Christmas '23)	
Archem TOT	NL	1/1/1996	30/4/2024	Deltares (1996-2017), FEWS (2022-2024), Vechtstromen (Christmas '23)	Archem TOT is sum of Archem Regge and Archem Linderbeek, real-time FEWS contains many gaps
Dalfsen	NL	12/6/2012	13/5/2024	Drents-Overijsselse Delta	

B Thiessen polygons

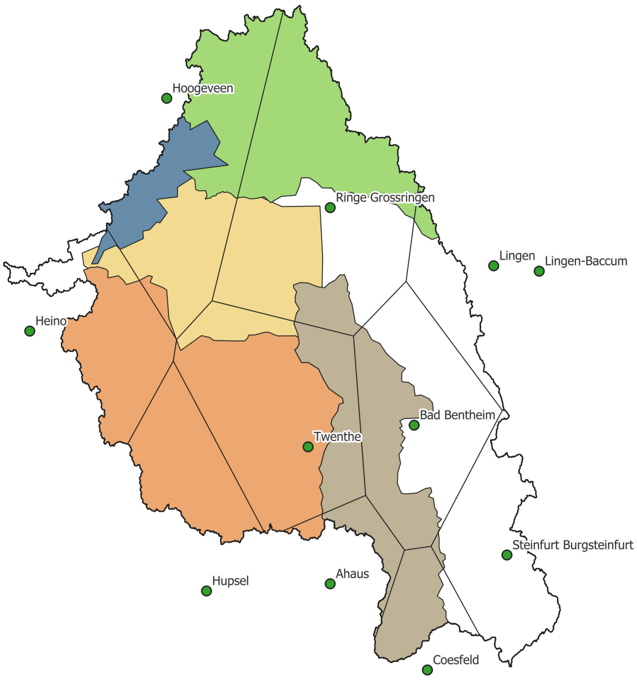


Figure 30: Thiessen polygons of the rain gauges drawn over the Dutch sub-catchments. The coverage over the catchment is used as weight to determine the amount of allocated precipitation.

C Calibration Results

C.1 PSO Figures

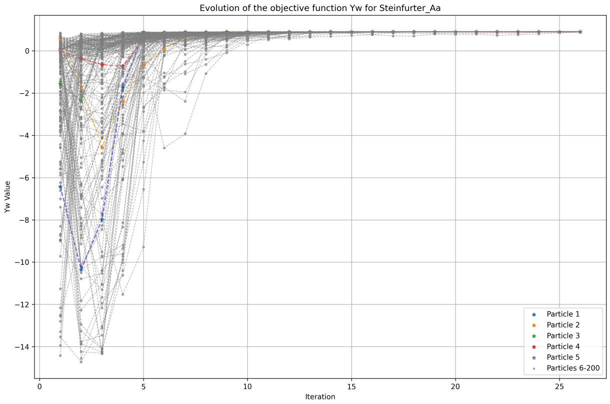


Figure 31: Evolution of the particles during PSO run 3 for HBV for sub-catchment Steinfurter Aa.

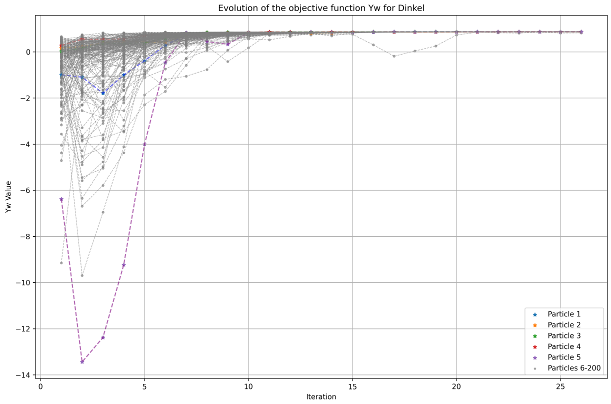


Figure 32: Evolution of the particles during PSO run 3 for HBV for sub-catchment Dinkel.

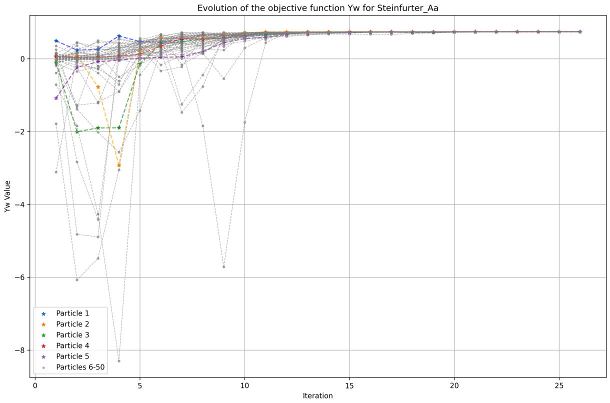


Figure 33: Evolution of the particles during PSO run 2 for GR4H for sub-catchment Steinfurter Aa.

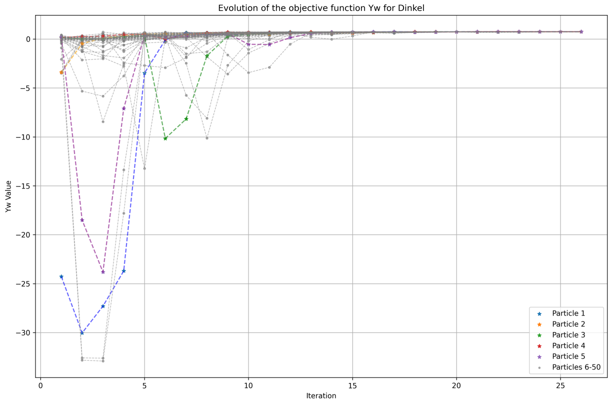


Figure 34: Evolution of the particles during PSO run 2 for GR4H for sub-catchment Dinkel.

C.2 Optimal Parameter sets

Table 17: Calibration results of HBV run 1 (50 particles and 25 iterations per catchment). In total 11.700 runs have been computed.

	FC	LP	BETA	CFLUX	ALFA	KF	KS	PERC	Y_w	NS_w	RVE
						(10ˆ-3)	(10ˆ-3)				(10ˆ-5)
Steinfurter Aa	564.18	0.85	4.78	0.02	0.1	16.1	1.9	0.0004	0.90	0.90	-0.83
Vechte A	389.92	0.85	2.45	0.03	0.1	7.3	7.6	0.017	0.86	0.86	-31.00
Dinkel	176.57	0.49	3.81	0.02	0.34	2	5.7	0.039	0.86	0.86	1.05
Afwateringskanaal	215.94	0.27	4.21	0.04	0.57	2	8.3	0.16	0.84	0.84	2.43
Ommerkanaal	118.89	0.37	1.68	0.02	0.34	2.7	4.5	0.003	0.85	0.85	0.07
Regge	549.8	0.71	3.23	0.02	0.46	3.7	7.5	0.12	0.82	0.82	-1.47
Vechte B	319.6	1	4.47	0.02	0.1	0.2	8.2	0.25	0.85	0.85	4.22
Vechte C	150.57	0.74	4.96	0.03	0.37	0.5	1.5	0.18	0.86	0.86	1.58
Radewijke+Itterbeek	520.16	0.34	1.19	0.03	1.97	0.4	0.02	0.25	0.91	0.91	-6.82
Stouwe	-	-	-	-	-	-	-	-	-	-	-

Table 18: Calibration results of HBV run 2 (50 particles and 25 iterations per catchment). In total 11.700 runs have been computed.

	FC	LP	BETA	CFLUX	ALFA	KF	KS	PERC	Y_w	NS_w	RVE
						(10ˆ-3)	(10ˆ-3)				(10ˆ-5)
Steinfurter Aa	363.67	0.70	4.39	0.03	0.1	14.75	3.48	0.01	0.90	0.90	0.17
Vechte A	507.59	1.00	5.35	0.02	0.1	9.10	7.34	0.08	0.87	0.87	-0.62
Dinkel	218.04	0.62	4.43	0.01	0.28	3.27	6.96	0.14	0.86	0.86	2.65
Afwateringskanaal	216.82	0.34	4.42	0.04	0.93	0.21	4.12	0.00	0.84	0.84	42.90
Ommerkanaal	484.47	1.00	6.00	0.03	0.1	8.58	2.79	0.00	0.85	0.85	0.47
Regge	445.65	0.49	1.86	0.04	0.1	8.71	7.30	0.17	0.83	0.83	-2.44
Vechte B	100.00	0.71	5.45	0.04	0.1	0.21	8.33	0.22	0.85	0.85	0.59
Vechte C	543.50	0.98	2.63	0.03	0.1	0.21	7.86	0.17	0.85	0.85	-0.69
Radewijke+Itterbeek	512.58	0.79	4.10	0.03	0.1	4.19	0.02	0.12	0.91	0.91	0.14
Stouwe	-	-	-	-	-	-	-	-	-	-	-

Table 19: Calibration results of HBV run 3 (200 particles and 25 iterations per catchment). In total 46.800 runs have been computed.

	FC	LP	BETA	CFLUX	ALFA	KF	KS	PERC	Y_w	NS_w	RVE
						(10ˆ-3)	(10ˆ-3)				(10ˆ-5)
Steinfurter Aa	247.08	0.47	4.24	0.04	0.11	12.57	6.87	0.06	0.90	0.90	0.72
Vechte A	457.33	0.98	5.31	0.01	0.10	9.05	6.14	0.11	0.87	0.87	2.00
Dinkel	213.39	0.50	2.32	0.02	0.21	3.78	5.77	0.03	0.87	0.87	1.93
Afwateringskanaal	222.58	0.29	3.51	0.04	0.10	8.22	5.71	0.03	0.84	0.84	3.34
Ommerkanaal	110.47	0.49	5.15	0.01	0.11	7.07	1.25	0.07	0.85	0.85	-2.83
Regge	447.95	0.53	2.09	0.02	0.10	8.79	4.56	0.10	0.84	0.84	2.11
Vechte B	323.44	1.00	4.39	0.03	0.10	0.21	8.33	0.25	0.86	0.86	-0.02
Vechte C	305.86	0.99	6.00	0.02	0.44	0.21	3.38	0.07	0.86	0.86	-0.02
Radewijke+Itterbeek	800.00	0.10	6.00	0.03	0.10	0.22	1.02	0.20	0.91	0.91	-0.004
Stouwe	-	-	-	-	-	-	-	-	-	-	-

Table 20: Calibration results of GR4H of PSO run 1 (25 particles and 25 iterations per catchment). In total 5.850 runs have been computed.

	X1	X2	X3	X4	Y_w	NS_w	RVE (10^{-5})
Steinfurter Aa	517.87	-1.36	84.34	12.00	0.75	0.75	1.33
Vechte A	450.07	-1.08	154.84	15.09	0.75	0.75	5.87
Dinkel	245.78	-2.06	220.48	15.53	0.72	0.72	5.65
Afwateringskanaal	108.53	-1.26	48.66	22.40	0.79	0.79	-20
Ommerkanaal	176.47	-0.15	10.00	54.66	0.68	0.68	-59
Regge	730.50	-0.80	27.69	14.35	0.82	0.82	7.73
Vechte B	371.54	-0.62	130.28	54.14	0.61	0.61	14.6
Vechte C	267.64	-0.38	46.61	30.20	0.72	0.72	2.88
Radewijke+Itterbeek	10.00	-6.73	46.74	12.48	0.73	0.73	1.28
Stouwe	-	-	-	-	-	-	-

Table 21: Calibration results of GR4H of PSO run 2 (50 particles and 25 iterations per catchment). In total 11.700 runs have been computed.

	X1	X2	X3	X4	Y_w	NS_w	RVE (10^{-5})
Steinfurter Aa	447.47	-1.62	108.88	12.00	0.74	0.74	-2.67
Vechte A	447.32	-1.07	153.63	17.63	0.76	0.76	3.15
Dinkel	455.19	-0.23	10.00	75.25	0.75	0.75	3.56
Afwateringskanaal	17.97	-3.67	153.99	15.81	0.79	0.79	-8.19
Ommerkanaal	46.81	-1.60	184.90	13.34	0.76	0.76	-4.41
Regge	271.50	-3.88	271.01	12.00	0.56	0.56	-52
Vechte B	523.93	-0.27	38.13	12.00	0.62	0.62	23.4
Vechte C	248.14	-0.38	47.12	47.87	0.75	0.75	4.29
Radewijke+Itterbeek	552.86	-5.67	59.12	45.06	0.73	0.73	-2.27
Stouwe	-	-	-	-	-	-	-

C.3 Calibrated hydrographs

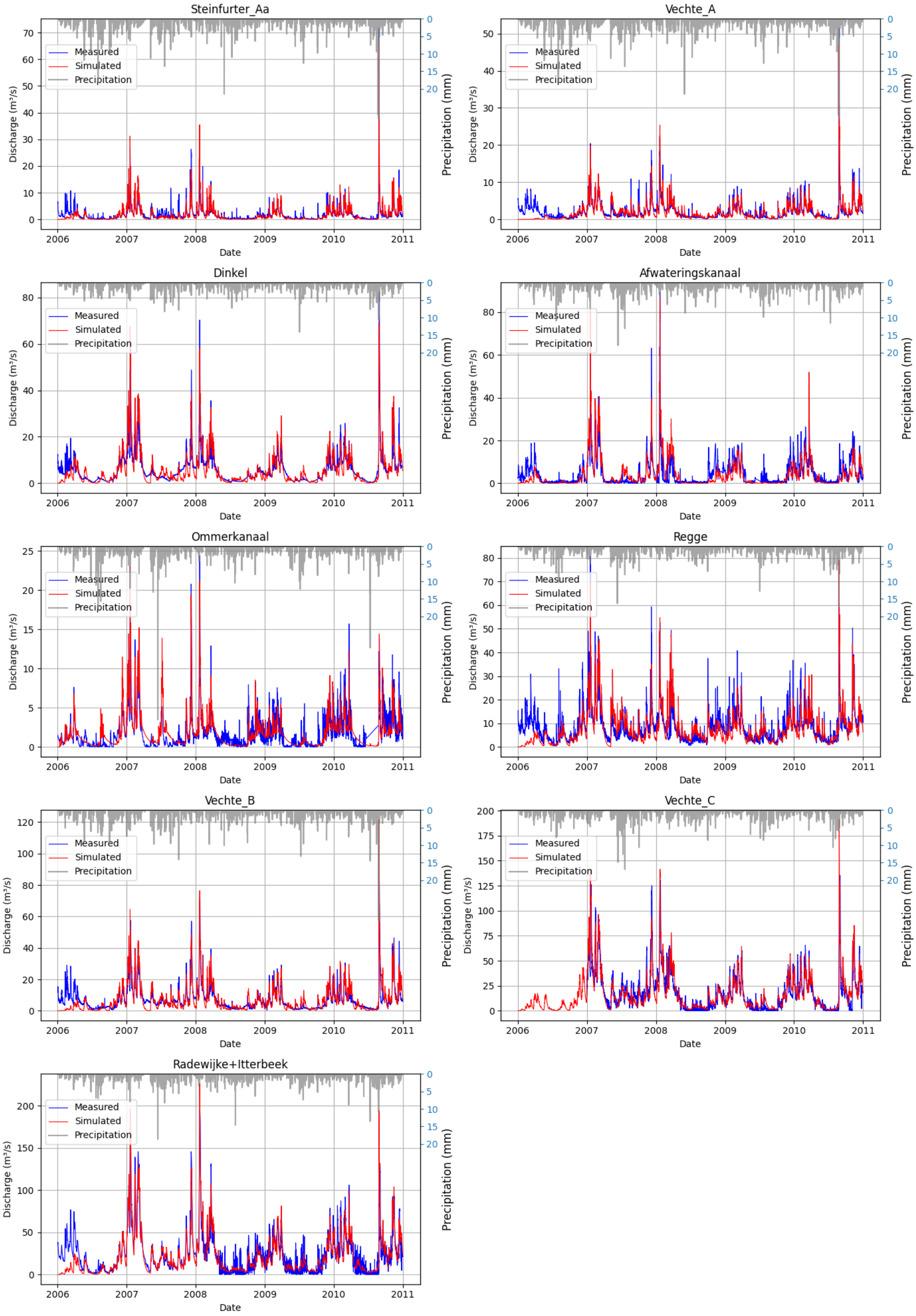


Figure 35: Observed and simulated hydrographs after calibration of HBV. The used parameter set is the best performing combination out of three PSO runs. The simulated discharge resembles the observed discharge closely, however, most peaks are slightly underestimated.

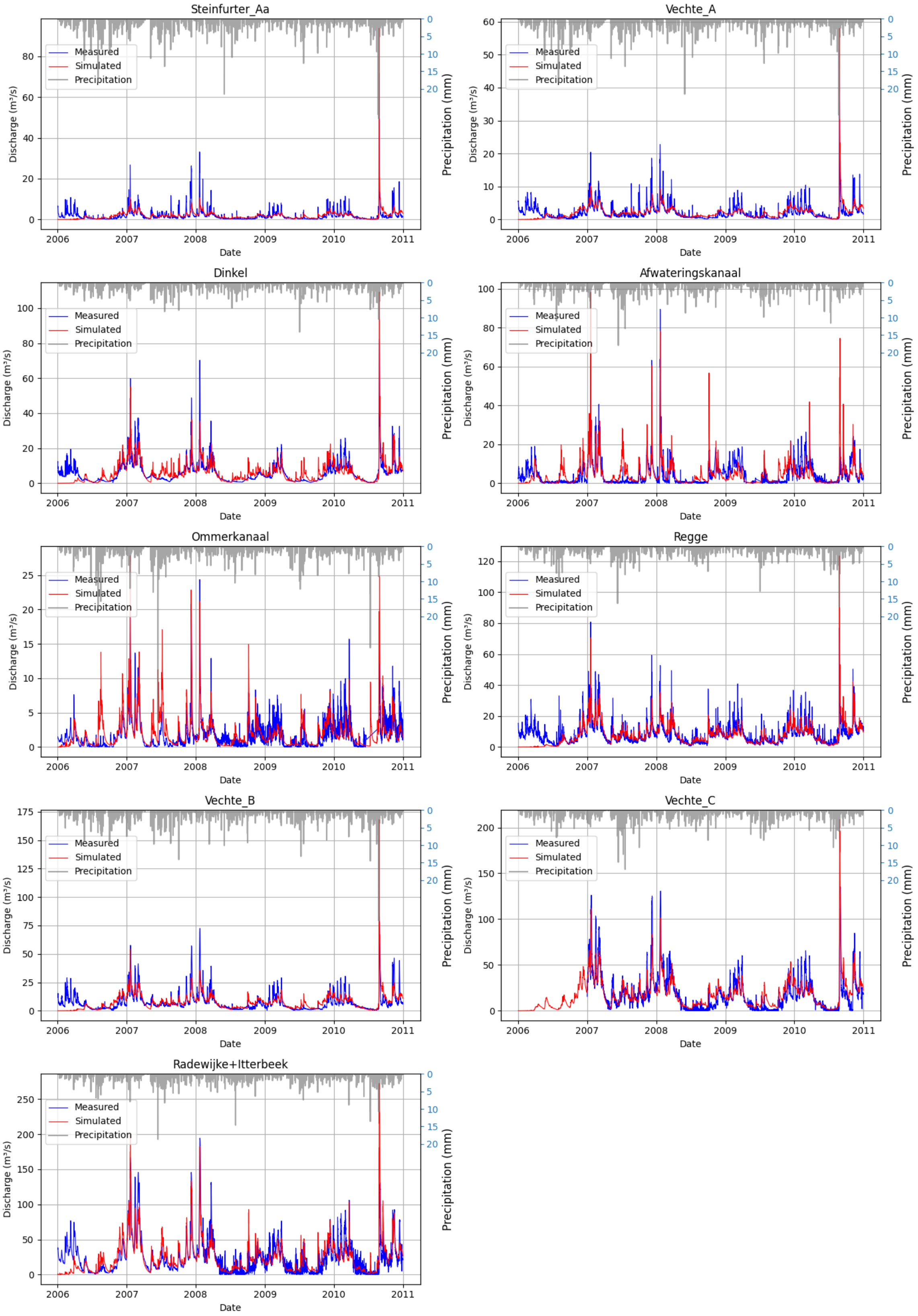


Figure 36: Observed and simulated hydrographs after calibration of GR4H. The used parameter set is the best performing combination of PSO run 1 and PSO run 2. Clearly, the largest peak is often overestimated, while baseflow and small peaks during winter are often underestimated.

D Validation Results

D.1 Objective Function Values

Table 22: Objective function values for validation 1 and 2

	Validation 1 (2023-2024)						Validation 2 (2013-2016)					
	HBV			GR4H			HBV			GR4H		
	Y_w	NS_w	RVE	Y_w	NS_w	RVE	Y_w	NS_w	RVE	Y_w	NS_w	RVE
Steinfurter Aa	0.77	0.78	-0.02	0.36	0.43	-0.21	0.61	0.73	-0.20	0.45	0.46	-0.02
Vechte A	0.52	0.68	-0.31	0.24	0.31	-0.29	0.65	0.78	-0.19	0.40	0.42	-0.07
Dinkel	-	-	-	-	-	-	0.83	0.88	-0.07	0.66	0.67	0.03
Afwateringskanaal	0.51	0.55	0.08	0.49	0.58	-0.19	0.53	0.72	0.35	0.42	0.63	0.52
Ommerkanaal	0.70	0.79	-0.11	0.55	0.67	-0.21	0.78	0.85	-0.09	0.86	0.86	0.01
Regge	-0.26	-0.28	0.09	0.49	0.49	0.01	0.52	0.57	-0.11	0.56	0.61	-0.08
Vechte B	0.66	0.73	-0.12	0.37	0.45	-0.21	0.85	0.91	-0.07	0.64	0.66	0.03
Vechte C	0.72	0.82	0.14	0.83	0.84	0.01	0.74	0.84	-0.13	0.67	0.70	-0.04
Radewijke+Itterbeek	0.79	0.80	0.01	0.61	0.67	-0.09	0.74	0.83	-0.13	0.82	0.82	-0.01
Stouwe	0.64	0.74	0.15	0.80	0.84	0.05	0.87	0.88	-0.01	0.76	0.84	0.11

D.2 Validation Figures

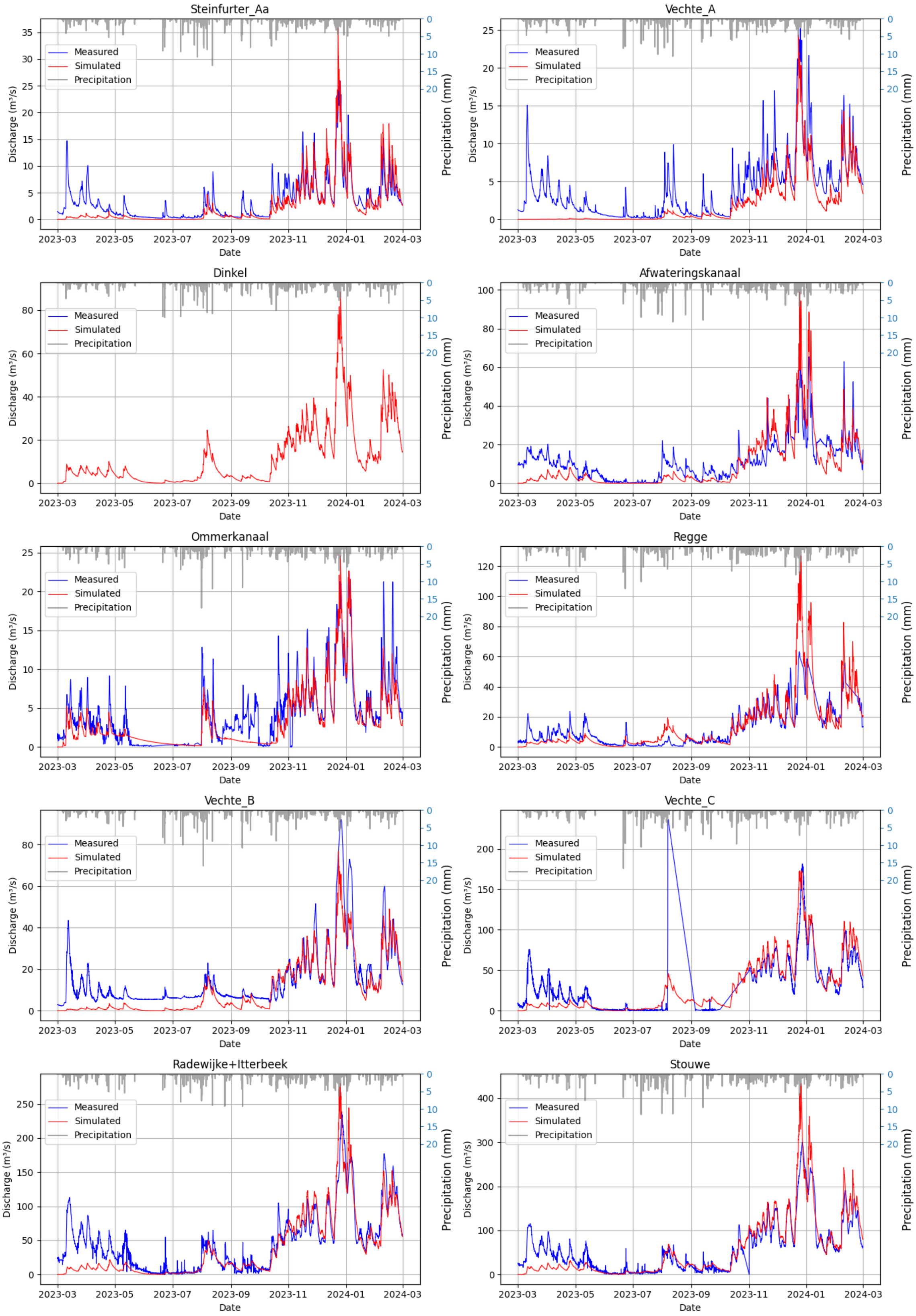


Figure 37: Validation hydrographs of HBV for validation period 1 (Mar 2023 - Mar 2024)

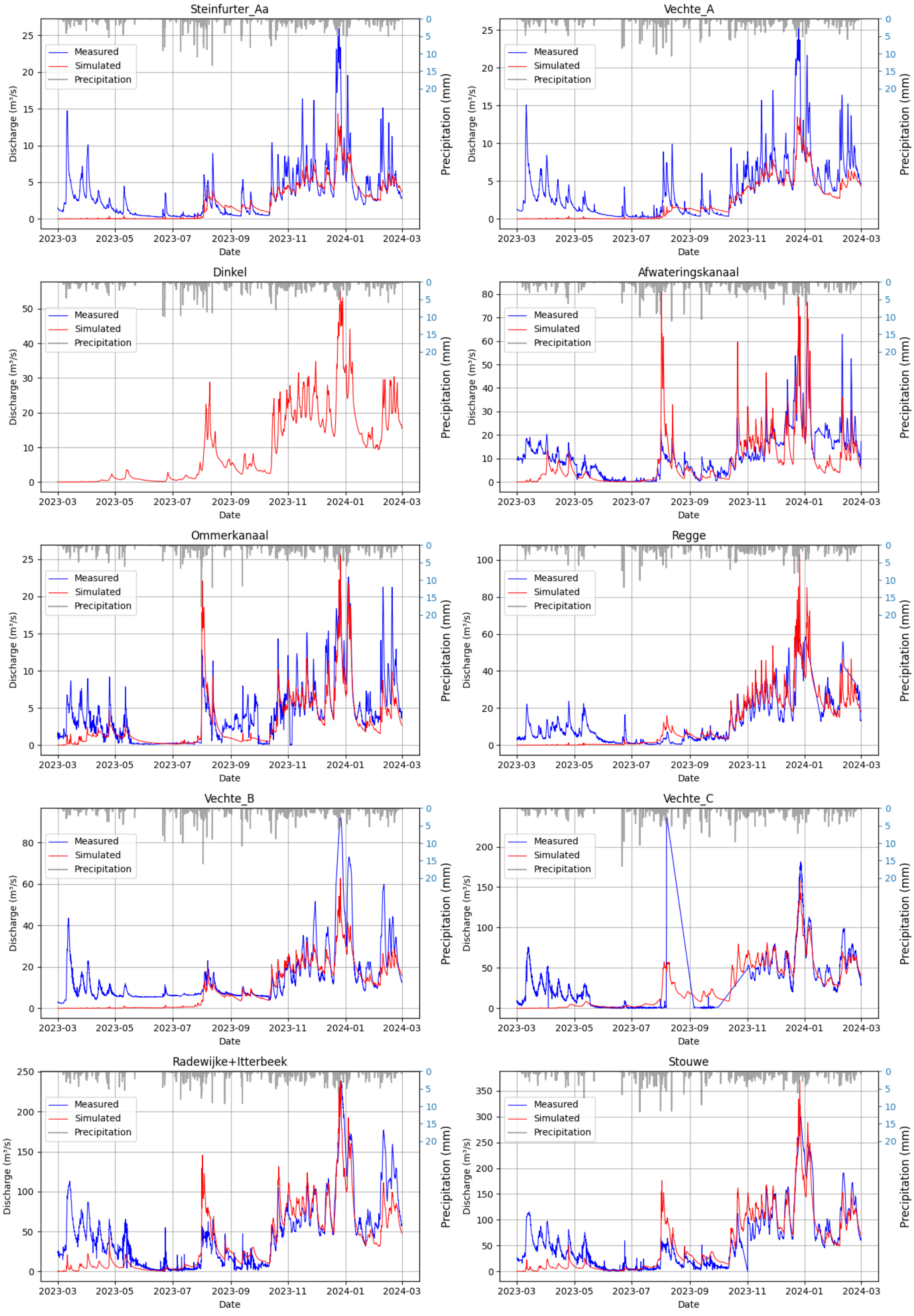


Figure 38: Validation hydrographs of GR4H for validation period 1 (Mar 2023 - Mar 2024)

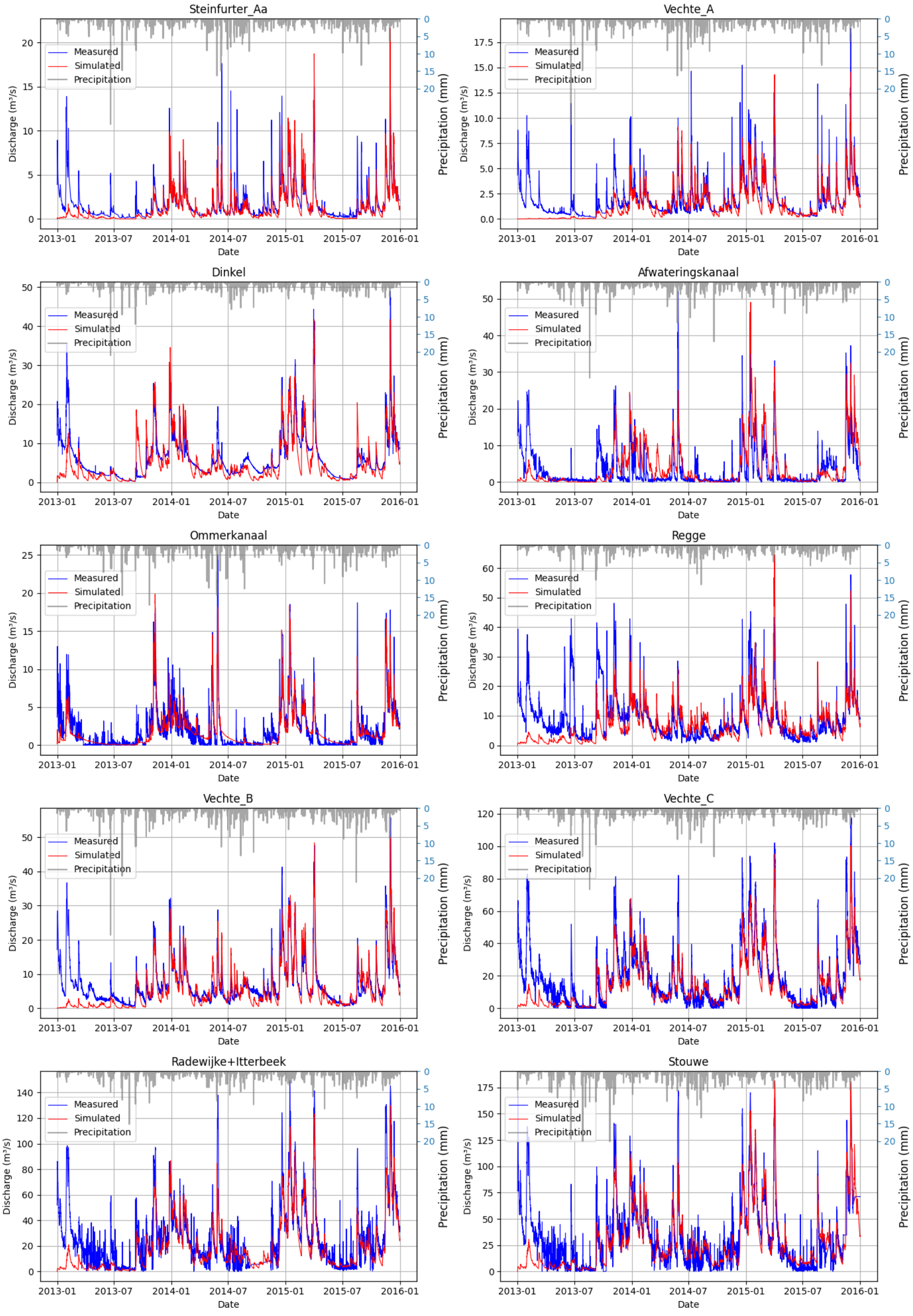


Figure 39: Validation hydrographs of HBV for validation period 2 (Jan 2013 - Jan 2016)

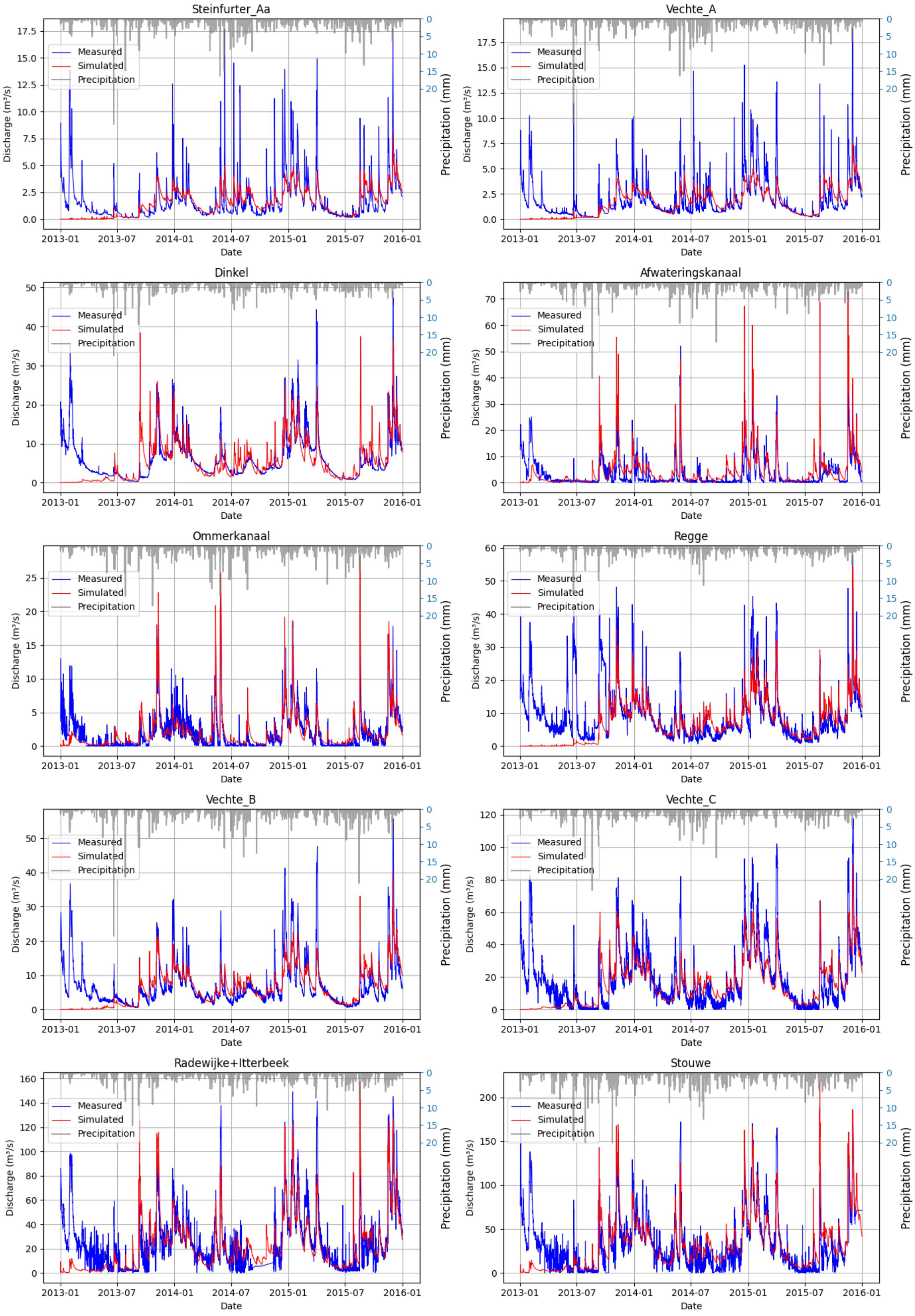


Figure 40: Validation hydrographs of GR4H for validation period 2 (Jan 2013 - Jan 2016)

E Qh-relation station Dalfsen

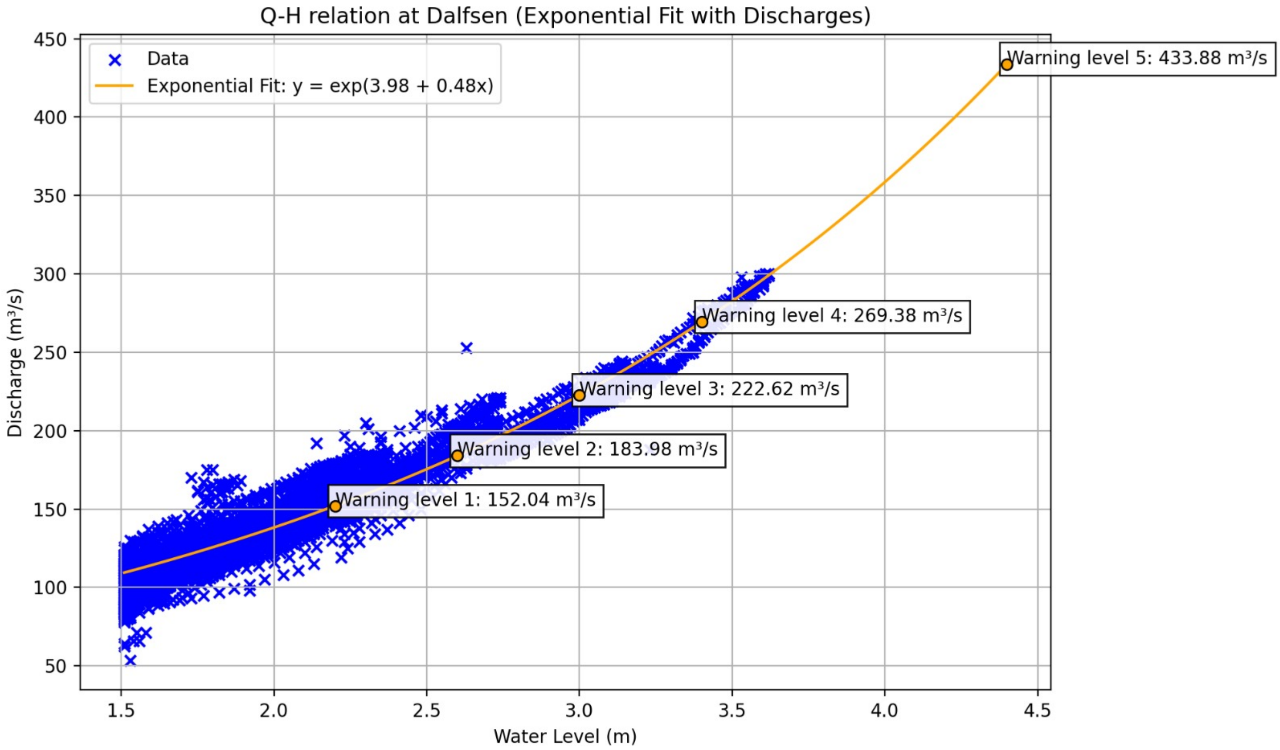


Figure 41: Qh-relation of discharge station Dalfsen. Discharge warning levels have been derived from an exponential fit, based on fixed water level warnings.

F Forecast evaluation (additional material)

F.1 Spaghetti plots

Figures 42 and 43 show the spaghetti plots for sub-catchments Vechte C and Ommerkanaal (forecasts issued at 23-12-2024 13:00 and 08-03-2024 13:00) for each forecast period for all model combinations. Vechte C is one of the worst performing catchments for the Christmas event, mainly due to poor discharge updating at the forecast time (lead time of 0 hours). At the first forecast time the observed discharge from Vechte B, Dinkel and forecasted discharge in Vechte C are used to determined the forecasted discharge. However, during the Christmas 2023 event there was no observed discharge available for Dinkel. It was therefore impossible to update the storages of the models on the last observed discharge. This could have led to the overestimation of discharge for Vechte C, but also for the catchments further downstream. During the March 2023 event, discharge data from all sub-catchments was available, hence a better forecast at a lead time of 0 hours. The multi-model forecasts do not clearly outperform the individual models, except for increasing the ensemble spread.

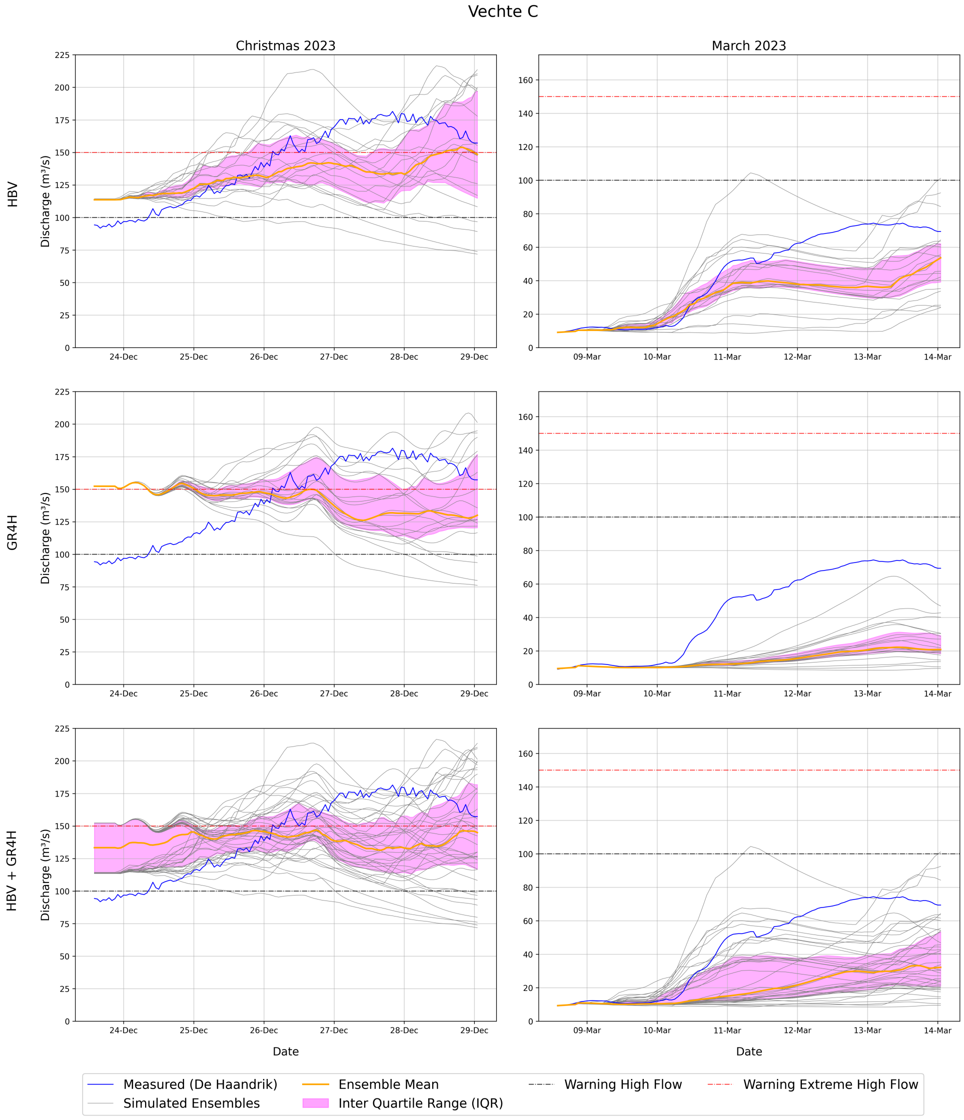


Figure 42: Spaghetti plots of the ensemble forecasts (grey), measured discharge (blue), ensemble median (orange) and inter quartile range (pink) for sub-catchment Vechte C for both forecast periods for HBV, GR4H and the multi-model. The dashed horizontal lines represent the warning levels at which the water authorities increase the state of alertness. Note that the y-axes are not similar between the events, due to the large difference in peak discharge.

The spaghetti plots for sub-catchment Ommerkanaal show accurate forecasts at a lead time of 0 hours. For all six upstream catchments the storage updating procedure resulted in accurate discharge forecasts at the first forecast time. The multi-model shows clearly the different individual model behavior at the start of the forecast. HBV underestimates the observed discharge, while GR4H slightly overestimates. The ensemble median of the MHM shows hence a more accurate forecast compared to the individual model ensemble medians. At all other lead times, for both forecast periods, all models show an under-dispersion of forecasted discharge ensembles. This has been observed for most catchments for most forecasts, which is remarkable because the COSMO-LEPS weather prediction model has shown over-dispersion of precipitation compared to rain gauges and radar.

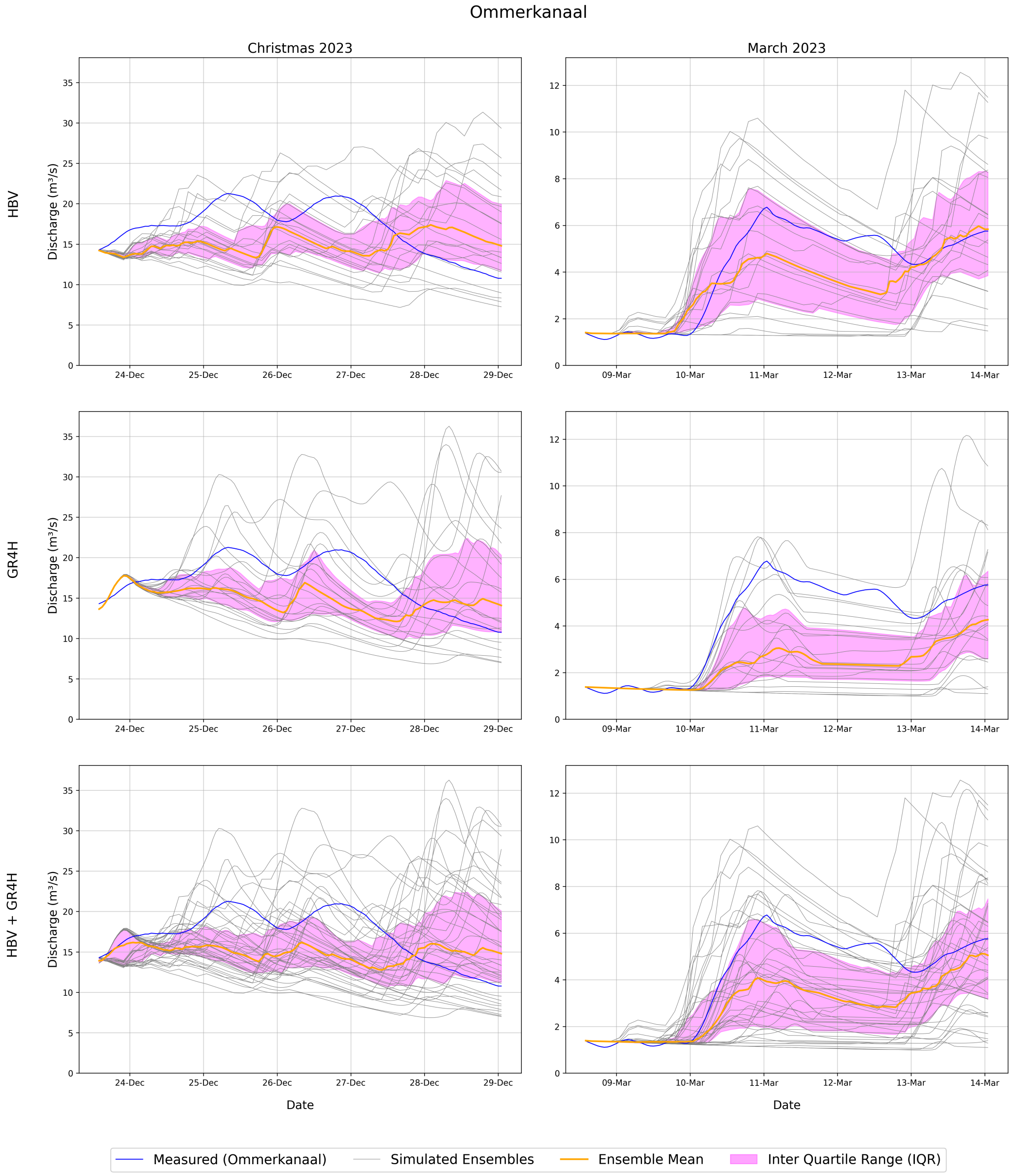


Figure 43: Spaghetti plots of the ensemble forecasts (grey), measured discharge (blue), ensemble median (orange) and inter quartile range (pink) for sub-catchment Ommerkanaal for both forecast periods for HBV, GR4H and the multi-model.

F.2 RMAE

Figure shows the RMAE values for the multi-model forecasts for both forecast periods. Compared to the individual model forecasts, the multi-model forecasts do not clearly reduce the RMAE. Extreme outliers of some individual forecasts are reduced, but the general maximum RMAE and spread remain more or less the same. For the downstream catchments the spread for short lead times (<24 hours) seems to decrease by using the multi-model forecasts, indicating more certainty in the forecasts. The multi-model does reduce the error of the worst performing model, but can also increase the error of the best performing model.

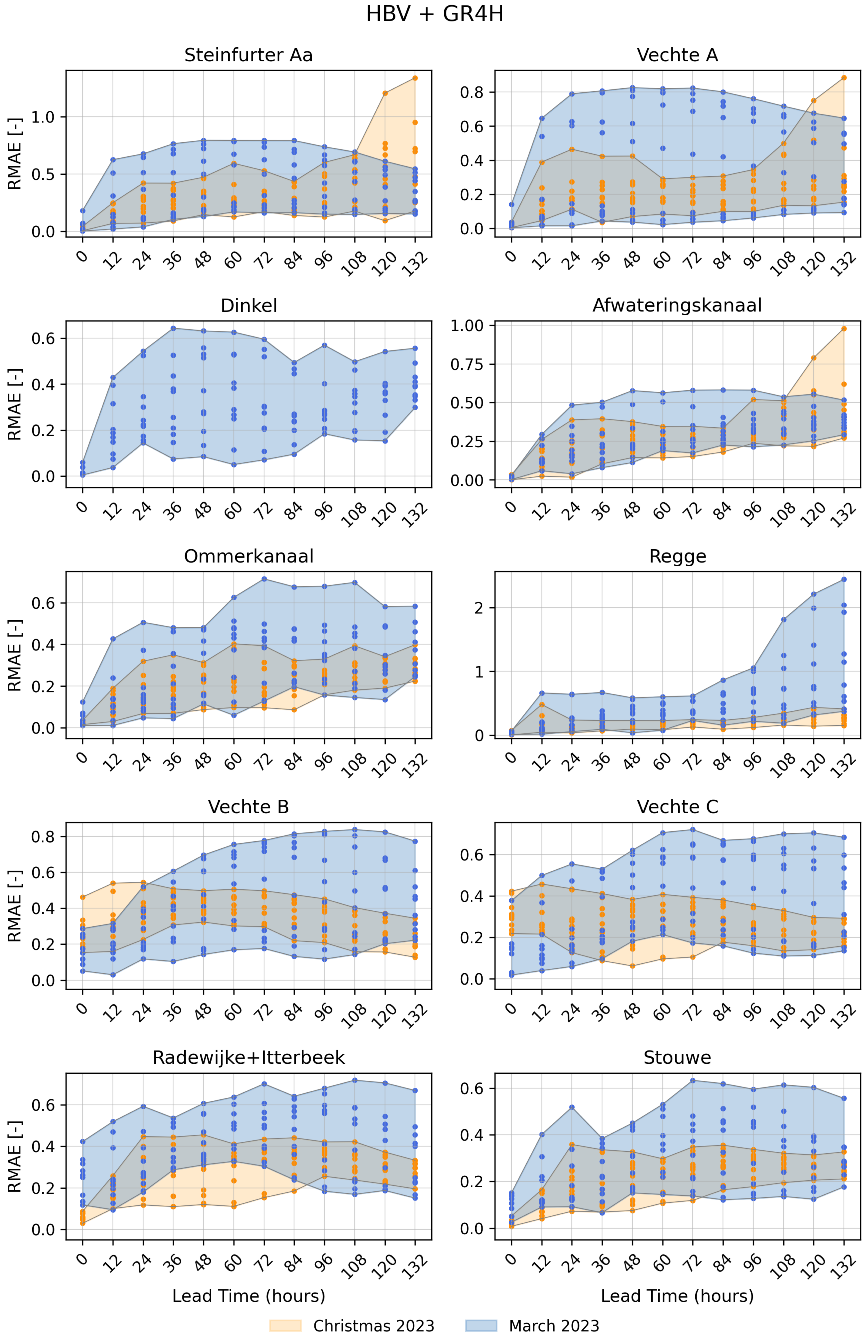


Figure 44: RMAE of the ensemble median across the 10 forecasts (10 values per lead time) for the multi-model forecasts for all sub-catchments. In blue the spread of the March event and in orange the spread of the Christmas event.