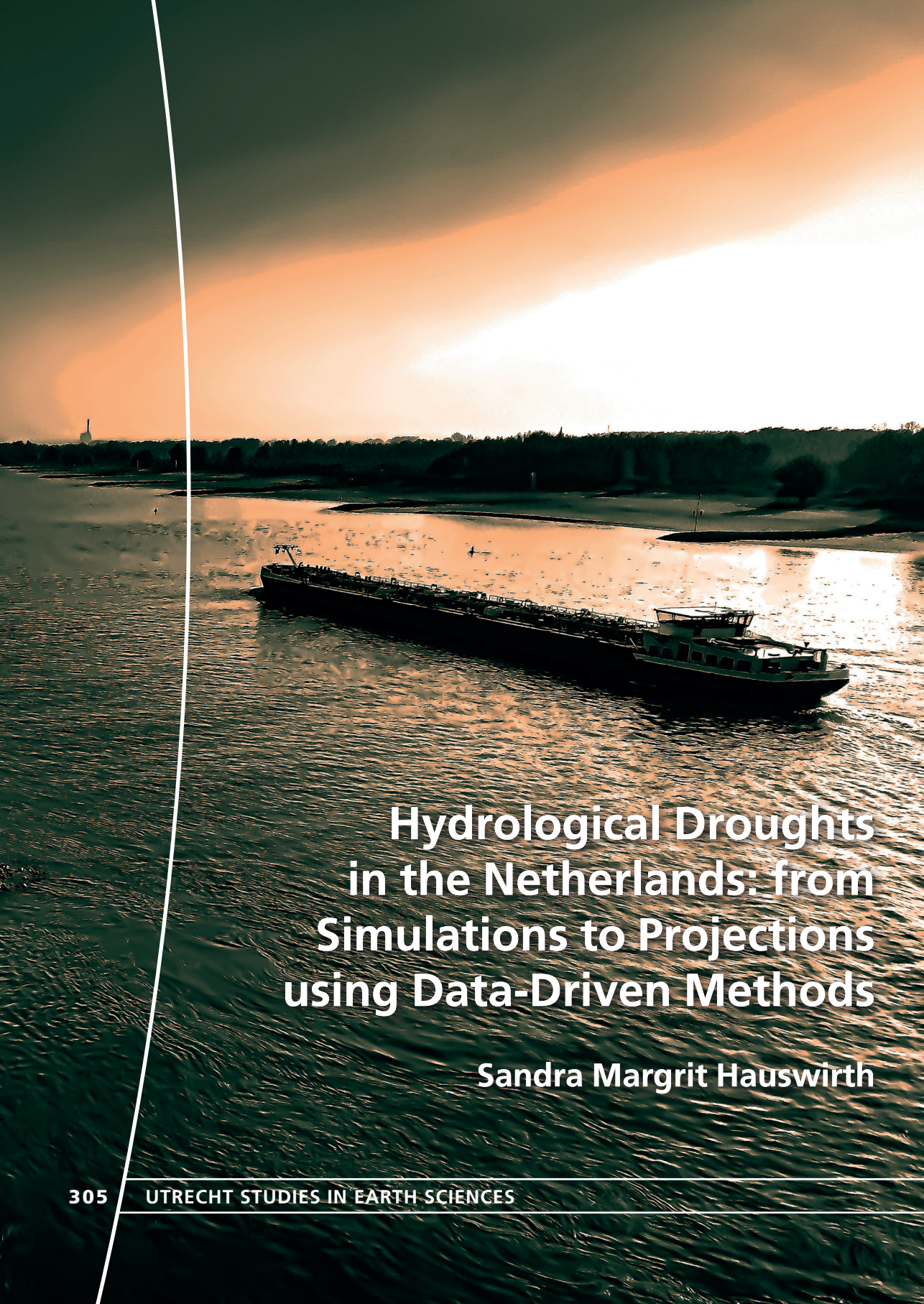


Hydrological Droughts in the Netherlands: from Simulations to Projections using Data-Driven Methods

Sandra Margrit Hauswirth

305 UTRECHT STUDIES IN EARTH SCIENCES

Hydrological Droughts in the Netherlands: from Simulations to Projections using Data-Driven Methods

Hydrologische droogte in Nederland: van simulaties naar projecties met behulp van data-gedreven methoden

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties in het openbaar te verdedigen
op 12 april 2024 des morgens te 10.15 uur

door

Sandra Margrit Hauswirth
geboren op 28 januari 1994 te Altdorf, Uri, Zwitzerland

Promotoren:

Dr. ir. Niko Wanders

Prof. dr. ir. Marc F. P. Bierkens

Beoordelingscommissie:

Dr. Gemma Coxon

Prof. dr. Derek Karssenberg

Prof. dr. ir. Remko Uijlenhoet

Prof. dr. Bettina Schaefli

Prof. dr. Louise Slater

Hydrological Droughts in the Netherlands: from Simulations to Projections using Data-Driven Methods

Promotoren:
Dr. ir. Niko Wanders

Prof. dr. ir. Marc F. P. Bierkens

Examination committee:
Dr. Gemma Coxon
University of Bristol, United Kingdom

Prof. dr. Derek Karssenberg
Universiteit Utrecht, the Netherlands

Prof. dr. ir. Remko Uijlenhoet
Technische Universiteit Delft, the Netherlands

Prof. dr. Bettina Schaefli
Universität Bern, Switzerland

Prof. dr. Louise Slater
Oxford University, United Kingdom

ISBN 978-90-6266-678-2

Published by Faculty of Geosciences, Universiteit Utrecht, The Netherlands, in:
Utrecht Studies in Earth Sciences (USES), ISSN 2211-4335

Typeset using XeLaTeX

Cover: The river Waal close to Nijmegen (2018), Sandra M. Hauswirth and Niko Wanders

Cover design: Margot Stoete

Photography chapter covers: Chapter 1-6: Sandra M. Hauswirth and Niko Wanders, Appendices: same photos edited with PIXLR Photo Editor including AI Image Generator and AI Design tools

Printed by Ipskamp Printing, Enschede, The Netherlands

Correspondence to: s.m.hauswirth@uu.nl

This work was financially supported by the Cooperate Innovation Program and the Department of Water, Transport and Environment at the Dutch National Water Authority, Rijkswaterstaat.

Utrecht Studies in Earth Sciences 305

Hydrological Droughts in the Netherlands: from Simulations to Projections using Data-Driven Methods

Sandra Margrit Hauswirth

Utrecht 2024

Faculty of Geosciences, Utrecht University

Contents

"Every accomplishment starts with the decision to try"

John F. Kennedy

Summary

Droughts in Europe have become more frequent and can be highly impactful due to their multidisciplinary character. Recent examples such as the droughts in 2003, 2015, 2018, and 2022, highlight this. Drought related losses throughout various sectors currently amount to EUR 9 billion annually for the EU and UK, however, drought losses in the future are expected to increase up to EUR 12.2 billion under climate change.

The past few extreme drought events showed that even countries that are usually less affected by droughts, due to their large natural water availability, experience increasing challenges related to them. One prime example includes the Netherlands, which is usually known for its abundance of water and advanced water management system. While this system was originally designed to prevent floods and flood impacts, the past extreme drought events showed that it was not optimally designed for water management during drought events. This made it clear that a shift in water management strategies and improving drought preparedness for future events are necessary. Especially, as unlike floods, droughts progress slowly and require long-term planning strategies for impacts and recovery. This requires early action and mitigation strategies to limit drought impacts throughout the country in short and long term drought adaptation strategies and policies.

The diverse drought impacts in the Netherlands affect shipping, agriculture, industry, energy production, drinking water, dike safety, saltwater intrusion, land subsidence, as well as nature and biodiversity. Water management actions usually follow a priority sequence during drought periods to limit impacts to the most vulnerable sectors or infrastructures. However, ideally most of the preventative measures would have been taken earlier on in the season when they are most effective, such as increasing water storage in the IJsselmeer lake. The increasing frequency and intensity of extreme events due to climate change highlights the need to better understand where the weaknesses are in the current system and what actions could be taken to mitigate drought impacts throughout the country. The key challenge is enhancing preparedness for future droughts through improved modelling, forecasts, climate change assessments, and through changes in water management strategies or adaptations that are needed.

In this thesis, which was a joint collaboration with the National Water Authority of the Netherlands (Rijkswaterstaat), the current challenges of water management regarding drought events in the Netherlands are addressed by exploring the potential of data-driven techniques for these aspects: simulating, forecasting and projecting of hydrological droughts and the effect of water management on potential drought impact mitigation. Central challenges to be addressed included having efficient and flexible models for various hydrological variables and availability of local hydrological information. Based on that, the ultimate goals of increasing the preparedness for upcoming droughts as well as the exploration of water management decisions to mitigate drought impacts could be addressed. All-together, this lead to a support modelling framework that increases information for informed decision-making for water management and could enhance the drought resilience of the Dutch hydrological system in the future.

The chapters throughout this thesis highlight the development and testing of the support modelling framework, to see whether data-driven techniques were indeed suitable for addressing current challenges water managers are facing:

In **Chapter 2**, the base of the modelling framework was developed and tested. Different machine learning techniques were tested in terms of their performance for simulating various hydrological variables for various locations in the Netherlands. All based on a simplified model and input data

setup with the intention of potential transfer to operational settings in later stages. Hydrological variables such as discharge, surface water levels and surface water temperatures, as well as groundwater levels were simulated and compared to local observations. The different machine learning techniques showed good overall performance for simulating this set of hydrological variables without large differences between the different methods. However, the delicate aspect of simulating extreme events with only limited observations and the setup with data driven models showed the complexity of simulating hydrological droughts. Slightly lower performance values were found for drought events, and adding additional basic water management information lead only to minor improvements. This is because the management information is inherently already in the observations that are used to train the machine learning techniques. Nevertheless, the power of the data-driven models tested in simulating various target variables based on a simple input dataset was promising and the modelling framework was put to another test in later chapters.

To assess the water management aspects in more detail, a separate focus was given to water management actions linked to potential drought impacts in **Chapter 3**. The modelling setup in this chapter was one of the first for the Netherlands, combining machine learning techniques, water management scenarios and drought impact functions to explore and simulate the potential of water management in mitigation of drought impacts. While the necessity for this information became increasingly clear through the past extreme events, the data availability for both water management as well as drought impacts were challenges that needed to be overcome. For water management scenarios, the historic operations from observations were taken into account. In terms of drought impacts, three main rivers were selected to evaluate potential reductions in drought impacts, based on documented thresholds from water management reports. While this preliminary experiment showed that there is potential in mitigating drought impacts through water management, there is still potential for further exploration, be it with more data on impacts and management but also regarding optimisation processes. Furthermore, the question remains whether water management alterations and adaptations will be enough to address the challenges that come with climate change as drought severity will often surpass the limits where mitigation strategies can still prevent damage.

To address the aspect of increasing preparedness for drought events, seasonal forecasting and climate change information was taken into account by changing the input data set from observations to seasonal (re)forecasting information (**Chapter 4**) and large climate ensembles (**Chapter 5**). The machine learning models, originally trained on historical observations, were used again, this time with input data from large-scale physically-based models, moving the modelling framework into a hybrid modelling format.

In terms of seasonal forecasting (**Chapter 4**), the modelling framework was tested in a hindcast setting, allowing for validation on historical forecasts. The hindcasts showed promising skill up to 1-2 months, with an increased skill during spring months when the initial hydrological conditions play a critical role. The data-driven models learned to distinguish general discharge signals, such as the spring snow-melt signal that is driven by upstream snow conditions. The hindcast experiment further showed that this hybrid format allowed to create seasonal hindcasts, with comparable skill to currently used large-scale forecasting models.

The incorporation of large climate ensembles into the modelling framework in **Chapter 5** gave the chance to project discharge under different warming levels at a more regional to local scale than usually available in climate change assessments. While the large climate ensembles enable to study extreme events in more detail due to the many realisations available compared to the ones in observational records, the latter also provides a challenge for the modelling framework as many of these extreme events were unseen before. To deal with this limitation, a separate post-processing step was introduced which takes advantage of the observational records of every location simulated separately and extrapolating the distribution tails of them. While this approach has its own

limitations, the simplicity regarding its implementation and utilization for local applications proofed to be beneficial as it essentially captured local characteristics, which could then again be observed in the local projections. The results from this chapter showed that droughts are likely to increase in the Netherlands with some distinct regional patterns. Furthermore, a shift is observed in the drought timing between the different regions in the Netherlands which likely also affects future drought impacts.

For both of these chapters, the benefit of the local modelling framework in a hybrid setting was found to be the additional feature that enabled the translation of large-scale input, observed or simulated data into localized simulation results. Basically a form of dynamic downscaling, however, different from conventional approaches and with a strong incorporation of the local knowledge and historic observational records. This method increases the availability of local information, which can be used for future decision making.

Throughout this thesis, various objectives and key challenges were addressed by combining practical insights from stakeholders with recent scientific developments, showing the prospect and strength of such collaborations. As a result this thesis highlights the potential of data-driven techniques and hybrid modelling approaches to improve hydrological modelling for drought monitoring, forecasting and projections and calls on water managers to advance the state-of-the-art of operational water management with the use of data-driven techniques to reduce future drought impacts.

Samenvatting

Droogte in Europa komt steeds vaker voor en kan door zijn langdurige en grootschalige karakter een grote impact hebben op de samenleving. Recente gebeurtenissen zoals de droogte van 2003, 2015, 2018 en 2022 zijn hier goede voorbeelden van. De economische schade door droogte in verschillende sectoren bedraagt momenteel 9 miljard euro per jaar voor de EU en het VK, maar deze schade zal in de toekomst naar verwachting toenemen tot 12,2 miljard euro als gevolg van de klimaatverandering.

De afgelopen extreme droogte hebben laten zien dat zelfs landen die gewoonlijk weinig last hebben van langdurige droogte, vanwege hun grote waterbeschikbaarheid, steeds meer met droogte te maken krijgen. Een goed voorbeeld hiervan is Nederland, een land dat bekend staat om zijn overvloed aan water en geavanceerde waterbeheer. Dit waterbeheer is met name ontworpen om schade door overstromingen te voorkomen en water te verdelen over het land. De afgelopen extreme droogte toonde aan dat dit ontwerp helaas niet optimaal is voor waterbeheer tijdens droogte. Met het oog op een toekomst met meer droogte is het duidelijk dat een verschuiving in het waterbeheer en een betere voorbereiding op droogte in de toekomst noodzakelijk zullen zijn. In tegenstelling tot overstromingen duurt een droogte weken tot maanden en zijn er lange-termijn strategieën nodig om de langdurige gevolgen te beperken. Dit vereist vroegtijdige actie en strategische maatregelen om de gevolgen van droogte in het hele land te beperken, zowel in het operationele beheer als het lange termijnbeleid.

De gevolgen van droogte in Nederland zijn te zien in de scheepvaart, landbouw, industrie, energieproductie, drinkwater, dijkveiligheid, zoutwaterindringing, bodemdaling, natuur en biodiversiteit. Maatregelen in het waterbeheer om droogteschade te beperken volgen tijdens droogteperioden gewoonlijk de verdringingsreeks. De verdringingsreek zorgt ervoor dat de meest kwetsbare sectoren het langst gespaard blijven. Idealiter zouden de meeste preventieve maatregelen echter zo vroeg mogelijk in het seizoen genomen worden, wanneer ze het meest effectief zijn. Denk hierbij bijvoorbeeld aan het vergroten van de wateropslag in het IJsselmeer.

De toenemende frequentie en intensiteit van extreme droogte als gevolg van klimaatverandering onderstreept de noodzaak om beter te begrijpen waar de zwakke punten in het huidige watersysteem zitten en welke maatregelen kunnen worden genomen om de gevolgen van droogte in het hele land te beperken. De belangrijkste uitdaging is om beter voorbereid te zijn op toekomstige droogte met behulp van verbeterde modellering, voorspellingen, klimaatprojecties en adaptieve waterbeheersmaatregelen.

In dit proefschrift, dat is ontstaan uit een samenwerking met Rijkswaterstaat, worden de uitdagingen voor het huidige waterbeheer tijdens droogte in Nederland onderzocht. Verder worden er data-gedreven technieken gebruikt om simulatie, voorspelling en projectie van hydrologische droogte en het mogelijke positieve effect van beheersmaatregelen te analyseren. De belangrijkste uitdagingen hierbij is om de modellen zo efficiënt en flexibel mogelijk te maken, zo dat ze meerdere hydrologische variabelen kunnen simuleren en getraind kunnen worden met lokale informatie. Na het overwinnen van deze uitdagingen werden de modellen getest om te kijken in hoeverre ze kunnen helpen om beter voorbereid te zijn op toekomstige droogte en de mogelijke positieve gevolgen van pro-actief waterbeheer te kwantificeren. Het eindresultaat was een modelsysteem dat extra informatie verschaft voor beter onderbouwde besluitvorming tijdens toekomstige droogtesituaties.

De hoofdstukken in dit proefschrift beschrijven de ontwikkeling en validatie van het modelsysteem voor beter onderbouwde waterbeheersmaatregelen, om te zien of data-gedreven technieken inderdaad geschikt zijn om de huidige uitdagingen waarmee waterbeheerders worden geconfronteerd te adresseren:

In **hoofdstuk 2** is de basis van het modelsysteem ontwikkeld en getest. Verschillende machine-learning modellen zijn getest op hun nauwkeurigheid bij het simuleren van diverse hydrologische variabelen voor verschillende locaties in Nederland. Alle technieken zijn gebaseerd op een simpele modelopzet en met minimale invoergegevens. Met het doel om dit later relatief makkelijk om te kunnen zetten naar een operationeel systeem. Hydrologische variabelen zoals rivierafvoer, waterdiepte, watertemperaturen, en grondwaterstanden zijn gesimuleerd en vergeleken met beschikbare metingen. De verschillende machine-learning modellen lieten over het algemeen goede resultaten zien voor deze selectie van hydrologische variabelen, waarbij er weinig verschillen zaten tussen de verschillende modellen. De nauwkeurigheid van de modellen ging achteruit voor de simulaties van hydrologisch extremen, zoals droogte, omdat er voor deze situaties relatief weinig observaties beschikbaar waren. Deze minder goede simulaties voor de droogte konden deels verbeterd worden met extra informatie over het historische waterbeheer. De toename in nauwkeurigheid was wel beperkt doordat een groot deel van informatie over het waterbeheer al indirect in de historische metingen zit die gebruikt werden om de modellen te trainen. Desalniettemin is de kracht van de geteste data-gedreven modellen bij het simuleren van verschillende hydrologische variabelen op basis van een eenvoudige model ontwerp veelbelovend en kan het toegepast worden in de overige hoofdstukken van dit werk.

Om de potentie van proactief waterbeheer beter te bestuderen, is in **hoofdstuk 3** aandacht besteed aan maatregelen die verband houden met het beperken van de gevolgen van droogte. De simulaties in dit hoofdstuk waren een van de eerste waarbij technieken voor machine learning, waterbeheersmaatregelen en droogteschadefuncties werden gecombineerd om te onderzoeken in hoeverre waterbeheer die droogteschade kan beperken. Hoewel de noodzaak van deze informatie steeds duidelijker wordt door de extreme gebeurtenissen vanuit het verleden, vormde de beschikbaarheid van gegevens van zowel het waterbeheer als droogteschade een uitdaging die moest worden opgelost. Voor de waterbeheersopties is er gekeken naar historische activiteiten op basis van metingen. Wat de gevolgen van droogte betreft, is er gekeken naar de drie grote rivieren en berekend hoe de gevolgen van droogte beperkt kunnen worden op basis van beschikbare informatie uit documenten van Rijkswaterstaat. Deze studie heeft aangetoond dat er mogelijkheden zijn om de gevolgen van droogte te verminderen door middel van adaptief waterbeheer. Ook is gebleken dat er nog ruimte is voor verbetering als er meer informatie beschikbaar komt over het historische beheer en de mogelijke droogteschade. Bovendien blijft het de vraag of wijzigingen en aanpassingen in het waterbeheer voldoende zullen zijn om de uitdagingen van klimaatverandering de baas te kunnen blijven. Dit omdat de ernst van de droogte steeds vaker grenzen zal overschrijden waarbij maatregelen nog schade kunnen voorkomen.

Om beter voorbereid te zijn op droogte is er ook gekeken naar de middellange en lange termijn simulaties met behulp van seizoensvoorspellingen en projecties voor toekomstige klimaatverandering. Hiervoor werden de historische observaties in de modelopzet vervangen door seizoensvoorspellingen (**hoofdstuk 4**) en projecties uit een groot klimaatensemble (**hoofdstuk 5**). De machine learning modellen, die oorspronkelijk waren getraind op historische waarnemingen, werden opnieuw gebruikt, deze keer met gegevens van grootschalige fysisch-gebaseerde hydrologische modellen, waardoor de modelopzet in een hybride model veranderde.

Op het gebied van seizoensvoorspellingen (**hoofdstuk 4**) is het model getest met behulp van historische seizoensvoorspellingen, die gevalideerd zijn met observaties. Deze historisch voorspellingen toonden dat de modellen droogte wel één tot twee maanden vooruit kunnen voorspellen. De kwaliteit van de voorspellingen was het beste in het voorjaar wanneer de initiële hydrologische omstandigheden een belangrijke rol spelen. De data-gedreven modellen leerden algemene afvoersignalen te onderscheiden, zoals de piek van de sneeuwsmelt die met name wordt bepaald door de hoeveelheid sneeuw bovenstrooms. Een andere conclusie was dat dit hybride model

het mogelijk maakte om seizoensvoorspellingen te maken, met een vergelijkbare kwaliteit als de operationeel gebruikte fysische voorspellingsmodellen.

De implementatie van grote klimaatensembles in de modelopzet in **hoofdstuk 5** gaf informatie over de veranderingen in rivierafvoeren onder verschillende klimaatscenarios op een meer regionale tot lokale schaal dan gebruikelijk. Hoewel de grote klimaatensembles het mogelijk maken om extreme gebeurtenissen gedetailleerder te bestuderen, door de vele realisaties die beschikbaar zijn in vergelijking met die in de waarnemingen, vormt dit ook een uitdaging omdat veel van deze extreme hydrologische gebeurtenissen nog niet eerder zijn waargenomen. Om hier mee om te gaan, is er een statistische nabewerking ontwikkeld die gebruik maakt van verdeling van extremen in de historische gegevens vervolgens is die informatie gebruikt om schattingen te maken van toekomstige hydrologische extremen. Hoewel deze aanpak beperkingen heeft, bleek de eenvoud van de implementatie en het gebruik voor lokale toepassingen toch voldoende voordelen te hebben. Mede omdat de lokale karakteristieken gebruikt kunnen worden voor het maken van lokale projecties die een toegevoegde waarde hebben voor het waterbeheer in de toekomst. Dit hoofdstuk laat zien dat droogte waarschijnlijk zal toenemen in Nederland, met een aantal duidelijke regionale patronen. Verder vinden er verschuivingen plaats in de piek van het droogteseizoen tussen de verschillende regio's in Nederland, wat waarschijnlijk ook gevolgen heeft voor de toekomstige impact van droogte.

Voor zowel hoofdstuk 4 als 5, is het grote voordeel dat het hybride model de mogelijkheid biedt om grootschalige gegevens en simulaties te vertalen naar lokalen analyses. In principe is dit een vorm van een dynamische verfijning van de gegevens, die echter wel verschilt van conventionele benaderingen en waarbij de lokale kennis en historische waarnemingen beter worden geïntegreerd. Daarmee vergroten deze methodes de beschikbaarheid van lokale informatie, die gebruikt kan worden voor toekomstig beleid en beheer van droogte.

In dit proefschrift zijn verschillende doelstellingen en belangrijke uitdagingen aangepakt door inzichten uit het werkveld te combineren met recente wetenschappelijke ontwikkelingen, wat het potentieel van een dergelijke samenwerkingen aantoont. Dit proefschrift laat de mogelijkheden zien dat data-gedreven en hybride modelbenaderingen voor droogtemonitoring, droogtevoorspellingen en droogteprojecties. Tegelijkertijd toont dit werk aan hoe waterbeheerders hun operationele waterbeheer kunnen verbeteren met de implementatie van deze data-gedreven technieken om de gevolgen van toekomstige droogte te verminderen.

1

Chapter 1 | Introduction

1.1 Context

Droughts are multifaceted hydrological extremes, which develop slowly, and affect large regions for prolonged periods of time. They can even have severe impacts in areas that are usually recognized for their abundance of water: such as the Netherlands, a country known for extensive water management and water availability - however, recent droughts have shown that these extreme events pose an increasing challenge. Especially, with the extreme drought events hitting Europe more frequently in the last decade (2003, 2015, 2018 and 2022), the Dutch hydrological system and water management ran against its limits, making it abundantly clear that the traditional water management targeting floods and flood impacts needs to be adapted to facilitate more proactive drought management. Especially challenging is the delicate balance between having the ability to quickly discharge but also store enough water in the system to be able to accommodate both hydrological extremes.

Compared to floods, the water management required for droughts calls for long term planning due to the slow progression of the drought process, drought impacts as well as the recovery from droughts. Drought impacts in the Netherlands are manifold, affecting many different sectors such as shipping, agriculture, industry, energy production and drinking water supply. Droughts also negatively impact dike safety, land subsidence, salt water intrusion, and ecosystems. The slow drought recovery makes that many of these impacts persist after the hydrological system has (partly) recovered, with increased vulnerability to subsequent drought events as a result.

As extreme events will likely increase in intensity as well as in occurrence due to climate change, it is important to understand a) how the hydrological system of the Netherlands will be affected by current and future extreme events, and b) which management decisions need to be taken to mitigate impacts as best as possible. Modelling frameworks that can map out future drought events and impacts provide valuable decision support tools for water managers in mitigating drought impacts. However, currently these modelling frameworks are often based on large-scale physically-based models which are complex and high in computational demand. Running them for seasonal forecasts or exploring a large space of possible water management decisions can therefore be slow and inefficient and the large-scale nature of these models makes their results too inaccurate for local applications.

In this thesis, which was a joint collaboration with the National Water Authority of the Netherlands (Rijkswaterstaat), the potential of data-driven techniques to simulate, predict and project hydrological droughts was explored. These techniques have been tested for their suitability as a support modelling framework to simulate and predict hydrological variables relevant to water managers, from historical, to subseasonal and projection timescales and bridging the gap between large-scale and local information in doing so. Developing such a modelling framework, lifts the constrains of previously used models by lowering computational demands, increasing flexibility and removing restrictions as these new models are purely based on data and therefore sets an interesting example for potential future water management. Especially in establishing and revisiting water management plans in light of extreme events.

1.2 Droughts

Droughts are multifaceted phenomena, occurring over a large time span from flash droughts to multi-year droughts and often over a great spatial extent. The general definition of droughts includes the lack of water, which can occur in different parts of the hydrological cycle, compared to the normal conditions that are commonly encountered in these components (Van Loon et al., 2016; Tallaksen and Lanen, 2004). Typically, distinctions are made between different types of droughts: meteorological drought (represented by a rainfall deficit), soil moisture drought (represented by soil moisture levels below the norm), and hydrological droughts that can be divided into below normal surface water and groundwater availability. Several years of research have been done to unravel and define climate-induced droughts but also human-induced droughts - the influence of human actions on the hydrological cycle and different types of droughts (Wanders and Wada, 2015a; Van Loon et al., 2016; AghaKouchak et al., 2021). The latter often occur in unnatural flow regimes due to dams, increased groundwater abstraction, and increased water demands, leading to unsustainable water use and larger recovery times for the hydrological system, with water management playing a critical role in all of them.

Besides human influence on droughts, climate change has been affecting drought occurrences and intensity over years and will only continue in the coming decades. Research has shown that extreme events such as floods and droughts will increase in their intensity, return periods and extent (Samaniego et al., 2018; Thober et al., 2018).

The multifaceted nature of droughts can be seen in the range of impacts that arise during and after the events, often accumulating slowly and persisting even when the drought is over (Wilhite, 2000). These impacts are often divided into direct and indirect, cascading impacts (Logar and van den Bergh, 2013). An example of direct impacts of meteorological droughts is the impact on vegetation and nature in general. Soil moisture droughts have negative impacts on crop production which is more easily measured in terms of loss of monetary value. Direct impacts of hydrological droughts are limitations to shipping, problems with supplying insufficient drinking water, and irrigation water. Indirect impacts are often a result of a combination of cascading drought impacts. For example precipitation deficit over weeks can lead to soil moisture droughts, both impacting vegetation and, for example, agriculture. This would results in direct impacts on yield except for irrigated crops that initially can relay on irrigation water from surface water or groundwater. However, if the drought propagates throughout the hydrological cycle and below-normal streamflow or groundwater levels occur, irrigation measures must be stopped or limited to not completely deplete the water resources. Yield losses occur, which also leads to an economic response with higher prices for consumers and farmers and potential food shortages in some regions of the world.

Depending on the type of drought, different variables and information sources are considered for drought identification and prediction. For meteorological droughts variables such as temperature, precipitation and atmospheric evaporation potential are taken into account. For soil moisture droughts, remote sensing observations of soil moisture, vegetation health and surface temperature are considered. Hydrological drought can be observed in groundwater, reservoir levels or streamflow records. In case observations are lacking which is often the case for soil moisture and hydrological droughts, hydrological models are used to fill this gap by simulating how precipitation affects the soil moisture, groundwater and discharge.

While observations are a great way to monitor the current state of the system, to identify historic or ongoing drought events, they lack the potential to make reliable outlooks of future drought conditions. For these forecasts, climate and hydrological models serve a purpose by predicting, among other variables, future precipitation, evaporation, soil moisture, groundwater and discharge. The predictive skill of these models comes from legacy effects that can be derived from the current state of the ocean, atmospheric and hydrological system. For example snow cover in winter can

provide information about potential snow melt in spring. Seasonal climate forecasting including weather patterns and climate conditions on a seasonal scale, such as El Nino and La Nina events, sea surface temperature and other climatic indicators, can give a longer outlook on potential drought developments in the atmosphere. This information then feeds into hydrological simulations with the aim to provide reliable and skillful forecast of future soil moisture and hydrological drought (Wanders et al., 2019).

Often drought indices are used to identify areas that are subject to drought and to forecast severity and duration of drought events. Examples of drought indices include the Standard Precipitation Index (SPI), Standardised Precipitation Evapotranspiration Index (SPEI), the Palmer Drought Severity Index (PDSI), Streamflow Drought Index (SDI), and Normalized Difference Vegetation Index (NDVI) to just name a few. These indices are basically a combination of meteorological, hydrological and vegetation data, depending on which drought type is assessed.

To be able to predict and prevent drought (impacts), not only observations, models and drought indices are needed but also a broader system understanding. Modelling and forecasting different parts of the hydrological system is needed to understand water availability and potential use, and link meteorological, agricultural and hydrological droughts to impacts at different spatial and temporal scales. Blauhut et al., 2022 showed the extent of drought impacts on the natural and human system seen in the previous years with extreme drought events in Europe, highlighting furthermore the need for unified drought governance approaches at larger scale than just regional or national.

1.2.1 Droughts and drought impacts at the European scale

Focusing on Europe, there have been several large scale extreme drought events in the past two decades: 2003, 2015, 2018, and 2022. For every event unique conditions in terms of severity, spatio-temporal extent, and associated direct and indirect impacts on human and natural resources were reached (Stahl et al., 2016; Blauhut et al., 2022; Rakovec et al., 2022; Tripathy and Mishra, 2023).

As an example, the extreme drought event of 2018, which was a compound event with both rainfall deficits and high temperatures (Rösner et al., 2019), hit large parts of Europe, even the more humid regions in the North (Bakke et al., 2020). Large impacts were recorded in north-central and northeastern Europe, especially for agriculture, farming and forestry (Bakke et al., 2020; Beillouin et al., 2020; Rösner et al., 2019; Schuldt et al., 2020; Blauhut et al., 2022). While drought-related losses in the EU and UK on average are about EUR 9 billion annually (for agricultural, energy and public water supply sectors, Cammalleri et al., 2020), drought losses in the future under global warming could increase up to EUR 12.2 billion annually at 2°C global warming levels (GWL, Naumann et al., 2021; European Commission. Joint Research Centre., 2020).

In terms of large-scale drought monitoring, the European Drought Observatory (EDO) is an example including different indicators for whole Europe. Also on the national level different drought monitoring systems exist that often use some form of standardized drought index. To name a few, in Germany droughts are monitored by the Helmholtz Centre for environmental research (UFZ) using the anomalies in the soil moisture saturation, in the Netherlands the Royal Netherlands Meteorological Institute (KNMI) monitors drought using a precipitation deficit as well as the SPEI and in Switzerland there is an information platform available from the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) including drought monitoring for different drought types.

Extreme events such as the droughts of 2018 and 2022 will likely increase in the future due to climate change. Especially snow melt in the alpine regions will likely have a big influence on spring and summer streamflow, which is an important component for many downstream regions to improve their drought preparedness and supplement irrigation in summer (van Tiel et al., 2023; Stahl et al., 2022). Furthermore, human water use and its effect will also change in the future due to increased

water demands for agriculture and industry. Countries that are situated in downstream regions, such as the Netherlands, will likely feel the strongest combined climate and human impacts and have to adapt their water management for these extreme events.

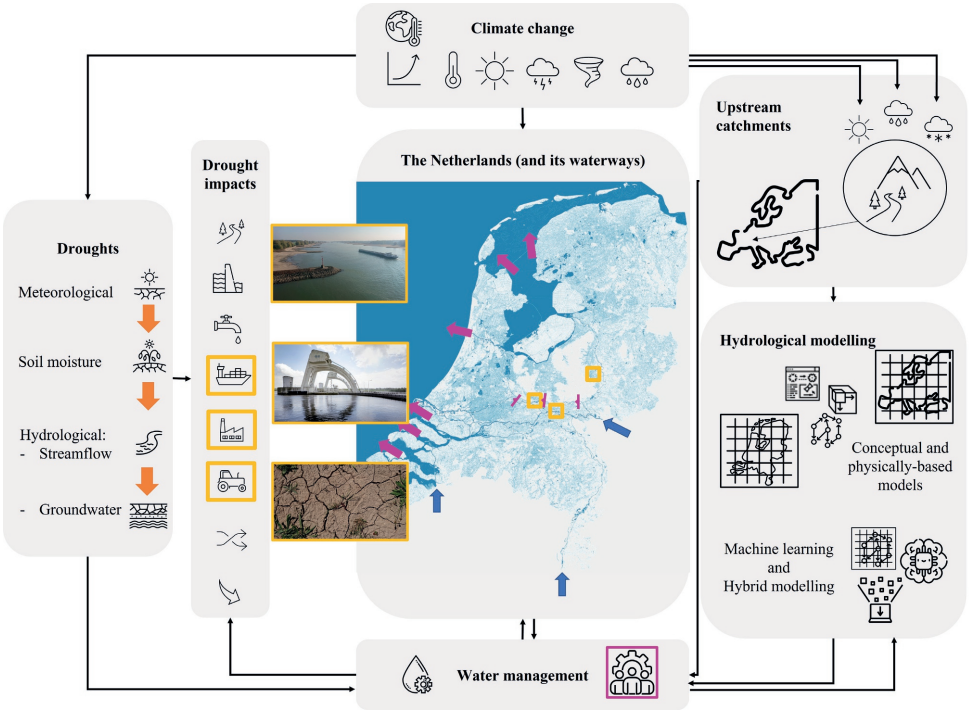


Figure 1.1 Overview of topics, methods and their connections touched upon throughout the introduction. Various aspects and connections will be addressed and explored in a new way with this thesis: ranging from simulating, forecasting and projecting of hydrological droughts using data-driven techniques and hybrid approaches to explore water management and the potential to mitigate a selection of drought impacts in the Netherlands.

1.2.2 The Netherlands and droughts

The Netherlands, usually known as a country with an abundance of water and a history of water management focused on flood protection, has been challenged increasingly by droughts. One of the most extreme droughts on record goes back to 1976, however in recent years an increase drought frequency has been observed. Already in the last 20 years summers like 2003, 2015, 2022 and especially 2018, with it's record breaking intensity and scale throughout whole Europe, challenged the Netherlands and its water systems and management immensely.

The 2018 drought for example led to significant impacts throughout various sectors such as agriculture (due to irrigation limitations), shipping, drinking water, while extra attention had to be given to dike safety, salt water intrusion, affecting drinking water provision and increased land subsidence rates in the peat soils of the Netherlands. The list of impacts in this case also highlights the complexity of the drought challenge in the Netherlands: while an initial precipitation deficit over a long time period led to a soil moisture drought impacting agriculture and the irrigation water needed to combat that situation was not available without causing other impacts, as during that time span the surface water levels were too low and the management regulations very strict. However, this lead to

intensified groundwater use, which was also restricted later in summer due to rapidly declining groundwater levels. Groundwater in the Netherlands is partly used for drinking water, irrigation and also plays an integral part in many industries. Extensive groundwater pumping can not only lead to groundwater depletion, which can take several years for the groundwater levels to recover to normal, but also to salt water intrusion in dune areas which are used for drinking water and to land subsidence in certain areas of the country, which in turn can lead to infrastructure damages. Low surface water levels were not only limiting irrigation water use but also impacted shipping in the larger rivers, such that less goods could be transported. Furthermore, reduced river flows also resulted in increased salinity, following the river network upstream as result of tidal incursions not being pushed by river discharge. Increased salinity values and in general lower water quality due to less dilution can also impact ecosystems and limit drinking water uptake. As a results of all these factors, drought conditions and the resulting increased water demand required strict and intense water management decisions.

The succession of extreme droughts in the Netherlands had led the water authorities to realize that a shift is needed from the current water management focused on flood protection to one that also accommodates droughts. This was deemed especially necessary because climate scenarios for the Netherlands project that climate change will lead to shifts in precipitation (decrease in summer, increase in winter), increased rainfall variability and more extreme events (KNMI, 2021; KNMI, 2023; Masson-Delmotte et al., 2021; Cook et al., 2020; Bartholomeus et al., 2023).

Current water management practices during drought periods follow the line of the 'Verdringingsreeks', a priority sequence which lists the priorities and importance regarding management decisions for different sectors and key points to limit drought impacts. The shift in water management would include measures to increase storage of soil, surface and groundwater during the wetter winter periods for use during the subsequent drier periods in the summer. However, long term planning with options of increasing water storage early on in the year can be challenging as there needs to remain a safety margin for preventative flood management. Furthermore, seasonal or short term forecasts would be needed to pro-actively manage this water storage, and they are often only available on larger spatial scales. In addition, compared to other countries, the Dutch hydrological system is comprised of an intricate, heavily managed river network that is optimised for fast drainage of excess water and therefore limits the ability to store large quantities of water.

Key to projecting the impacts of future droughts and investigating structural and operational measures to mitigate drought impacts in the Netherlands is the National Water Model (NWM). This model is one of the most complete physically-based integrated hydrology and water resources models for the Netherlands, covering not only surface but also groundwater interactions at 250m scale (De Lange et al., 2014). While this model captures most aspects of the Dutch water system, the computational demands make it challenging to use in operational settings, that require adaptive and optimized management scenarios, and monthly seasonal forecasts, except for selective case studies.

The previous drought events stress the need for early warning about the imminent occurrence of droughts and the subsequent evaluation of drought management and mitigation options. In this context, the trade off in terms of having too much water earlier in the year which would also increase flood risk and higher groundwater levels, which would affect agriculture, is still a challenge. To be better prepared, new modelling systems that are fast, flexible and computationally efficient could support and help to optimize water management decisions not only in the short but also long-term planning and could be used for drought and drought impact forecasting.

1.3 Modelling

1.3.1 Hydrological modelling

Hydrological modelling has undergone substantial development, progress and refinement over the past decades, encompassing a spectrum of different model types and approaches. With the rise of computational power, increased processes understanding and the wealth of new observations coming from for example remote sensing, the hydrological community has been able to develop new modelling concepts which are better constrained by observations. Three major modelling concepts can be distinguished that have been applied by scientist around the world.

Conceptual models use simplified representations of physical processes (e.g. empirical relationships, Nash and Sutcliffe, 1970), while physically-based models include these processes through physical laws that are captured by differential equations or numerical solutions (Freeze and Harlan, 1969). Furthermore, models can also be split in terms of complexity, be it process complexity or spatial complexity (Clark et al., 2017). In more recent times, data-driven approaches have gained prominence in hydrological modelling. While these models in general do not rely on physical laws, they do take advantage of observational data to estimate relationships between different variables. The selection of a particular model concepts depends on the objectives, constraints (including temporal and spatial considerations, data availability, and computational resources), and the desired level of complexity. All these factors determine which modelling concept is most suited for hydrological modelling.

Physically-based models, be it small or large scale, grounded in fundamental hydrological principles, typically necessitate extensive data for parameterization. Despite this requirement, they offer a comprehensive depiction of the hydrological system and its intricate interactions. These models can be configured at various spatial extents, ranging from global to regional and local. Furthermore, a distinctive advantage of physically-based hydrological models lies in their capacity to simulate entire hydrological systems, encompassing water balance considerations and water management aspects (Bierkens, 2015). Not only do they facilitate the simulation of historical, present, and future hydrological variables, spanning near-future, sub-seasonal, and seasonal time frames, as well as projections contingent upon diverse external forcings but they can be instrumental for intervention and scenario analysis, enabling the exploration of diverse forcings, such as Shared Socioeconomic Pathways (SSPs, O' Neill et al., 2014) and Representative Concentration Pathways (RCPs, van Vuuren et al., 2011) that represent distinct socioeconomic and climate change scenarios.

However, challenges remain regarding tradeoffs between process and spatial complexity, domain and ensemble size as well as simulation periods as listed by Clark et al., 2017. Despite the ongoing advances in computational capacity, computational demands remain high, as developments of increased model resolution and process complexity keep pace. These aspects are particularly relevant when employing physically-based models in large-scale applications and for operational simulations.

1.3.2 Machine learning in hydrology

Data-driven techniques, such as machine learning, have been a fast growing field in hydrology over the past years with major scientific developments (Shen, 2018; Shen et al., 2018). Especially, the large amount of available data, through decades of observation and remote sensing, made the use of machine learning more interesting as these techniques provide more opportunities to explore the full range of observational data, discover new insights and potentially increase prediction capabilities (Xu and Liang, 2021). While physically-based models are constrained by "known" physical relations and equations, machine learning techniques apply a purely data-driven approach which allows large amounts of data to be handled and autonomously learn the relationships between different observed variables. Therefore, relationships between different variables that might not have been detected in empirical studies before

can be discovered and implemented. This data-derived information can be used to inform scientists on potentially missing relationships that then can be incorporated in physically-based models.

List of studies in this field are long and cover many different interests and almost all parts of the hydrological cycle (streamflow, groundwater and other subsurface processes, precipitation, evaporation, and soil moisture as examples Shen et al., 2021) and cover various spatial extents (global, regional, local). But especially in the past few years, rainfall runoff studies based on recurrent neural networks have brought the power and potential of machine learning in hydrology to a broader audience as they can include long-term dependencies, compared to more traditional or conventional machine learning techniques which do not necessarily have a mechanisms that represents different processes and their temporal evolution (Xu and Liang, 2021).

While machine learning models for hydrological problems such as rainfall-runoff modelling can be found already in the 1990s (e.g. Hsu et al., 1995), recurrent neural networks, such as Long Short Term Memory Models (LSTMs) in particular, have been increasingly used for rainfall-runoff modelling and streamflow forecasting in the past years (Kratzert et al., 2018; Li et al., 2021; Feng et al., 2020). Most recognizable are the studies where LSTMs are not only used for daily streamflow simulations using meteorological forcings in comparison to commonly used hydrological models (Kratzert et al., 2018), but also the exploration of LSTMs in terms of regionalization (Kratzert et al., 2019b), and predictions in ungauged basins (Kratzert et al., 2019a; Ma et al., 2021; Feng et al., 2021). While this is only a selection, many more machine learning studies have shown since the potential benefits of recurrent neural network but also other machine learning techniques (Xu and Liang, 2021).

Although machine learning models benefit from reduced computational time and computational demands, they also suffer from limitations related to data handling. As machine learning models learn from data, large amounts are needed for training, testing and independent validation purposes. Therefore, machine learning models are often regarded only as good as the data they can use for training purposes and otherwise will have limitations regarding extrapolation to new data. Training with limited data and samples being one of the challenges listed by Xu and Liang, 2021, others also include the development of interpretable models. Especially, if only data is used and little to no process information, machine learning models can often feel like black boxes with little information on how results are derived, making it a challenge to use them and understand the outcomes. Apart from that, machine learning models often do not necessarily preserve physical principles such as mass, momentum and energy. By ensuring physical consistency, scientists can have a better characterization and potential propagation of uncertainty, as well as allow for predictions under non-stationary conditions. This will be feasible when physical consistency is incorporated or by combining the advantages of machine learning with those of physically-based models, moving into a more hybrid setup.

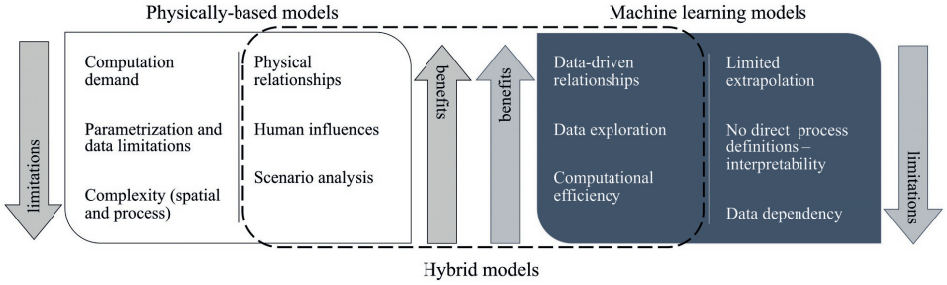


Figure 1.2 Brief overview of main benefits and limitation of the physical-based and machine learning modelling approaches and the interface where hybrid modelling combines the best of both.

1.3.3 Hybrid modelling in hydrology

More recently data-driven techniques have also increasingly been used in combination with physically-based models to support and overcome past and current challenges and restrictions for both of these model approaches, however still many opportunities for this interdisciplinary setup are open (Shen et al., 2021).

Currently, a combination of these approaches are used for replacing parts of a physically-based model by a machine learning component. For example replacing slow modelling components with machine learning counter parts to reduce the computational demand. Machine learning can also be used to create so called surrogate models for optimization purposes (Sun et al., 2023; Tsai et al., 2021). Surrogate models are machine learning based replacements of physical based models and can help to create a large number of simulations with significantly reduced computational time to better understand model response to different parameter settings. To obtain a surrogate model, the machine learning model is trained on simulation data from the physically-based models to mimic the response of the physically based model to changes in input parameters. An added benefit of these models is that their computational demand is significantly lower than the original physically-based model and they have additional constrains when it comes to the conservation of mass, momentum and energy (Tartakovsky et al., 2020; Zhu et al., 2019; He et al., 2020). Furthermore, machine learning models can also be used for error correction of physically-based models (Shen et al., 2022; Magni et al., 2023) or also be part of process discoveries (Tsai et al., 2020). Under these conditions the machine learning model is used to understand the relationship between errors in the simulations from the physically-based models and observations. These machine learning model can then be used as a post-processing application to reduce uncertainties in simulated hydrological variables. In terms of machine learning model development and training, augmenting observations with physically-deterministic model output can also further improve machine learning model simulations. Overall, the benefit of combining both modelling approaches, machine learning and physically-based, allows to combine their complementary strength in many ways (Figure 1.2) and reduces some of the limitations of both individual approaches (Xu and Liang, 2021).

As an example of a field where we see significant developments in terms of hybrid modelling: hybrid forecasting has seen a growing interest due to advances in the meteorological and climate predictions systems, the convincing performance of machine learning and computational resources. While the development of hybrid forecasting has gone fast, the term hybrid modelling can remain vague. Slater et al., 2023 made one of the first attempts to define hybrid forecasting by dividing the modelling process into three groups including statistical-dynamic, serial and coupled approaches between hydroclimate and data-driven models. These hybrid methods still present the strengths of machine learning models, such as improved speed and higher performances. Furthermore, they are able to easily integrate and combine large datasets of multiple predictability sources. However, new challenges such as combining and merging different models, assimilation of human influence, such as water management aspect, or climate change effects and representation of hydrological extreme events remains an ongoing challenge (Slater et al., 2023; Shen et al., 2021).

While machine learning has been used for both large and small scale studies, a comprehensive hybrid modelling framework not only for basic hydrological simulations but also forecasting, climate change assessments or simulating water management influence has not been explored yet, especially not in the case of the Netherlands. Such a framework could show potential and close the gap that would come with more traditional approaches that might also be used in current operational settings, compared to the scientific field.

1.4 Research question and aim

As highlighted previously, the Netherlands has been confronted increasingly with challenges arising from extreme events, particularly in all aspect related and connected to its hydrological system. Historically, Dutch water management primarily focused on flood risk prevention and mitigation. However, there has been an increasing recognition of the need for a systemic shift of water management (infrastructure and practice) to also address the impacts of droughts.

Droughts, in contrast to floods, unfold gradually, demanding and allowing proactive management measures taken weeks to months in advance to mitigate drought impacts as effectively as possible. These impacts are multifaceted and intricately linked to human-induced activities. The extensively managed Dutch hydrological system relies on numerous surface water networks, be it larger streams to smaller ditches, dikes, dams, sluices, and other infrastructures. Therefore, effective management of these infrastructures is critical to alleviate and limit multisectoral impacts, such as for shipping, agriculture, drinking water supply, dike safety, recreation, nature and more. Not only now but also long-term under climate change.

Meeting these challenges requires the availability of models of the Dutch water system that are accurate enough to inform water managers from national to local scales, and that are computationally efficient, fast enough to explore many possible scenarios and solutions and allow for real-time prediction and optimization. The overarching challenge then becomes the development of modelling tools that meet these requirements: accurate, fast and accurate at multiple spatial and temporal scales.

In this thesis, which was done in collaboration with the National Water Authority (Rijkswaterstaat), this overarching challenge was addressed and the following key components were defined and explored in various sub-projects conducted throughout the PhD trajectory:

- **Efficient and Flexible Models for Various Hydrological Variables (Chapter 2)**
  To effectively assess various management options, perform scenario analysis, create seasonal forecasts, and conduct climate change assessments with extensive datasets, the necessity of an **efficient** and **adaptable modelling framework** is evident. Traditional practices often rely on large-scale physically-based models, which can pose computational, handling, and data storage challenges. In recent years, there has been growing interest in **machine learning models** due to their demonstrated promise across diverse fields, extending beyond hydrology. These models offer the potential to address limitations inherent in current practices. However, it is essential to empirically evaluate their reliability in reproducing hydrological information and their applicability in the context of drought-related water management.

- **Water Management Decisions to Mitigate Impacts (Chapter 3)**
  To effectively **mitigate drought impacts**, it is essential to assess the consequences that may arise from various **water management decisions** and the legacy effect of these decisions for future water availability and drought impacts. Understanding how to minimize these impacts by **optimizing** water management actions in both the short-term and sub-seasonal timeframes is key for enhancing preparedness for future drought events and **reducing** the resulting **impacts**.

- **Availability of Local Information, from Large to Local Scale (Chapter 4 and 5)**
  Information regarding meteorology and hydrology is often disseminated through large-scale models, which aid water managers in their duties. However, decisions and their consequences mostly impact **local scales**, notably in the Netherlands with its extensively managed hydrological system. Therefore, the development of a modelling framework **capable of integrating both large-scale and local information** is essential for informed decision-making at local scale.

- **Preparation for Upcoming Droughts (Chapter 4 and 5)**
  A central concern in present-day water management revolves around improving preparedness for upcoming drought events. This requires the ability to **forecast** and **project** hydrological conditions, not only in the **near future** but also **longterm** under **climate change**. These aspects are critical for simulating management decisions and optimizing strategies to mitigate drought impacts across various sectors. The key question is whether emerging technologies such as **machine learning, hybrid modelling, and large climate ensemble data** can be paired into a **unique, unified framework** to address these challenges effectively and provide relevant information at the **local scale**.

By combining practical insights from stakeholders with scientific advancements, this thesis addresses various objectives (Figure 1.1): developing an efficient modelling framework, bridging the gap between large-scale information and local-scale applications, enhancing seasonal forecasting potential, conducting climate change assessments, and testing water management scenarios for potential drought impact mitigation.

1.5 Thesis outline

In essence, the thesis explores the utilization of machine learning techniques in the domain of water management in order to arrive at fast and accurate model approaches. It combines innovative approaches such as machine learning, hybrid modelling, seasonal forecasting, and climate change assessments, bridging the gap between large-scale data and local-scale applications and the interface between research and application.

The thesis aims to address the outlined challenges through a series of chapters:

- Chapter 2: Development and validation of a local machine learning modelling framework based on historical observations to simulate various hydrological variables at the local scale
- Chapter 3: Exploration of a local-focused experiment for potential drought impact mitigation, centered on water management scenarios
- Chapter 4: Evaluation of the modelling framework's suitability in a hybrid hindcast setting to assess its potential for hybrid seasonal forecasting
- Chapter 5: Integration of the machine learning framework with large climate ensembles to enable local-scale climate change assessments and the evaluation of hydrological extreme events

Chapter 6 will include the synthesis of findings, including broader discussions and future research prospects.

Chapter 2 | The potential of data driven approaches for quantifying hydrological extremes

Abstract

Recent droughts in Europe have shown that national water systems are facing increasing challenges when dealing with drought impacts. Especially the Netherlands has seen an increasing need to adapt their water management to improve preparedness for future drought events. Ideally, the necessary information needed for operational water management decisions should be readily available ahead of time and/or computed flexibly and efficiently to ensure sufficient time to evaluate the various management actions. In this study, we show that in addition to physically based hydrological models, the upcoming and promising trend of incorporating machine learning (ML) in hydrology can provide the basis for future efforts in supporting national operational water management by providing the needed information efficiently and with the required accuracy. As a precursor for their use in a forecasting system, we assessed the ability of five different data driven methods to simulate hydrological variables at a national-scale. We developed a unified workflow where we use limited information on hydro-meteorological variables and general water management policies to simulate historic timeseries of discharge, groundwater levels, surface water levels and surface water temperatures. We find that all ML methods, ranging from very simple to more complex ones, showed a generally good performance for stations and target levels which are closely linked to the input data and location (e.g. stations along main river network). For downstream stations and small rivers, the Random Forest method outperforms the other methods both for discharge and surface water levels. For surface water temperature no location dependency was observed and for groundwater levels, all methods were performing comparable with most stations ranging in nRMSE 0.2-0.3. Generally, the best performances were reached by the more advanced Random Forest and LSTM methods, which was also seen when simulating high and low flow events. High flow events were slightly better captured than low flow events but overall simulating extreme events based on a simple input data set remains challenging. Specific training sets, including event related information and additional input variables, could improve future assessments. Including the feature importance of the methods allowed us to detect how and where water management influence played an important role. The addition of information on water management in the ML routines increases overall performance, although limited. We conclude that ML and other data driven approaches have potential in predicting different hydrological variables. We were able to capture and incorporate water management aspects in our analysis, creating a base for future experiments where scenario analysis might reveal ML based mitigation strategies. The combination of limited input data requirement and short computation times makes this new framework suitable for forecasting purposes.

Published as: Hauswirth, S. M., Bierkens, M. F., Beijk, V., and Wanders, N. (2021). The potential of data driven approaches for quantifying hydrological extremes. *Advances in Water Resources, 155, 104-017*, https://doi.org/10.1016/j.advwatres.2021.104017

2.1 Introduction

Droughts are prolonged periods with below normal water availability and can have a severe impact on nature and society (Wanders and Wada, 2015a; Van Loon et al., 2016). Over the last two decades, many European countries experienced severe droughts, with notable examples in the years 2003, 2005, 2010 and 2015 (Stahl et al., 2016; Van Lanen et al., 2016). More recently, the consecutive years 2018 and 2019 with record breaking droughts and heatwaves posed a big challenge for many Western European countries (Bakke et al., 2020). These consecutive drought years have had a severe impact on agriculture, water management, shipping, ecosystems and other water dependent sectors. Consecutive years with severe and long-lasting drought events do not only pose a challenge during the event, but also lead to difficulties for nature to recover from these extreme situations, which makes it important to not underestimate them (Ruiter et al., 2020; Hari et al., 2020).

All these droughts were caused by a lack of precipitation in combination with an increased evaporative demand that propagated through the hydrological cycle, resulting in deficits in soil moisture, surface water and groundwater (Herrera-Estrada and Diffenbaugh, 2020; Ionita et al., 2017). Both drought periods in 2018 and 2019 further showed how drought can propagate throughout the hydrological system, not only in time but also spatially. Studies have shown that both the frequency and severity of drought events are expected to increase under climate change (Hari et al., 2020; Hattermann et al., 2018; Marx et al., 2018; Philip et al., 2020; Samaniego et al., 2018; van der Wiel et al., 2019b). This further indicates how important it is to advance monitoring and prediction of drought periods and their associated drought impacts (Sutanto et al., 2019).

Despite recent efforts to improve the monitoring and forecasting of drought conditions (e.g. Wanders et al., 2019), the recent drought events have shown that the challenges for water management planning and water distribution remain high (Witte et al., 2020). In order to mitigate upcoming drought impacts and water shortages as much as possible, water management decisions that rely on upstream water storage, on groundwater or reservoirs, need proper lead times to be effective. Especially for the larger rivers systems, this would require forecasts of droughts and drought impacts with lead times up to several months or longer (Samaniego et al., 2019).

Ideally, the information necessary to make adequate decisions should be readily available when decisions need to be made. To fulfill this goal, the hydrological and water management models that support decision-making should be computationally efficient, reliable, adaptive and able to ingest the latest observations available. This would allow a scenario analysis to find the optimal near-future management strategies to be part of day-to-day operational water management, e.g. through dynamic programming (Bertsekas, 2011) or model predictive control (Camacho and Bordons, 1999; Aydin et al., 2019). This can be challenging as national scale hydrological models are often too time and computationally demanding to be used for large scale ensemble simulations to support scenario analysis with uncertainty estimations. To overcome the obstacle of computational efforts in real-time forecasting and scenario analysis, data-based modelling methods such collectively called Machine Learning (ML) methods (e.g. Alpaydın, 2020) have proven to be a valid alternative in providing hydrological simulations as accurate as hydrological models at a fraction of the computational demand (Kratzert et al., 2018).

Machine learning in hydrological analysis is interesting as it allows to handle large amounts of observational data, to analyse and detect relationships between them and to predict or recreate the outcome based on the established relationships in an efficient manner. Thus, machine learning might not only be helpful in finding answers to hydrological research questions, but also provide additional advantages for operational water management. Studies have shown that various ML algorithms, ranging from simple regression to more advance neural networks (Deep Learning), can be effectively used for rainfall-runoff modelling (Kratzert et al., 2018) or for assessing groundwater levels (Koch et al., 2019) while using similar input data as common hydrological models. Other studies indicate

the potential in capturing and forecasting different drought indicators such as the Standardized Precipitation Index and Standardized Precipitation Evaporation Index amongst others (Sutanto et al., 2019).

Even though a large number of studies have shown the usefulness of machine learning for predicting individual hydrological variables, such as time series of catchment-scale discharge (Kratzert et al., 2018) or basin-scale terrestrial water storage (Sun et al., 2020), to our knowledge, the suitability of different machine learning methods for predicting several interlinked hydrological variables on national-scale based on a simple input dataset has not been tried. Furthermore, in the context of using forecasting for operational water management, the impacts of water management measures on hydrological states needs to be included, which has only been done in a limited number of studies (e.g. Majumdar et al., 2020; Zhang et al., 2018). Being able to predict hydrological parameters with and without management strategies included, creates new opportunities to not only analyse the hydrological situation in the present and near future but to also include scenario analysis, which can be more efficient and flexible than large hydrological models used so far.

Before ML methods are used in a forecasting system, it is necessary to first test their ability to reproduce historical observations. Ideally, these ML-based predictions should rely on only a few input variables that can also be obtained as forecasts from current forecasting systems. Once the accuracy in reproducing historical observations is known, it is then possible to attribute lack of forecasting skill in to model errors and forecasting errors of the input variables. The aim of this study therefore is to evaluate the ability of ML methods in reproducing a large set of inter-related time series of multiple hydrological variables for a human influenced national-scale hydrological system. To do this we use a small set of meteorological and hydrological input variables (precipitation, evaporation, upstream discharge and sea level) that are readily available from current forecasting systems. We will test the accuracy of different ML algorithms for predicting multiple hydrological variables (discharge, groundwater levels, as well as surface water levels and surface water temperatures) at multiple locations, that are usually computed by a (large-scale) hydrological model. Additionally, we will incorporate water management variables as input to our analysis to see whether ML can indeed be suitable for operational water management, especially during periods of drought. Our goal is to contribute to the current knowledge in the field and set the basis for the potential use of this approach for operational long-range forecasting and water management, including the selection of optimal water management options.

In the following Section 2.2, we will describe the study area and the data, introduce the ML methods used as well as the experiment setup to test the predictive capabilities of these methods. In Section 2.3, the accuracy of the different ML based simulations will be evaluated for the different hydrological variables, followed by more in-depth assessments of the performance of specific algorithms, both for average, high and low flow periods. The results will be followed by a discussion (Section 2.4) on the findings and placing them into context of other current research efforts in the field. The paper will be closed by a conclusion (Section 2.5) and outlook on possible following studies.

2.2 Material and Methods

The following sections present information on the study, on the general experiment setup, and on the data sets and -sources used, as we provide and give an introduction of the selected ML algorithms and describe the final pipeline of our experiment.

2.2.1 Study area and experiment setup

The Netherlands, located in Western Europe, was used as a study area. On the one hand, the location of the Netherlands directly at the sea, its topographic characteristics (e.g., 26% of the area lying below

sea level and over 50% sensitive to flooding) and the elaborate and complex water management system create an interesting case with unique hydrological challenges not only during wet but also dry periods. On the other hand, the last few summers have shown that water management, which is historically known for reducing flood impacts, has to adapt and further specialise on buffering and storing water prior to upcoming drought periods (Witte et al., 2020). Approaching this as a case study allows to develop and test out methods in a "confined" setting. Furthermore, the Netherlands has long observational records of various hydrological variables, making it possible to assess large amounts of data, constituting a good foundation for ML methods. Lastly, the Netherlands has an intricate and extensive water management system, which makes it interesting to combine and assess the effect of human influence on the national water distribution during drought episodes.

To explore the potential of machine learning in this hydrological setting, we have selected five frequently used data-driven statistical and ML techniques (hereafter generally referred as ML methods). The set of machine learning algorithms that will be used ranges from simple linear methods, as a benchmark, to recurrent neural networks with the purpose to identify the optimal trade-off between complexity and performance and includes the following: Multi-linear Regression, Lasso Regression, Decision Trees, Random Forests and the recurrent neural networks algorithm Long Short Term Memory (LSTM). Multi-linear and Lasso Regression are more classical statistical methods, Decision Trees and Random Forest represent regression algorithms that use simple decision rules to make an optimal regression of the variable of interest and finally, LSTM as recurrent neural networks and time dependent deep learning to provide simulations of the target variable.

The general experimental setup used to test for the suitability of ML methods in predicting hydrological variables is as follows: a simple set of input data (or features) is defined per type of analysis (e.g. natural vs water management run, type of ML algorithm and hydrological target variable); the time series of input and target variables to be predicted are split up into training and validation data, after training the ML algorithm, the predictive performance of the ML method is evaluated using the validation data. Each method will be evaluated using frequently used performance metrics like the Pearson Correlation (R), the normalized Root Mean Squared Error (nRMSE) and Kling Gupta Efficiency (Gupta et al., 2009, KGE). Additionally, information on the input sensitivity (also known as feature importance) will be gathered to investigate the influence of different input parameters and specifically the inclusion of water management plans in the ML simulations. Where possible, more information on ML method and feature-specific behaviour will be provided.

2.2.2 Data sources and preparation

We focused on gathering national hydrological and meteorological data for the period 1980-2019 to ensure a sufficiently long record of historic observations for our assessment, including several drought years. The aim was to minimize data pre-processing efforts to keep the methodology applicable for real-time applications; however, a minimal quality control is performed as the quality of the input variables is essential for a successful ML application. This generally includes removing outliers by using the 1 and 99% percentile and aggregating the data to either daily or weekly temporal resolution, depending on the target variable.

We additionally include water management plans used during previous drought periods to investigate the suitability of ML methods in reproducing those decisions regarding water distribution and to create opportunities to develop scenario analysis for operational water management settings in the future. We obtained water management guidelines through the National Water Management Planning Office (Rijkswaterstaat).

The following sections will give more insight on the target variables, input variables and water management data considered.

Target variables

The data availability for the Netherlands is very good, with an abundance of observations of (sub)surface hydrological variables. We focused our analyses on hydrological variables (or target variables) that are of interest to water managers and informative for operational water management under drought conditions. These include, besides observations on discharge, surface water levels (main river network), surface water temperature, and groundwater levels. For the first three variables, the measurements (depending on the location and the observation years) have a daily to 10min temporal resolution. For the analysis, the data was prepared by removing outliers using the 1 and 99% percentile and aggregated to daily temporal resolution. In the end, only stations with more than 10 year of observation records were retained. This resulted in the following number of stations per variable: 69 for discharge, 97 for surface water level, 105 for surface water temperature.

Groundwater observations were aggregated from daily to weekly temporal resolution, when needed, to ensure consistency between the records. As groundwater levels usually show a slow response time, this will have no major impact on the results and it has a better resemblance with the temporal resolution required for operational water management decisions. The same approach of removing outliers was taken as for the other hydrological data sets and again only stations with an observation length of 10 years were taken into account to ensure sufficient data coverage. This resulted in 3984 locations with groundwater level observations.

Input variables

The input variables chosen include a selection of five meteorological and hydrological variables (location of each input variable can be seen in Figure 2.1). The selection of input variables was intentionally kept simple and easily replaceable, as our aim is to test potential of the data-driven approaches and their skills in retrieving and combining information, even if only limited input parameters are used, and to be able to easily replace input forcing with forecasts of these variable in later subsequent work.

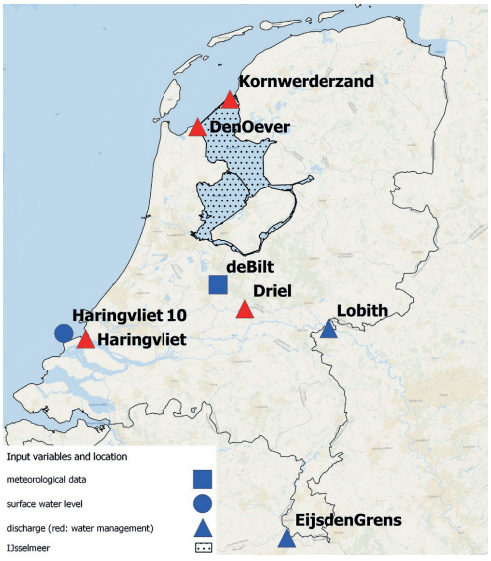


Figure 2.1 Overview of the Netherlands and stations locations of the different input data (incl. water management stations in red).

Regarding meteorological data, daily time series of precipitation and potential evaporation were obtained from the national weather service (KNMI). Contrary to gathering and preparing all possible stations available as it was done with the hydrological observations, it was decided to only take the main observational site in De Bilt, in the centre of the Netherlands. This was done consciously with the idea that the meteorological data will be part of the very simple set of input data used for the ML routine and with the goal to keep the input data simple. A second advantage is that the use of only this location will also allow to run the routine in forecast modes in the future, as weather forecast and projections are by default generated for De Bilt weather station.

From the whole set of hydrological variables, we selected the border stations of two major rivers to be part of the simple set of input data for the ML routine. These inflow points into the national system were selected as they are important reference stations for the national water authority and common

output from medium-range and seasonal forecasting models. The selection includes the discharge observations at Lobith and Eijsden, which are representative for the discharge values for the Rhine and Meuse at the border, covering the main rivers feeding the river system of the country. Additionally, a station measuring the water levels of the North Sea close to the Haringvliet sluice was chosen as a reference of the tidal influence and the major free outflow outlet points of the Dutch hydrological system near the delta of the river Rhine and Meuse.

Water management

Information on water management practices was gathered with the help of water managers at Rijkswaterstaat. We additionally included water management plans used during previous drought periods to investigate the suitability of ML methods in reproducing those decisions regarding water distribution and to create opportunities to develop scenario analysis for operational water management settings in the future. Most of the management decisions are based on the Rhine discharge at Lobith, which affects the water distribution and operation of the locks along the river network towards the sea and the IJsselmeer lake (also see Figure 2.1, red locations). The main infrastructures addressed in those management plans that we included in our analysis were the locks at Driel and at Haringvliet. The lock at Driel is used to steer the water distribution between the rivers Nederrijn and Waal, which subsequently guides the water into different regions. The Haringvliet locks determine how much water is released to the North Sea at the west coast. As no observations exist of the operating policies, we reconstructed the time series of the discharge at Driel and Haringvliet, using the default operating rules based on the operational plan and the Rhine discharge observation at Lobith (Appendix A.1, Figure A.1 and Table A.1). We further decided to include the locks Den Oever and Kornwerderzand at the IJsselmeer lake, which represent the amount of water released from the IJsselmeer lake to the Wadden Sea, by including time series of their management, as the IJsselmeer is currently used as one of the main water bodies to store water in case of droughts. The reconstructed time series combined with the historical observations of the discharge at Den Oever and Kornwerderzand were used as an additional set of input data for the ML routine for all the stations and target variables.

2.2.3 ML algorithms

For the experiment, a selection of different ML algorithms was used, covering simple linear to more complex techniques: Multi-linear Regression, Lasso Regression, Decision Trees, Random Forests and the recurrent neural networks algorithm Long Short Term Memory (LSTM). The analysis was done using Python3.6 and most methods used were setup by either using the libraries scikit learn or keras. To find the best parameter setup for most methods, the hyperparameter tuning process GridSearchCV was used and run over all discharge stations (with min 10 years observation record). The best options that were counted most were used in the setup. The following sections provide a brief summary of every method used and how they were set up.

Multilinear regression

Multilinear regression (MReg) is a very common type of linear regression, used to predict a target variable based on using several explanatory/input variables. Due to its simplicity it is used here as a benchmark method to compare more advanced methods to.

Lasso Regression

The Lasso (acronym Lasso standing for Least Absolute Shrinkage and Selection Operator) Regression (LASSO), which was first introduced by Tibshirani (1996), is a type of linear regression, which includes an additional benefit compared to other linear regression types by eliminating variables which are less informative than others (Bardsley et al., 2015). This is achieved by incorporating a regularization to

the quadratic error term that penalizes large regression coefficients representing input variables that add little information. A tuning parameter λ defines a threshold below which a regression weight will be optimised or forced towards zero, meaning the variable will have little information (Bardsley et al., 2015). The optimal value for λ was found to be $\lambda = 0.0001$.

Decision Tree and Random Forests

Decision Trees (DT) are a type of regression model build up in a tree structure, where the data is split multiple times into subsets until no further subsets can be made. Random Forests (RF), originally proposed by Breiman (2001), are an ensemble consisting of a number of decision tree predictors. The concept of RF includes the setup of decision trees containing elements of randomness by bagging and selecting a fixed number of variables randomly while setting up the decision tree (Koch et al., 2019; Tyralis et al., 2021). Hyperparameter tuning, using the GridSearchCV function, was taken into account for both one DT and one RF including the following parameters and results; the maximum depth (DT: 10, RF: 100), which defines the maximum depth of the decision tree, the maximum features (DT: Auto, RF; sqrt), which describes the number of features considered to find the best split, the minimum sample leaf (DT: 100, RF: 1) and sample split (DT: 100, RF: 2) to set the minimum number of samples required for a leaf node and the minimum number to split an internal node, as well as the number of trees in the forest (RF: 1000).

Long Short-Term Memory

A more complex method considered in this study includes the Long Short-Term Memory (LSTM) network, first introduced by Hochreiter and Schmidhuber (1997). LSTM belongs to a special class of artificial neural networks called recurrent neural networks. The advantage of this specific type is that compared to classical feedforward approaches seen in other neural networks it has the ability to invoke a sort of memory by using their internal state, allowing to learn long-term dependencies. A detailed explanation of the LSTM method is given in the work by Kratzert et al., 2018.

The LSTM architecture chosen for this study was build using the Keras library and consisted of two LSTM layers with Dropout layers in between. The end of the architecture further included a layer to flatten the outputs and a Dense layer. The number of neurons, epochs and batch size were determined by running several combinations of different setups, varying the number of neurons, epochs and batch sizes, for all the discharge stations. The best setup of each station was selected and the most recurrent number of neurons, epochs and batch size throughout all the discharge stations was chosen for the overall LSTM setup. The number of input neurons N_h to be tested was chosen by following a rule of thumb ($N_h = \frac{N_s}{\alpha * (N_i + N_o)}$, where N_i = number of input neurons, N_o = number of output neurons, N_s = number of samples in training data set, α = an arbitrary scaling factor (usually 2-10)): This resulted in testing [14, 40, 44, 50, 57, 67, 80, 100, 133, 199], to estimate the possible number of suitable neurons. Number of epochs and batch size was tested over [1, 10, 50, 100]. Based on the calculated RMSE score for the training period the following setup appeared to be most suitable: two LSTM layers (neurons:199, epochs:100, batch size:1) with two dropout layers (dropout=0.4) in between, one layer to flatten the outputs and a dense layer at the end.

2.2.4 Experimental Setup (Pipeline)

The target data, in our case for example discharge observations for each station individually, was predicted based on the input data set (natural run or water management included). Before the input data was incorporated in the standard training and testing routine, we evaluated lagged correlations that could help explain the historic patterns. Using the partial autocorrelation function (PACF) we identified the lagged input with significant information content. For each input time series, the PACF was computed and the timeseries corresponding to the first three lags were considered as additional

Table 2.1 Example of target and input dataset set used for the ML methods, including the lagged timeseries. For water management the data set was extended the same way including the additional timeseries.

target variables	**predictor variables**			
discharge,	**natural: P, EVAP, QRhine, QMeuse, WL Haringvliet***			
surface water level,	**water management: natural + QDriel, QHaringvliet,**			
surface water temperature,	**QDenOever, QKornwerderzand**			
groundwater levels,				
(at location xi, yi)	**for each predictor variable:**			
	obs	lagged input	lagged input	lagged input
$Y_{xi,yi,t}$	X_t	X_{t-1}	X_{t-2}	X_{t-3} *
$Y_{xi,yi,t+1}$	X_{t+1}	X_t	X_{t-1}	X_{t-2} *
...	...	...	...	...
				*X_{t-25},X_{t-24}, ...

timeseries in the input data set (Table 2.1). Taking into account this lagged input helps explain second order statistics that provide valuable information in timeseries analysis.

The input and target data were split into training and testing/validation data (60% training, 40% testing/validation) by using random segments (of min. one year length). Even though the traditional approach of splitting timeseries in two consecutive independent sets is used for this type problem, the random segments approach was done to keep the timeseries aspect but preserve the chance of including drought periods of the recent years depending on the random selection of the segments. Otherwise these recent drought events would have fallen away if splitting the timeseries traditionally with the first 60% as training and last 40% as testing/validation data. Furthermore, the 60/40% splitting ensures that the individual ML models are not overtrained and can only use a certain amount of the whole dataset to establish the relationship between the different variables.

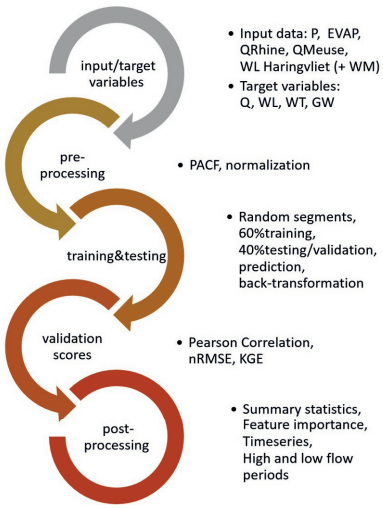


Figure 2.2 Pipeline schematization of experiment routine.

All target values except groundwater levels were trained on the daily dataset. Groundwater levels were computed by training on the difference ($y_k - y_{k-1}$, further also referred as deltachange) of weekly groundwater observations, which were added to the observation of the previous time step to compute back to the actual groundwater levels in prediction mode. The 40% testing/validation data set aside was then used to evaluate the predictions computed with the trained model using Pearson Correlation, nRMSE and KGE (Gupta et al., 2009).

An overview of the pipeline can be seen in Figure 2.2, while the following lines describe each step in more detail:

1. input/target data: general input data set the **natural run**: precipitation (P), evaporation (EVAP), discharge of Rhine (QRhine) and Meuse (QMeuse) at the borders as well as water level observation close to the coast and the Haringvliet sluice (WLHaringvliet).
   Water management (WM): input data natural run + reconstructed discharge time series of the main water infrastructures like lock Driel and Haringvliet, and observations regarding two locks at the IJsselmeer lake
2. calculating PACF (using timeseries corresponding to the first three lags as additional input time series)
3. for each station:
   the target variable was selected and we ensured that the time periods of both input and target align (Table 2.1).

   a) normalization: input and target data were normalized using the quantile transformer by scikit learn

   b) splitting data into training and testing sets: input and target data were defined into segments of min one year length and randomly selected into training and testing/validation sets, overall using 60% of the segments for training and 40% for testing/validation

   c) training and prediction per algorithm: every algorithm is first trained with the same subset of training data per variable of interest. The obtained trained models are then used to predict the variable of interest based on the testing/validation data.

   d) back-transforming: the results are back transformed into the observations units of the original observation.
4. validation scores: to analyse the predicted results and the goodness of fit of the historical time series we computed the Pearson Correlation, normalized Root Square Mean Error (nRMSE) and Kling Gupta Efficiency (KGE) for the training, testing/validation and total period.
5. post-processing:

   a) summary statistics: the results of all station simulations are used to get an overall understanding by including summary statistics like CDFs, averages, etc.

   b) feature importance: the feature importance of all stations were considered to get a better understanding of the input data sensitivity, also regarding water management influence.
   The feature importance for Random Forest and Decision Tree was computed by using the the feature importance property by scikit-learn, also known as Gini importance, and for LASSO and Multi-linear Regression the property model coefficient from scikit-learn was used to get a crude type of feature importance

   c) high and low flow periods: to assess high and low flow periods for extreme events, the highest/lowest 30% of the observation records and the corresponding simulations results of each station were selected, the nRMSE was computed to assess the performance of each method

2.3 Results

In our analysis we explored the suitability of five different Machine Learning (ML) methods to simulate various hydrological observation by: (1) developing a data preparation routine, (2) training of different ML models, (3) validating our simulations based on historical observations. We extended our base (natural) scenario by additionally incorporating water management influence to assess the effect of additional data on human water interactions on the method's predictive skill. Further, we evaluated the model performance for high and low flow events, to see whether we can accurately capture drought events with our methods. The following sections will highlight the main results for these three main aspects: reproduction of historical observations, water management influence, and high and low flow events.

2.3.1 Simulation and reproduction of historical observations

The routine developed for the different ML algorithms was used to predict and reproduce discharge, surface water level, surface water temperature and groundwater level observations. The input data was defined into segments, 60% of the segments were randomly chosen and used to train the different models, while the remaining 40% of the data was used to test/validate the simulation performance of the methods. For validation purposes the normalized RMSE (nRMSE) score was computed, which gives an indication on how close the predicted values are to the observed ones, while the normalization makes it possible to compare stations with different mean discharge.

Comparing the training and validation nRMSE scores per method and target variable by plotting the results as a cumulative distribution function (CDF) highlights the differences in prediction skills (Figure 2.3). Generally, a slight decrease in predictive skill is expected for the validation set. We observed that both training and validation scores for the daily simulations (discharge, surface water level and surface water temperature) show a similar pattern, especially for the simpler methods including Multi-linear regression (MReg), Lasso Regression (LASSO), and Decision Tree (DT), with slight decrease in prediction skill for the validation set. The most noticeable decrease in predictive skill was seen for both the Random Forest (RF) and LSTM models, which ultimately brings their validation score closer to the more simple models. While the nRMSE scores for the surface water temperature lay within a range of 0-0.2 and with a steep CDF, the surface water level and discharge scores and CDFs indicate that prediction of these target variable is more challenging (nRMSE ranging from 0-0.45). Looking at the CDFs of training and validation scores for discharge, there is a small plateau effect recognizable, followed by a similar slope as seen for the surface water level scores. The plateau effect is likely due to the information included in the input data set combined with the proximity of some of the discharge stations to the two main discharge time series included in the input data set (Rhine and Meuse).

The nRMSE scores for groundwater levels show a slightly larger variation between the different methods as well as a plateau effect for almost all methods, both for the training and validation set. There is a distinct cut-off in the nRMSE around 0.1 to 0.25, which is related to stations that have no nRMSE around or equal to the climatological mean. In case the input data and the target variable have a very low correlation, most algorithms will default to a prediction that is around the long term mean, resulting in a nRMSE between 0.0-0.5. Computational time regarding training the different ML methods for all the target variables and all their stations was ranging between a few hours (MReg, LASSO and DT), a few days (RF and LSTM) to a week (groundwater levels by RF and LSTM models). While training the models can be time consuming (depending on the method and the number of station considered), the simulations of the target variables based on the trained models and the whole input data was done within a few hours up to one - two days. These calculation would be even shorter (order of hours) when used in forecasting mode, with expected maximum lead times up to 3-6 months at maximum. Having a relatively short computation time can be of benefit, especially

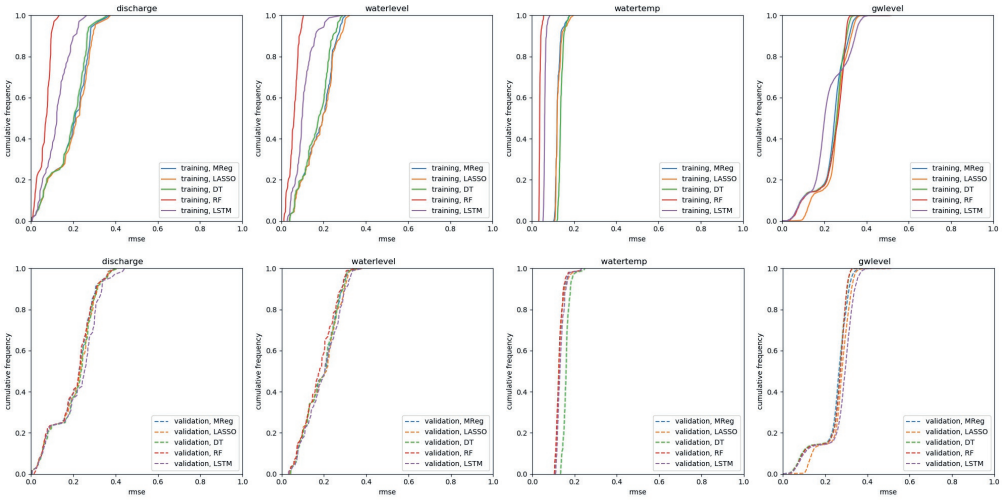
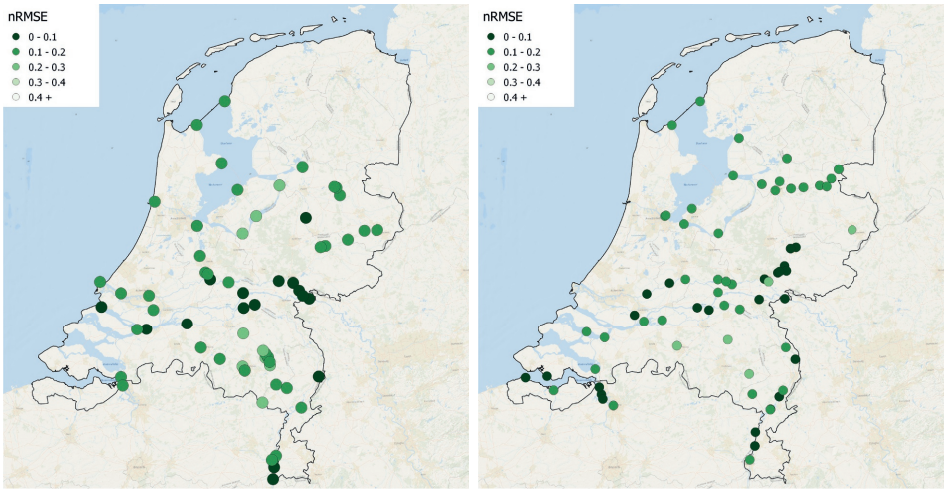


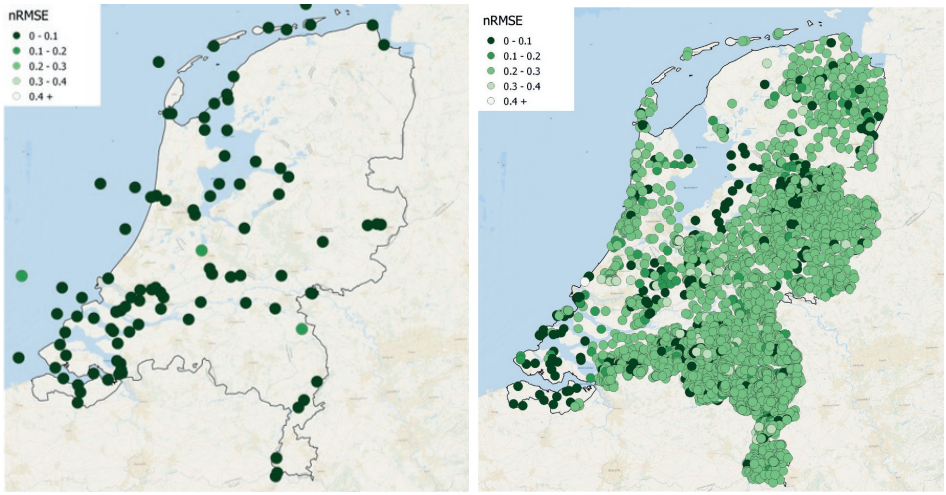
Figure 2.3 CDF of nRMSE training (top) and validation score (bottom) per method and target variable for the natural run.

regarding water management aspects, as the information necessary (timeseries, scenarios, etc.) to make decision can be delivered reasonably fast compared to large scale hydrological models, which can require large computation times.

Figure 2.4 provides a spatial picture of the different stations considered per target variable and the nRMSE scores of the predictions from the RF model, which is chosen as it was giving the best results overall. The scores were computed based on the training and testing data together, where the whole data for the specific station was used. The darker the colour of the station, the lower the nRMSE score, indicating a good performance of the RF model. For discharge and surface water level it can be seen that upstream stations along the main river network (Rhine and Meuse) show low errors, while the prediction skill decreases further downstream or in smaller side rivers. This is likely due to the strong influence of the input features which is still visible on upstream locations, especially for discharge stations. Downstream stations and smaller river areas are often heavily influenced by local water management infrastructures and management decisions, making it more difficult to provide accurate simulations, resulting in a higher nRMSE score. Other factors like soil moisture, land cover and catchment characteristics could potentially influence the model performance for these stations. As the focus in this study was laid on testing out a simple input dataset, assessing all possible influencing factors was out of scope. This general trend and spatial pattern is also observed for the other ML models, with minor differences (Appendix A.2.2). While counting the number of stations falling in the range of nRMSE used in Figure 2.4 for the different target variables (Table A.2), it can be seen that for discharge all methods have roughly the same amount of stations (16 & 17) in the lowest nRMSE range (0 - 0.1). A large difference can be observed for the following ranges, 0.1-0.2 and 0.2-0.3, where RF and LSTM models include more stations that are having a better performance (e.g. 12 stations (MReg, LASSO and DT) vs 41 & 20 stations for RF and LSTM). For surface water level prediction a similar distinction between the more simple and advanced methods can be made, where the largest difference is seen again in the range of nRMSE 0.1-0.2. Compared to discharge, the lowest range is reached by more stations of the RF (31) and LSTM (21). A very clear pattern regarding performance skills was observed for surface water temperature: all models for MReg, LASSO and DT fall in the nRMSE range 0.1-0.2, while LSTM shows almost equal numbers for 0-0.1 and 0.1-0.2, and RF having the most number of station in the lowest nRMSE range. In contrast to the clear distinction between



(a) discharge **(b)** surface water level



(c) surface water temperature **(d)** groundwater level

Figure 2.4 nRMSE results for discharge, surface water levels, surface water temperature and groundwater levels simulations using the RF model for the total timeseries. The darker the colour, the better the score, indicating a only minor differences between predicted and observed values (nRMSE ranges form 0-1).

the different methods and number of models per nRMSE range seen before, the results of the groundwater level models show a relatively comparable order. Of the ∼4000 station models, around 400-450 fall into the range 0-0.1, 150-200 in the range 0.1-0.2 and the majority (2300-3200) into range 0.2-0.3 for all methods, except LASSO where a large amount of stations can already be found in 0.1-0.2 but also 0.3-0.4 nRMSE range (Table A.2). Overall, it can be observed that for the simulation of groundwater levels, no particular method is doing better than the others. This might be due to the input data not having a direct influence on the simulation as the Rhine discharge for downstream stations as an example. Minor differences between the methods, regarding average nRMSE scores (Table A.3), were also seen for surface water temperature. While RF and LSTM show the lowest

errors, overall the location of the stations and the simulation does not influence the results. This is likely due to the strong link to the evaporation data, and with that connected to air temperature, included in the input data set.

Simulating full timeseries of the different target data and comparing them to the observed records further shows how close the ML methods like RF and LSTM come in reproducing the historical observations (Appendix A.2.3, Figure A.5). Please note that in these figures we combine training and validation data for the purposes of visual inspection. We see that surface water temperature timeseries are well predicted by all methods and especially well by the RF model, which is in agreement with the low nRMSE scores. For discharge and surface water level simulations we see that that the prediction skill deteriorates depending on the location of the station within the national river system, as already observed in Figure 2.4. Discharge stations which are along the main river network and close to the geographical location of the border discharge input data show a higher prediction skill (low nRMSE scores but also KGE ranging from 0.7-0.985) than stations that are located further downstream (KGE between 0.6-0.5, lowest KGE value 0.4). The timeseries of upstream stations have a stronger correlation with the input timeseries (not shown) and ML models for these locations are therefore expected to show a good performance.

To get a better understanding of the way the input data was used by the different models, we had a look at the feature importance (or model coefficients as a crude type of feature importance score for MReg and LASSO models) for all methods except the LSTM due to its complexity. The strong influence of the different input variables was observed by plotting the feature importance per station as seen in Figure 2.6 for discharge and surface water level (both natural and water management run) using the RF method. Each station is represented by a diagram, showing the feature importance (lags are summed up) of the different input variables with a color code, which got extended for the water management input data set.

We observed that the Rhine and Meuse discharge play an important role in the prediction of discharge and surface water levels, which was expected for some stations as these rivers are the main two tributaries to the Dutch hydrological system. For discharge stations, the Rhine influence is mostly seen along the main river network leading to the sea, while the Meuse plays a strong role for stations in the south of the Netherlands but surprisingly also in some stations in the North East. An increase in importance of the information gained from the surface water level station close to Haringvliet can be seen for the surface water level simulations of stations close the sea, while the main influence of the Rhine and Meuse discharge in the main river network remains. Stations around the IJsselmeer and to the East show a more balanced incorporation of the input data. For surface water temperature, the main influence stems from evaporation information, which is ultimately also connected to air temperature. While the RF model includes the different input variables in a more balanced way, the DT focuses on a few main variables with Meuse discharge being more prominent than in the RF model. The model coefficients used for MReg and LASSO further show which input variables have an inverse relationship in these models (e.g. additional lagged timeseries or certain input variables depending on station location, Figure A.6, A.8 and A.10).

2.3.2 The impact of water management

To assess the impact of including water management as an explanatory (input) variable in our simulations, we included reconstructed timeseries of water management infrastructure based on operational plans and the discharge at Lobith. These timeseries were added to the general input data set and used in the ML routine to predict the discharge, surface water levels and surface water temperatures, as we hypothesize that the impact on groundwater levels is negligible. As the previous nRMSE scores for the normal scenario were already good, including water management as input variable generally leads to a small added benefit (Figure 2.5).

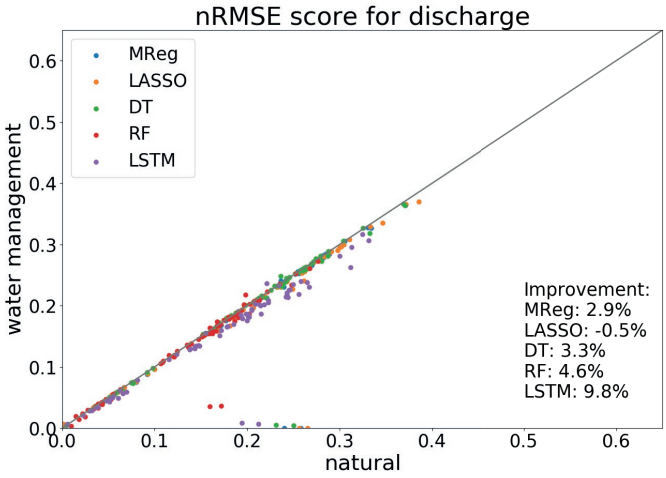


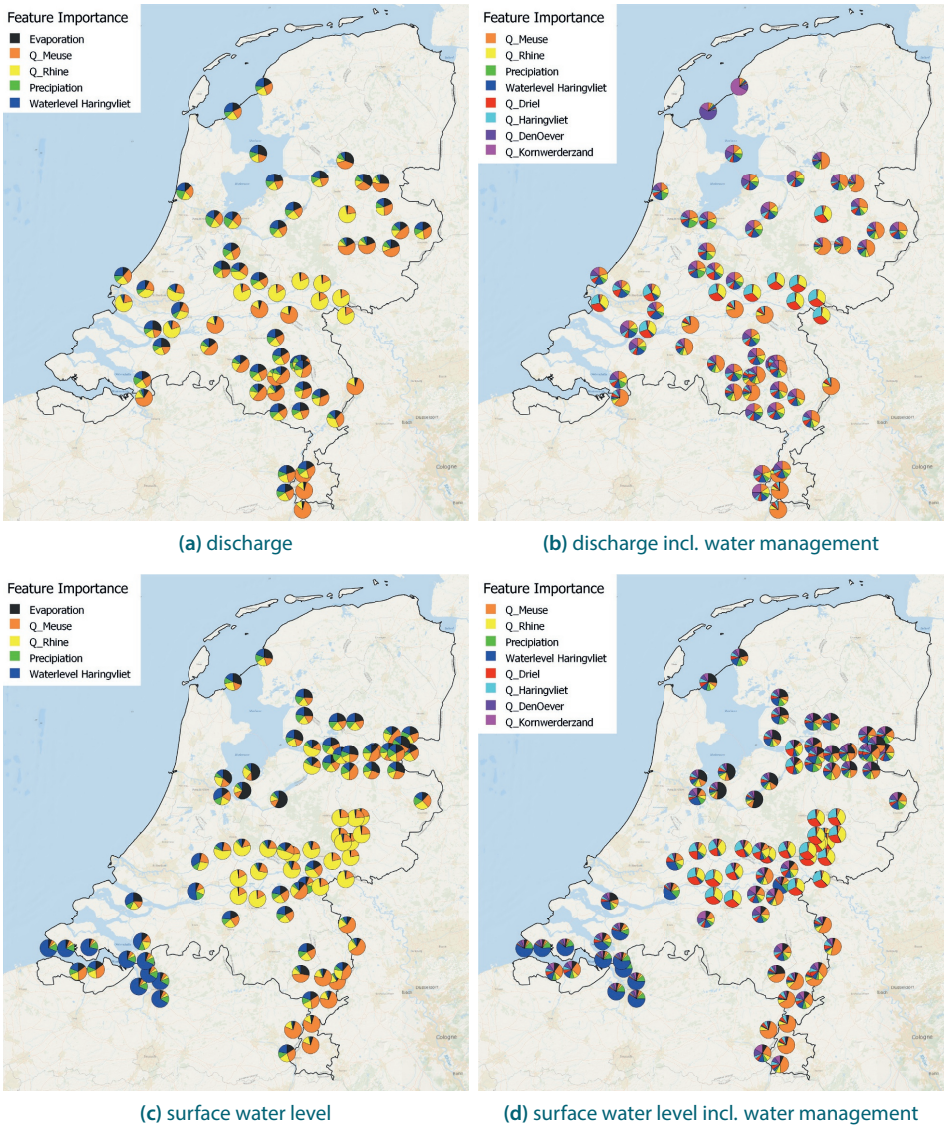
Figure 2.5 nRMSE scores of natural and water management influenced run plotted against each other with every point representing one discharge station. Stations which lie below the 1:1 line indicate that water management influence improves the predictive skill for that station.

The simple models included in our assessment display a similar behaviour as the RF model, where the included water management information only improved the nRMSE scores minimally: average improvement for MReg: 2.9 %, DT: 3.3%, RF: 4.6% and for LASSO even a slight decrease: - 0.5% compared to the natural run. The slight decrease for the LASSO method might be due to the random segments splitting approach for defining training and testing data, which might have lead to segments being less informative than the one of the natural run. The LSTM model however notably benefits from the additional information on management, with almost all stations showing an improved validation nRMSE score (average improvement by 9.8%).

This is likely due to the internal memory that LSTM architecture includes, enabling the additional information to be included over more calculations steps and improving simulations of stations further downstream, which are more affected by water management decisions than the upstream ones. For every method, the two stations showing the largest improvements in the water management influenced run are the two stations at the IJsselmeer lake which were used as additional input information to represent the water management influence.

Looking again at the feature importance of the different methods for the water management scenarios highlights to which extent the added water management information impacts the simulations even though the improvement in nRMSE score was only minor for most methods. For the run including water management information, the influence of the added reconstructed operational timeseries can be seen in red (lock at Driel), azure (lock at Haringvliet), pink and violet (two locks at the IJsselmeer). The influence of all these additional input variables can be seen throughout all the discharge and surface water stations in Figure 2.6 with a varying strength. For both of these target variables, the influence of the lock Driel, which is the main infrastructure to change the discharge distribution and located only a few km downstream Lobith, can be seen especially at stations which are close to the main river system. Observed influence upstream is caused by backwater effects during times when the lock Driel is (partly) closed. The reconstructed time series of lock Haringvliet have a comparable importance across most of the stations as found for the lock at Driel. As both of those reconstructed timeseries depend on the discharge at Lobith, they are all linked together and show a similar importance. The influence of the two

locks at the IJsselmeer appears to be less strong in general with some exceptions close to those locks. Compared to the more balanced feature importances of the RF model, the DT model (Figure A.7 and A.9) focuses again mostly on one major input series for each station, while the simple methods only include the water management to a limited degree. The predictions of surface water temperature barely take the additional water management information into account, with evaporation remaining the most influential input feature as seen in the natural run.



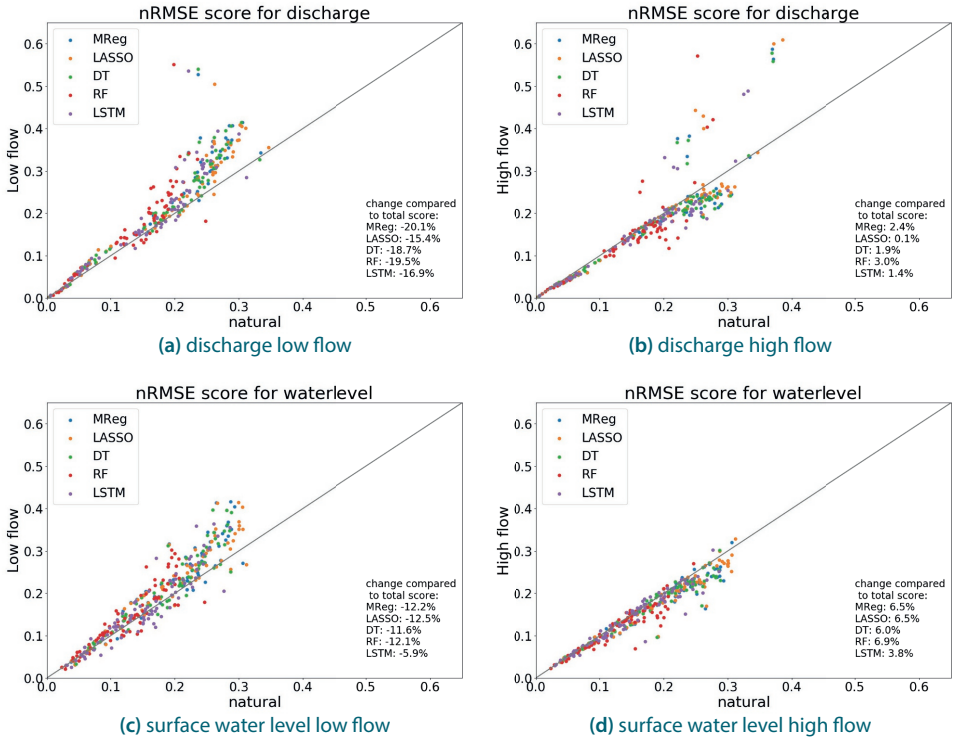
(a) discharge **(b)** discharge incl. water management

(c) surface water level **(d)** surface water level incl. water management

Figure 2.6 Overview of the feature importance (RF model) for every discharge (top row) and surface water level (bottom row) stations included in the natural (left column) and water management (right column) run.

2.3.3 High and low flow simulations

In order to test the ability of the ML methods to reproduce discharge and surface water levels during high and low flow periods, we selected the highest and lowest 30% of observed records of every station and calculated the nRMSE score from the observed and predicted timeseries for these timesteps. In Figure 2.7 we show the nRMSE scores for the high and low scores plotted against the natural run scores for every method and target variables discharge and surface water levels. For both target variables it can be observed that the prediction skill for low flow periods is lower compared to the total natural run scores. In average the skill for low flow periods are 15-20% and 5-12% lower for discharge and surface water levels, respectively.



(a) discharge low flow **(b)** discharge high flow

(c) surface water level low flow **(d)** surface water level high flow

Figure 2.7 nRMSE score for low and high flow scenario compared to the total natural run for discharge (top) and surface water level (bottom) simulation.

While discharge low flows appear to be more challenging to simulate using the currently trained models, the surface water level low flows are closer to the total scores (with LSTM showing the smallest loss in prediction skill and lying closer to the 1:1 line). The same observation can be made for high flows, where the nRMSE results for surface water level stations are generally closer to the total natural nRMSE scores, showing even a slightly higher skill of 3-6%. Compared to the latter, the scores for discharge also indicate a relatively good performance for high flow periods, with most stations having a similar to slightly better score but also a few stations which struggle to reproduce the high flow events. Table A.3 gives a further overview of the average prediction skill for every method and target variable.

Figure 2.8 further shows the simulated timeseries of surface water levels for 2015-2019, highlighting the difference in prediction skill for low flow periods per method for one station lying half way on the main river network. While the simple methods like MReg and LASSO struggle in general to reproduce

low flow periods for the time period shown, the other methods especially struggle during the drought of 2018, when low flows occurred over several months while previous low flows are captured relatively well. Besides the model properties also the choice of using random segments for training can have an impact on the simulation performance, as extreme low events (e.g. the drought 2018) might not have been selected for the training phase.

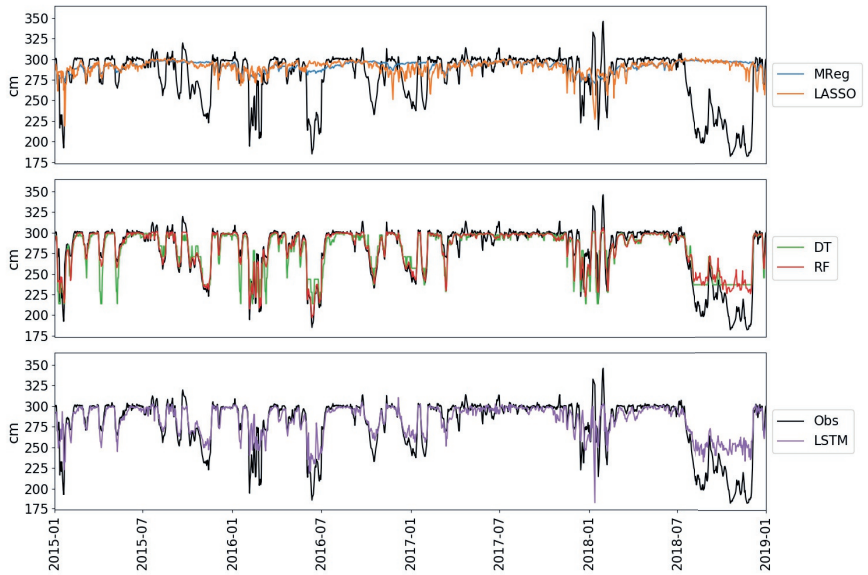


Figure 2.8 Time series of surface water level predictions at station Hagenstein Boven from the natural run using the different ML methods for 2015-2019, including the extreme drought in summer 2018 (black line representing the observations, Obs).

2.4 Discussion

In this study we investigated the suitability of five different Machine Learning (ML) methods in simulating various hydrological variables in a national-scale managed hydrological system based on a simple input dataset. In addition, we assessed the impact of including water management information on the ML methods' predictive skill, by incorporating operational rules of major water infrastructures. Finally, we looked at high and low flow periods separately to check the ML methods' capability of capturing for example drought periods, which provided the most challenging conditions for operational water management during the last few drought periods.

2.4.1 Machine learning models: differences, difficulties and advantages

The selection of the models was done considering different levels of complexity. The Multi-linear regression (MReg) method was chosen as a benchmark representing the simplest method in our methods set. The Lasso regression (LASSO) was included due to its simple composition but with the benefit of variable selection. The inclusion of the Decision Tree (DT) and Random Forest (RF) method were considered for their buildable complexity and LSTM as a representation for a recurrent neural network, which is promising for hydrological assessments (e.g. Kratzert et al., 2018; Kratzert et al., 2019b).

We saw that the more advance methods like RF and LSTM used in this study in combination with a very simple input data set of 4-8 main features outperform more simple methods, especially if target variables are directly connected to the input data. For example discharge of Rhine and Meuse was observed as one of the main features having a strong influence on discharge and surface water level simulations, primarily along the main river network. The RF models were further outperforming the other methods for stations further downstream for both target variables. Compared to the other target variables, the groundwater level simulations appeared to be more challenging for the different ML methods (with most stations of all methods falling into nRMSE range of 0.2-0.3), as the simple input dataset used did not have a dominating feature which some of the methods could have taken advantage off. Furthermore, with the change in groundwater levels being relatively slow and small, the input variables might have been less informative and relationships harder to extract compared to discharge observations, which fluctuate more. With all methods performing comparable and no clear method outperforming others, computation time could be added to the factors influencing the choice of which ML methods to use in this case. As there are roughly ∼4000 stations included, the simple methods like MReg, LASSO and DT are preferable if a lower computation time is a requirement, as these methods were able to simulate all stations within a few hours up to a day.

While ML methods appear to become a respectable choice in hydrological assessments and we saw that even with a small set of input features some methods can already perform well, some aspects can limit or enhance their full potential. First, since ML methods as used here cannot rely on preconceived rules of system behaviour, their dependency on sufficient data of good quality and time series length is key. Second, ML predictions are difficult to interpret owing to their black to grey box nature. Depending on the ML method, different aspects of the method can be analysed or visualized (e.g. tree structure for DT or RF, feature importance and model coefficients) but for more complex methods as neural networks it can be difficult to fully identify which drivers and processes drive accurate prediction. Recent efforts to remedy this are based on reporting and analysing cell states (Kratzert et al., 2018) which can help to unravel some of the black-box characteristics. Furthermore, finding the right hyperparameter setup for specific methods (e.g. LSTM) can be challenging. We tried to include a coherent approach between the different methods but the best architecture of the LSTM neural network is often found by trial and error as there are no clear guidelines on finding the perfect setup.

A strong advantage of ML methods is their flexibility in learning new relationships, handling large amounts of data efficiently and often being less computational demanding than large-scale hydrological models. In our case, simulating the time series of a single variable takes a few seconds to minutes depending on the method and simulating all stations is done within a few hours. Note that these are serial computation times based on a standard desktop PC. Simulating many stations simultaneously could be efficiently parallelized on a high-performance computer. Finding the best hyper parameterisation can be more time-consuming, especially if the GridSearchCV or a similar setup is run for many parameters and for example for more complex ML methods like RF and LSTM. However, these analyses can also be speeded up by performing them in parallel. Regardless, once the model is setup with the most promising parameters and trained, the simulation run time is often shorter than for physical-based hydrological models.

2.4.2 Impact of input timeseries, feature importance

Our analysis shows that different ML methods based on a very few input features can perform well in simulating hydrological variables, especially if the there is a direct connection to the target variable. We could observe that there is a location dependency of prediction accuracy for discharge and surface water levels, related to the proximity of the station to the input locations. Slight deterioration of the nRMSE score could be observed if the station of interest was further away from the different input variable locations, which is in line with the general expectation. This is also recognizable by the feature

importance analysis, specifically looking at discharge and surface water level stations following the main river network (input variable regarding Rhine discharge), showing the decreasing influence of the Rhine discharge on the simulation results if the station is closer to the coast and thus further from Lobith (Figure 2.4). Other examples of distinct use of the input features were also seen in specific areas, e.g. the use of the water level Haringvliet input feature for stations close to the sea (target variable surface water levels), as well as the use of Meuse discharge input feature for stations South-East and the use of the Rhine discharge for around main river network following from Lobith. Areas especially in the northern part of the Netherlands were incorporating the different features more equally or even showing a higher importance of a feature that was not necessarily expected (e.g. Meuse discharge for discharge stations). Overall, the incorporation of the feature importance shed more light on the importance of various input variables in the simulations and how this relates to location in the hydrological system. As such, it provided a limited peak inside the black boxes that most ML methods are.

2.4.3 Impact of human water management

As the nRMSE of the different ML methods already showed a strong performance without it, adding additional information on water management practices led to only a small improvement, except for the LSTM model where the improvement was considerable (around 7% and 9% for surface water levels and discharge, respectively.). Nevertheless, the feature importance analysis revealed that including management information is being picked up as a valid predictor by some of the methods (Figure 2.6). The main water management infrastructures, which are operated based on the Rhine discharge, appeared to have a large influence on upstream stations and stations close to the infrastructures when looking at the RF results, while no impact was found for the more simple methods like MReg and LASSO. The total influence of the reconstructed time series (Driel and Haringvliet) and the Rhine discharge are equal to the original influence of the Rhine discharge in the normal run. Even though there is a clear dependency on the Rhine discharge, the influence of the different management infrastructures depends on the location of the stations. Although we did not compute the feature importance or a similar measure for the LSTM model due to its complex setup, we observed that the LSTM model was the one method that profited most from the inclusion of water management information (Figure 2.5). As the additional water management input timeseries are based on operational rules that are linearly or quasi-linearly related to discharge, we hypothesize that methods like the RF are especially good in learning these management operations from discharge alone, while probably having lower predictive skill if the management rules would be more complex. We would expect the LSTM model to be more suitable for these due to its characteristics and using the internal state as a sort of memory. To substantiate this assumption and to assess the sensitivity to different types of management rules further experiments would be needed, which is beyond the scope of this study. Nevertheless, by including human water management, the ML algorithms now provide a representation of the hydrological system that is closer to the reality of a managed system. In addition, this would allow us in future studies to explore the potential of water management changes, with the clear limitation that the rules and impacts are assumed to be stationary in time.

2.4.4 Performance under droughts conditions

In the last part of the analysis we focused on high and low flow periods in more detail. We showed that more advanced methods like RF and LSTM are generally better in capturing low flow events compared to the simple methods but struggle to recreate extreme events like summer 2018. Simple methods like MReg and LASSO are not able to capture any of the peaks, both low and high, due to the stronger non-linear relationships that occur during these extreme conditions. We observed that simulating high and low flow periods can be challenging, especially if only limited data is given to the ML models. Including more information regarding catchment characteristics and soil moisture, which are also important aspects for drought development, could improve future simulation

experiments, similar to Kratzert et al., 2018; Kratzert et al., 2019a; Kratzert et al., 2019b. Furthermore, the choice of using random segments to split the data into training and testing data further affects the prediction skill as extreme drought periods like 2018 might likely not be in the selected training data, resulting in a less good performance for such events as the model has not seen this data before. A possibility to improve the predictability of low flows could be to include additional training sets including long term droughts, while at the same time using a wider set of input data that inform on the water storage in upstream areas, e.g. snow cover data or total water storage from GRACE. Compared to the low flow periods, the scores for the high flow periods are generally closer to the nRMSE scores of the total simulations and in most cases slightly better (e.g. for surface water level stations).

2.4.5 Opportunities for improvement

The study's focus is on simulating hydrological variables for stations of the monitoring network of the National Water Management Planning Office (Rijkswaterstaat). With the choice of the input data and only working with existing stations, in accordance with future forecasting at these locations, the assessment assumes stationarity and changes regarding land use or climate change are not directly included. Climate change could be included in future studies by changing the input data with projections from global climate models in combination with large-scale hydrological models (van der Wiel et al., 2019b). Such an approach could be tested by training the ML models on data from e.g. the early 1970s and evaluate the trained methods using observations from the last 10 years. Out of the different methods, the LSTM model is the only model which has a sort of memory effect, which would be interesting to further assess the benefit of it in future studies touching on this aspect of changing climate. By focusing on existing stations from the monitoring network, the study does not touch upon the problem of prediction in ungauged basins which is an important issue in hydrological modelling. For future studies it would be interesting to branch out and test out different setup, including adding additional input data as well. For example, recent studies by Kratzert et al., 2018; Kratzert et al., 2019a; Kratzert et al., 2019b have shown the promising use of LSTM models for regional rainfall-runoff modelling. In comparison to their work, we did not include any location-specific static information, such as surface elevation, soil type, proximity to power plants along the river etc. It is likely that including such additional information would improve our general predictive skill and would need to be tested in additional experiments. Furthermore, adding additional information regarding soil moisture, land use and vegetation might be useful, especially for simulating drought period as the current setup appeared to only give limited information for this purpose. A particularly interesting result from Kratzert et al., 2019b is that training a single LSTM model on all timeseries using location-specific static information as additional input variables performed better than the LSTM models trained on each timeseries separately. Performing such an experiment on the groundwater timeseries used in this study would be an interesting follow-on study.

Furthermore, regarding incorporating ML methods for groundwater level prediction, Koch et al., 2019 recently showed their efforts to combine the RF method and a physically-based model to simulate the shallow water table at a national-scale for Denmark. It would be interesting to see if using e.g. the long-term average groundwater depth from the National Hydrological Model (De Lange et al., 2014) as location specific feature would improve groundwater level predictions in our case.

We also realized that the incorporation of water management information in predicting groundwater levels can be challenging, as this depends on regional to even local water management by (drinking) water companies, water boards, and farmers, with diverse and difficult to obtain specific management rules. As groundwater is an important aspect of the hydrological system in the Netherlands, it would be an interesting addition to also include information on regional water management and local groundwater pumping in future experiments, if the data availability allows it.

Given that we have shown that the trained ML methods are successful in reproducing observed time series of various hydrological variables at many locations relevant for national-scale water management, a next step would be to use the trained ML-models in a forecasting experiment and evaluate the forecasts for past periods of drought and water shortage. In this context, with water management as one of the input variables, the ML methods with their short computation times can also be used for on the fly implementation of operational management scenarios (e.g. operating scenarios) and evaluating their ability to mitigate imminent hydrological droughts.

2.5 Conclusion

In this study, we explored the suitability of five different ML methods for simulating hydrological variables, i.e. discharge, surface water levels, surface water temperature and groundwater levels, that are deemed relevant for water management at a national-scale based on a very limited set of input data that are also available from existing weather and large-scale hydrological forecasting systems. We additionally explored the possibility of incorporating water management practices and the ML methods' capability in capturing high and low flow events. This in light of future efforts to devise a computationally efficient seasonal forecasting framework that allows operational (on the fly) water management scenario analyses to mitigate imminent droughts.

We find that all ML methods, ranging from very simple to more complex ones, showed a generally good performance for stations and target levels which are closely linked to the input data and location (e.g. stations along main river network). For downstream stations and small rivers, the Random Forest method outperforms the other methods both for discharge and surface water levels. For surface water temperature no location dependency was observed. Generally, the best performances were reached by the more advanced Random Forest and LSTM methods, which was especially the case when simulating high and low flow events. High flow events were slightly better captured than low flow events but overall simulating extreme events based on a simple input data set remains challenging. Specific training sets, including event related information and additional input variables, could help to improve future assessments. For simulating groundwater levels, all methods were performing comparably with most stations ranging in nRMSE 0.2-0.3. Including the feature importance of the methods allowed us to detect how and where water management influence played an important role. The addition of information on water management in the ML routines increases overall performance, although limited.

We conclude that ML methods proved to be successful in simulating a large set of inter-related time series of multiple hydrological variables for a human influenced national-scale hydrological system. There are of course limitations and drawbacks of such methods that should not be ignored. Having enough data of good quality is an important prerequisite for being able to develop an ML routine with high skill. Also, the black-box nature of ML methods makes it hard or even impossible to follow or understand the information flow, implicit assumptions and the causes of prediction accuracy. Depending on the ML method it is possible to trace back some of the decision steps and we tried to get an insight on how the input data is used by inspecting the feature importance.

Nevertheless, ML methods in general and the routine developed in this study show a large potential for various types of hydrological assessments. Not only simulating current or historic observations but potentially also for forecasting experiments and scenario analysis in heavily managed river systems around the world. The ML routine developed in this study shows a high level of agreement with the observations and can thus act as an additional decision support tool to be implemented in operational water management.

Chapter 3 | Exploring and Optimizing Water Management Strategies for Mitigating Local Drought Impacts

Abstract

Recent drought events in Europe showed that these type of extremes can have large scale impacts throughout Europe. Having a better understanding of potential drought impacts, their development and how humans can mitigate these is needed to increase preparedness for future events. Current challenges lie in assessing and modelling of drought impacts at various spatial scales. Furthermore, it is also important to understand how human responses, including water management decisions, can either alleviate or intensify drought severity and drought impacts. An interesting case to study these aspects is the Netherlands, a country well known for its intensive water management and recent challenges with extreme drought events. To be able to assess and evaluate the potential of optimizing the water management to mitigate drought impacts, a modelling framework was setup, based on a machine learning model simulating discharge, water management scenarios and drought impact functions. The discharge simulations were based on a multi-target Long Short-Term Memory (LSTM) model simulating three main river branches linked to a major water infrastructure further upstream. Water management scenarios for optimisation were established based on the previous observations of operational behaviours. Combining the multi-target LSTM in forecasting mode over different lead times (maximum 30days) with operational management scenarios lead to a multitude of discharge simulations of the different river branches and their response to the management actions. For every river branch, which had a dedicated drought impact function based on its main purpose (e.g. shipping), the resulting drought impacts for each operational management scenario was computed. To find the optimal scenario for past drought years, such as 2003, 2015 and 2018, the lowest cumulative total impact was computed for every day for a 30-day window during these years. The modelling framework's evaluation showed that the multi-target LSTM is able to simulate the discharge for different river branches well, with sufficient sensitivity to changes in the operational water management settings. With the impact functions, the water management could be translated into potential mitigation strategies, showing that the total impact could be reduced especially in early lead days (about 2.5% for 2018 and 2015 and 6% for 2003). While the current setup was build and tested on a historical time frame and with a perfect hindcasting setting, it allows for a first exploration of potential drought impact mitigation. To make the modelling setup more realistic, future steps could include the introduction of uncertainty through adapting the hindcasting input. This would likely create more room for exploration of water management decisions and their consequences. Furthermore, a more iterative optimisation setup could be implemented to actively evaluate management decisions especially on the onset periods of drought events.

Based on: Hauswirth, S.M., Bierkens, M.F., Beijk, V. and Wanders, N. Exploring and Optimizing Water Management Strategies for Mitigating Local Drought Impacts. In preparation

3.1 Introduction

Droughts being multifaceted problems bring along many complications and impacts. Impacts differ per drought type and length, however the longer the drought, the stronger the implications and damages for both terrestrial and aquatic ecosystems. Longer and higher intensity droughts, likely also result in longer recovery time of the system, leading to an increased vulnerability for future occurrences.

In the last decades, an increasing number of extreme drought events have occurred in Europe affecting large areas throughout, including regions usually less prone to droughts, e.g. Western Europe and Scandinavia. These drought events and their related drought impacts have been extensive, having both multi-national and multi-sectoral dimensions (Blauhut et al., 2022). Impacts were seen for agriculture, drinking water, industry and shipping, as well as energy sector next to general impact on nature for various countries (Bakke et al., 2020). Most of these impacts are related to human water uses, which also alter the hydrological system and change the natural development of droughts for example by intensifying hydrological droughts or enhancing prolonged groundwater droughts.

These recent extreme droughts have shown that drought impacts and drought management aspects are highly linked to perceptions, and vary strongly across countries and organisational level and their operational scale (Blauhut et al., 2022; Stahl et al., 2016). Challenging factors include drought impact definitions (Wilhite et al., 2007; Blauhut et al., 2015; Bachmair et al., 2016), be it direct and indirect impacts. While there have been efforts to build and extend European drought impact databases (Stahl et al., 2016), a lack of the detection, recording and understanding of the effect of local management on drought processes and drought impacts can be a limiting factor for future studies on drought management and impact mitigation (Bachmair et al., 2015). Furthermore, quantification of the actual drought impacts and their relationship to vulnerability and exposure during past drought events is another challenge. While these impact relationships or impact functions would be exceptionally helpful in drought assessments, not only for the past put also future events under increasing climate change, these are often not yet studied in detail or not available.

Current approaches of drought impact monitoring often include linking impacts to drought indices such as SPI and SPEI (Vicente-Serrano et al., 2010). These are however more precipitation and evaporation related. For certain drought impacts, hydrological forecasting could help to estimate drought conditions and potential impacts. However, to fully assess the influence of human interactions on drought development and impacts, water management should be included into these simulations as well as potential drought impact relationships or functions (Bachmair et al., 2017; Wendt et al., 2021). This information can be extremely difficult to obtain, however it is important to understand the local as well as the large scale interactions between water management and drought impacts. Especially in heavily managed multi-country basins these processes might be highly important and result in inequalities between upstream and downstream water availability (Van Lanen et al., 2016; Van Loon et al., 2016). For smaller cases, studies have used agent based modelling to understand the aspect of human decisions on drought risk and drought impacts. However, it can often be difficult to detangle and label impacts as droughts being a slow developing phenomena that can also lead to long lasting impacts. Therefore, recovery times might be too long to accurately link impacts to actions. Oftentimes the relation between water management and actual damage can only be assessed on short timescale events such as floods, where one event can have a rapid and devastating impact. Furthermore, the interdisciplinary aspect of drought impacts and the difficulty of prioritizing human actions to limit impacts relevant to the specific area is also difficult to understand (Di Baldassarre et al., 2021) or find information on as often expert knowledge is guiding current ways of planning and operation.

An interesting case area for drought management and drought impact mitigation in light of the past extreme droughts in Europe is the Netherlands. Usually known for its heavily managed water system targeted towards floods, the recent droughts of 2003, 2018, 2020 as well as 2022 have been challenging

the Dutch hydrological system (Bartholomeus et al., 2023). While the intensive water management system is one aspect that makes it interesting to study drought management, the large list of possible drought impacts as experienced in a densely populated area as the Netherlands is another. It only highlights the importance of an optimised water management strategy that can handle both floods and droughts. The list of drought impacts in the Netherlands includes reduced crop production, problems with drinking water supply, restricted intake of cooling water for power plants and discharge thresholds for hydro-power, limitations to shipping (reduced tonnage), as well as increased land subsidence rates due to altered groundwater levels and saline intrusion along to coast. Furthermore, prolonged drought conditions can also have negative impacts on the stability of dikes, which are a crucial infrastructure for flood prevention. These drought impacts, some of which are regional and some national, form the basis of a priority sequence which lists the priorities of supplying water to sectors and functions during a drought. Here, avoiding irreversible damage to infrastructure comes first, drinking water and energy supply second, expensive crops and industry third, and the remaining agriculture and shipping last.

Surprisingly, the operational water management decisions during a drought, complex as they are with a distributed water system and the priority sequence, are not aided by water resources models. While the complete Dutch water system is represented in a physically based integrated groundwater-surface water model (NWM, De Lange et al., 2014), the computational demands of this model are too high to allow its use for real-time forecasting and operational water management optimisation under droughts (Mens et al., 2021). Furthermore, detailed information of how water management decisions are made is limited. While a few operational plans for major infrastructures are available, these are based on information and estimates during the construction and development phase of these major water works. In reality these plans are not always followed in crisis situations such as major droughts. More often expert knowledge of Dutch water managers is incorporated, especially during drought extremes which were unprecedented.

In this chapter, the aim is to address these challenges and to create a pilot study to assess the potential of optimising water management practices to mitigate drought impacts in the Netherlands. This will be done by developing a three-part system: 1) a fast data-based model for hydrological simulations for a specific case area, 2) the definition of operational water management scenarios, and 3) defining specific impact functions. As an alternative to the current computational demanding modelling practices, a new approach using machine learning techniques will be tested and used. The application of machine learning in hydrology has been increasing over the past decade with very promising results. Benefits of using machine learning not only lie in the computational efficiency but also in the potential of exploration and handling of large amounts of data. These aspects make machine learning an interesting contender for efficient and accurate predictions as well as in regards to scenario exploration. As the Netherlands has an extensive monitoring network, numerous observations are available to study hydrological droughts. While observations regarding discharge are numerous, water management information is unfortunately limited especially in terms of operational handling. To be able to study the direct impact of water management operations on drought impacts, a set of water management scenarios will be created based on the variability of past observed operational management. These scenarios are used in a multi-target Long Short-Term Memory (LSTM) neural network model, which is specifically setup for three major river branches and one of the major infrastructures that affects the distribution of water between these rivers. These three branches are also known for different primary functions (e.g. shipping, water supply) in the Dutch hydrological system, therefore impact functions that correspond with these purposes were created that describe the potential drought impact. The impact functions are related to the discharge in these river branches, which can be simulated and altered through the water management scenarios used to run the multi-target LSTM model. The developed setup will explore past drought events and with a separate optimisation setup, the potential drought impact mitigation will be assessed.

The following Section 3.2, describes the case area, the setup of the machine learning model, the water management scenarios, the impact functions as well as the optimisation process. Results on the model performance, a first impact assessment based on the baseline scenarios and impact functions as well as the optimisation and mitigation of drought impact simulations follow in Section 3.3. Findings, realisations and suggestions are discussed in Section 3.4, followed by the conclusions in Section 3.5.

3.2 Material and Methods

In the following sections, details on the case area as well as the model and optimisation design can be found, covering model setup, water management scenarios and impact functions.

3.2.1 Case area and general idea

To be able to assess the potential of machine learning to optimise water management a specific use case in the Netherlands was chosen, which was previously already part of studies looking at the national scale (Hauswirth et al., 2021). It includes the main river branches fed by the Rhine and an associated water management infrastructures (Figure 3.1). More specifically, the focus area is looking at the river branches Waal, Nederrijn/Lek and IJssel. One of the first and most important sluices along the Nederrijn, a few kilometers away from the bifurcation of the Rhine into IJssel and Nederrijn/Lek, is the sluice Driel which is operated to divert the water to the different river branches and therefore different parts of the country. If the sluice is closed for a longer period, backwater effects can influence the discharge in the IJssel and have a minor effect on the Waal. The largest amount of the Rhine discharge however goes directly into the Waal without being influenced by the sluice at Driel. Water management reports list the discharge distribution over these three rivers as 2/3 going to the Waal, 1/3 to the Pannerdensch Channel which splits up into 2/3 to the Nederrijn/Lek and 1/3 to the IJssel.

While the Waal is important for shipping, the IJssel is crucial in supplying the IJsselmeer with freshwater. The IJsselmeer ultimately acts as a large water storage facility to prevent long-term drought impacts by providing enough water for irrigation purposes for agriculture for all its bordering provinces. While the Nederrijn/Lek also has minor shipping activity, a small hydro-power plant further downstream is dependent on enough discharge. Furthermore, combating saline intrusion further downstream at IJmuiden is also linked to discharge from the Nederrijn/Lek via the Amsterdam-Rhine channel. As these rivers cover different aspects regarding drought impacts in the

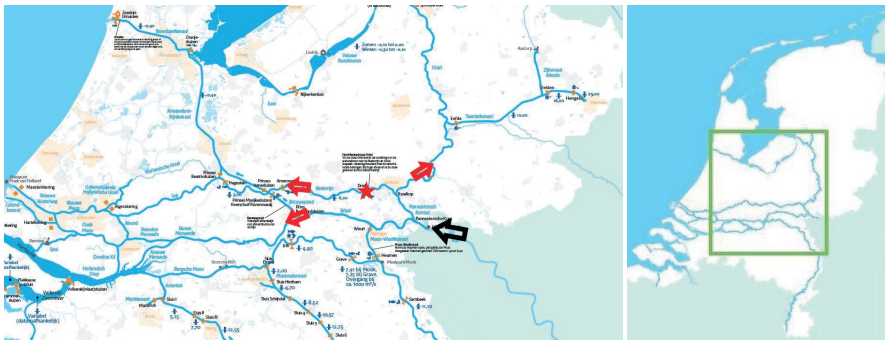


Figure 3.1 Focus area with overview at national scale. Red arrows highlight the different river branches included, while the star indicates the location of the sluice Driel. Incoming discharge of the Rhine is indicated by the black arrow. Figure taken and adapted from Rijkswaterstaat, ©Rijkswaterstaat, WMCN. September 2020.

Dutch system and are additionally linked to one of the main water management infrastructures, the exploration of machine learning and water management optimisation strategies to mitigate drought impacts was dedicated to this case area.

3.2.2 Multi-target model setup

To assess the potential of water management to reduce drought impacts, first a data-driven model was developed that is capable of simulating discharge in the three major river branches. Instead of a more classical approach that uses a recurrent neural network type Long-Short Term Memory (LSTM) trained on one single target variable, a multi-target approach was used here to train and later simulate the dependencies of the rivers to water management decisions. Target variables were discharge of the Nederrijn/Lek and the IJssel, while input variables included the discharge of the Rhine at Lobith and the discharge at Driel (sluice where operational decisions are implemented). To always ensure full closure of the water balance, it was decided to compute the Waal based on the discharge simulation of the other two rivers and use the inflow via the Rhine as the total sum of all three rivers. Here we assumed that no significant in or outflow takes place between the border at Lobith and the bifurcation points of these three major river branches. Observation records, ranging from 1985 to 2020, were split into training, validation and testing sets using a timeseries split approach to be able to evaluate the models performance prior to introducing water management scenarios. A separate loss function (using normalized Root Mean Square Error, RMSE) for training purposes was introduced, where the sensitivity of different weights for both rivers were tested. However, the best training results were obtained by keeping the weight equally distributed across the rivers. Evaluation criteria for the total framework for testing and validation included normalized RMSE (nRMSE), Kling Gupta Efficiency (KGE), and Pearson Correlation (R).

The input data was specifically chosen to only include Rhine discharge and observations from the location close to sluice Driel, as previous operational water management reports indicate the close link to operational decisions based on the Rhine discharge at the border. Therefore, to be able to implement water management the input for Driel is the only variable that needs to be changed.

3.2.3 Water management scenarios

The observational records from the location Driel were used to assess and develop the water management scenarios for the optimisation part of this study. In Figure 3.2 the observations at Driel are presented in relationship to the Rhine discharge at Lobith. The latter aspect is key, as current operational plans are guided and depend on the amount of discharge coming in at the border of the Netherlands. The scatter plots shows the relationship between Rhine discharge and discharge at Driel, which can ultimately be translated to the operational setting, closing or opening the sluice. For low flow periods (below 1500 $m^3 s^{-1}$ Rhine discharge), the sluice is operated at more conservative setting with trying to keep up a minimum flow. However, between 1500-2000 $m^3 s^{-1}$ the operational range varies as can be observed by the spread of the observation for Driel. With incoming discharge above 2000 $m^3 s^{-1}$, the sluice are operated at an open setting. Interestingly, the observations show a substantial difference compared to the operational plan, highlighted in red. Furthermore, changes in water management regulations can also be seen for different drought years in the right panel of Figure 3.2. One could argue that through the past drought years, the operational settings have slightly changed with the more recent drought years highlighting closing off the sluice at Driel already earlier while the Rhine discharge is still above 1500 $m^3 s^{-1}$. Earlier closures for example help with guiding the water to the IJsselmeer, which acts as a storage and a buffer to decrease drought impacts in the agricultural sector in the Northern part of the country.

For the water management scenarios, the operational range as seen in Figure 3.2 on the left side was taken into account. For every discharge value observed in the historical timeseries of the Rhine, 100 management options for Driel were picked out from the corresponding observation cloud. The

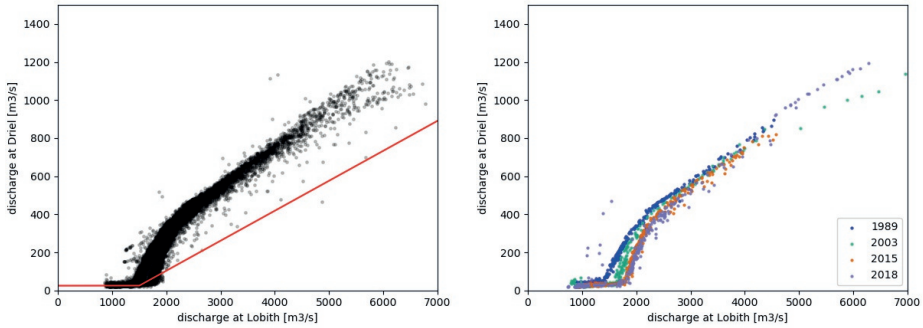


Figure 3.2 Observed discharge range at sluice Driel over the years in relation to the Rhine discharge entering the country at the border. The observed discharge translates to the operational range of the water management there . The red line represents the operational plan of the sluice taken from old water management reports (left panel). Observed discharge range at sluice Driel throughout different drought years 2003, 2015 and 2018 (right panel).

options were selected by first ordering the Rhine discharge and the corresponding observations at Driel. Second, a moving window with the size of $\pm$ 5 $m^3 s^{-1}$ Rhine discharge was used to narrow down the observations of the water management. The scenarios were then created by taking the lowest and highest value of the water management observations and picking out 100 values from a uniform distribution out of that specific range. This was repeated for every discharge value observed for the Rhine, which led to a input dataset with 100 management options for the location Driel per observed discharge value at Lobith.

3.2.4 Impact functions

Different impacts were defined on the river branches and their use to be able to create impact functions. The Waal is an important river branch for shipping, while the IJssel ensures that the IJsselmeer has enough storage available for agricultural use. The Nederrijn/Lek is connected to a small hydropower plant but more importantly also ensures combating saline intrusion further downstream at IJmuiden and the Hollandse IJssel. Various water management reports were consulted to see whether there were impact functions defined based on the previous drought events. However, concrete information regarding drought impacts on different sectors is limited and therefore basic threshold were used to define linear impact functions, ranging between 0 and 1 with 1 representing maximum damage. For the Nederrijn/Lek and IJssel previously defined thresholds were taken into account, such as a minimum flow requirement for the Nederrijn/Lek of 25 $m^3 s^{-1}$ and 285 $m^3 s^{-1}$ for the IJssel (Rijkswaterstaat, 2011; Rijkswaterstaat and van Waterschappen, 2019). For discharge values below these thresholds, a maximum impact value of 1 was given, while higher discharge values for higher discharges were linearly decreased until the value of Q_{50} of each branch. Above the Q_{50} impacts were assumed to be zero. For the shipping impact, the information from a water management report of potential shipping damage in relation to low flow was considered for inspiration and translated to a linear impact function that is based on a threshold of minimum water levels required (de Jong, 2019). To be able to translate the simulated discharge of the Waal to the shipping impact, the discharge was translated to water levels based on the observational relationship between those two variables. The linear impact functions to the corresponding rivers can be found in the Appendix B.1 in Figure B.1.

The impacts for every river were calculated based on the discharge simulations from the multi-target LSTM. For the impact assessment later on, the total impact sum of all rivers was considered and depending on the cumulative sum of the total impact was taken into account.

3.2.5 Optimisation process

For the exploration of the optimal water management to mitigate most of the potential drought impact, the total impact of all three river branches was taken into account. The impacts were calculated based on the discharge simulation results from the multi-target LSTM, which was run for every day for a few selective drought years. In terms of input data this meant that observations were included until the specific starting date of the optimisation window, starting from that date the timeseries was switched to the 100 scenarios (see Figure B.8 in Appendix B.2). The optimisation window was set to the following 30 days, resulting in 30-day discharge simulations of the different rivers and their impacts, and ultimately for every of these days the lowest cumulative impact was recorded with the corresponding scenario. This setup was repeated for every day of the selected drought years 2003, 2015 and 2018, resulting in an 'archive' of potential water management options and their corresponding impacts for the past few extreme drought years. For any day of interest, the specific simulation and impact results can be retrieved and assessed how the water management could have been adapted to mitigate potential drought impacts. The simulated discharge and impacts based on the historical observations over the total available period were used as a baseline for comparison. These simulations allow for an assessment of the highest potential drought damage reduction, given a perfect river discharge forecast. Furthermore, no retro-active forecast is used where the management decisions of previous days affects the options for the current day. While this would be closer to reality, the validation of such scenarios would be difficult and especially for longer lead times the initial starting time could have a significant impact on the final result.

3.3 Results

3.3.1 Multi-target LSTM: discharge simulations and model performance

The multi-target LSTM was evaluated on the testing set using evaluation metrics such as nRMSE, KGE and R. The evaluation scores were computed for both simulated rivers Nederrijn/Lek and IJssel separately but also as an average. The same evaluation scores were additionally also computed on the total simulated timeseries. The evaluation scores indicated a relatively good performance with nRMSE of 0.07 and 0.09, KGE of 0.95 and 0.92 and R of 0.98 and 0.95 for Nederijn/Lek and IJssel, respectively (Table B.1, Appendix B.2). The performance is considered good even though the hydrograph shows slightly lower performance for low flows likely due to having seen less occurrences of these in the training dataset (Figure B.2, Appendix B.2).

3.3.2 Water management scenarios

The scenarios created with the approach described in Section 3.2.3 give the opportunity to simulate alternative water management decisions based on the observed Rhine discharge. Figure 3.3 highlights the input timeseries for 2018, for both discharge at Lobith and the 100 scenarios at location Driel. While it seems the management options have a low impact for high flows and low flows, the operational range in the intermediate flow regimes shows a higher variation and bandwidth of options, which is ideal for studying the optimisation potential before heading into a low flow period for example. The smaller bandwidth of options is due to the selection process of the values for the highest and lowest scenarios, as these correspond to the limits of the operational range observed through the discharge observations at Driel. For high flow regimes the sluice is normally wide open, while for low flow it is normally fully

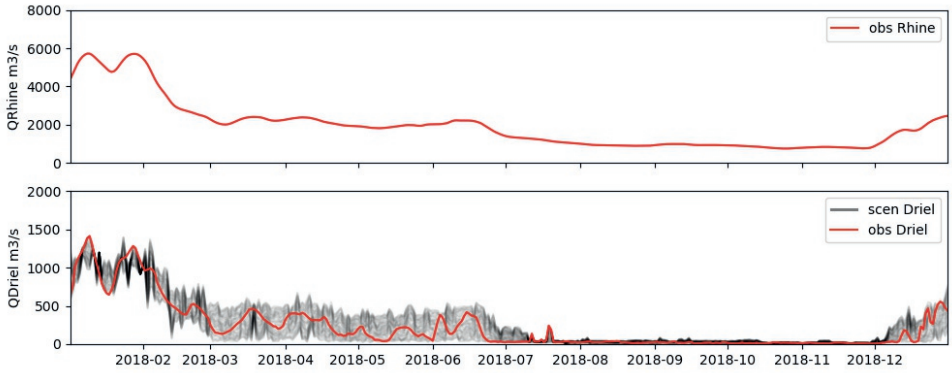


Figure 3.3 Input data example for the year 2018. Top panel shows the Rhine discharge, while the bottom panel represents the 100 timeseries created for the water management scenarios at Driel (black lines) and the actual observations at Driel (red). Every scenario, together with the Rhine discharge, are used to simulate one management version. Ultimately leading to 100 different discharge scenarios for every downstream location.

closed. This limits the potential management options as these are retrieved from the bandwidth in the observational record of the sluice operations.

The multi-day memory of the multi-target LSTM was evaluated in a synthetic experiment by altering the input timeseries of the observations with the highest and lowest scenario values for two consecutive days for a test period, to see whether and how much a rapid and substantial change of water management would influence the system and the simulation results. The simulation results for the three different river branches showed that the model has multi-day memory when it comes to these changes and that there are different behaviours of these river branches in terms of response time, ranging between 5-7 days until the system was back to a unified signal (Figure B.3). The model showing sensitivity to water management changes indicates that there is potential for optimisation at a later stage, when looking for impact mitigation.

Figure 3.4 in addition highlights the highest and lowest discharge values that could be simulated through all the scenarios, compared to the middle scenario and the observed discharge for the three different river branches for the drought year 2015. The lowest panel highlights the same information for the discharge at Driel, which is used as one of the input features. Especially for the Nederrijn/Lek it can be observed that the observations are often lying within the the highest and lowest scenario, indicating that there is room for potential improvement of the water management. This is linked also to the location of the observation station for that river, which is further downstream of the sluice Driel.

3.3.3 Impact functions and first impact assessment exploration

Based on the discharge simulations produced by the multi-target LSTM, the impact functions were used for the corresponding rivers to compute the potential impact, ranging between 0 and 1 for each river individually. Figure 3.5 shows the highest and lowest impacts calculated based on the corresponding simulations from Figure 3.4. The lowest panel presents the total impact over all three river branches.

Most impacts could be mitigated through alteration of water management in the case of the Nederrijn/Lek, while the other two river branches seem to be less severely and less frequently impacted by drought impacts. Most optimisation potential lies in the days/weeks right before a longer low flow period occurs, while during optimisation efforts are limited (see case IJssel). This is mostly due to the fact that in the progression from no drought impact (enough water) to maximum drought impacts, there is a transition period in which management could redistribute the water flow over the

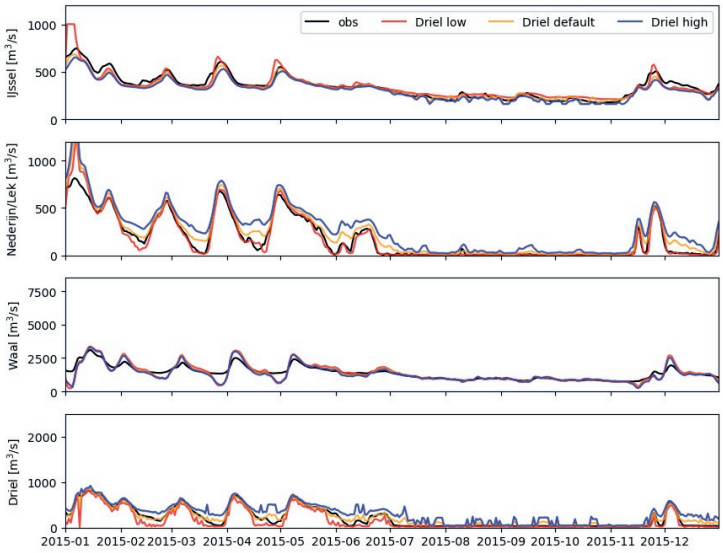


Figure 3.4 Baseline discharge simulations for 2015 based on the multi-target LSTM model for the focus area. Shown are the highest, default and lowest scenario results for the three different river branches but also the input at Driel used for the simulation (high corresponding to open management, low corresponding to more constrained management.

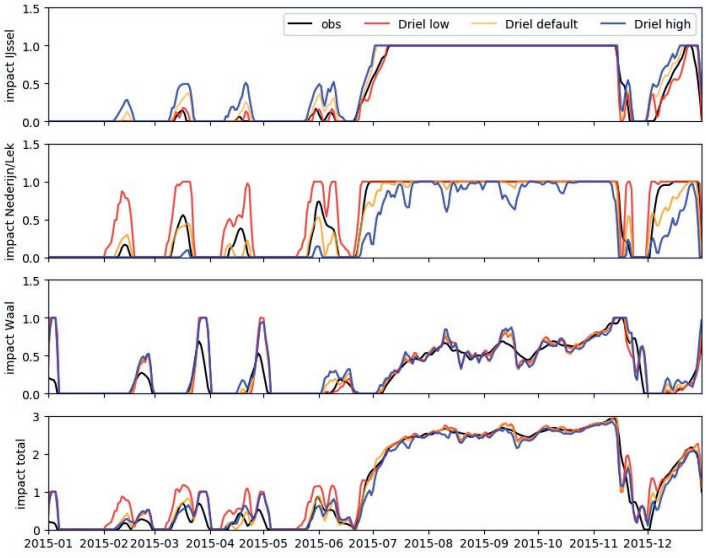


Figure 3.5 Baseline impact calculation for 2015 based on simulation output of the multi-target LSTM model for the focus area. Shown are the highest, default and lowest scenario results for the three different river branches but also the total impact in the lowest panel.

rivers in a way that reduces total drought impacts. While if the drought is at its worst, the potential for mitigation is limited to non-existing as there simply is not enough water to mitigate any impacts and all major river systems are affected. While the scenario with Driel being open shows high impacts for both Waal and IJssel, the impact results for Nederrijn/Lek indicate that the closing Driel will lead to the highest impacts there. This can be explained by the management at Driel having a direct impact on the Nederrijn/Lek, therefore by closing Driel and limiting the discharge, the impact further downstream will increase. While this might lead to some backwater effects where the IJssel can profit from, also the Waal responds positively to this management. In contrast, having an open management at Driel shows potential for high impacts for both of these rivers. In terms of the total impact, a switch between opening/closing sluice scenarios at Driel can be observed throughout the year 2015. While the shorter impact periods in the beginning of the year indicate a clearer distinction between scenario choices, once hitting a longer period of low flows the management options seem to be limited.

3.3.4 Optimisation of water management to mitigate drought impacts

To analyse the optimisation potential of water management decision to limit drought impacts, the optimisation design described in Section 3.2.5 was as followed for the selected drought years 2003, 2015 and 2018. This optimisation setup lead to the creation of an archive of past water management options and potential improved management decisions.

In terms of how the optimisation simulation and selection of scenarios per lead time differs, Figure 3.6 gives an insight on a optimisation window for mid May 2015. While the grey lines represent the different impact simulations based on the scenarios, the coloured lines indicate the default scenario (orange), observed simulated (dotted black) and the optimal scenarios (red) depending on lead days (the different scenarios and lead days are represented by separate lines that stop at 1, 7, 14, or 30 days). The same description applies to the lower panel of the figure including the cumulative total impact. The optimised water management is represented with the red dotted line, which is both lower than the default scenario but also the observed simulated impact. The difference between the latter and the optimised management indicates the mitigation potential.

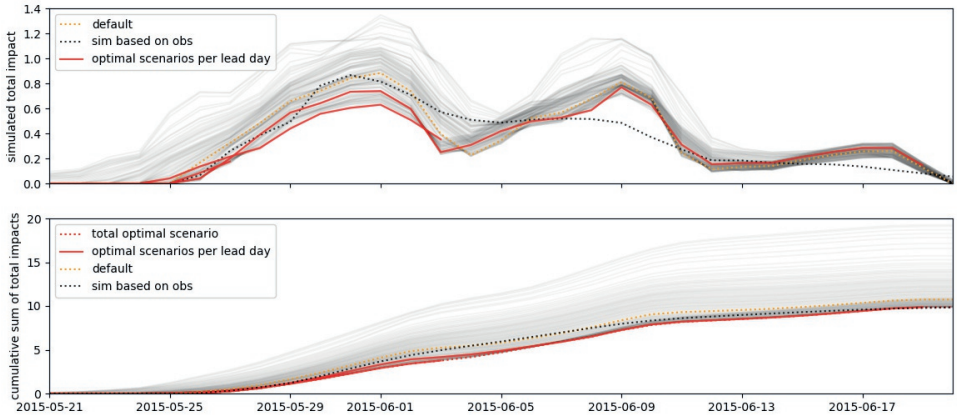


Figure 3.6 Optimisation example for impact simulations mid May 2015. Top panel shows the simulated total impact for the next 30 days in grey, the default scenario in yellow, the simulated impact based on observations in black as well as the optimal scenarios in red for different lead times (the shorter the lead day, the earlier the red line stops). Same for the cumulative total impact on the lower panel.

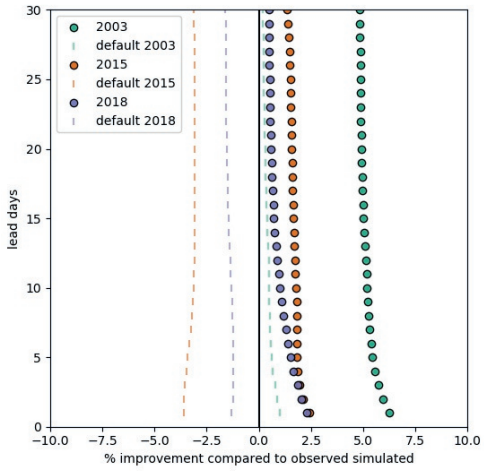


Figure 3.7 Total impact per lead days for different drought years. The percentage of mitigation potential compared to the observed simulated impact is highlighted for both the optimised simulations and default scenario.

With the optimisation window of 30 days and reporting the best scenario with the lowest cumulative total impact for every day, the total impact per lead day for the selected drought years could be assessed. The potential mitigation/improvement compared to the observed (simulated) impacts is shown in Figure 3.7, for both the optimized management and the default management scenario.

The different coloured dots represent the percentage of improvement per lead day for different drought years 2003, 2015 and 2018. What can be seen is that for most of the simulated impacts, the optimal scenarios show a reduction in total drought impact compared to the observed drought impact. Especially in the early lead days (starting with 2.5% for 2018 and 2015 and 6% for 2003), the use of adaptive optimal water management has benefits over default operations. It is important to note that for the longer lead days the optimal management scenario is fixed over the entire simulation window. This makes that longer lead days suffer from more static water management and thus it can be expected that reduction in impacts might be sub-optimal. Nonetheless, they still show a reduction in drought impacts compared to the observed management scenario. If the average management (in this case scenario 50/default scenario) would have been followed during these years, the improvement would be drastically less (2003) and even deteriorating for some years (2015 and 2018). This shows that to truly reduce the impacts, adaptive management is required for an optimal gain. Interesting to observe is that the initial reduction in drought impacts for the optimised water management is lowest for the more recent years indicating less potential for improvement. This might be the result of changes in the water management since the extreme drought event in 2003. Figure 3.2 already showed that the water management has already changed since 2003 and in addition some preventative management actions might have been taken earlier for the more recent droughts.

3.4 Discussion

In this chapter a first attempt was made to assess and explore the potential of water management scenarios in a machine learning model to potentially mitigate drought impacts in the Netherlands. This case study is characterized by three major river branches, all connected to the river Rhine and affected by one important infrastructure distributing river water throughout the country. In terms of modelling setup, the commonly used LSTM was taken as main model but contrary to the general approach of one target variable, a multi-target variable approach was chosen. This enabled the model to learn the relationship for both target variables/rivers in regards to observed Rhine discharge and past water management behaviour, represented by the observations at the sluice infrastructure at Driel. Water balance closure was ensured by simulating the rivers (IJssel and Nederrijn/Lek) and computing the discharge of the Waal by subtracting this simulated discharge from the Rhine discharge. Given the size of the average discharge at Lobith of 3000 $m^3 s^{-1}$ and the limited contributing catchment area between the Rhine entering the Netherlands to the locations studied it was assumed that there was no significant in or outflow occurring throughout the case area of these three rivers and that the water balance could be closed in this way.

The initial model evaluation showed good performance based on the evaluations scores, however some limitations regarding very high or low flow periods were observed, which are likely linked to limited data samples available for training of the model. The memory component, which is one of the characteristic of the LSTM compared to other machine learning methods, showed to have an impact on the longer term discharge changes as a result of water management decisions.

These water management decisions were represented by a newly developed set of scenarios, which were based on the observed operational decisions following the inflow of the Rhine, that have been performed at the sluice Driel over the past 30 years. By focusing only on the observations from the past, extreme scenarios that might not have occurred yet are currently outside of the scenario range. Furthermore, in the initial setup for this chapter, no additional source of uncertainty or perturbation to the scenarios has been included. For example the forecast uncertainty has been neglected due to the use of perfect inflow forecasts. This allows for a quantification of the maximum potential gain from using this technique. Would the forecast uncertainty be included this might be reduced, but for a fair comparison the model should then be exposed to the exact operational information available to the water managers at the time. As this has proven to be difficult, it was decided to use perfect forecasts. Furthermore the retroactive aspect of the forecast is not included due to the high computational demand and the difficulty for a direct comparison with the observed operations. Including these aspects in a future research would make the water management aspect more realistic and introduce a larger pool of scenarios that could be used to analyse drought impact mitigation.

Defining drought impact functions to simulate and assess mitigation strategies proved to be more difficult than expected. While the past few extreme drought events lead to a long list of reported impacts that gained large attention from media and society, quantification of individual impacts, in monetary terms or otherwise, proved to be difficult or non-existent. This is a pity because actual reported damages could be used to construct the damage or impact functions greatly needed when optimizing operational water management for drought mitigation. In addition, many of the drought management adjustments are not documented in publicly available archive making it hard to infer the drought mitigation actions that have been taken and use these in drought impact models. Therefore this study relied on standard water management reports from the last few decades and the focus was put on potential discharge thresholds and potential rudimentary impact functions that were available from management plans. The three river branches all serve a unique purpose before and during drought events, therefore the impact functions created in this study were tailored to these aspects (shipping on the Waal, additional water storage of the IJsselmeer via the IJssel for agricultural purposes and hydropower production as well combating saline intrusion via the Nederrijn/Lek). The

current impact functions are simplified, linear and equally weighted, which is likely not the actual representation of the impacts and their development. However, these assumptions and simplification were necessary to establish a first example study, as updated information and understanding of recent events is limited and to stress the need for such information to actually improve understanding and preparedness of the whole system for upcoming droughts.

While many studies have been dedicated to drought analysis and drought impacts, a common finding is that the drought impact definitions and drought impact functions can be difficult to define be it due to perception, limited data but also local differences. Even though drought impact monitoring has been getting increasing attention over the years, it can still be difficult to define distinctions between direct and indirect impacts and capturing them adequately (Blauhut, 2020; Blauhut et al., 2022). Yield loss, drinking water deficits, shipping and production limitation in industry sector, thresholds for nature, are all very localized and not necessarily unified or streamlined throughout Europe or linked to a specific drought indicator (Blauhut et al., 2015). Furthermore, drought being a slow developing phenomena can also lead to long lasting impacts. Therefore, recovery times might be long and the actual damage can only be assessed with time. In addition, the interdisciplinary nature of drought impacts and the difficulty of prioritizing human actions to limit impacts relevant to the specific area is also challenging to obtain information about. Having an extensive drought impact inventory (e.g. European-Drought-Centre, 2013, Stahl et al., 2016) might also help to address and detangle human influences by looking at it coming from the impact side.

In terms of optimisation, the current approach allows to study the optimisation potential of past (drought) years by having a 30-day simulation of a 100 scenario ensemble for every day during selected drought years. While the average management scenario showed minor improvement compared to simulated impacts based on observations for 2003, middle of the road water management would have led to even more impacts following the same approach. In comparison, the optimized water management simulation results showed a mitigation potential of 2-6% for the first few lead days. This improvement is can only be obtained during the transition periods from normal to severe drought conditions as during normal conditions drought impacts are zero and during severe drought there is a maximum impact in all river branches. What has to be kept in mind is that during these extreme drought events, the actual water management actions are normally determined by expert knowledge from the drought water management committee (in Dutch Landelijke Commissie Waterverdeling, LCW) which renders normal operations rules useless. This could also be seen in the optimisation example, where the observed simulated impact lied outside of the scenario ranges for some of the days in the historic record (e.g. May 2015 as seen in the optimisation example). For future assessments, a retroactive approach could be interesting to look at the continuation of water management during drought events and its long-term effects. A big difference then is that the best scenario for the next lead period for a given day may depend on the scenario choices (settings) of the previous days, which would more resemble model-predictive control (Aydin et al., 2019). However, this would ultimately result in a much more computational intensive setup, especially if the input data would also be extended to include uncertainties or perturbations to allow for a more realistic operational setting and decision process.

Nevertheless, this study has shown that LSTM based models pose an interesting alternative or supplement to current modeling approaches, especially if scenario simulations are involved. LSTMs have the added benefit that they are capable of simulating the long-term memory and impact of ad-hoc water management decisions over more simple machine learning based models. Due to their capabilities to handle large amounts of information and their reduced computational efforts, using these techniques for testing various potential settings and outcomes for water management interests is beneficial. Computational demands and model restriction for water management assessments were

also listed as one of the main limitations in Miro et al., 2021, where they combined groundwater modeling, machine learning and robust decision making to improve groundwater management.

One of the most challenging aspects here, but also in general for machine learning studies, is the data availability. Especially for this study, where water management is an essential part of the features included, the limited detailed information available creates a simultaneously constraining and challenging situation if one wants to fully address the potential of water management operations for impact mitigation.

Nonetheless, by combining the aspects of machine learning, water management and drought impact functions, although simplified, in this study, the potential for simulating drought impact mitigation through management changes could be provisionally quantified. With increased understanding from the past extreme events, hopefully more data collection and quantitative assessment of drought impacts, and feedbacks from human responses (e.g. water management), the potential for forecasting and use in operational setting could be even more increased and facilitated. These additional types of informed forecasting next to the current use of more broadly available drought forecasting methods (e.g. SPI and SPEI, Sutanto et al., 2019), could create additional, more informative and supportive information at scales more relevant for operational management and ultimately increase the preparedness for future drought events.

3.5 Conclusion

With droughts being multifaceted phenomena, leading to a large variety of severe impacts, as seen through past events in Europe over the last decades, increasing the preparedness for future drought events and related impacts is more important than ever. Especially as the frequency and intensity of drought events are likely to increase due to climate change. In order to increase the preparedness at a local scale, there is a need to understand the interaction between drought development, operational water management decisions and drought impacts. This study addressed the potential of mitigating drought impacts in the Netherlands, which is known for its extensive water management, being increasingly challenged by drought events and a large variety of drought impacts related to these events. By developing a data-driven hydrological modelling framework, including explicit water management scenarios, the direct influence of water management options was assessed for three different rivers for which drought impact functions were obtained to evaluate the potential impact mitigation. With the modelling framework being sensitive to water management changes, the optimal water management scenarios could be assessed for past drought events. An improvement of 2-6% could be observed compared to observed impacts. The observations showed an improvement of the actual operational water management responses after each subsequent major drought event. This was highlighted by a relatively large improvement of the optimised water management over the observed management for 2003, but smaller improvements for 2015 and 2018 in that order. Additional inclusion of uncertainty in the forecast experiment and potential retroactive optimisation techniques would extend the framework and creating a more realistic management and mitigation scenario for future work. While the current assessment already hints at the potential of combining the three pillars, machine learning, human action and drought impact functions, for optimal drought impact mitigation, a major challenge remains the data availability of water management information and the derivation of drought impact functions from past drought events and the limited impact observations. To further improve the potential of drought impact mitigation more quantitative data on drought impacts are required.


4

Chapter 4 | The suitability of a seasonal ensemble hybrid framework including data driven approaches for hydrological forecasting

Abstract

Hydrological forecasts are important for operational water management and near future planning, even more so in light of increased occurrences of extreme events such as floods and droughts. Having a forecasting framework, which is flexible in terms of input forcings and forecasting locations (local, regional or national), that can deliver this information in fast and computational efficient manner is critical. In this study, the suitability of a hybrid forecasting framework, combining data-driven approaches and seasonal (re)forecasting information from dynamical models, to predict hydrological variables was explored. Target variables include discharge and surface water levels for various stations at national scale with the Netherlands as focus. Five different machine learning (ML) models, ranging from simple to more complex and trained on historical observations of discharge, precipitation, evaporation and sea water levels, were run with seasonal (re)forecast data including European Flood Awareness System (EFAS) and ECMWF seasonal forecast system (SEAS5) of these driver variables in a hindcast setting. The results were evaluated using the evaluation metrics Anomaly Correlation Coefficient (ACC), Continuous Ranked Probability (Skill) Score (CRPS and CRPSS), and Brier Skill Score (BSS) in comparison to a climatological reference hindcast. Aggregating results of all stations and ML models revealed that the hindcasting framework outperformed the climatological reference forecasts by roughly 60% for discharge predictions (80% for surface water level predictions). Skilful prediction for the first lead month, independently of initialization month, can be made for discharge. The skill extends up to 2-3 months for spring months due to snow melt dynamics captured in the training phase of the model. Surface water levels hindcasts showed similar skill and skilful lead times. While the different ML models showed differences in performance during a testing and training phase using historical observations, running the ML framework in a hindcast setting showed only minor differences between the models, which is attributed to the uncertainty in seasonal forecasts. However, despite being trained on historical observations, the hybrid framework used in this study shows similar skilful predictions as previous large scale forecasting systems. With our study we show that a hybrid framework is able to bring location specific skilful seasonal forecast information with global seasonal forecast inputs. At the same time our hybrid approach is flexible and fast, and as such a hybrid framework could be adapted to make it even more interesting to water managers and their needs, for instance a part of a fast model-predictive control framework.

Based on: Hauswirth, S. M., Bierkens, M. F., Beijk, V., and Wanders, N. (2022). The suitability of a seasonal ensemble hybrid framework including data-driven approaches for hydrological forecasting, *Hydrology and Earth System Sciences, pp. 501-517*, https://doi.org/10.5194/hess-27-501-2023

4.1 Introduction

Forecasting in combination with local system knowledge plays an important role in increasing the readiness for imminent extreme events such as floods and droughts. Especially over the last few years, where the effects and impacts of climate change have become more and more distinct with an increasing recurrence of extreme events require adaptive planning based on skilful forecasts. For instance, knowledge of upcoming water surplus or shortage is important to limit damages to infrastructure and impacts on society during floods as well as to increase and sustain the water availability prior but also during droughts.

Over the past years, platforms of openly available forecasting services and data sets that deliver meteorological and hydrological predictions have increased. These data sets can differ in leadtime (e.g. short- to medium range, sub-seasonal, and seasonal time scales) and include uncertainty by consisting of various numbers of ensemble members. Examples of openly available seasonal forecasting system are the operational European Flood Awareness System (EFAS, Thielen et al., 2009; Arnal et al., 2018) and the Global Flood Awareness System (GloFAS, Alfieri et al., 2013), as well as ECMWF latest seasonal forecasting system SEAS5 Johnson et al., 2019. Forecasting systems like these are run by large scale, physically based models (e.g. Lisflood Van Der Knijff et al., 2010; De Roo et al., 2000 in case of EFAS), which require a lot of information regarding parametrization, can be slow and computational intensive as well as require large data storage facilities. Another example of forecasting systems facing similar challenges are multimodel ensemble systems, which combine several general circulation models and hydrological models (Wanders et al., 2019; Samaniego et al., 2019).

Even though continuously improved in terms of ease of use and interoperability, the main challenges of handling such large, computational and data intensive systems remain. Furthermore, Samaniego et al. (2019) highlighted that in case forecasting systems are used for decision making, prediction horizons, spatial scales, model choices, storage and computational requirements and reported variables can limit the applicability of forecasting systems to local water management. We hypothesize that incorporating data-driven approaches to support seasonal forecasting systems can be beneficial not only in terms of reducing computational requirements but also their flexibility and data use. Especially, if the forecasting system can be kept simple, for example regarding input forcings or the complexity of forecasting setup, the threshold of applying it on various spatial and temporal scales would be even further lowered. This would for example bridge the gap from large scale to local forecasting systems and make it more readily applicable to create efficient operational settings and support local water management. In this study we explore these opportunities by incorporating data driven approaches in a seasonal forecasting framework, combining both local and global information.

Data driven approaches, including machine learning (ML) models, have been explored and tested out increasingly in hydrological assessments over the last few years (Shen, 2018; Shen et al., 2021), either as standalone models or also in hybrid-settings (coupled with physically based models, Kratzert et al., 2018; Koch et al., 2019). ML can be used for any spatial and temporal scale study, as long as there is sufficient data available for training and validation. Besides using local observations and remote sensing information an upcoming trend is also to incorporate knowledge based learning (Koch et al., 2021), where ML models are also trained with information provided by physically based model or in hybrid model setups. ML has shown to be promising in simulating hydrological variables such as discharge and groundwater levels but also in contributing to operational water management.

However, most of the previous research focused on successfully simulating past observations or current hydrological states but incorporating ML in a seasonal forecasting framework has only scarcely been explored in the hydrological field. Work by Hunt et al., 2022 being one of the most recent examples, where LSTMs were explored in a hybrid forecasting setup to predict discharge for short term scale.

A substantial issue as to using ML for seasonal forecasting is often the limited amount of samples for training. This is often resolved by including long climate model simulations for training as an example, however depending on the scale and resolution these might not always be ideal for more local studies. Nevertheless, if sufficient local data are available, it is worthwhile investigating how one can exploit the assets of ML for seasonal forecasting (limited complexity of forecasting setup, computational demand and handling of large data amounts) to increase the support for water management. This can be especially useful for floods but also drought occurrences, where local information has to be available and updated within a short time frame (floods) and changes in water management planning have to be reassessed both ahead and in time of an event to optimise water availability (droughts).

To be able to have such a forecasting system that can support water managers as an example, a first step is to build a framework that can be explored in a so called hindcasting experiment, where the forecasting framework is tested based on historical observations in near real-time. Once this is successful, the forecasting framework could be switched to seasonal forecasting. The aim of this study is to explore the first step: test the suitability of a hybrid forecasting framework in a hindcast experiment. The framework will build on an existing ML model framework based on a previous study by Hauswirth et al., 2021 in combination with (re)forecast information as input variables. In this study we want to test the suitability of these models based on their historical performance to forecast discharge and surface water levels in a hindcast setting.

The ML models (trained on historical observations), will be run with seasonal (re)forecasting data replacing the previous input dataset consisting of discharge, precipitation, evaporation and sea water level observations. Running the models with seasonal (re)forecasting information creates an ensemble of the target variables for each station of interest. The focus of the target variables is laid on discharge and surface water levels, the latter including rivers, streams and lakes. These ensembles will be analysed and compared with historical observations to assess the skill of the ML framework for seasonal forecasting for general hydrological conditions but also extreme events such as droughts by computing different skill scores common to evaluate seasonal forecasting frameworks. Furthermore, the benefits but also challenges of such a simple setup will be explored and listed to assess whether the framework is suitable for current practices and whether it opens up possibilities for future assessments.

In the following sections the study's approach will be laid out, followed by an evaluation of the performance of the hindcast experiment by assessing skill scores both for general and droughts. Thereafter, the findings will be summarized, discussed and put into a bigger context of the field, followed by the main conclusions.

4.2 Material and Methods

This section is divided into subsections covering the general concept of the hybrid hindcast framework used in this study, the seasonal (re)forecast data and its preparation, the data-driven model setup as well as the skill scores used to evaluate the forecasting skills of the hybrid framework.

4.2.1 Hybrid hindcast framework

The hybrid hindcast framework used in this study can be described as a simple three block system including: pre-processing of the input data, a main model block including a selection of ML models, as well as post-processing of the target variable and skill score calculation (Figure 4.1). The main model block, which is based on the ML model study of Hauswirth et al., 2021, was run on seasonal (re)forecasting information, which has first undergone the pre-processing block. The hindcast results were evaluated based on different skill scores included in the post-processing block to assess how

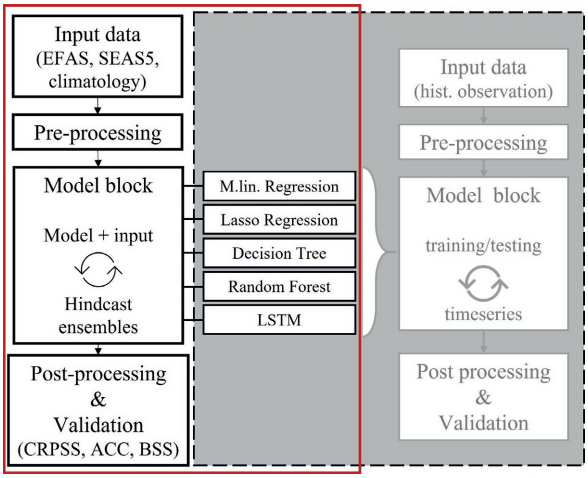


Figure 4.1 Schematization of hybrid hindcast framework highlighted in red frame, grey box indicating model framework previously developed by Hauswirth et al., 2021 and used as base for the model block.

skilful the hybrid framework is in hindcasting historical observations. The spatial and temporal setting of the hindcast experiment is focusing on the Netherlands and the time period 1993-2018. Target variables include discharge and surface water level in freshwater bodies for selected stations (69 for discharge, 97 for surface water level) throughout the observation network of the National Water Authority.

Data and pre-processing

The input dataset based on seasonal (re)forecasting information covers the period 1993-2018, including a lead time of 7 months (215 days) and consisting of 25 ensemble members (50 from 2017 on). The input dataset replaces the previously used historic data of the ML framework, which consists of a simple set of variables including discharge, precipitation, evaporation, and sea level observations at specific locations of the case study area for the time period 1980-2018 (Figure C.1). These variables were chosen as they are part of the observational network of the National Water Authority and readily available. Furthermore, the main model block as defined in Hauswirth et al. (2021) was designed with the idea of being flexible in the sense that input datasets could be easily exchanged by datasets representing the same variables (e.g. seasonal (re)forecasting data). The seasonal (re)forecasting data was obtained from the forecasting systems SEAS5 and EFAS, accessed via the openly available data platform Copernicus Climate Data Store. SEAS5 is ECMWF's fifth generation seasonal forecasting system, providing predictions on atmosphere, ocean and land surface conditions (Johnson et al., 2019). Meteorological information including precipitation, u and v wind components, 2m temperature, surface net solar radiation, and mean sea level pressure were gathered from SEAS5 and were taken from the grid cell that included the original observation location. The latter three variables were used to calculate the Makkink reference potential evaporation (de Bruin and Lablans, 1998). Seasonal reforecasting information on discharge was obtained from the European Flood Awareness System (EFAS, (Thielen et al., 2009; Wanders et al., 2014; Arnal et al., 2018)), a pan-European seasonal hydrological forecasting system which is based on the Lisflood model (5x5km resolution), with SEAS5 as meteorological forcing. The same approach of selecting the grid cells including the stations location of the original input data was done. Having SEAS5 as a forcing for both the EFAS reforecasts of discharge and as source for the meteorological input data for the ML framework enables a consistency in terms of model forcings.

For the sea level data, the water level for the historic period was simulated by using the pytide python module. The difference between the observed and pytide predicted sea level fluctuations (including only tidal components) was computed to get the anomalies. A simple multi-linear regression model was then used to compute the final sea level ensemble set based on the u and v wind speed and the previously computed sea level anomalies.

The input dataset was processed in a similar manner as done by Hauswirth et al., 2021, for example by including the lagged time series of every input variable. This was done in the previous study by using the partial autocorrelation function (PACF) to identify and incorporate significant information content that could explain the historic patterns and be additionally fed to the machine learning in its training phase (Hauswirth et al., 2021). In this case the seasonal (re)forecasting data in a first step was bias corrected using cumulative density function (CDF) matching approach before extending the input variable by incorporating the lagged times series approach (Wanders et al., 2014).

We additionally prepared an input data set including water management influence using the same approach as in Hauswirth et al., 2021. This simulation includes operational rules of main infrastructures which are related to the Rhine discharge at Lobith (one of our main input variables) for two specific input locations and two additional observation records of locations based at smaller infrastructures (Figure C.1). We were therefore able to use the same approach regarding the operational rules for the main infrastructures, as these are based on the Rhine discharge we obtain from the EFAS dataset. For the two other additional timeseries climatology was used as operational plans were not available.

For every ML model (representing a station of interest) the input data set after data pre-processing finally consists of a set of ensembles, including ensemble members of all input variables. The models were run with every set of ensemble members (e.g. input dataset based on first ensemble members of discharge, precipitation, evaporation and sea level information).

Data-driven model setup

We applied here a recently developed ML model framework by Hauswirth et al., 2021. This framework has a focus on a simple setup using only readily available input data to simulate target variables such as discharge, surface water level (including rivers, streams and lakes), surface water temperature, and groundwater levels for several stations at a national scale in the Netherlands. The station information and observational records were taken from the national monitoring network and covered 69 discharge, 97 surface water level, 105 surface water temperature, and 4000 groundwater stations (Figure C.1 for discharge and surface water level stations). For every station of interest a ML model was trained on historic observations of the target variable and the input dataset, consisting of the five variables: precipitation and evaporation from the deBilt, Rhine discharge at Lobith, Meuse discharge at Eijsden, sea-level observations close to Haringvliet Dam (Figure C.1). Furthermore, the framework is able to incorporate the influence of water management aspects. This was done by expanding the input dataset with discharge timeseries of the most important infrastructure, based their operational rules which is linked to one of the main input variables. Different ML methods were trained and tested based on a 60/40% train-test split including timeseries segments which were chosen randomly. The ML methods incorporated in the study range from simple to more complex methods including: Multi-linear Regression (MReg), Lasso Regression (Lasso), Decision Tree (DT) and Random Forest (RF), as well as Long Short Term Memory (LSTM) models. For more information regarding the specific models, model setup steps and evaluation, as well as data pre-processing we refer to Hauswirth et al., 2021.

For this study the pretrained models were rerun based on a prepared seasonal (re)forecast input dataset. We decided to test out all the original ML models to see, whether similar observations regarding their performance and differences could be made. The input dataset is made up of the same variables as in the previous study but taken from seasonal (re)forecasting datasets such as EFAS and

SEAS5. The models were not retrained, so the input data was used for an extensive validation of the simulation of seasonal forecasting skill. Using the pretrained models has the benefit of saving computational time, which would have otherwise been needed for testing and training the models. Secondly, this study aims to test the suitability of this ML framework for seasonal forecasting in an operational setting. In such an operational setting one would like to keep a consistent modelling framework that has been validated on an extensive hindcast archive. On the other hand, not retraining the models on ensemble datasets limits the potential improvement the model could experience by seeing forecasting data in the training phase. As we want to test the suitability of the developed ML framework for hindcasting, we are putting a focus on the pretrained model in combination with the seasonal (re)forecast input dataset. This allows us to test the performance of the models based on information that the models have definitely not seen before. Running the model based on a seasonal (re)forecast input dataset, consisting of several ensemble members, creates an ensemble of time series for the target variables discharge and surface water levels. Same amount of ensemble members (25 members, 50 members from 2017 on) and same lead time (215 days) as the seasonal (re)forecast input data were generated for the target variables. These ensemble simulations of the target variables were then analysed by computing frequently used skill scores.

4.2.2 Evaluation - Skill scores

To evaluate the performance of the hybrid hindcast framework, the target variables were compared to observations at daily, weekly and monthly temporal scale using different skill scores. These skill scores are shedding a light on various aspect of hindcast skills, for example overall performance, accuracy and reliability. Skill scores that are common in the forecasting community and are used here include: Continuous Ranked Probability (Skill) Score (CRPS and CRPSS), Brier (Skill) Score (BS and BSS), as well as Anomaly Correlation Coefficient (ACC). In this study, the hindcasts (including 7 months of lead time) will be addressed by their lead time, where lead one equals the first months of the forecast in which it was initiated, lead two equals the second month after initialization, etc. For the results, the focus will be on weekly and monthly averages, time scales which would also be of interest to water managers for mid- to long-term planning.

CRPS and CRPSS

The CRPS, which is one of the most common used evaluation benchmarks used in ensemble forecasting studies (Pappenberger et al., 2015), was used to assess the overall performance of the hindcasting framework. It compares the differences in the hindcast and observed Cumulative Distribution Functions (CDF) and ranges from 0 to infinity. The lower the computed score, the better the performance of the hindcasting framework (Arnal et al., 2018; Pappenberger et al., 2015). Equation 4.1 taken from Hersbach, 2000 (where P(x) is the cumulative density function of the hindcast and P_a observation probability) is computed) was used and the CRPS computed over all ensemble members for each lead day of every hindcast before aggregating it to other temporal scales.

As a skilful benchmark (baseline) we also compare the hindcast framework with a forecasts based on the historical distribution of observations. In other words, for each forecast day we look at the historical observations for that day and select values for all the years of this historical observations to generate an observation based climatological hindcast ensemble. Both the hindcast CRPS and the baseline CRPS were used to compute the CRPSS (Eq. 4.2). The CRPSS range lies between 0 and 1, with 1 indicating the forecast giving the best performance compared to climatology and 0 having no skill compared to climatology.

$$CRPS = CRPS(P, x_a) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 dx \tag{4.1}$$

$$CRPSS = 1 - \frac{CRPS_{hindcast}}{CRPS_{baseline}} \tag{4.2}$$

BS and BSS

To determine the accuracy and the performance of the hindcasts for simulating high and low flow periods the BS (Brier, 1950) and BSS can be used. To assess these specific categories, thresholds can be defined e.g. the lowest 20^{th} percentile data to account for droughts similar to other studies (Van Loon and Laaha, 2015). This allows one to analyse events which are either higher or lower than the usual observations for a given month (Candogan Yossef et al., 2017; Wanders and Wood, 2016). This threshold was used for both hindcasts and observations, before aggregating the data to the temporal scale of interest and computing the BS. The BS is calculated by Eq. 4.3, where N equals the number of hindcasting instances, f and o are the hindcast and observed probability of exceeding a threshold, respectively (Candogan Yossef et al., 2017). Score values range between 0 and 1, whereas 0 is indicating the best performance.

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \tag{4.3}$$

$$BSS = 1 - \frac{BS_{hindcast}}{BS_{ref}} \tag{4.4}$$

Furthermore the BSS (Eq. 4.4) can be used to compare the accuracy and performance of the hindcasting framework compared to a reference system, climatology in this case. Same range and interpretation can be used as for CRPSS.

ACC

To measure the quality of the hindcasting framework, the Anomaly Correlation Coefficient (ACC) is computed by using Eq. 4.5, where (f_t) represents the hindcasts, (o_t) the observations, while $\overline{f}$ and $\overline{o}$ are the longterm averages.

$$ACC = \frac{\sum_{t=1}^{N}(f_t - \overline{f})(o_t - \overline{o})}{\sqrt{\sum_{t=1}^{N}(f_t - \overline{f})^2 \sum_{t=1}^{N}(o_t - \overline{o})^2}} \tag{4.5}$$

The hindcasts and observations were first aggregated to the temporal scale of interest before computing the ACC. The ACC helps to verify the hindcast and observed anomalies, compared to the normal correlation where seasonality can influence the calculation results. Therefore, the ACC can also be seen as skill score in comparison with the climate. The ACC score ranges from -1 to 1, with 1 representing a perfect correlation between observations and forecast. For representation purposes the significance level was computed based on the number of observational years (in general) and only stations with less than 10 missing observation months were considered (same criteria of station selection was used for the other scores).

4.3 Results

The results will be presented such that first an overview of the general performance of the hindcast framework for one target variable, the discharge hindcasts, will be given. Subsequently the focus will be directed towards one model (Random Forest, RF) and an example station to provide a more in-depth insight into the different evaluation scores and differences in temporal resolution. The scores were calculated for all the initialization months of the hindcast and for different temporal resolutions (daily, weekly, monthly). For demonstration we highlight the performance of the hindcast framework

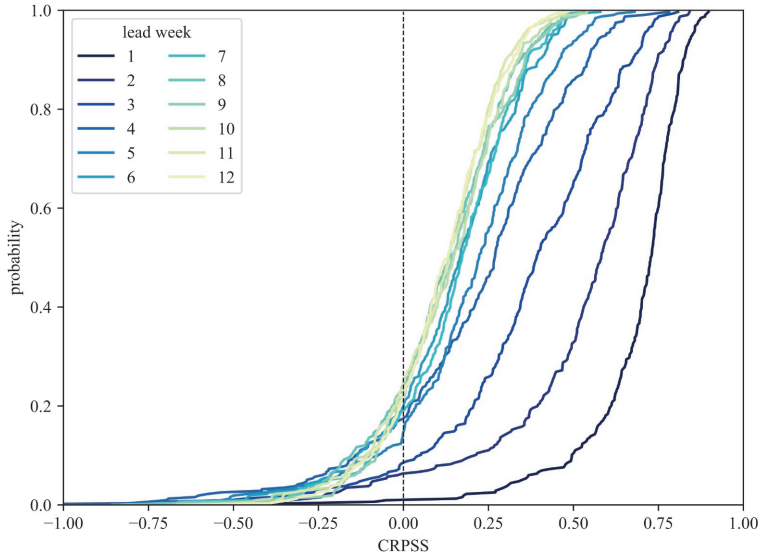


Figure 4.2 CDFs of weekly CRPSS shown for different lead weeks with CRPSS being aggregated over all models and station for discharge hindcasts. CRPSS is decreasing with increasing lead weeks but even up to 12 weeks roughly 60% of all stations and models show a better performance than the reference hindcasts.

for selected months providing weekly to monthly scores, which are temporal scales of interest for long term water management decisions. Further background information on evaluation score results based on different initialization months, temporal scales, target variables and ML models can be found in the Appendix.

4.3.1 General performance

To obtain understanding of the overall performance of the hybrid framework for discharge hindcasts, the CRPSS aggregated over all hindcasts for all stations and all ML models was computed. This was done by computing all the individual daily CRPSS results of all the hindcasts of every station and methods. The CRPSS results were additionally aggregated by lead day for the different temporal scales before averaging. The average CRPSS (weekly temporal scale) was used for Figure 4.2, where the CDFs for the first 12 lead weeks are highlighted, with CRPSS ranging from -1 to 1, with values above 0 indicating the hindcast framework outperforming the climatological reference. As expected, the CRPSS decreases with increasing lead time (with CDF lines moving up and the zero line being crossed earlier) and naturally converges after 7 weeks. However, up to lead week 12 roughly 60% of all stations and models show a better performance than the climatological reference. Even better results can be seen for surface water levels (Figure C.2), where up to 80% show a better performance. Even though the separate evaluation scores can vary slightly between target variable and station (locations) (shown later on), the overall hindcasting framework shows a positive tendency compared to hindcasts solely based on climatology.

We also observe that the hindcast framework shows higher performance compared to the bias-corrected EFAS seasonal (re)forecasting data (Figure 4.3), indicating the added benefit of having a hybrid framework that includes locally trained models. In Figure 4.3, the CRPSS results for two stations (Lobith and Eijsden) based on the bias-corrected EFAS (for the grid cell where the station is located) and the hybrid framework are highlighted. Differences in skill for larger lead times but also

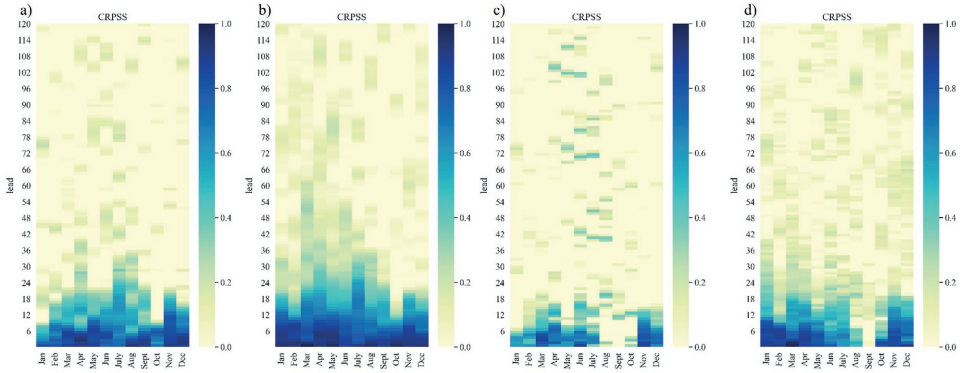


Figure 4.3 CRPSS (daily) shown for EFAS input data (bias-corrected) and hindcasts computed by the hybrid framework for stations Lobith and Eijsden (a) EFAS at Lobith, b) hybrid framework at Lobith, c) EFAS at Eijsden, d) hybrid framework at Eijsden). Differences in skill for lead times and also throughout different initialization months can be observed with the hybrid framework indicating a higher skill for local predictions than the large scale forecasting system.

throughout different initialization months can be seen for both stations. Furthermore, despite the EFAS data being bias-corrected for the specific locations, the skill for example station Eijsden is relatively low for some months indicating that forecasts based on climatology showing similar skill. The improvement in local skill is a result of the local training and the ability of the ML model to also use other information (e.g. precipitation, temperature) to further improve its forecast compared to the EFAS system.

During the analysis of the evaluation scores for all the different ML model hindcasts only a minor differences between the models are noticeable, which is seen throughout most stations along the main river network, especially for discharge hindcasts (Figure 4.4). Minor differences between methods are observed for surface waterlevel depending on the station location (Figure C.3), where for the example stations shown the more simpler methods show a slightly better skill. The minor differences are likely due to the limited impact of the model selection compared to the inherent uncertainty (represented by the ensemble spread) in the dynamical meteorological and hydrological forecast data. The forecasting skill is likely more dependent on the skill by which the input variables are forecasted, which apparently make the differences in skill between the ML models insignificant in comparison. In addition, the high temporal aggregation (monthly) and post-processing of the results before calculating the different evaluation scores, smoothing out the original differences in hindcasts results reduce the differences in performance between models.

Shifting the focus from the whole hindcast framework to a more detailed exploration of the evaluation metrics, the following paragraphs focus on results of one ML model and later on one example station. As hindcast results indicate that the differences between the models are minor, we will focus on the RF model, which previously already showed a promising performance Hauswirth et al., 2021. Figure 4.5 shows the weekly Anomaly Correlation Coefficient (ACC) for the discharge hindcasts at various stations (each represented by a pie chart) throughout the Netherlands for initialization months a) January, b) April, c) July and d) October. The ACC values per week are indicated by the pie slices arranged clockwise and their colour, dark blue indicating a high correlation coefficient while light yellow slices show weeks with a lower coefficient (note only significant values are shown). Looking at the results for the different months in Figure 4.5 indicate that for all months shown, the ACC decreases with increasing lead weeks. However, for all months the first few weeks (min. 3-4 weeks) show a high and significant score. This can be observed for all stations along the

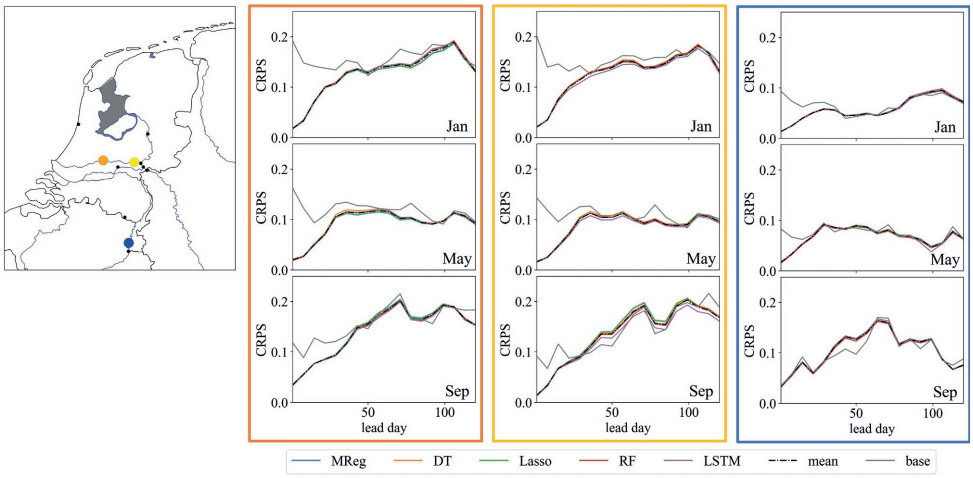


Figure 4.4 Overview of weekly CRPS values for month January, May and September. Different ML model scores, average of ML models score (dashed line) as well as climatological reference (grey) are shown for three discharge stations (Hagenstein Boven (orange), Driel (yellow) and Borgharen Dorp (blue)). For most months shown and first few lead weeks, the CRPS of the hindcast framework shows a lower score than the climatological reference. However, only minor differences between the ML models were observed. Maps were created using the python package Cartopy (Elson et al., 2022), which uses basemap data from Made with Natural Earth and © OpenStreetMap contributors 2022. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

main river networks, both Rhine and Meuse, while stations which are located at smaller streams or channels, which are strongly influenced by water management, can be more challenging (e.g. station close to the sea which is located at a shipping channel).

The observations from countrywide ACC analysis are supported by a more detailed analysis for station Hagenstein Boven located on the Rhine river network, roughly in center of the Netherlands and influenced by water management. We clearly see that there are differences in ACC per lead and initialization month related to the initialization month of the hindcast and the length of the forecast (Figure 4.6), also already observable in Figure 4.5 for the different stations and initialization months. In addition, Figure 4.6 shows that the differences on the ACC in temporal aggregation from daily, weekly to monthly temporal scale have a minor impact and that the skill assessment is robust. Significant ACC values can be observed throughout the first lead month for all initialization months. For early spring and summer months (March-July), significant ACC values for discharge predictions can be seen until two months in advance, in all temporal aggregation levels. The increase in significant lead time for the early spring months (March and April) is due to the snow melt dynamics in upstream catchment that were captured in the model training period (done prior to this study) and the physical model inputs from the EFAS system at the Lobith and Eijsden stations. The observation of more significant ACC values during the spring months due to the snow melt dynamic can be found throughout the stations along the Rhine, and less pronounced for the stations along the Meuse. Discharge predictions from late summer on show lower ACC values, likely due to the lower predictability in atmospheric weather patterns and reduced water storage in highly predictable stores like snow and groundwater. Unrealistic long lead times with significant values are likely due to lower observation records that can occur throughout the years, despite the selection of stations with limited missing records. Overall, ACC values for the discharge hindcasts show that hindcast anomalies are captured well for lead times up to one or two months for all initialization months compared to the

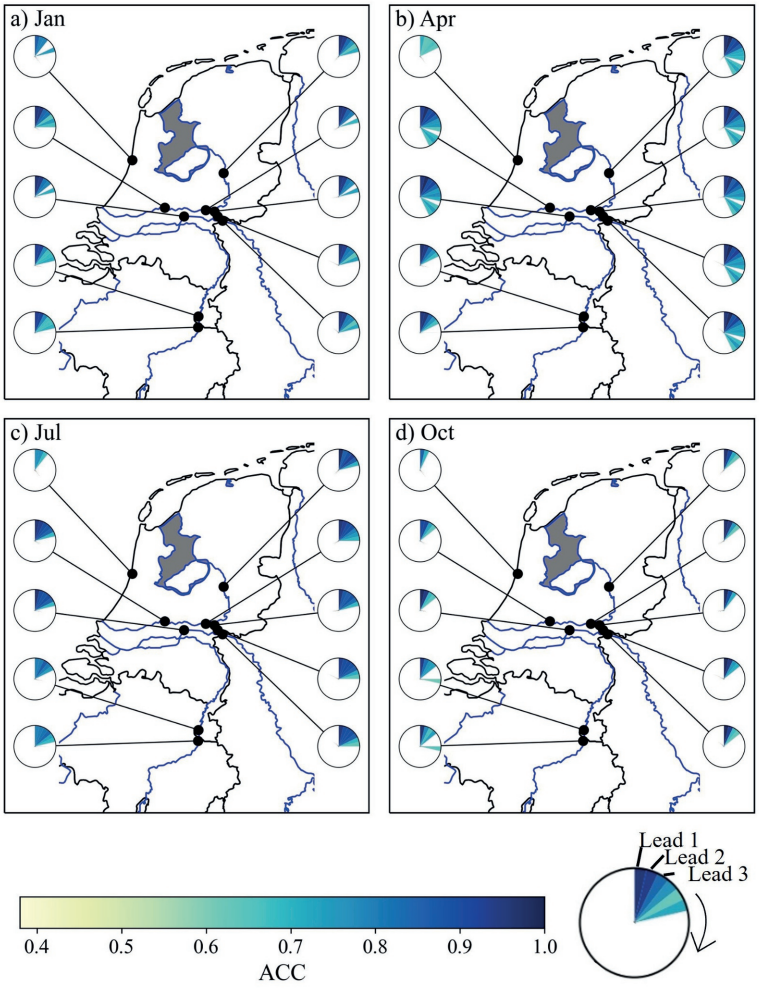


Figure 4.5 Anomaly Correlation Coefficient (ACC, weekly) for a) January, b) April, c) July and d) October for discharge hindcasts computed by the RF model showing the results for different stations (limited selection for visual purpose) in the national monitoring network. Only significant values are shown (indicated by range of 0.4 and higher). Furthermore, only major river network are shown, smaller streams or infrastructures are not highlighted (station along the coast is placed at a sluice along a stream). Maps were created using the python package Cartopy (Elson et al., 2022), which uses basemap data from Made with Natural Earth and © OpenStreetMap contributors 2022. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

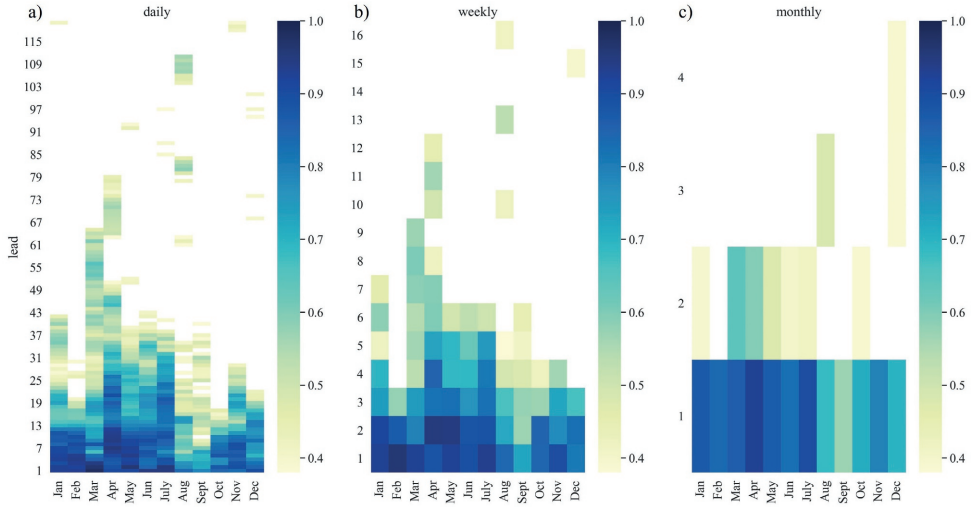


Figure 4.6 ACC results for the discharge hindcasts using the RF model of station Hagenstein Boven over different temporal scales (a) daily, b) weekly , c) monthly) and different initialization months. Only significant values are shown (indicated by range of 0.4 and higher).

observed anomalies for a complex station like Hagenstein Boven, which is affected by water management and upstream water reallocation. Furthermore, the hindcast framework was able to capture the general snow melt dynamic in early spring, resulting in significant values up to a lead time of two months at the onset of summer.

To check the robustness of the results we also analyzed the forecast performance with the CRPSS and BSS. The CRPSS was computed to assess the general performance in terms of spread and accuracy of the hindcast framework. Figure 4.7 represents the weekly CRPSS values for the same example station Hagenstein Boven in the center panel. The heatmap giving an overview of the skill score throughout the year with values ranging from 0 to 1, with values above 0 representing lead weeks where the hindcast framework outperforms the climatological reference. Similar to the ACC, the first few lead weeks show consistently good performance, indicating that the hindcast spread is close to the one of observations, while with increasing lead time the CRPSS decreases. This pattern can be observed by stations along the main river networks.

4.3.2 Hydrological extremes - low flows

To assess the hindcast frameworks capability of simulating low flow events, the BSS was computed using a threshold for the lowest 20^{th} percentile of discharge observations and hindcasts. This approach is similar to other studies using a threshold approach for drought definition (Van Hateren et al., 2019; Van Loon and Laaha, 2015). Figure 4.7 represents the weekly BSS on the right panel again for the example station Hagenstein Boven. Blue tiles on the heatmap indicate lead times where the hindcast framework outperforms the climatological reference strongly.

The BSS shows a less consistent skill pattern in the first lead month compared to the other scores. However, most of the skill that is seen shows a similar range of lead time, skill throughout long lead periods is decreases (e.g. after lead week 5). Compared to the other scores, tiles with lower BSS performance can be spotted in summer (June, Aug, Sept) and early spring months (Feb, March) for this station, in some cases for the first lead month or the following longer lead months. Some of these weeks appear to be more difficult to predict compared to early months in the year. This is likely due to

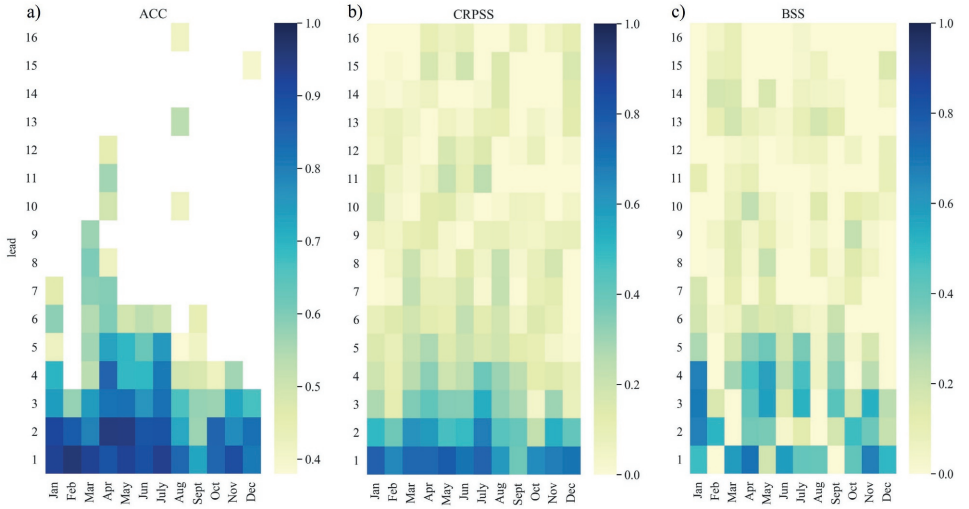


Figure 4.7 Heatmaps of weekly ACC (same as in Figure 4.6), CRPSS and BSS for discharge hindcasts (RF model) of example station Hagenstein Boven for different initialization months. a) ACC only significant values are shown (indicated by range of 0.4 and higher), b) CRPSS and c) BSS indicating a good performance (in dark blue and values above 0) compared to climatological reference.

unequal distribution of low flow occurrences throughout the year: where during summer months low flows can be more common and therefore chances to not fully capturing every event are higher, the low flows during winter are less common and captured relatively well with the snow melt dynamic as seen in previous scores. However, important months in terms of capturing drought development (April and May) show skill for up to 5 weeks. The described observations for the example station are in line with findings at other stations along the Rhine river network and less pronounced along the Meuse.

4.3.3 Difference between target variables and scenario runs

The result section so far has been focusing on the discharge hindcasts to be able to focus on the different evaluation scores in more detail. However, the hindcast framework was also used to hindcast surface water levels (Figure C.2, C.3, C.4, and C.5). Surface water level prediction skill for the national and model average (Figure C.2) show a slightly higher CRPSS skill than for discharge (Figure 4.2). In addition, similar patterns and trends regarding ACC for stations along the main river network (Figure C.3) and CRPSS results (Figure C.4 for an example station) were observed. Yet, surface water levels seem to be more challenging to hindcast with the evaluation scores being slightly lower especially for stations in smaller channels and further away from the main river network. Similar findings are found for BSS, where the performance of the hindcast for low flow periods was tested. As can be expected, stations which are not along the main river network and located downstream the main input variables (Rhine at Lobith, Meuse at Eijsden) show lower skills in capturing low flow periods compared to stations closer to input variables and the Rhine river. While ACC and BSS show a slightly lower performance, the CRPSS ranges are in the same range as for discharge hindcasts.

An additional discharge hindcast run was done including water management information representing some of the major infrastructures, based on the same approach previously explored by Hauswirth et al., 2021. Similar to the findings in Section 4.3.1, only minor difference between the different ML methods were observed. Furthermore, incorporating the additional water management

information only lead to insignificant improvements regarding the hindcast skill. The improved performance as well as the differences in ML model performance as seen in the previous study by Hauswirth et al., 2021 could therefore not be detected in this forecasting experiment.

4.4 Discussion

In this study we tested the suitability of ML models for seasonal predictions of several hydrological target variables at local scales throughout the Netherlands. This framework incorporated ML models over varying complexity, ranging from Multilinear Regression, Lasso Regression, Decision Tree to Random Forests and LSTM.

While the methods have shown differences in their performance during training and testing phase on historical observations (especially their ability to reproduce extreme events Hauswirth et al., 2021), interestingly applying the same subset of models on seasonal (re)forecasting information did not lead to large differences in model performance. We hypothesize that this is caused by the large uncertainty in the meteorological and hydrological input data, that outweighs the relative difference in performance by the different ML algorithms. In other words the forecasting skill is very much dependent on the skill by which the input variables are forecasted, which apparently make the differences in skill between the ML models insignificant in comparison. In addition, the minor differences seen between the ML algorithms in the original hindcasts were further smoothed out while calculating the evaluation scores on different temporal scales. The results in this hindcast experiment and the minor differences between the methods that were observed can be interesting in terms of model choices, in case computational demand is a key factor. With simple methods showing similar performance as more complex ones, which require more time regarding setup and training, the previous might appear more suitable. However, a more important factor limiting the model performance is the uncertainty introduced by incorporating seasonal (re)forecasting information. A subject for future research could be on how to incorporate and assess the way the different models deal with that additional challenge. Nevertheless, we observed that the ML modelling framework used here, which is based on locally trained models, allows for the opportunity to make hydrological forecasting more locally relevant by being able to forecast based on the station specific characteristics.

While the ML models were previously trained on direct observations, the seasonal (re)forecasting information from SEAS5 and EFAS introduces additional uncertainty from their forecasting system. We deliberately decided not to retrain the ML models on the forecasting information, as this more closely mimics the normal operational setting where an already trained model is used to produce forecasts. However, this provides an additional challenge, as we add another source of potential uncertainty as the ML models might not be well tuned to the forecast information. Retraining the models would also open up the opportunity for overfitting the ML models on the forecast data, which is something that should be avoided. Therefore, we preferred to use the more realistic operational scenario and use ML models trained on historic observations only, over a setup that uses ML models specifically trained on forecast data. Assessing the approach of additionally retraining the models for different cases, e.g. focus on extreme events or climate change trends are opportunities for future projects.

We extended our runs including water management, in line with the approach previously explored by Hauswirth et al., 2021. However, incorporating variables that represent water management settings in the ML models lead to negligible improvement and the improved performance as seen in the previous study could not be detected. We think that the uncertainty included in the seasonal (re)forecast input data is having a larger influence than the one of added water management information and therefore the strength of incorporating the additional information as seen in the previous study could not be observed. For future research it might be interesting to explore what

additional steps would be beneficial in terms of model framework and data, to be able to capture and simulate the details of water management setups also in a forecasting setting, as this would create the opportunity for scenario simulations.

The evaluation metrics show that for hindcasts of discharge and surface water level stations initialized in the early spring, skilful predictions for the first lead month one can be made. For early spring and summer month the skill increases up to 2-3 months, due to the snow melt dynamic being captured by the models in their training phase and the presence of this signal in the seasonal reforecasts of the discharge at Lobith used as input to the ML models. The skilful prediction for the first few lead months are comparable with other studies which have evaluated physically based systems (Wanders et al., 2019; Arnal et al., 2018; Girons Lopez et al., 2021; Pechlivanidis et al., 2020). However, contrary to the large scale physically based forecasting systems, hybrid frameworks such as the one presented in this study show to skilfully forecast target variables at specific locations which would not be feasible and at a fraction of the computation demand. While training ML models can range from a few minutes to hours, depending on the method and setup (Hauswirth et al., 2021), running the hybrid framework as used in this study only takes a few seconds to minutes per station and ensemble member. This can be interesting for water management needs at smaller scale or scenario analysis. Besides the fast running times of the models an additional benefit for the current framework is the input data set, which can be easily replaced by other input sources regarding precipitation, evaporation, discharge and surface water level as it was done in this study with EFAS and SEAS5 data. It is however important to realise that the original computation time required for large scale seasonal (re)forecasting information such as the latter two is still required but outside of the hybrid framework presented here. However, in case of scenario simulations, where local information would be required and tested in a more quick setting but based on one large scale input data set, the hybrid framework is still of benefit as it can compute the different hindcasts more efficiently.

A further point to realise is that this framework only focuses on time series at existing stations and therefore does not address the challenge of predicting at ungauged basins. However, recent advances in deep learning methods show that forecasting at ungauged sites may be a possibility if auxiliary geographically distributed variables (elevation, soil, river network topology) are incorporated (Kratzert et al., 2019a).

Similar to Hunt et al., 2022 we show that a hybrid forecasting system can provide added benefits compared to physical forecasting system. In addition to Hunt et al., 2022 this work confirms that the benefits of hybride forecasting can also be obtained for long-term forecasting. In this study we also show that these hybrid forecasting systems have the ability to provide more local information compared to large-scale physically based systems. As with other models, using ML models in hydrology comes with benefits and drawbacks. While the data availability can be a limiting factor for effectively train a model, the flexibility and low computational demand compared to large scale physically based models is an advantage. We think that with the right data available, ML models like the ones used here can easily be (re)trained for more specific studies and cases as well. Additional training on low flow periods for example could enhance drought predictions while incorporating climate change aspects and land use could help to assess future trends regarding water availability under increased human influence.

4.5 Conclusion

In this study we explored the suitability of a hybrid hindcasting framework, combining data-driven approaches and seasonal (re)forecasting information to predict hydrological variables locally for multiple stations at national scale for the Netherlands. Different ML models, previously trained on historical observations, were run with a simple input data set based on forecast data from EFAS and SEAS5 and evaluated using the evaluation metrics Anomaly Correlation Coefficient (ACC),

Continuous Ranked Probability (Skill) Score (CRPS and CRPSS), and Brier Skill Score (BSS). The hindcast framework's skill was compared to the skill of a climatological reference hindcast. Aggregating the hindcasts of all stations and ML models revealed that the hindcasting framework was outperforming the climatological reference forecast by roughly 60% and 80% for discharge and surface water level hindcasts. ACC results further show that independently of the discharge prediction's initialization month, a skilful prediction for the first lead month can be made. For spring months the skill extends up to 2-3 months due to stronger link to snow melt dynamic and temperature related impacts on the hydrological cycle that were captured in the training phase of the model. CRPSS and BSS show a similar pattern of skilful predictions for the first few lead weeks compared to the climatological reference forecasts. Skilful discharge predictions are particularly observed along the main river networks, Rhine and Meuse, which can be linked to the close proximity of the discharge input variables. This distribution of performance is also observed for surface water level hindcasts. We also observed that the difference between different ML models in the hindcast results are only minor, contrary to the differences observed when reproducing historical timeseries. This reduction in differences in performance between ML models is attributed to the relatively large uncertainties in seasonal (re)forecast data, reducing the relative impact of the model uncertainty in the total hindcast uncertainty. Even though the current hindcast framework is trained on historical observations, the hybrid framework used in this study shows similar skilful predictions as previous large scale forecasting systems. With the focus on creating a hindcast framework that is simple in its setup, fast and also locally applicable, challenges that can come with large scale operational forecasting systems for local users can be lowered. In addition, the ML hindcast framework also significantly reduces the computation demand and allows decision makers to explore more options and better quantify forecast uncertainty using a variety of ML models and inputs. Adapting the framework to special interests, e.g. droughts or climate change trends, by retraining the original ML models for specifically this purpose could further increase its performance. We conclude that the ML framework as developed in this study provide a valuable way forward, to making seasonal (re)forecast information more accessible to local and regional decision makers in the field of operational water management. In this study we purposely used publicly available seasonal forecast information which is globally available. This allows us to deploy this framework around the world and potentially provide relevant forecasting information for water managers and decision makers outside of the study area.

5

Chapter 5 | Simulating hydrological extremes for different warming levels - combining large scale climate ensembles with local observation based machine learning models

Abstract

Climate change has a large influence on the occurrence of extreme hydrological events. However, reliable estimates of future extreme event probabilities, especially when needed locally, require very long time series with hydrological models, which is often not possible due to computational constraints. In this study we take advantage of two recent developments that allow for more detailed and local estimates of future hydrological extremes. New large climate ensembles (LE) now provide more insight on the occurrence of hydrological extremes as they offer order of magnitude more realizations of future weather. At the same time recent developments in Machine Learning (ML) in hydrology create great opportunities to study current and upcoming problems in a new way, including and combining large amounts of data. In this study, we combined LE together with a local, observation based ML model framework with the goal to see if and how these aspects can be combined and to simulate, assess and produce estimates of hydrological extremes under different warming levels for local scales. For this, first a new post-processing approach was developed that allowed us to use LE simulation data for local applications. The simulation results of discharge extreme events under different warming levels were assessed in terms of frequency, duration and intensity and number of events at national, regional and local scales. Clear seasonal cycles with increased low flow frequency were observed for summer and autumn months as well as increased high flow periods for early spring. For both extreme events, the 3°C warmer climate scenario showed the highest percentages. Regional differences were seen in terms of shifts and range. These trends were further refined into location specific results. The shifts and trends observed between the different scenarios were due to a change in climate variability. In this study we show that by combining the wealth of information from LE and the speed and local relevance of ML models we can advance the state-of-the-art when it comes to modelling hydrological extremes under different climate change scenarios for national, regional and local scale assessments providing relevant information for water management in terms of long term planning.

Based on: Hauswirth, S. M., van der Wiel, K., Bierkens, M. F., Beijk, V., and Wanders, N. (2023). Simulating hydrological extremes for different warming levels - combining large scale climate ensembles with local observation based machine learning models, *Frontiers in Water, Volume 5, 2023*, https://doi.org/10.3389/frwa.2023.1108108

5.1 Introduction

Throughout the last decades it has become increasingly evident that climate change has a significant impact on the hydrological cycle, which is expected to increase even more in the future (IPCC report, Pachauri et al., 2015; Masson-Delmotte et al., 2022). Climate change leads to more extreme weather events, which can translate to more extreme flood and drought events as an example (Milly et al., 2005; Samaniego et al., 2018; van der Wiel et al., 2019b). These extreme events create challenging conditions for both nature and society, where for example drought conditions can lead to agricultural (crop loss and increased irrigation requirements), industrial (energy supply due to cooling water being too warm) and domestic impacts (drinking water shortages, heatwaves, health) and negatively impact ecosystems (e.g. wildfires, ecosystem collapse, Vörösmarty et al., 2010). Research has shown that the occurrence of extreme events is changing due to climate change, leading not only to a higher frequency of events (e.g. leading to multi-year events as observed by van der Wiel et al., 2023) but also more severe events (Wanders and Wada, 2015b). To get a better understanding and insight for future adaptation measures and planning, being able to analyse such extreme events under climate change is essential not only on the large but also local decision making scale.

Modelling of extreme events and the influence of climate change on the hydrological cycle has been done in various ways. Regarding extreme events and their return period, statistical approaches are a commonly used method, where the sample statistics of extreme events are extrapolated using an assumed probability distribution of extremes such as Gumbel or Generalized Extreme Value. However, observational records are often short or already affected by trends, thus making it difficult to provide accurate extrapolations of the probability of extreme events and their return periods in the future. In addition, this approach underlies the assumption that extreme events follow a single given probability distribution, which is not always the case for extreme events.

Regarding modelling the influence of climate change based on physical-based models and the effect on extreme event, climate models are used including simulations of different climate change scenarios. One of the simplest method including Global Climate Models (GCMs) as input data, includes using the projected change from GCMs as input data, which is used and applied to local timeseries (Fowler et al., 2007; Graham et al., 2007). This method has the advantage that it is computationally efficient (excluding the GCM runtime), however it can be difficult to reproduce the higher order statistics (e.g. rainfall intermittence or auto-correlation) and inter-dependencies between different variables. Directly using GCM output can also be done, as these models provide physically consistent and constrained simulations of a future climate. The downside of this approach is that GCM simulations are hampered by biases and the strong climate change signal in transient GCM simulations make it difficult to get a large enough sample size to study changes in the tails of the distribution. Coupling GCMs and Global Hydrological Models (GHMs) enables to simulate the influence of climate change throughout the hydrological cycle based on physically based models. Regional Climate Models (RCMs), which are complementary to GHMs, can be used to bridge the gap of GHMs studies for more regional climate assessments, as these models provide more detailed information and clear advantages in terms of modelling regional scale impacts (Rummukainen, 2010; Giorgi, 2019). These type of climate models can be linked to regional hydrological models as well (Graham et al., 2007). The opportunity here is to have processes connected and directly responding to the corresponding climate change model forcing. A drawback regarding this method to assess and model climate change impact is the computational intensive demand for running such large scale physically based model combinations. Furthermore, both the statistical and combined model approach only produce a few realisations of a timeseries including an extreme event, which makes robust statistical inferences difficult, if not impossible.

Large Climate Ensembles (LE) have been used more recently to address and assess the influence of climate change and its impacts on various aspects of Earth's physical climate (Deser et al., 2020; Maher et

al., 2021). Including LE in modeling studies allows to assess extreme events more specifically, as these include many realisations of possible events. Because of their time-slice experimental design, these data have no forced trend, setting them apart from other transient climate change simulations. LE data suffer from the same biases as other GCM output (e.g. Kelder et al., 2022), however they do provide the opportunity to empirically study the tails of the distributions and thus look at (hydrological) extreme events, i.e. without having to use statistical extrapolation methods for the tail of the distributions (van der Wiel et al., 2019a; van der Wiel et al., 2021). Nonetheless, in cases like these where GCMs and GHMs are combined for assessing different scenarios, large computational infrastructures are required. Furthermore, these assessments are on a coarse spatial scale (10-50km), while for certain interest finer, more local information is necessary which would ideally be computed in a fast and flexible manner, especially for national-local water management aspects.

Machine Learning (ML) provides the chance to facilitate data intensive modelling with more efficient and computational less intensive efforts. In the past years, several studies have shown that ML models such as Long Short-Term Memory Models (LSTM) or Random Forests (RF) are suitable for predicting hydrological variables such as discharge or groundwater levels (Kratzert et al., 2018; Koch et al., 2021; Hauswirth et al., 2021), often even outperforming the conventional physically based models (Mai et al., 2022). However, even when considering the potential of machine learning in hydrology, the challenge regarding simulating extreme events remain (Hauswirth et al., 2021; Hauswirth et al., 2022). ML models are generally good in simulating what they have seen during training, however extrapolating to ”unseen” events is not possible and as such makes it also difficult to apply to the climate change signal. A focus on extreme events can be achieved through specific choices in training and testing, however this method faces the same challenges as more traditional statistical approaches (see above) as observational records are often limited and unbalanced in terms of the availability of extreme events. Another option would be using simulated data for training and testing (Felsche and Ludwig, 2021), though these introduce other types of uncertainty compared to observational data (see above, e.g. GCM limitations), in general the use of observations is preferred.

As ML provide many computational advantages it makes it very attractive to apply ML to make projections of future extremes. This would reduce computational demand and make projections more locally relevant as reliable information can be provided at the local scale (Hauswirth et al., 2021). However the future is likely going to contain a lot of ”unseen” and extreme events providing challenges for ML to make accurate projections. Therefore, the objective of this study is to test if and how climate change modelling information from LEs can be combined with ML models to assess the probability of extreme events under climate change. Furthermore, the potential of this setup assessing the effect of climate change as well as providing estimates on discharge low and high flow events at national, regional and local scale will be assessed. We do this by using a ML modeling framework, including several location specific ML models that are trained on historical observations, developed for the case study of the Netherlands, and GCM LE simulation data to provide outlooks for future warming scenarios. The ML modeling framework used was developed especially for this region in previous studies (Hauswirth et al., 2021; Hauswirth et al., 2022), and has shown to be suitable for local predictions and seasonal forecasts, using both local but also larger scale input data. We approach the problem of simulating extreme events by introducing a post-processing step at the end of the ML framework, where we use the characteristics of the low flow distributions, taken from historical observations, based on the assumption that the tail of the distribution serves as baseline for extreme event extrapolation. This general post-processing step can be implemented for every station of interest, which allows us to include location specific characteristics. This in turn allows us to assess the influence of global climate change at national, regional and local scale.

Information regarding the modelling framework and the post-processing approach, as well as the use of the LE used in this study will be further elaborated in Section 5.2. Section 5.3 will first include

the ”proof of concept” of the post-processing step based on historical observations and simulations. Afterwards the focus will be shifted towards the scenario results and spatial trends covering national, regional to local scales. The approach, findings and challenges of incorporating climate change aspects in ML will be further along discussed in Section 5.4, followed by a conclusion (Section 5.5).

5.2 Material and Methods

Information regarding the data used, including historical observations, LE and the combined input data set used for the ML framework can be found in Sections 5.2.1, 5.2.1 and 5.2.1. The ML framework will be explained in Section 5.2.2, including the new post-processing step (Section 5.2.2) and its evaluation (Section 5.2.2). Explanations on the analysis of extreme events, such as droughts and floods, will be given in Section 5.2.3.

5.2.1 Data

Historical observations

Historical observations of discharge, tidal information, precipitation and evapotranspiration were gathered from the national monitoring network of Rijkswaterstaat (National Water Authority of the Netherlands) and the KNMI (Royal Netherlands Meteorological Institute) for different locations throughout the Netherlands for the time period 1980 to 2019. Station selection and data processing steps were done according to Hauswirth et al. (2021), which describes the original development of the ML modeling framework applied here. The input dataset for training and testing this framework consisted of a reduced set of 5 variables including precipitation and evapotranspiration at a central measuring location called deBilt, discharge of the Rhine (at station Lobith) and Meuse (at station Eijsden) at the border of the Netherlands, as well tidal observations close to one of the biggest dam infrastructure Haringvliet along the coast (Figure S1 in the Appendix). For further information regarding historical input data and pre-processing of the observations see Hauswirth et al. (2021).

Large Climate Model Ensembles

Climate change information was included by incorporating large ensembles of climate model data as input data for this study. These large climate ensembles consist of different warming scenarios and are a simulation product of the GCM EC-Earth v2.3, which combines an atmospheric, an ocean, a land surface, and a sea ice model (Hazeleger et al., 2012). The scenarios are based on different levels of global warming (i.e. global mean surface temperature values, GMST): the ”present-day” scenario includes a GMST equal to GMST observed between 2011-2015, the ”2°C warming” and ”3°C warming” scenario a GMST of observed pre-industiral temperature +2°C warming and +3°C warming, respectively van der Wiel et al. (2019b). The large climate ensembles are made by first running 16 long transient simulations with historical forcing (based on period 1860-2005) and RCP8.5 (for years 2006-2100). In a following step, from each of the 16 simulations 25 ensemble members were re-initialized. Re-initialization was done with perturbed physics that were matching the observed GMST (for more details we refer to the Supplementary Material of van der Wiel et al., 2019b). Every scenario consists of 400 ensemble members (16x25) of a time period of 5 years, creating a 2000 year ensemble dataset for each warming or GMST level. The large climate ensembles have been previously incorporated in various studies such as by van der Wiel et al. (2019b), van der Wiel et al. (2020), and van der Wiel et al. (2021) and van der Wiel and Bintanja (2021).

Input data for ML models

For the climate change scenario simulations, the same variables that were incorporated in the training and testing of the ML framework were used (Section 5.2.1). The precipitation was directly taken from

the LE. Reference potential evaporation was calculated using Makkink (de Bruin and Lablans, 1998) based on the variables temperature, surface incoming solar radiation, and mean sea level pressure provided by the LE. Discharge ensembles were created by running PCR-GLOBWB 2 (Sutanudjaja et al., 2018), a global physically based water balance model at 5 arcmin spatial resolution, for the Rhine and Meuse catchment using the LE as meteorological forcing.

For sea-level timeseries, the base approach used in Hauswirth et al. (2022) was incorporated. In the latter study, the historic tide level was simulated using the Pytides Python module. Anomalies between observed and predicted sea level fluctuations were computed and combined with multi-linear regression model to compute tidal information input dataset based on anomalies, u and v windspeed. In this case the wind speed data from the LE were used to compute the tidal information using the same multi-linear regression model.

The LE input data was bias corrected based on the observational records for the five input variables (precipitation, evaporation, discharge at two locations, and tidal information): linear bias correction was used for discharge input time series, same for wind speed data which was used to compute the tidal timeseries. For precipitation the fraction of dry days was calculated, which was then used to correct the precipitation data from the large climate ensembles by imposing a threshold to the LE precipitation below which precipitation was assumed zero. Next, the remaining annual total precipitation values for the full LE were compared and corrected against to the total annual precipitation values of the observations. Evaporation was bias corrected using mean and standard deviation assuming a normal distribution. The same principles were used for all different scenarios, for more information see van der Wiel et al. (2019b). In line with previous work by Hauswirth et al. (2021) and Hauswirth et al. (2022), the input data was extended in size by using a lagged timeseries approach, which includes additional timeseries of the input variables corresponding to the first three lags to the input dataset. Hauswirth et al. (2021) incorporated the lagged input identified by using the partial autocorrelation function (PACF) to extend the input data set and to help explain second order statistics that provide valuable information in timeseries analysis that the modelling framework was trained on. To fill the missing values that were obtained through including the lagged input data, the climatological mean of the present-day LE were used.

5.2.2 Modelling framework

The modelling framework used in this study includes different ML methods, ranging from simple linear regression methods to more complex methods such as neural networks. In terms of linear regression models, Multi-linear Regression (MReg) and Lasso Regression (LASSO) were included. The latter including an additional benefit of being able to eliminating variables which carry less information than others (Bardsley et al., 2015). Furthermore, regression types which are build on tree like structures, such as Decision Tree (DT), and ensembles of tree structures such and Random Forest (RF, Breiman, 2001) are part of the modelling framework. The most complex method included is the Long Short-Term Memory (LSTM), a recurrent neural network first introduced by Hochreiter and Schmidhuber (1997). LSTMs have the ability invoke a sort of memory provided by their internal state, setting them apart from the classical feedword networks, and therefore allowing them learn and simulate long-term depencies.

The modelling framework including these methods was developed and used to simulate historical hydrological timeseries in previous work by (Hauswirth et al., 2021). For every location of interest from the national monitoring network, a separate ML model for all these methods was trained and tested on historical observations. The different models were tested out for their suitability to simulate hydrological target variables such as discharge, surface water levels and surface water temperatures, and groundwater levels (Hauswirth et al., 2021). For this study however, we will focus on discharge predictions and use all the different ML methods.

The input dataset was of the original modelling framework was kept simple and replaceable by only incorporating five different variables including precipitation, evaporation, discharge at two locations and tidal information (see Section 5.2.1 and for locations Figure S1 in the Appendix). Hauswirth et al. (2022) used the same framework combined with seasonal reforecasting information in a hindcast setting. In this study the input data set will be replaced by the LE dataset (Section 5.2.1). The ML models, will be used with this new input data without retraining, which is possible because of the bias corrections applied. Using the set of ensemble data will create a set of discharge ensembles for the corresponding climate change scenarios. Introducing the LE into the modelling framework without retraining creates the challenge of the different models not having seen the additional information regarding extreme events. ML models experience difficulties to extrapolate data points out of what they have seen in training. Therefore, we are introducing a new separate post-processing step, which supports and corrects the simulations regarding extreme values of their distribution.

Post-processing

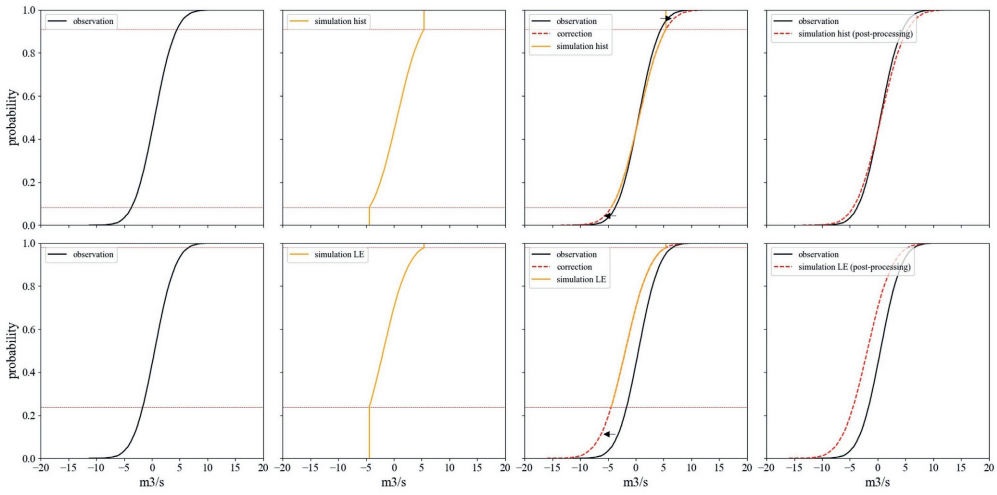


Figure 5.1 Schematic of post-processing approach based on information of historical observation CDF (black), which is transferred to simulation results (orange) of historical simulations (top) and single ensemble member of simulated target variable based on LE data (bottom). The information regarding tail of the distribution is transferred to simulation results by shifting and extrapolating the simulated tail to the shape of the historical observation one (indicated by red dotted line), resulting in a corrected target value (red, most right panel).

The post-processing approach introduced in this study is done to correct for the fact that the ML modeling framework struggles to simulate extreme events that were not represented in the training dataset (due to limited event records), which was seen in a previous study by Hauswirth et al. (2021). Furthermore, the step of transforming (normalizing) and back-transforming the data in ML routines and the choice in transformer can influence the simulations results. In this case, the ML model framework includes the quantile transformer, normalizing the data between 0-1. Running the same framework with the input data based on the LE leads to simulation results, where values which are lower than in the observational training data are computed as 0 before back-transformation, whereas after the latter the simulation results for discharge show a distinct cutoff for low values. This can for example be seen if the simulation results are compared to historical observations looking if plotted as cumulative distribution curves (CDFs), also see Figure S2 in the Appendix.

Figure 5.1 is summarizing the general idea of the post-processing approach in a schematic way, representing the CDFs of the historical observation record in black, the simulated historical simulation and the simulation ensemble member based on the LE in orange. The cutoff for low and high values described above is highlighted by the red horizontal line. To correct for these cutoffs we propose the post-processing step, which is based on the distribution of the historical observations of a given station and the extreme events represented in the distribution tail. This non parametric approach is based on the assumption that the form of the distribution tail for extreme events (droughts) remains the same. The key assumption here is that, although the distribution may shift or becomes wider or narrower, the form of the tails above the cutoffs percentiles (determined by the maximum and minimum values found in the training data set) does not change. Therefore allowing for more complex distributions to be corrected rather than distributions that can only be parameterized and thus brings the advantage of not assuming a distribution a-priory. The information from these extreme events probabilities will be used to extend the ensemble simulations of the target variable, which is indicated by the dotted red lines for both the historic (top) and the LE simulation case (bottom) in Figure 5.1 resulting in corrected simulation results which portray similar but extended tails.

An example in more detail: while the CDF of the observations (black) would represent values for the probability ranging from 0-1, the simulation (orange) would only indicate values corresponding to probabilities of 0.23-0.98, as an example for LE simulation in Figure 5.1. For the post-processing we take into account the shape of the tail of the observation distribution. This is done by taking the form of the CDF of the observations below the cumulative probability value of the simulated value that represents the cutoff of the lowest observation (0.23 in this case, indicated by the horizontal line in red), if we want to correct for low flows. Furthermore, the number of simulation values that need correction is determined by using the bin count for the lowest simulation value. For the correction the full observation record as well as the full 2000 year simulation data was considered.

The correction step then consists of assigning to each simulated value that needs correcting a value by extrapolating back following the tail of the observed distribution, where the order of assigned values is determined by the order of the simulated values before back transformation (on the 0-1 scale). If the extrapolated tail yields negative values, it is squeezed to fit between zero and the value of the minimum observed value. The same procedure is followed for the upper extreme values of the simulations. The form of the CDF of the observations for range larger than the CDF cutoff value of the simulations is used to extrapolate the simulated values larger than the maximum observed value.

This general post-processing step can be implemented for every location of interest, which allows to include the location specific characteristics by using the station specific observation records to extract the information regarding the distribution tail. This in turn gives the possibility to assess climate change influence at national, regional and local scale. The limitations of the general post-processing step will be discussed in Section 5.4.

Evaluation

Evaluation of the modelling framework, including the new post-processing steps is validated on historical simulations before applying it to the ensemble simulation results. In this evaluation we specifically focus on the ability of the post-processing routine to reproduce ”unseen” events. For this the same modelling framework as in Hauswirth et al. (2021) is used to the test the historic simulations results with and without the post-processing step. The models are trained and tested on a shorter sample of the historical records (1980-2019), with the 20^{th} and 80^{th} percentile of observations representing low and high flows being withheld and thus can be regarded as ”unseen” events. The 20^{th} and 80^{th} percentile serve as an example, in other cases the cutoff may be at other percentiles. This example is provided to imitate the situation the models will come across when incorporating the LE for future scenarios (where extreme events will be included that were clearly out of the training set

range or outside of the distribution of the observational record). The post-processing step is tested on the periods including most "unseen" events to see whether these low/high flow periods can be recreated by the model, having not seen them in the training phase. The CDFs of the observational records for the validation period, the simulations with and without the additional post-processing step will be assessed to demonstrate the potential of the suggested post-processing approach.

5.2.3 Analysis of extreme events

For the analysis of extreme events we looked at the change in frequency, average duration, average intensity and number of events for national, regional and local scale. To define low and high flows, the 20^{th} and 80^{th} percentile of the annual discharge was chosen. Regarding frequency, the percentage of discharge that falls into these categories was computed for every month. For average duration, average intensity and number of events a threshold of minimum 7 continuous days was introduced to classify time steps as low or high flow events. The average duration for events in a given month was calculated as the average number of days of extreme events that started in that month (and potentially continued into the following months). The average intensity was computed by including the mean discharge of the extreme events (thus > 7 days above or below a threshold) for the station of interest. The number of events was calculated over the whole ensemble simulations. Furthermore, percentage difference between the scenarios means, 20^{th} and 80^{th} percentiles and standard deviations were computed to assess whether the simulation results were underlying a change in climate variability.

5.3 Results

The results section will lead from the proof of concept for the post-processing approach to the results of the different climate change scenario simulations. The latter will include a focus on one ML model and the general trends and patterns regarding the different climate change scenarios and their influence on extreme events. Furthermore, the results will be reported first on a national scale, before breaking them down to regional and local results, highlighting a few stations separately. Information and results from different ML models will be listed in the Appendix.

5.3.1 Proof of concept

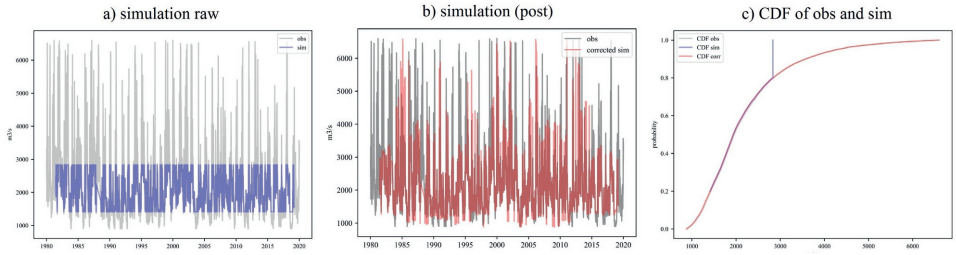


Figure 5.2 Example station Lobith (MReg model) including the a) discharge observations (grey) and the simulated discharge (withholding $20^{\text{th}}$and 80^{th} percentile in training, blue), b) post-processed discharge simulation (red) and c) discharge CDF including observations (obs) and simulations (sim), both raw and post).

The proof of concept for the suggested post-processing step is highlighted in Figure 5.2, where the observations as well as the simulated timeseries of one ML model and one station is shown. The model simulation in blue (left panel), which did not undergo the separate post-processing step, is not able to simulate the low and high flows for this station. This is due to a) having excluded the 20^{th} and 80^{th} percentile in training but also b) the transformer used for the input data is not able to back-transform

values that were outside the range of values included in the input data (used to fit the transformer initially). Including the post-processing step, which incorporates information gained from the tail of the observation distribution, corrects the simulation results for the low and high flow periods as seen for the timeseries in red (center panel). The correction for the discharge simulation can furthermore be seen in the right panel including the CDF curves for observations and simulations (with and without correction). The CDF curve of the corrected simulation is following the one of the observations, both in terms of range but also shape. The effect of the post processing on the simulations including the LE are further highlighted in Figure S2 in the Appendix for the different scenarios.

5.3.2 Scenario results

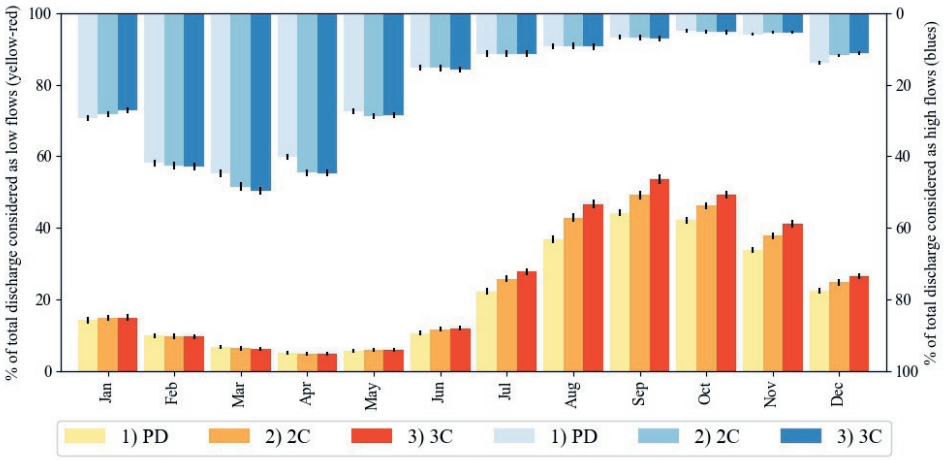


Figure 5.3 Seasonal cycle of the percentage of time in a month of low and high flows under different scenarios including present day, 2°C and 3°C warmer climate (PD, 2C and 3C). Low flows are highlighted in yellow (PD), orange (2C) and red (3C), high flows in light blue (PD), blue (2C) and dark blue (3C). Results are averaged over all models for all stations of the study area. Whisker represent 20^{th} and 80^{th} percentile of the uncertainty analysis based on bootstrap calculation of the mean.

The focus of the different scenario results was put on extreme events such as droughts and floods, which were analysed in terms of their frequency, duration and intensity. Regarding frequency, the discharge percentage that fell into the 20^{th} (low flow) and 80^{th} (high flow) percentile for every month was computed. Figure 5.3 presents the results based on the average of all the stations and all ML models, therefore representing a national and model average. The low flows are indicated in yellow (present-day, PD), orange (2°C, 2C) and red (3°C, 3C) for the different warming scenarios, the high flows are represented in light blue (PD), blue (2C) and dark blue (3C). The white spacing between the high and low flow bars represents the normal flow (between the 20^{th} and 80^{th} percentile) during the different months.

For every scenario, it can be seen that low flows occur in all months with varying frequencies. A seasonal cycle can be observed in all scenarios, represented by an increase in low flow frequency throughout summer and autumn (Jun-Sep) with a max of 54% (3C), 49% (2C) and 44% (PD) in Sep compared to earlier in the year. Lowest frequency ($\sim$ 5% for all scenarios) in low flow events are showing in early spring, which is due to snow melt dynamics influencing stations along the Rhine, and was also observed in the previous forecasting study by Hauswirth et al. (2022).

Differences between scenarios for the low flow are largest during the second half to the year (Jul-Dec, up to 6% between scenarios), while Jan-May show comparable percentages of discharge falling into the low flow category (up to 0.5%).

Shifting the focus to the seasonal pattern of the high flow events, an increase in high flow frequency throughout Dec-Mar, followed by a decrease during the following months is observed for all scenarios. Highest percentages are seen for Mar with 50% (3C), 48% (2C) and 45% (PD). As can be seen, not only the present day variation in the frequency of high and low flows complement each other, as expected, but also the changes in high and low frequencies under climate change are complementary, showing opposite trends where e.g. lower low flows during Feb and Mar but increased high flow percentages due to shifts in snow melt dynamics.

Smaller changes between the different scenarios are observed in the high flow frequency. Feb, Mar and Apr represent months in which 2C and 3C lead to a higher frequency, however the differences between the latter two are relatively small (1-4%) compared to the ones observed for the low flows (4-6%). Summer and autumn months which show a decline in high flow frequency only indicate minor differences between the scenarios.

The trends and patterns seen for both low and high flows in the different scenarios are due to a change in climate variability, which was found by looking into the mean, 20^{th} and 80^{th} percentile, and standard deviation (also shown in Figure D.4). While the changes in the mean discharge under 2C and 3C lies below 5% compared to the PD scenario for the large majority of the stations, the changes regarding the low flows (20^{th} percentile) range between 10-20%. Analysis of the simulation results of the different ML models (Appendix Figures D.5, D.6, D.7 and D.8) indicates that the direction and magnitude of trends and changes is similar across the methods. The absolute frequency between the different ML methods differ but the direction of change is similar for all models. Due to the large amount of data provided by the LE the changes seen are statistically significant. We additionally incorporated an uncertainty analysis based on the bootstrap approache, where we included the national and model average data and calculated the mean after randomly dropping 10% of the data for each month and scenario. This was repeated 1000 times. From the selection of means, the 10^{th} and 90^{th} percentile were chosen for the whiskers, which are represented in Figure 5.3 in black. Regarding model forcing, it was not possible to assess the GCM uncertainty as only one LE with this magnitude was available.

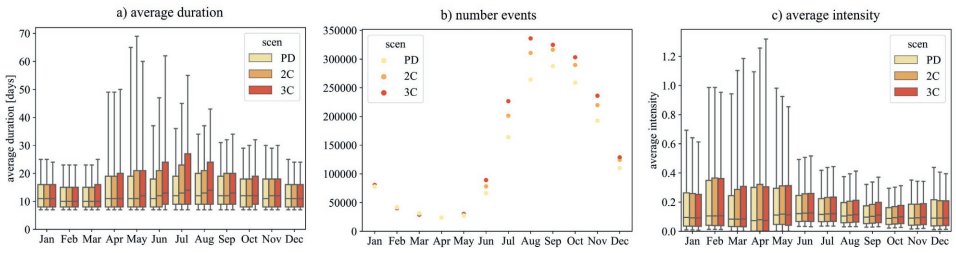


Figure 5.4 A) average low flow duration, b) number of events and c) average intensity based on different months in which low flow periods started for PD, 2C and 3C scenario (based on national and model average). Low flow events are defined by the 20^{th} percentile threshold, lasting a minimum of 7 days. The number of events is total over the full ensemble in a scenario. Average intensity is normalized by the annual mean discharge of the stations included. Whiskers represent 10^{th} and 90^{th} percentile.

The increasing severity of low flows in different months was also evaluated in terms of duration, number of events and average intensity. Figure 5.4a) presents the average event duration based on the starting month of the low flow events. Early spring to autumn show an increase in average duration distribution for different months but also between the scenarios, with the 3C indicating the largest

variation (Jul, 9-27 days). Differences in median duration between scenarios are minor during winter months and start to increase in May-Sep (14 days, Jul, 3C). Largest variations are observed in the 90^{th} percentile (May, 70 days - Nov, 30 days), with largest differences between scenarios in month Jun and Jul. The increase in average duration (seen in the distribution but also the 90^{th} percentile) leads to events which are overlapping with the following months, which corresponds to the higher low flow percentage observed in Figure 5.3 for the summer and autumn months. The number of events seen in Figure 5.4b) show the same seasonal pattern as observed in Figure 5.3. While Jan-May show similar number of events for all scenarios, clear differences are observed for summer and autumn months with 3C showing the highest number of events. In terms of average intensity (Figure 5.4c), differences in median intensity are small throughout all months and scenarios. However, the largest variation of average intensity is found for months Feb-May, which are also the months with the lowest number of events.

Focusing on the national scale for assessing the differences in scenario results show that both high and low flow percentages represent a seasonal cycle with increased low flow frequency during summer and autumn months, while high flow frequencies are increasing during (early) spring months.

5.3.3 Spatial characteristics - regionally

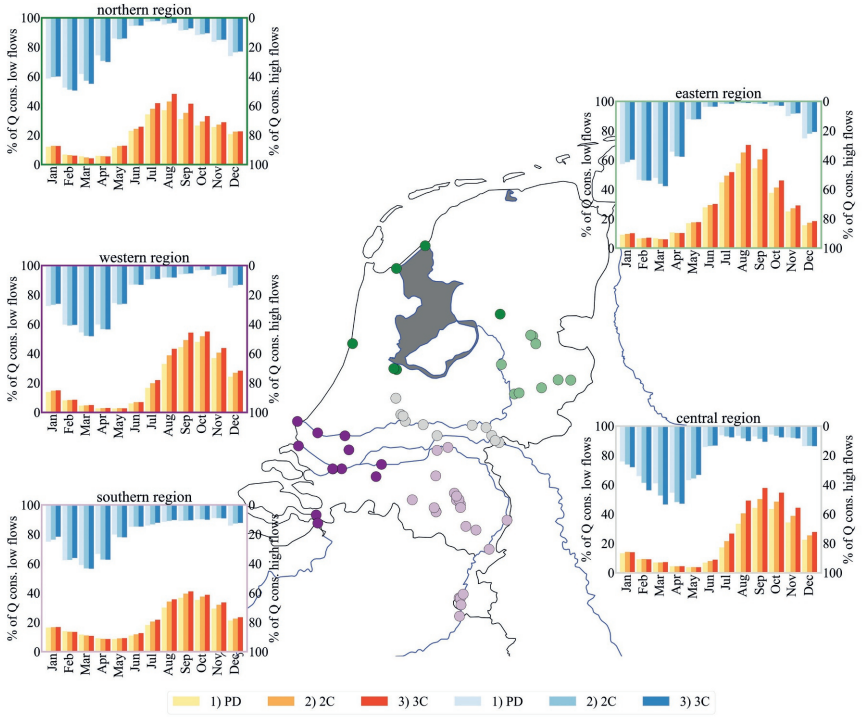


Figure 5.5 Seasonal cycle of percentage of low and high flows under different scenarios PD, 2C and 3C, averaged over different regions (indicated by the group of dots with same colour) based on the LSTM model results. Regions are defined into northern (dark green), eastern (light green), central (grey), southern (light pink) region and western (purple) region. Major rivers are presented by blue lines, while the dark grey area represents the IJsselmeer lake. Low flows are highlighted in yellow (PD), orange (2C) and red (3C), high flows in light blue (PD), blue (2C) and dark blue (3C).

Moving from the national scale to a more regional assessment, Figure 5.5 shows the averaged results for stations of 5 different regions based on the model input of the LSTM model and the same data analysis used as in Figure 5.3.

Focusing on low flow frequency first, a similar underlying pattern in terms of seasonal low flows can be observed as for the national average (Figure 5.3). Compared to the latter, shifts in terms of an earlier increase in low flow frequency (northern and eastern region), as well as higher maximum percentages (central:∼58%, eastern: ∼70%, northern: ∼48% and western region: ∼55% for 3C) is observed. Furthermore, the differences between the scenarios are more pronounced, especially for the summer months for which increased low flow percentages are recorded.

The central region is strongly influenced by incoming Rhine discharge, which was also seen in previous work by Hauswirth et al. (2021). Compared to the national average, spring months including Apr and May show a lower percentage of low flows for these months, likely due to the snow melt signal captured in the Rhine discharge. However, regions which are also linked to the Rhine (eastern, northern and western) show a stronger and earlier increase in low flows as well as decrease in the winter (Nov, Dec, Jan, Feb).

The southern region, which is predominantly fed by the Meuse discharge, also shows an increase in low flow frequency during the summer months. However, the maximum percentages observed are lower (∼41% in Sep, 3C) and furthermore, a relatively high percentage of ∼15% is already observed in Jan-Mar, which might be due to the Meuse being a rain water system where the snow melt dynamic signal is missing.

The differences observed between the regions and the main Rhine discharge can furthermore be seen in more detail in Figure S3, where the regional difference are presented relative to the ”base” station at Lobith (Rhine observations which were used as input variable in previous studies).

If the focus is shifted towards the high flow percentages in Figure 5.5, similar observations regarding shifts and differences between the regions can be made as for low flow periods, as well as the underlying seasonality which corresponded to the national average in Figure 5.3. The central, eastern, northern and western region indicate a stronger and earlier increase in high flows during the first few months of the year with Mar and Apr showing the highest percentages (53%, 58%, 50% and 48% for 3C). For the northern and eastern regions high flow percentages are lowest during summer months, while in other regions they vary around ∼3%. The southern region shows a less pronounced seasonal cycle than the other regions, with a lower maximum high flow percentage around ∼43% for both warming scenarios in Mar.

In general, the largest increase in frequencies with warming are observed for the low flows, while the changes in high flow frequency are less pronounced and much closer to the PD for most months. More details regarding the differences of the regions compared to the ”base” station Lobith can furthermore be seen in Figure S3.

The results shown in Figure 5.5 and supported by Figure D.4 showed that we are able to capture regional differences with locally informed models. These regional differences can also be seen for the other ML models (Appendix Figures D.5, D.6, D.7 and D.8).

5.3.4 Spatial characteristics - locally

Using the ML model framework based on local models allowed to assess the influence of climate change on extreme events also on a local scale, next to the national and regional already previously elaborated. Figure 5.6 shows the result for high and low flow frequency for the different scenarios (same approach as in previous Figures 5.3 and 5.5) for a selection of stations throughout the Netherlands. Some stations are located along the main river network, others at smaller streams (not drawn on the map).

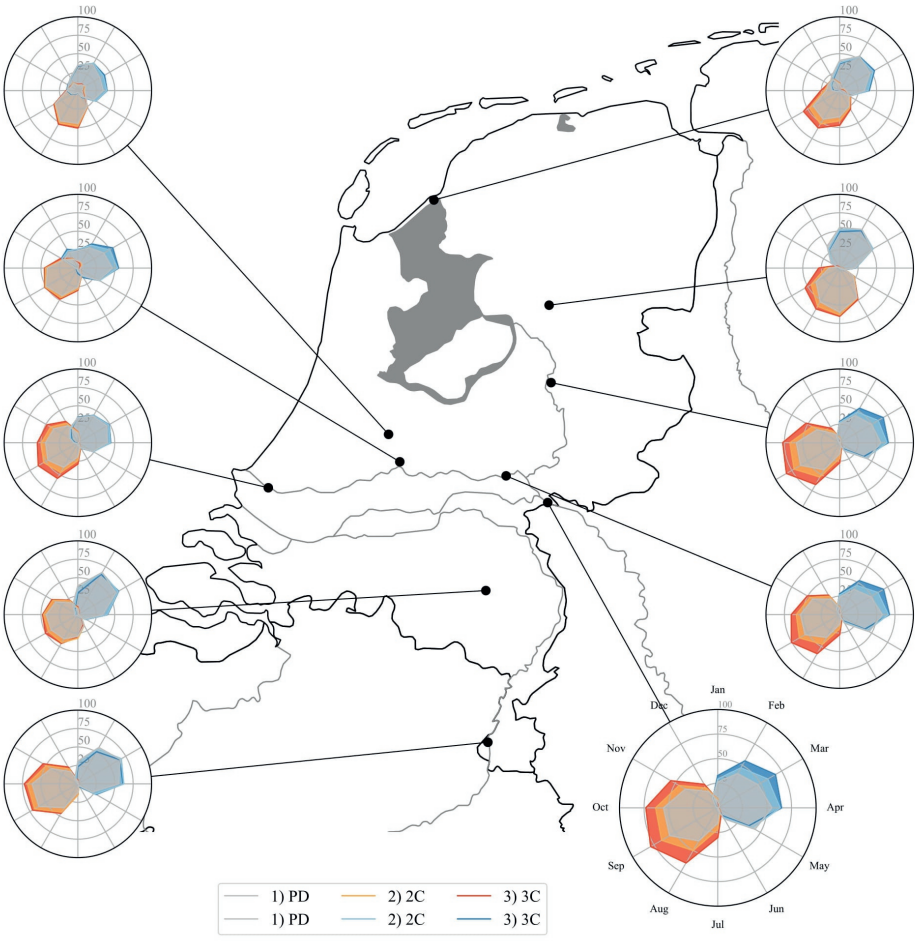


Figure 5.6 Seasonal patterns in percentage of low and high flows for different stations and scenarios PD, 2C and 3C, based on results from LSTM model. PD values for both high and low flows are represented by the grey shaded areas. Low flows for 2C are highlighted in orange and for 3C in red. High flows for 2C are shown in blue and 3C dark blue (3C).

Looking at the station represented by the bottom right, enlarged radar plot as an explanatory example, the percentages for both high and low flows are represented by the shaded shape for each scenario in corresponding colors throughout the year (PD in grey, 2C orange/light blue, 3C red/blue).

For both changes in low and high flows we observe similar trends as on the national and regional scale, albeit with more local details. For the example station at Lobith (bottom right) we see that the frequency of low flow events for all scenarios increases from Jul-Sep, while high flows increased in the period Jan to Apr.

Other stations highlighted in Figure 5.6 show similar patterns as the example station in terms of the shape of the different scenario results. Some station vary in the shift of these shapes (most northern ones), indicating that the high and low flow periods are starting earlier in the year, or also in terms of frequency. For example, stations located in the southern region show lower frequencies in terms of maximum high flows and furthermore the differences between the scenarios are smaller. Minor differences between scenarios but larger deviations in terms of frequency compared to other stations are also observed (top two left stations). Most of these observations regarding shifts and differences in frequencies were already found in the regional assessment, however more detailed patterns and exceptions were now able to be explored.

The shifts observed for the different stations are also observed in the mean, the 20^{th} and 80^{th} (Figure S8), which indicates that the trends and patterns seen for the different scenarios are due to a change in the mean climate and its variability. Shifts in the mean range between $\pm$ 10% for most stations compared to the PD scenario, with the largest shifts seen for stations with low mean annual discharge. Shifts in low flow extremes range up to 20%, while high flow events show a smaller range ($\pm$ 10%). This indicates that the changes in climate variability will have the most significant impact on low and high flows in the Netherlands.

Overall, for both high and low flows the differences in shifts for the different stations correspond well to the shifts found in the respective regions the station is located in. However, analysing the results of the stations also showed that frequencies and changes in frequencies can be locally slightly different from the regional average. This shows that having a locally trained ML framework is useful for bringing large scale information, such as the large climate ensembles, to local scale and therefore making it possible to assess future changes under changing climate and the adaptation requirements more location specific.

5.4 Discussion

This study combined a locally trained ML model framework with global LE with the goal to see if and how they can be combined to assess extreme events under different warming levels (PD, 2C and 3C) to potentially create valuable local estimates of climate change impacts from large-scale inputs.

Due the setup of the ML model framework including a simple, exchangeable input dataset, the incorporation of the LE was straightforward, similar to the previous study using the same model framework for seasonal forecasting by incorporating seasonal reforecasting data as input (Hauswirth et al., 2022). The locally trained models inherit information on the discharge characteristics for each station, translating large scale information to regional and even scales which are deemed relevant for adaptation measures. Especially in terms of climate change assessment, this can be of big interest as most of these assessments are based on large scale trends. The LE used in this study create a new opportunity to assess climate change and extreme events for different warming scenarios, compared to the more traditional approach including stochastic or GCM+GHM models where only one non-stationary timeseries is given. The ensemble approach provides orders of magnitude more realisations of extreme events, compared to more classic CMIP6 type transient climate simulations (van der Wiel et al., 2019b). At the same time, LE simulations allow to identify the hydrological impact of specific warming levels. LE have already proven to be relevant for impact assessment in for example energy production or agriculture (van der Wiel et al., 2019a; Goulart et al., 2021). In this study we showed that by combining LE with a local ML modelling framework enables us to project local hydrological effects of climate change which gives water management a more detailed outlook on potential direct impacts on flood risk and available water resources in the future.

To make extrapolation in a ML model framework possible, a new post-processing approach was introduced. The suggested non parametric post-processing step in this study underlies the assumption that the tails of the distribution (representing extreme events) of the historical

observations for each station remain the same for future events. This approach is relatively simple, can be implemented readily without needing high computational demand and brings the added benefit of being location specific and therefore creating the opportunity to scale down possible effects of climate change. However, it is dependent on having a long observational record for locations of interest. Due to extreme events having large return times and observational records not always being available or not long enough (in our case 30 years) this of course can lead to the post-processing step not correcting for the full scale of the extreme events in the LE. Furthermore, the form of the tails of the probability distribution of future events might be changing and therefore our assumption might not hold. However, this simple post-processing step is a first step to a fully data-based approach to simulating future extreme events. Assuming that the form of the tails above and below the training data values remains the same has its limitations, but is not better or worse than assuming some parametric form for extrapolation. Also using the statistical information of modelled data from physically based models has its limitations as they inherently also make assumptions and as such are also biased in their reproduction of the tail (and with a significantly higher computation demand). Using ML models trained on historic local observations gave us models that are closer to the original observations, compared to having models trained and tested on data obtained from future projections with physically-deterministic model forced with the LE. It remains to be seen if a physically based model is closer to the unknown future than our ML method with post-processing. The proof of concept showed that the approach taken in this work provided a reasonable solution for the inherit extrapolation problem in ML based future projections. Even though the most extreme events might not be able to be reconstructed with this method, the general trend and patterns, especially for long term events such as drought, can still be simulated and therefore gives a base for climate change assessments on a local scale.

The trends observed in this study are in line with earlier work projecting changes in hydrological extremes for the river Rhine. However the approach used here allows for further downscaling of the results, bringing the relevant information and insights to a more local level and making them more applicable for decision makers. This can be of interest to local authorities who are in need of locally specific information, as studies regarding climate change are often on a larger scale. Changes observed between the different scenarios were likely due to a change in climate variability. The discharge simulation (based on LE and PCR-GLOBWB 2) incorporated in the input data for the ML framework did include a shift in the extreme events and the mean, which was also present in the input data after bias correction. We observed that this information was also found in simulated discharge for the different stations. While the shifts in the mean for the stations were between $\pm$ 10%, shifts in the extreme events (especially low flow periods) could reach up to 20%. This is in line with previous assessments at the national scale suggesting that climate change will be more prominently felt in shifts in the extremes than in the average climate.

5.5 Conclusion

In this study, large climate ensembles (LE) and local, observation based hydrological machine learning (ML) models were combined to assess extreme events under different warming scenarios (present-day (PD), 2°C (2C) and 3°C (3C) warmer climate). The incorporation of LE, which in our case consisted of 2000 years of global data for every scenario, provided the opportunity to be able to empirically assess extreme events due to the large number of realizations of future weather. Combining this information with local ML models allowed for detailed and local, regional and national estimates of future hydrological extremes, therewith creating locally specific information that is of interest to local water managers.

A new post-processing approach based on historical information was introduced to enable the projections of extreme values outside the observed data range and incorporate the LE for local assessments of extreme events. The application of the post-processing step was tested on historical simulations first before being implemented for the different warming scenarios.

Extreme event analysis at the national scale in terms of discharge low and high flows for the different climate change scenarios and spatial scale showed a clear seasonal cycle with increased low flow periods from summer till the end of autumn (∼45% average for August-October) and increased high flow periods for early spring (∼43% average for February-April). Highest frequencies of low flow periods were reached with the 3C climate scenario showing the highest percentages for both type of events (53% and 50% respectively). Regional differences were seen in terms of shifts (low flows occurring earlier in the year) and ranges (higher/lower percentages). These trends, albeit with slightly different values, were further detangled into location specific results.

Differences in extremes between the different scenarios were predominantly due to a change in climate variability, which was seen by analysing the mean, the 20^{th} and 80^{th} for different discharge stations. Largest shifts regarding the mean were observed for stations with low annual mean discharge, while shifts in extreme events such as low flow were seen for all.

In this study we show that by combining the wealth of information from large climate ensembles, local characteristics captured locally with observation-based ML models and a suitable post-processing method for tail extrapolation allows the projection of future extremes under climate change. The local modeling framework thus provides important information for local to regional water management to be used in long term planning.


6

Chapter 6 | Synthesis

6.1 Main conclusions

Extreme hydrological events such as droughts and floods are challenging for societies and often result in significant damages. Over the past few years, large parts of the European continent have been hit by several extreme drought events with varying impacts and extents. While the root cause of these events can be found in the physical meteorological and hydrological system, the impacts are often strongly affected by societal decisions and management responses. Despite growing interdisciplinary research in the field of droughts, bridging the gap between scientific understanding and societal response remains imperative. One specific example, where most can be won from joint collaboration is the design and development of water management strategies: extreme drought events in the past have made it obvious that current water management strategies are under pressure of the continuous hydrologic change, that will persist over the next few decades due to climate change, and therefore need careful reconsideration.

The Netherlands serves as a prime example with its intricate and heavily managed hydrological system. While originally developed mostly with the focus of flood protection and moving water efficiently throughout the country, the past few years with increasing drought events showed the necessity of adapting the water management to store water for drier periods. These adaptation measures are not only important for the short - but also critical for the long-term planning, considering the evolving climate and socio-economic landscapes.

The pressing need to enhance preparedness for future drought events has sparked broad interest among society and water managers, fostering collaborative efforts with universities. This thesis, a joint endeavor involving the National Water Authority (Rijkswaterstaat) of the Netherlands and Utrecht University, explores key challenges in strengthening drought preparedness, which include improved modelling, forecasts, climate change assessment and water management explorations or adaptations that might be needed. Leveraging advanced technologies like machine learning and hybrid modelling, along with cutting-edge climate data, this thesis assesses their potential as decision support tools for water managers.

The thesis furthermore emphasizes the urgency of adapting water management strategies to address the increasing challenges posed by shifting hydrological patterns and climate change, as well as the critical role of collaborative efforts and innovative technologies in increasing the preparedness for future drought events. Every key challenge and objective introduced in the introduction will be addressed in the following sub chapters, combining the insights gained throughout the different project phases (which are also highlighted in Figure 6.1).

6.1.1 Efficient and Flexible Models for Various Hydrological Variables (Chapter 2)

To effectively assess various management options, perform scenario analysis, create seasonal forecasts, and conduct climate change assessments with extensive datasets, the necessity of an ***efficient*** *and* ***adaptable model framework*** *is evident. Traditional practices often rely on large-scale physically-based models, which can pose computational, interoperability, and data storage challenges. In recent years, there has been growing interest in* ***machine learning models*** *due to their demonstrated promise across diverse fields, extending beyond hydrology. These models offer the potential to address limitations inherent in current practices. However, it is essential to empirically evaluate their reliability in reproducing hydrological information and their applicability in the context of drought-related water management.*

In this thesis, the development of a modelling framework was focused and tailored to the spatial and temporal requirements of hydrological drought analysis and impact assessment in the Netherlands. Specific locations within the the National Water Authority's monitoring network served as the base components of the setup. In Chapter 2, various machine learning models were systematically assessed for their capability to simulate hydrological variables over historical periods and in the later Chapters 3 and 5 also for their suitability for hybrid modeling, encompassing seasonal hindcasts, forecasts, and climate change projections. The modelling framework in Chapter 2 was deliberately designed with a low level of complexity, featuring a selective, interchangeable set of input variables that accommodated both full machine learning and hybrid approaches. The training and testing of the machine learning models using historical observations was time-intensive, however significantly smaller than the development of some state-of-the-art physically-based hydrological models. In addition, once established, running the simulations with the pre-trained models for all three applications proved significantly computationally more efficient than relying on large-scale physically-based models typically employed in current water management planning. The drawback is of course that it is more difficult to understand the physical processes simulated by the machine learning models as they do not follow classic hydrological concepts and empirically determine causalities in the data. Chapter 2 of this thesis reports promising performance of the machine learning models in replicating various hydrological variables under a consistent setup and input data configuration. Therefore, this framework was further used as basis in Chapters 3 and 5 for seasonal hindcasts, forecasts and climate change projections.

6.1.2 Water Management Decisions to Mitigate Impacts (Chapter 3)

To effectively ***mitigate drought impacts****, it is essential to assess the consequences that may arise from various* ***water management decisions*** *and the legacy effect of these decisions for future water availability and drought impacts. Understanding how to minimize these impacts by* ***optimizing*** *water management actions in both the short-term and sub-seasonal timeframes is key for enhancing preparedness for future drought events and* ***reducing*** *the resulting* ***impacts****.*

Chapter 3 delved deeper into water management and its potential impacts or component for mitigation. Focusing on the Rhine and three downstream river branches, this study served as a proof of concept to investigate the feasibility of combining machine learning with non-linear human water management responses and quantify the potential impact as well as evaluate potential mitigation strategies. By incorporating straightforward impact functions and examining a range of water management scenarios, the water management decisions could be explored for past drought events. While this setup acts as a first example of combining machine learning, water management and impact functions for drought impact modelling for the Netherlands, the potential for future setups including forecast uncertainties or different optimisation strategies is apparent. However, challenges in terms of adequately representing these processes still remain mostly in the data availability and impact understanding.

6.1.3 Availability of Local Information, from Large to Local Scale (Chapter 4 and 5)

Information regarding meteorology and hydrology is often disseminated through large-scale models, which aid water managers in their duties. However, decisions and their consequences mostly impact ***local scales****, notably in the Netherlands with its extensively managed hydrological system. Therefore, the development of a modelling framework* ***capable of integrating both large-scale and local information*** *is essential for informed decision-making at local scale.*

The original modelling framework developed in this thesis, as discussed in Chapter 2, demonstrated the potential of machine learning for simulating hydrological variables accurately at local scale. It not only operates effectively in an initial observational-based configuration but also in a hybrid setting where large-scale output data from physically-based models were integrated. During the initial learning phase, the machine learning models successfully identified relevant relationships, trends, and patterns from observational records. These relationships between large-scale information and local relevance are, without any further calibration, successfully reproduced as showcased in the seasonal forecasting setup outlined in Chapter 4. Chapter 5 further revealed localized patterns, including climate projections to simulate extreme events under varying warming scenarios. In contrast to conventional downscaling techniques applied to other large-scale models or datasets, the modelling framework offered the distinct advantage of directly translating large-scale hydroclimate forecasts and projections into locally relevant information based on non-linear observational relationships, thus enriching the pool of tools available for local assessments.

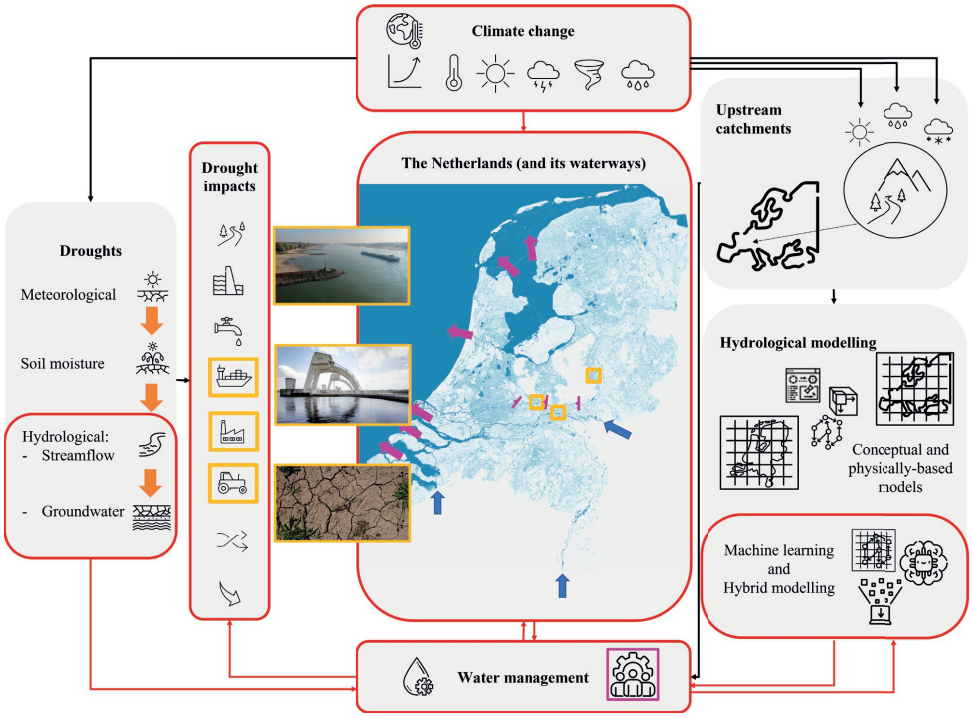


Figure 6.1 Overview of topics, key challenges and their connections touched upon throughout this thesis (red). Various aspects and connections were explored: simulating, forecasting and projecting of hydrological droughts using data-driven and hybrid approaches as well as exploration of water management and the potential to mitigate a selection of drought impacts in the Netherlands.

6.1.4 Preparation for Upcoming Droughts (Chapter 4 and 5)

*A central concern in present-day water management revolves around improving preparedness for upcoming drought events. This requires the ability to **forecast** and **project** hydrological conditions, not only in the **near future** but also **longterm** under **climate change**. These aspects are critical for simulating management decisions and optimizing strategies to mitigate drought impacts across various sectors. The key question is whether emerging technologies such as **machine learning, hybrid modeling, and large climate ensemble data** can be paired into a **unique, unified framework** to address these challenges effectively and provide relevant information at the **local scale**.*

In this thesis, the machine learning framework, developed based on historical observations as detailed in Chapter 2, was tested for its potential in seasonal forecasting (Chapter 4) and local climate change projections of discharge (Chapter 5). Leveraging its adaptable input features, the framework seamlessly transitioned into a seasonal hybrid modelling framework by incorporating new data sources, such as seasonal (re)forecasts, while retaining the same input features and without the need for retraining the machine learning models. The seasonal hybrid modelling framework showed slightly improved performance, with lead times of one to two months, in comparison to other large-scale physically-based forecasting systems. As added value compared to the traditional models, this information was now also available at more local scales. Chapter 5 demonstrated the hybrid framework's potential to simulate extreme events under various climate projections, effectively translating climate change impacts from large-scale data to localized trends and shifts. Unfortunately, it is hard to assess the accuracy of these projections, but the ability to downscale climate projections is promising for future applications. Despite the simplicity of the original machine learning framework, it now enables rapid creation of seasonal forecasts and climate change projections at the local scale for various locations. This not only expands possibilities for exploring developments, trends, and potential management adaptations concerning near-future extreme events but also provides water managers with an additional tool to address their specific interests to be better prepared for future drought events.

6.2 Key findings in a scientific and societal aspect

This thesis research and its close collaboration aspect with the National Water Authority gave the project a unique angle, with a direct focus on potential implementation and societal aspects. Therefore, this thesis includes both **scientific** and **societal** facets and the key findings for each of these will be presented in the following sections and are also listed in Figure 6.2.

6.2.1 Key findings and contributions to the scientific field

The interest in using machine learning techniques for exploring hydrological problems and the exploration of their limitations and boundaries has been increasing strongly over the past decade.

Development of an Operational Model in Real-World Context

Throughout this thesis, a specific focus was put on **developing** and **evaluation** of **machine learning** techniques regarding their **application for operational water management**. Utilizing various machine learning methods within a **low-complexity modelling framework**, their applicability for reproducing historic hydrological observations in the context of operational water management was explored (Chapter 2). These methods have then been tested for their ability to forecast seasonal (Chapter 4) and project future discharge values (Chapter 5) in the heavily managed water system of the Netherlands. While previous machine learning models have been developed for natural catchments, this is one of the first studies that **aims to simulate historic timeseries, provide seasonal forecasts and projections in a heavily managed water system**, like we find in the Dutch context.

Significantly, this research explicitly integrates **human-water interactions** into machine learning models, enhancing the understanding of the broader implications of water management, particularly during drought periods when water management plays a crucial role. A separate section of the thesis delves into water management practices (Chapter 3), building upon and refining insights gained from the original modelling framework. Key strengths of this modeling approach include its provision of **locally relevant information** through a streamlined and unified methodology that relies on a limited set of input variables.

Incorporating Human Decision-Making into Machine Learning Frameworks

The potential of including and simulating water management has shown to be an interesting aspect in Chapter 2. However, to comprehensively assess this aspect and connect it with impact and optimization strategies, a separate focus study was performed (Chapter 3). Given the constraints of limited data availability and historical records on water management strategies and drought impact relationships, the study developed water management scenarios based on the observed operational range throughout the observational record. Concerning drought impacts relationships, water management reports were considered, and basic thresholds for various rivers were outlined, providing initial insights into potential impact. Guided by input from the local stakeholders, the final model setup incorporated aspects of **water management** and the subsequent translation into **impacts**. Although the setup presented a simplified and focused on a small subset of national water management, it demonstrated the potential for **impact mitigation through water management optimization**, highlighting and confirming the efficiency of machine learning in facilitating **rapid scenario exploration**.

Enhanced Seasonal Forecasting with Machine Learning

Due to the streamlined, low complexity setup of the modelling framework, the transition to a **hybrid forecasting system** was feasible (Chapter 4). By introducing seasonal (re)forecasting information as input data, the modelling framework's performance and skill was explored in a hindcast setting. The seasonal dynamics observed in the original input data from the historical records could be detected again, despite moving from observational to simulation data as input for the hydrological and meteorological forcing. These modelled data do add additional uncertainty, however this did not limit the modelling framework in reproducing realistic timeseries for the historic events. Remarkably, the machine learning based modelling framework developed in this thesis exhibited slightly improved skill compared to seasonal forecasts produced by large-scale physically-based forecasting models, producing skilful forecasts for up to 1-2 months, depending on the season. This underscores the robustness of the hybrid modeling approach, despite its simplicity. In addition, the newly developed framework was capable of translating seasonal forecast information from large-scale physically-based models to the local scale, incorporating relevant local characteristics and providing added value. This research thus showcases the effective **enhancement of seasonal forecasting techniques through the integration of machine learning**.

Climate Projections Informed by Machine Learning

By incorporating new large climate change ensemble forcings as input data, this study explored the potential impacts and trends of future extreme events. Notably, this research represents one of the first attempts to **integrate large climate ensembles with machine learning for localized assessments** (Chapter 5). Overcoming a significant challenge in machine learning, which is data availability and training for extreme events, the study introduced an additional post-processing step into the existing modelling framework. This step, based on extrapolating distribution tails of observational records, allowed for the correction of extreme events not previously encountered. While acknowledging the

limitations of this approach, its simplicity of implementation and utilization of local observations proved beneficial in capturing essential **local characteristics** reflected in the simulation results. In addition it allows machine learning techniques to predict values that are outside of its statistical training range.

Additionally, the work presented in Chapters 4 and 5 highlights an advantageous feature of the modelling framework: the **ability to translate large-scale input, observed or simulated data into localized simulation results**. This process, similar to a form of downscaling, although distinctly different from conventional methods, effectively captured local nuances, statistics and temporal dynamics. This success was facilitated by the wealth of available observation records during the original model framework development phase. However, in regions with limited data availability, this approach will pose greater challenges.

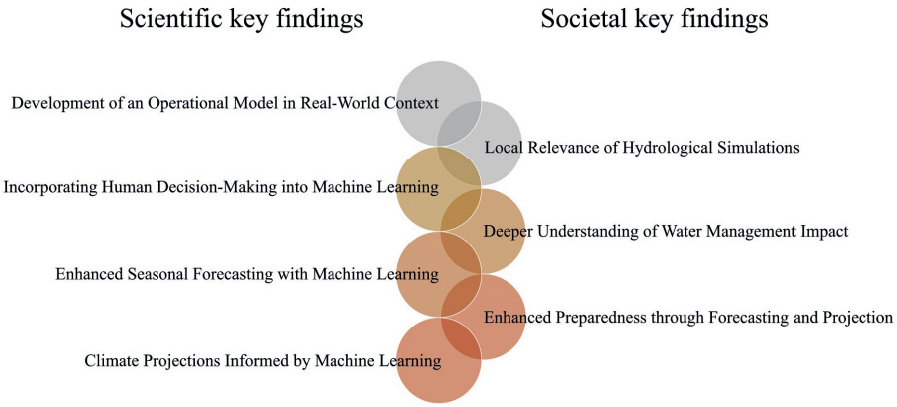


Figure 6.2 Overview of main scientific and societal findings through this thesis, connecting at many points at the interface of research and application.

6.2.2 Key findings and contributions to society

For society, the exploration of a machine learning modelling framework to support current water management showed several interesting outcomes, especially since with the challenges observed during past extreme drought events, the interest in developing and expanding current water management tools and practices has been growing.

Current water management practices often rely on large-scale physically-based models to simulate and understand the larger context and potential future development of the hydrological system of the Netherlands. In terms of water management decisions, operational plans from past decades which are often based on limited information for different infrastructures and expert knowledge are leading the decision process.

Local Relevance of Hydrological Simulations

With the developed modelling framework in this thesis, simplifying the accessibility and knowledge requirement for reliable drought information has increased. By developing the modelling framework primarily using **local observations**, the incorporated machine learning models had the opportunity to directly learn **local patterns and trends**. Furthermore, the extensive data available in the national monitoring network allowed for the inclusion of various hydrological target variables across the entire country. The modelling framework consistently demonstrated its ability to capture local trends in

different locations, whether applied to historical simulations, seasonal forecasting, or simulations under diverse climate change scenarios. This capability to generate **locally relevant information** provides water managers with a valuable additional tool for enhancing their drought preparedness.

Deeper Understanding of Water Management Impact

The thesis delved into an exploration of how **water management decisions** can **mitigate drought impacts** through a separate focus study. While the initial modelling framework showed responsiveness to additional information, such as the operational plans examined in the early stages of the research, the subsequent focus study specifically investigated the **direct influence of various management options** and their potential to alleviate drought impacts. This isolated model setup facilitated the **exploration of management scenarios** within the operational range observed in the past, revealing scenarios where impacts could have been mitigated, but also showing that actual water management had already adapted to more frequent drought situations.

Enhanced Preparedness through Forecasting and Projection

The transformation of the modelling framework into a hybrid setting facilitated the generation of **seasonal forecasting information and projections**, all available at the **local scale**. Currently, seasonal forecast information has limited usage in the Dutch water management, mostly due to the limited forecasting skill and the fact that forecast information is not locally available. Following the findings from this thesis a pilot has started where the results of this thesis are used to inform water managers on upcoming droughts. Local scale information also helps to accelerate the exploration of near-future trends at a regional scale, thanks to the reduced computational demands of the machine learning models integrated into the framework and give water managers the possibility to **enhance the preparedness** for future drought events. Locally relevant climate projections, tailored to the local setting have the potential to support decision making on increased climate resilience under different warming scenarios.

6.3 General discussion, limitations and challenges

This thesis has addressed various facets across its chapters, all united by a singular objective: enhancing the readiness of water managers to confront future drought-related extreme events. The topics encompassed machine learning techniques and the development of a modelling framework for operational water management, the creation of seasonal forecasting data and projections for the assessment of future extreme events and potential trends under diverse warming scenarios. Furthermore, additional emphasis was placed on understanding how water management practices influence drought impacts and how these impacts can be mitigated through the optimization of water management strategies. At the core of these efforts lies the modelling framework developed initially in Chapter 2.

Developing a modelling framework entails confronting numerous decisions and options, encompassing the selection of machine learning models, data gathering and selection, and, most crucially, the configuration and arrangement of these models. These decisions and options are also reflected in the increasing number of studies related to machine learning models in hydrological research in recent years. In terms of the framework's setup within this thesis, a deliberate choice was made to adopt a low-complexity and unified approach. This decision was driven by the aim of facilitating an initial comparison of various machine learning models and their performances, with the intention of later employing the framework in a hybrid context. However, this approach

occasionally necessitated trade-offs and compromises concerning parameter choices for models across different locations or the selection of training and testing datasets for various machine learning techniques. Consequently, the pursuit of a uniform setup may have curtailed the exploitation of specific models' unique strengths, however has increased the ability to compare the performance of different algorithms.

While the study focused on a national scale, guided by the interests and jurisdictional boundaries of water managers, no additional upstream information was taken into account. Nevertheless, the advantage of conducting model training and validation within a heavily monitored case area was evident. The outcomes demonstrated the suitability of different machine learning techniques, even those positioned at the simpler end of the spectrum, for simulating hydrological variables. Although it may be argued that more upstream information would have helped to improve the model performance, the upstream signals, such as the snowmelt signal, is observable in the modelled Rhine discharge and the modelling framework proved capable of capturing hydrological system patterns even in the absence of supplementary upstream data. In addition, including more information sources in an operational setting also increases the risk of failure when data sources are no longer available or not available in near real-time. The current modelling framework was created with this trade-off in mind, so that no additional barriers were created for a swift implementation in an operational setting.

While the modelling framework adeptly represented the general hydrological regime, it underscored the critical importance of choices related to model training and testing when endeavoring to employ machine learning for specific value ranges, particularly those relevant during extreme events, which are often extremely underrepresented in observational records. On the one hand the data availability plays a determining role but also the sampling of the training data. While extreme events are rare and might be one of the goals to explore, one also wants a model that is capable of capturing the overall hydrological regime. Striking this balance can be intricate, especially when the decision is made to rely solely on observational data for training, without introducing simulation data from other models.

The challenge of simulating unseen events was especially obvious while working on Chapter 5, were the hybrid modelling framework was used for discharge projections. This is a common limitation in machine learning approaches where models cannot predict values that are outside of their statistical training range. For this part of the thesis an additional post-processing step was introduced into the existing modelling framework. This step, based on extrapolating the distribution tails of records, allowed for the correction of extreme events not previously encountered. While acknowledging the limitations of this approach, its simplicity of implementation and utilization of local observations proved beneficial in capturing essential local characteristics reflected in the simulation results.

Furthermore, the findings discussed in Chapters 4 and 5 underscore a valuable aspect of the modeling framework, which is to provide information at local scales via transforming of extensive observations or outcomes from large-scale physically-based models into localized simulation results. The regional and local characteristics are capture through this kind of downscaling mechanism, which is slightly different from conventional approaches. The benefit of this process was made possible by the abundance of observational records accessible during the initial stages of modelling framework development. However, in areas marked by limited data availability, this approach may encounter more significant obstacles.

While vast amounts of observational records were available for different target variables, information regarding water management practices and specific handling of different infrastructures was challenging to gather. In this thesis, initially the water management was limited to a few operational plans which were all closely linked the discharge of the Rhine at Lobith, one of the main input target variables. While additional information seemed to improve the overall performance of some of the models, the improvement was less than expected. Then again, one has to remember that the heavily managed water system and the ongoing water management practices are likely already indirectly captured in the observational records over the past few decades. The range of water management actions and their influence was more specifically assessed in Chapter 3 in a more focused case study. However, the water management involved in this study was based on the previously observed operational range and therefore limited to current practices. This was done as machine learning models in general find it hard to predict behaviour they have not seen in their training data. While this prevented this work to exploring a full new set of operational strategies, the developed setup for the first time gave the chance to understand the management practices and resulting impact strategies in more depth and showed the potential of combining machine learning and human interactions.

In summary, the modelling framework, complemented by advancements such as machine learning, hybrid modeling, and large climate ensembles, emerged as a compelling configuration that could offer substantial support to water management decision-making processes. Its low-complexity setup, efficient processing times, and adaptability for diverse objectives constitute valuable attributes that could be incorporated into future practices. Comparable frameworks could conceivably be established for other variables of interest contingent upon data availability, with further adjustments to suit particular interests or event types. However, the incorporation of these modern techniques into existing practices necessitates a degree of systemic transformation, not only in terms of adapting modeling methodologies but also in data storage, accessibility, and, most critically, transparency, which may be challenging when dealing with perceived ”black box” models. Although certain machine learning techniques can be less intuitive and more intricate to comprehend, this thesis has underscored that even simpler algorithms and a straightforward setup can offer a valuable configuration for potential support. Nonetheless, when contemplating the future integration of modelling frameworks like the one developed and explored in this thesis into operational settings, the importance of establishing clear implementation goals and streamlining processes and steps along the way must be acknowledged as pivotal challenges for the long-term operational deployment.

Another related key aspect of the successful incorporation of machine learning models in an operational system is the trust that is put in them by water managers and decision makers. To move away from the idea that machine learning is just a ”black box” that provide no hydrological insights, to a modeling tool that is trusted and relied upon when making water management decisions takes time and cannot be expected to happen overnight. However, the focus on using these machine learning based models in parallel to ”proven” decision support systems can be a start. With careful evaluation of their weaknesses and strengths in a continuous assessment, trust can slowly be built and understanding on when and where to use these ”black box” models. This thesis provides a first attempt and assessment, together in collaboration with the National Water Authority, where these aspects and hurdles are faced and explored. It is however key that this ideally would be a continuous process, which gives water managers a close up view of the limitation and potential of machine learning until the desired setup for implementation for operational settings is fully developed.

Another aspect of machine learning modelling that has been touched up on in this thesis and has been increasing in the past is the exploration of hybrid setups. While the term hybrid can be interpreted in many ways, the common theme is the incorporation of either physically-based models and machine learning or simulation products from physically-based models with machine learning. The latter aspect was used in this thesis, for both seasonal forecasting but also projections. In terms of hybrid seasonal forecasting, the realisation that a simple framework as the one in this thesis based on only a few input variables was able to reach improved skilful predictions compared to commonly used large-scale physically-based forecasting systems was exciting. Especially, with the additional benefit of being able to add and highlight local characteristics and running simulations at a fraction of the simulation time to other large scale models. The projection results at local scale were also interesting as this type of information is usually only available at larger scale. Compared to the seasonal forecasting, which could be validated through the hindcast experiment in Chapter 4, the projections are more difficult to validate but could potentially be compared to other similar studies and setups in the future.

With the hybrid framework demonstrating promising performance and offering efficient forecasting capabilities, these machine learning based setups, also when coupled with water management scenarios assessed in Chapter 3, hold the potential for exploring the implications and mitigation options based on management decisions in the immediate future - an essential aspect of improving water management against upcoming drought events. While short-term decisions founded on near-future forecasts and adaptation strategies derived from current management plans can enhance preparedness for forthcoming drought events, a fundamental question remains: is this sufficient to counter the impact of future drought events? The results from Chapter 3 showed that exploring water management scenarios with machine learning models aid in potentially mitigating drought impacts, but it will only have limited benefit when the need to adapt to a changing climate in the coming decades increases past potential water management actions. The investigation into the impact of climate change on local extreme events has shown a shift in the potential drought landscape, encompassing changes in frequency, duration, and intensity. The localized challenges that have manifested in the Netherlands in connection with recent extreme events, coupled with the increasing influence of climate change, necessitate not only increased preparedness but also comprehensive adaptation measures on a larger scale. The hydrological modelling frameworks will play a role in preparing for both, upcoming and future drought events. From this thesis, it can be concluded that such modelling frameworks require speed and local accuracy, which can only be achieved by combining physically-based, hybrid and full machine learning modelling frameworks to aid water managers in drought decision making, both now and in the future.

6.4 Future research challenges and directions

In pursuit of a future-proof modelling framework equipped to accommodate adaptation measures or additional issues, approaches broader in scope and of a more **holistic** nature would be beneficial. Some of the issues mentioned below result from the challenges and limitations encountered and observed during this thesis.

Foremost among these challenges is the need to tackle the complexities of **understanding and defining impacts associated with the case area and the development of impact functions** capable of translating potential management practices into desired outcomes and assigning appropriate values to them. It is evident that while machine learning models have demonstrated their strength in efficiently and accurately simulating hydrological patterns, the challenge to properly quantify and predict impacts remains a pressing issue. The translation of these impacts into monetary values, although highly sought after, remains a gap in both research and practice. Furthermore, many current management decisions rely heavily on conventional large-scale forecasts, outdated operational plans, and expert knowledge. To advance, there is an urgent need **for more comprehensive and transparent information** concerning present practices, as well as a **thorough understanding of impact progression —both for direct and indirect impacts —within the intricate systems** characteristic of the Netherlands and other regions.

Nonetheless, drought events and their management extend far beyond localized problems. As evidenced by numerous extreme events in recent decades, these events can have extensive spatial coverage, impacting vast regions across Europe and affecting numerous sectors and ecosystems.

Moreover, the influence of **climate change** on future events is undeniable. For instance, shifts in snowfall and snowmelt dynamics are anticipated to **exacerbate downstream drought dynamics**. Capturing these dynamics in state-of-the-art modelling frameworks is key for responsible and adaptive decision making during droughts and to reduce future drought impacts. The combination of human influence, climate change, and the **potential non-recovery of a system** from preceding extreme events can present increased challenges, which, if not understood and potentially captured in our modelling frameworks, can have big consequences for society. Consequently, a **deeper comprehension of drought development - spanning onset, duration, and recovery** - is imperative for a more accurate **estimation of vulnerabilities and the identification of actionable strategies**.

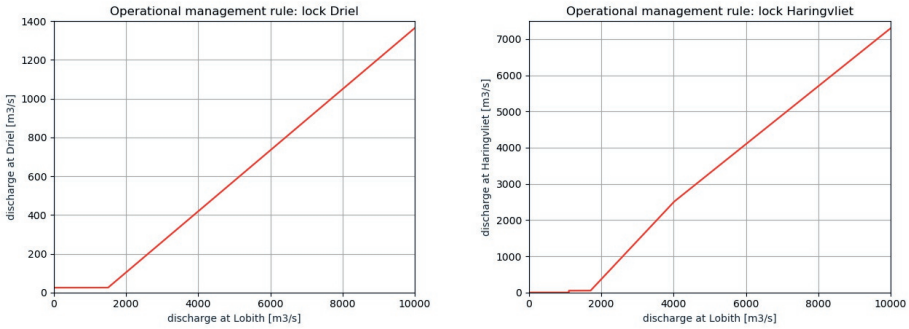
In order to investigate these large-scale dynamics, a **synergistic approach that combines machine learning and physically-based models could unlock considerable potential**. While machine learning has undoubtedly emerged as a formidable addition to contemporary hydrological modeling, challenges persist with respect to data availability, especially for extreme events and those that are expected to manifest with climate change. The inclusion of physically-based models in future research projects can help address some of the aforementioned issues. However, this hybrid approach introduces additional layers of complexity but at the same time increases the accuracy in model simulations. The combination of these two modeling paradigms could yield an effective means of **addressing challenges such as the incorporation of not only climate-related but also human and land-use changes**. These factors significantly influence both past and future developments and trends within a system. By incorporating all these aspects, as well as **potential feedback loops**, the development of comprehensive drought impact assessment strategies and support tools could be made even more complete. Ultimately leading to an improved understanding of drought processes and better quality information for water managers to aid their decisions during drought events.

Appendix A | The potential of data driven approaches for quantifying hydrological extremes

A.1 Material and Methods

Table A.1 Operational management plan for the locks at Driel and Haringvliet based on the Rhine discharge at Lobith (*reports "Stuw Driel - schematisatie en sturing" (1999) and "Lozings Programma Haringvlietsluizen (LPH)", Rijkswaterstaat (1984)*).

Driel		Haringvliet	
Rhine	**Measures**	**Rhine**	**Measures**
$\leq$ 1500 m3/s	Driel closed	$\leq$ 1100 m3/s	Haringvliet closed (0 m3/s)
$\leq$ 2330 m3/s	Driel in "use"	1100 - 1700 m3/s	Haringvliet 25 m3/s
$\geq$ 2330 m3/s	Driel open	1700 - 9500 m3/s	linear increase tot max discharge
		$\geq$ 9500 m3/s	Haringvliet max (7300 m3/s)



(a) Driel **(b)** Haringvliet

Figure A.1 Overview of operational water management scheme at a) lock Driel and b) lock Haringvliet based on the Rhine discharge.

A.2 Results

A.2.1 nRMSE scores

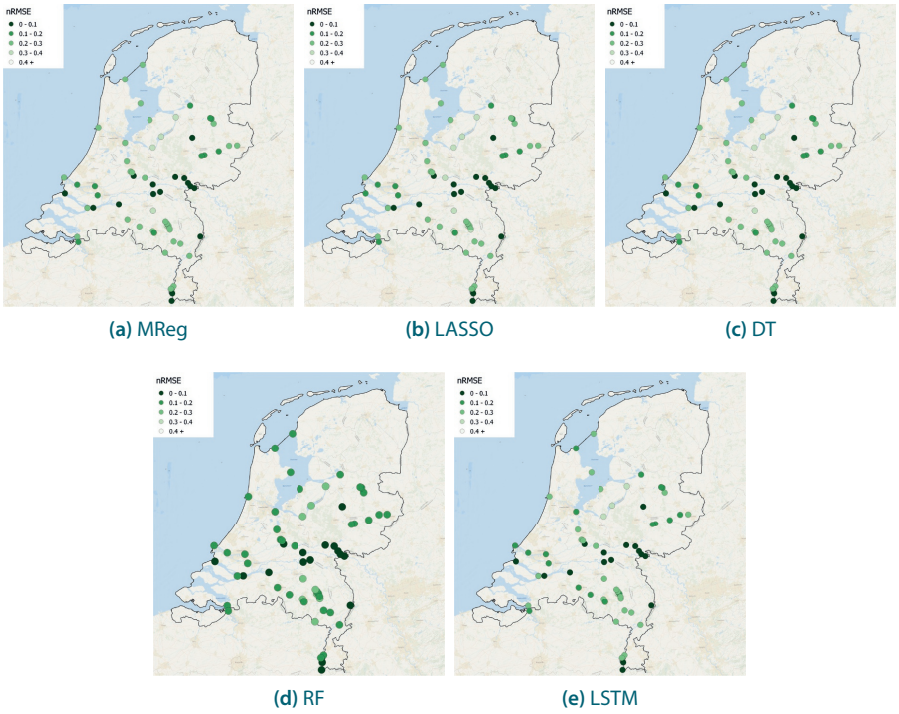
Table A.2 Overview of number of stations per ML methods that fall in same nRMSE range for the natural run and target variables: discharge, surface water levels, surface water temperatures and deltachange groundwater levels.

Discharge	nRMSE	MReg	LASSO	DT	RF	LSTM
	0-0.1	17	16	17	17	17
	0.1-0.2	12	12	12	41	20
	0.2-0.3	35	33	35	11	28
	0.3-0.4	5	8	5	0	4
	0.4+	0	0	0	0	0
Surface water level	**nRMSE**	**MReg**	**LASSO**	**DT**	**RF**	**LSTM**
	0-0.1	18	18	18	31	21
	0.1-0.2	28	28	30	49	45
	0.2-0.3	42	38	41	8	23
	0.3-0.4	1	5	0	0	0
	0.4+	0	0	0	0	0
Surface water temperature	**nRMSE**	**MReg**	**LASSO**	**DT**	**RF**	**LSTM**
	0-0.1	0	0	0	85	57
	0.1-0.2	100	100	100	15	43
	0.2-0.3	0	0	0	0	0
	0.3-0.4	0	0	0	0	0
	0.4+	0	0	0	0	0
Groundwater level	**nRMSE**	**MReg**	**LASSO**	**DT**	**RF**	**LSTM**
	0-0.1	447	20	452	417	421
	0.1-0.2	154	550	150	177	196
	0.2-0.3	2720	2571	2990	3222	2334
	0.3-0.4	648	828	377	143	1015
	0.4+	1	1	1	2	4

Table A.3 Average scores per target variable and method based on the total timeseries simulated as well as for high and low flow scenarios.

	Discharge			Surface water level			Surface water temperature			Groundwater levels		
MReg	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**
total timeseries	0.19	0.63	0.48	0.18	0.69	0.57	0.13	0.90	0.85	0.24	0.27	0.07
low flow	0.23	0.35	-0.85	0.21	0.37	-0.42	0.14	0.27	0.10	0.36	0.12	-17.03
high flow	0.20	0.56	0.46	0.17	0.60	0.53	0.12	0.82	0.82	0.33	0.18	-0.80
LASSO	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**
total timeseries	0.20	0.60	0.50	0.19	0.68	0.56	0.13	0.89	0.86	0.25	0.24	0.11
low flow	0.23	0.33	-0.83	0.21	0.35	-0.43	0.14	0.28	0.08	0.35	0.08	-16.07
high flow	0.21	0.55	0.44	0.17	0.60	0.54	0.12	0.82	0.81	0.34	0.22	-3.73
DT	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**
total timeseries	0.19	0.65	0.53	0.18	0.73	0.63	0.15	0.86	0.80	0.24	0.22	-0.02
low flow	0.23	0.34	-0.83	0.20	0.37	-0.36	0.15	0.24	0.01	0.37	0.10	-14.47
high flow	0.20	0.57	0.46	0.16	0.64	0.58	0.15	0.75	0.74	0.33	0.14	-0.26
RF	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**
total timeseries	0.14	0.85	0.72	0.12	0.89	0.78	0.09	0.95	0.89	0.24	0.20	-0.02
low flow	0.17	0.50	-0.32	0.14	0.55	0.09	0.09	0.58	0.45	0.37	0.12	-16.56
high flow	0.14	0.75	0.57	0.11	0.83	0.77	0.09	0.91	0.90	0.33	0.10	-0.86
LSTM	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**	**nRMSE**	**R**	**KGE**
total timeseries	0.17	0.75	0.67	0.15	0.81	0.75	0.10	0.94	0.90	0.24	0.32	0.19
low flow	0.19	0.41	-0.60	0.16	0.46	-0.10	0.10	0.49	0.39	0.35	0.12	-17.11
high flow	0.17	0.65	0.43	0.14	0.73	0.69	0.10	0.89	0.87	0.33	0.18	-2.14

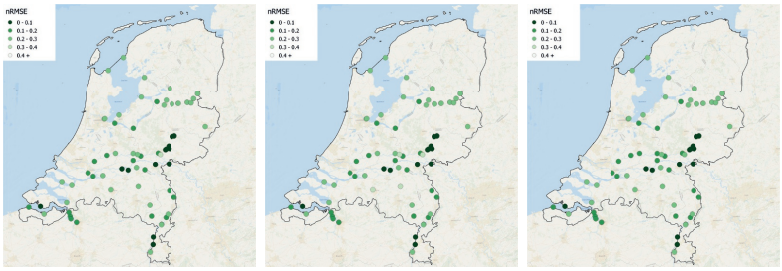
A.2.2 nRMSE scores - maps



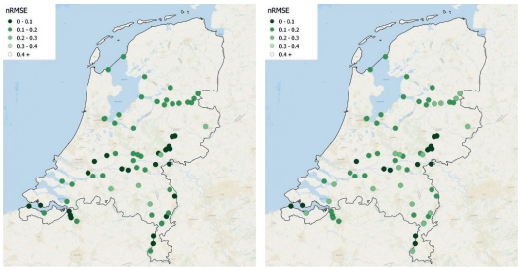
(a) MReg **(b)** LASSO **(c)** DT

(d) RF **(e)** LSTM

Figure A.2 Normalized RMSE results for discharge prediction and methods: MReg, LASSO, DT, RF, and LSTM. The darker the colour, the better the score (nRMSE ranges form 0-1).

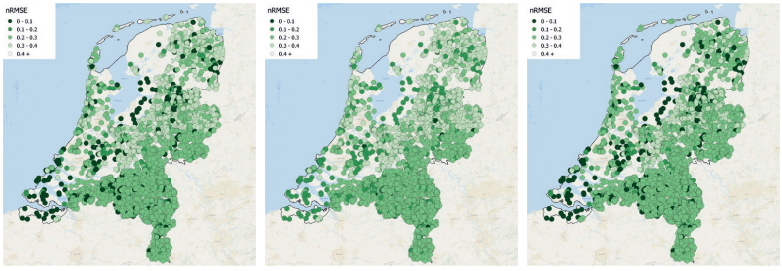


(a) MReg **(b)** LASSO **(c)** DT

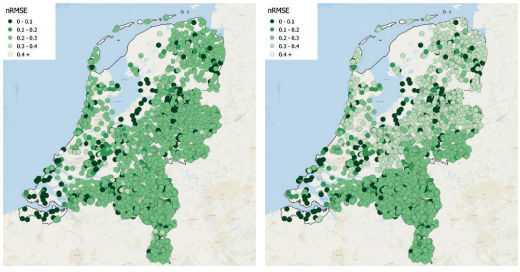


(d) RF **(e)** LSTM

Figure A.3 Normalized RMSE results for surface water level prediction and methods: MReg, LASSO, DT, RF, and LSTM. The darker the colour, the better the score (nRMSE ranges form 0-1).



(a) MReg **(b)** LASSO **(c)** DT



(d) RF **(e)** LSTM

Figure A.4 Normalized RMSE results for groundwater level prediction and methods: MReg, LASSO, DT, RF, and LSTM. The darker the colour, the better the score (nRMSE ranges form 0-1).

A.2.3 Time series - natural run

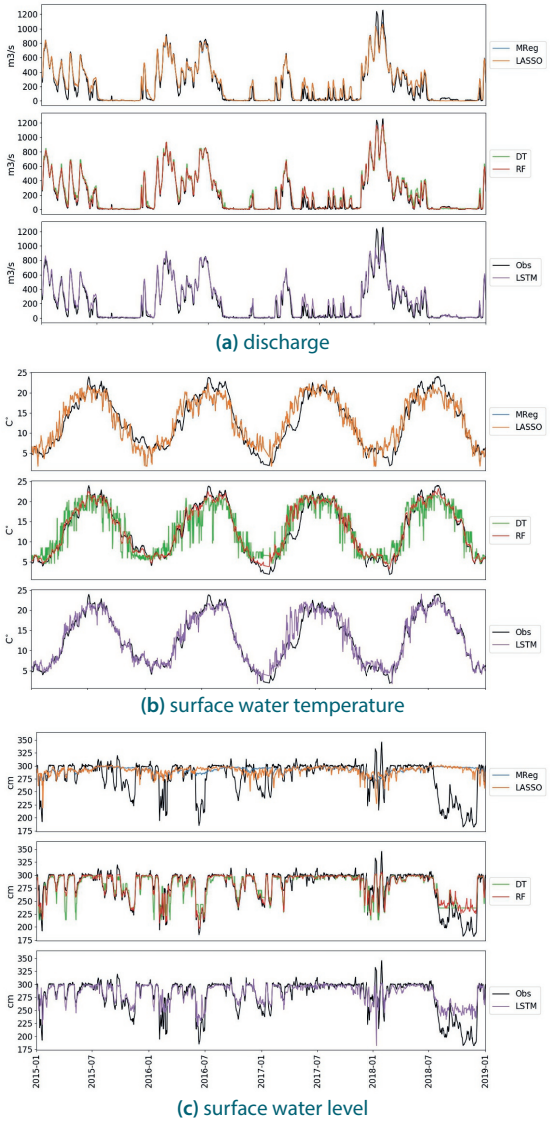


Figure A.5 Time series of predicted discharge, surface water temperature and surface water level (from top to bottom) for station Hagenstein Boven for 2015-2019, presenting the predicted observation per ML methods and the observations (obs).

A.2.4 Feature importance

The feature importance of the stations are ordered according to their distance from Lobith station (which acts as an input timeseries), which enables us to observe the different incorporation of the input data following the river downstream. Furthermore, the main input series (e.g. precipitation, evaporation, Rhine discharge, etc.) are colour coded including the additional lagged timeseries in a lighter colour.

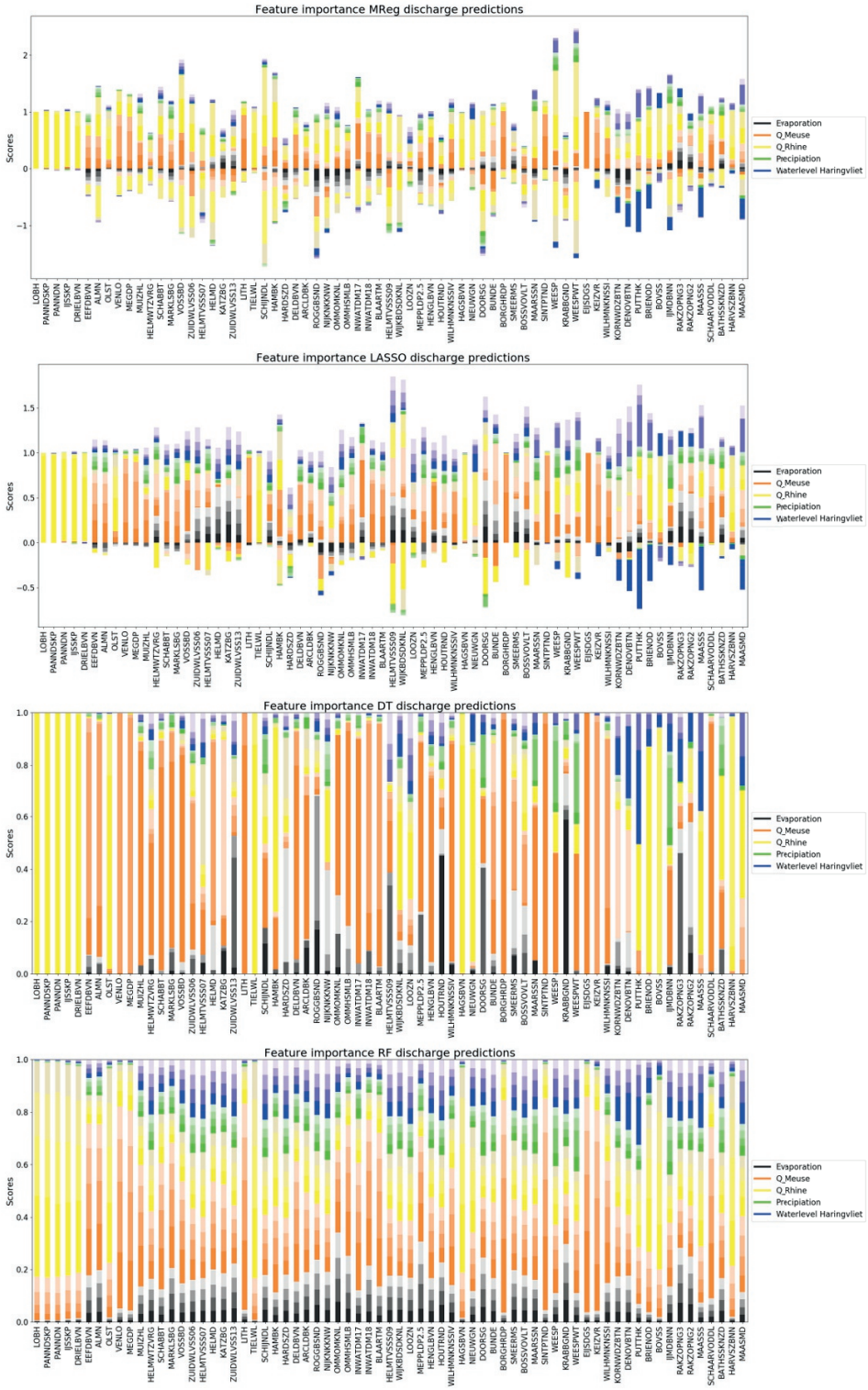


Figure A.6 Feature importance of the MReg, LASSO, DT and RF (top to bottom) for discharge prediction.

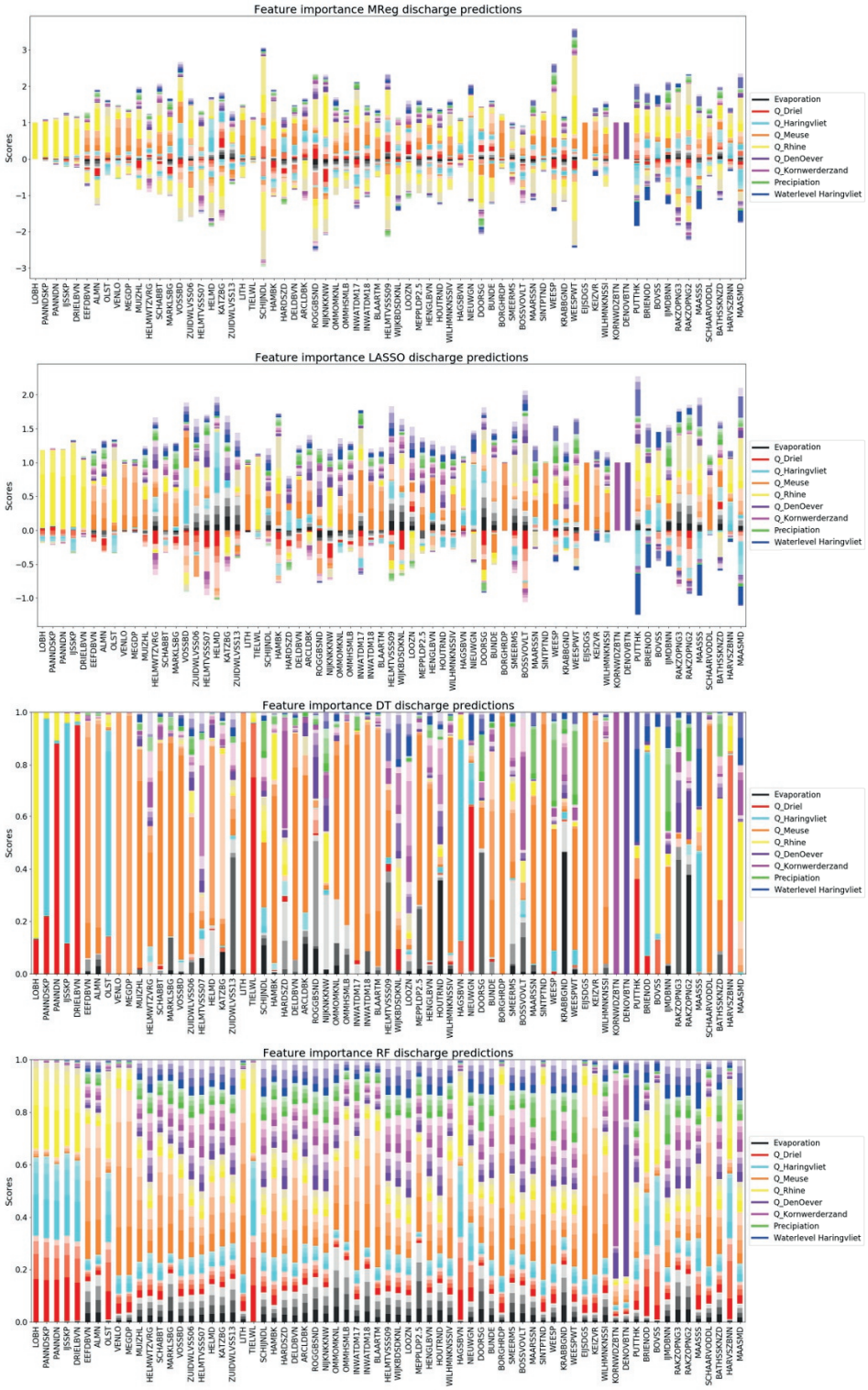


Figure A.7 Feature importance of the MReg, LASSO, DT and RF (top to bottom) for discharge prediction including water management influence.

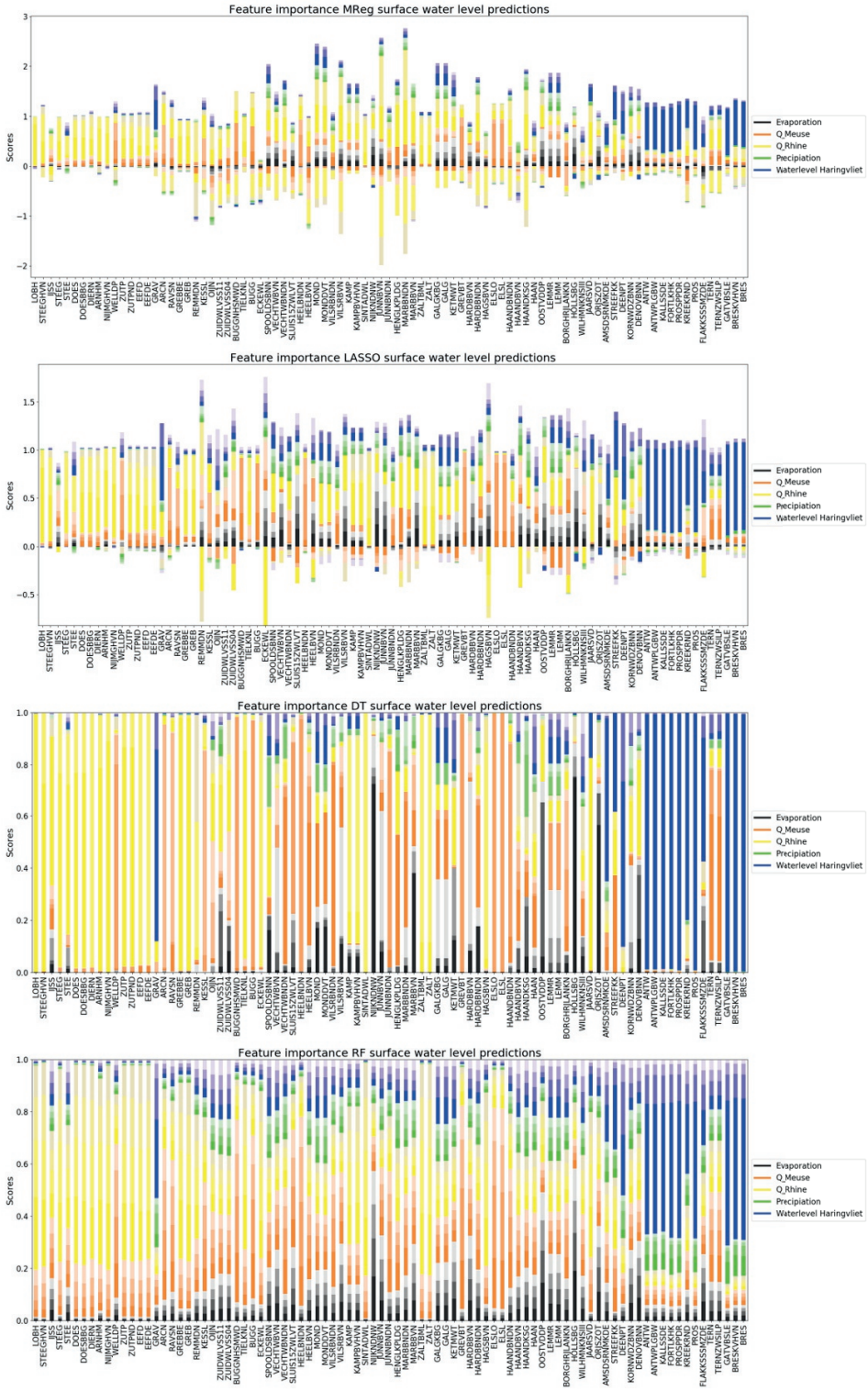


Figure A.8 Feature importance of the MReg, LASSO, DT and RF (top to bottom) for surface water level prediction.

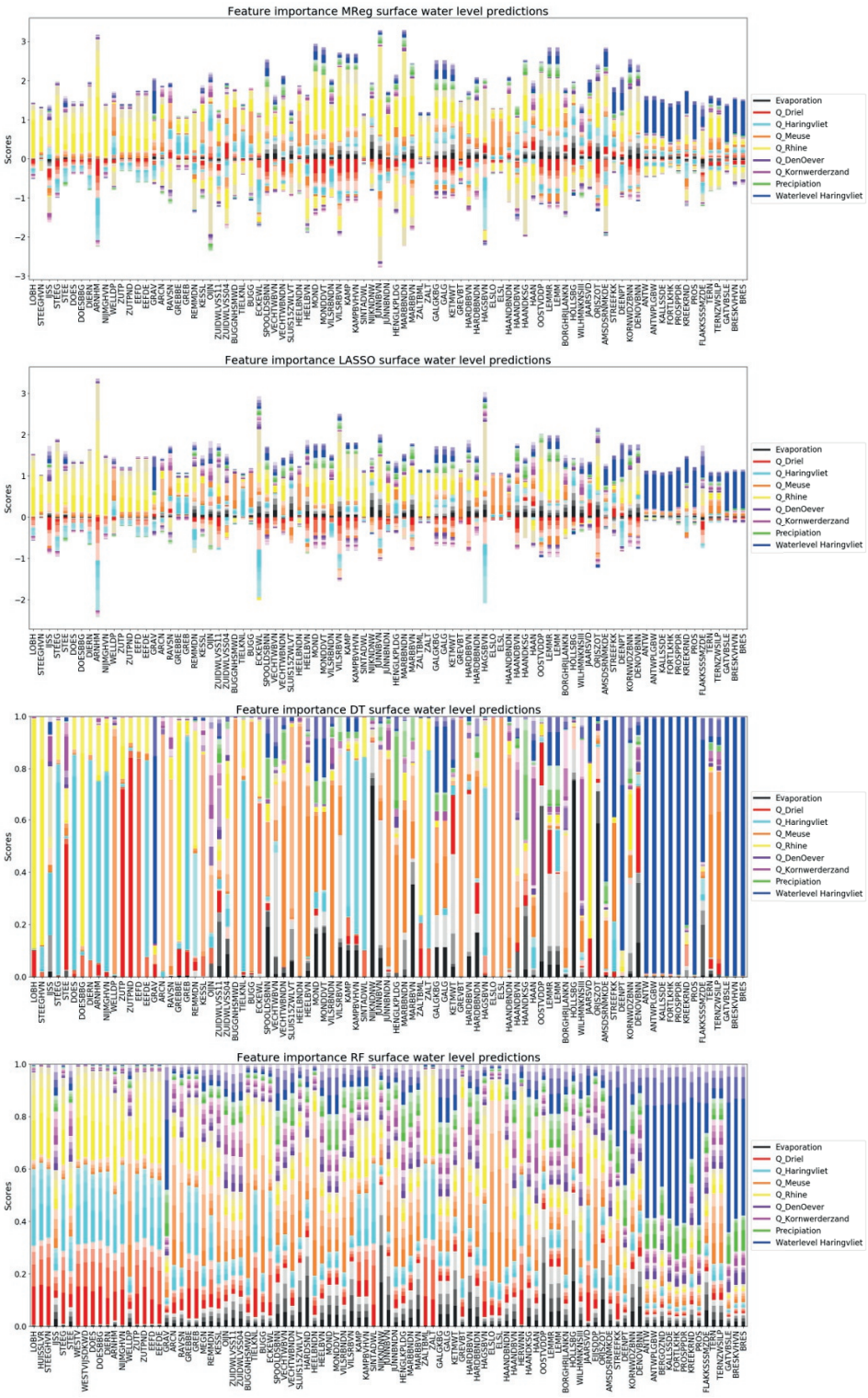


Figure A.9 Feature importance of the MReg, LASSO, DT and RF (top to bottom) for surface water level prediction including water management influence.

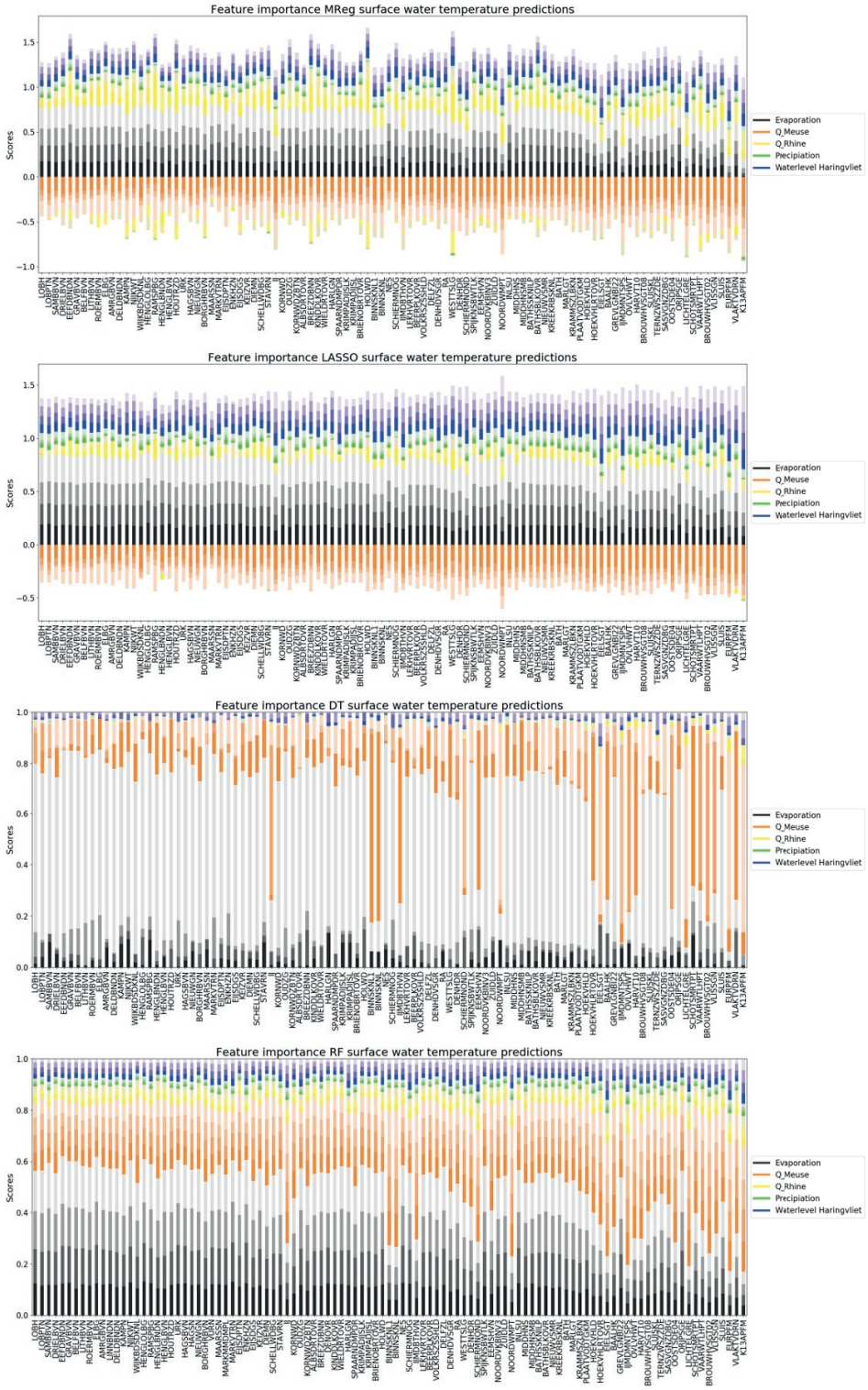


Figure A.10 Feature importance of the MReg, LASSO, DT and RF (top to bottom) for surface water temperature.

Appendix B | Exploring and Optimizing Water Management Strategies for Mitigating Local Drought Impacts

B.1 Material and Methods

B.1.1 Impact functions

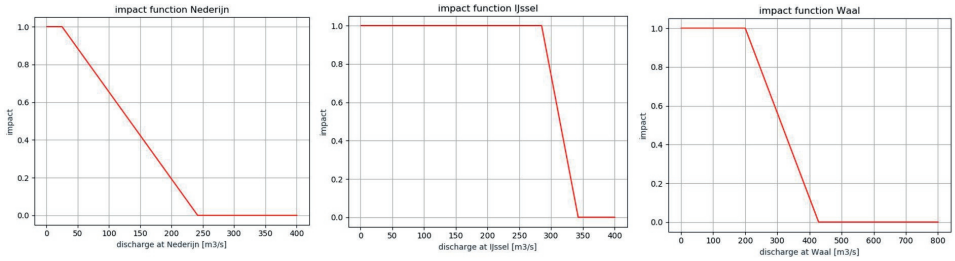


Figure B.1 Overview of impact functions used for the three different river branches in the focus area based on discharge observations and threshold information gathered from national management reports.

B.2 Results

B.2.1 Multi-target LSTM: discharge simulations and model performance

Table B.1 Evaluation scores of the multi-target LSTM for the rivers IJssel and Nederrijn/Lek and the average. Additionally the same evaluation scores for the total simulated timeseries is given.

	Testing data	Total Timeseries
RMSE average	0.077	0.078
RMSE IJssel	0.078	0.090
RMSE Nederrijn/Lek	0.077	0.066
R average	0.976	0.965
R IJssel	0.969	0.952
R Nederijn/Lek	0.982	0.977
KGE average	0.927	0.969
KGE IJssel	0.901	0.918
KGE Nederrijn/Lek	0.871	0.946

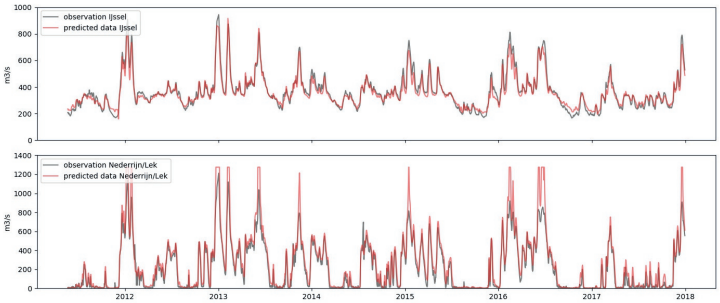


Figure B.2 Multi-target LSTM: comparison of simulated and observed discharges of the rivers IJssel and Nederrijn/Lek for the training data.

B.2.2 Water management scenarios

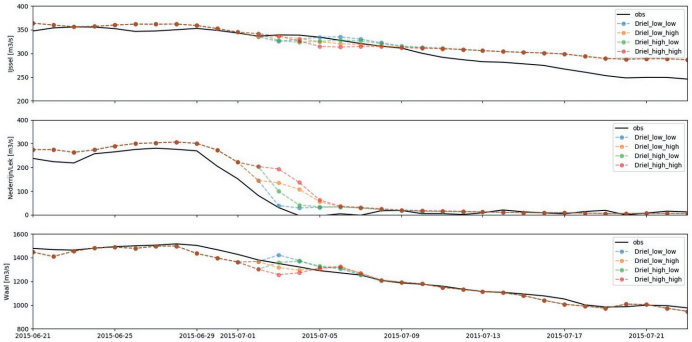


Figure B.3 Sensitivity of the model to different scenario switches on consecutive days and the response time.

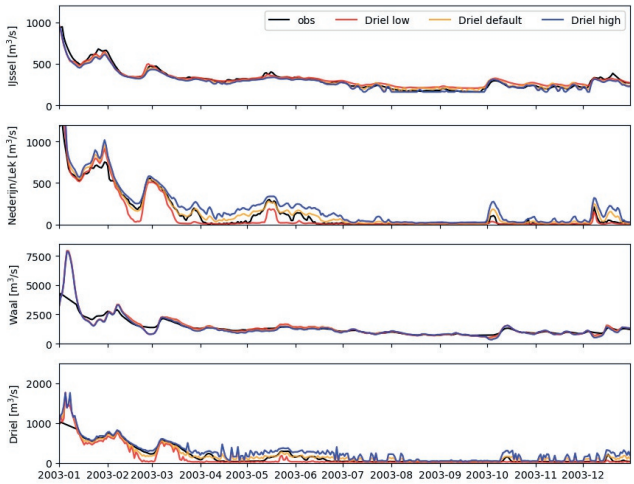


Figure B.4 Baseline discharge simulations for 2003. Shown are the highest, default and lowest scenario results for the three different river branches but also the input at Driel used for the simulation (high corresponding to open management, low corresponding to more constrained management.

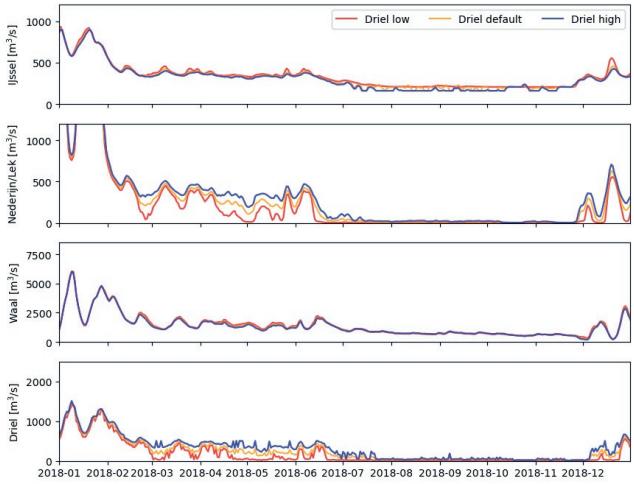


Figure B.5 Baseline discharge simulations for 2018. Shown are the highest, default and lowest scenario results for the three different river branches but also the input at Driel used for the simulation (high corresponding to open management, low corresponding to more constrained management.

B.2.3 Impact functions and first impact assessment exploration

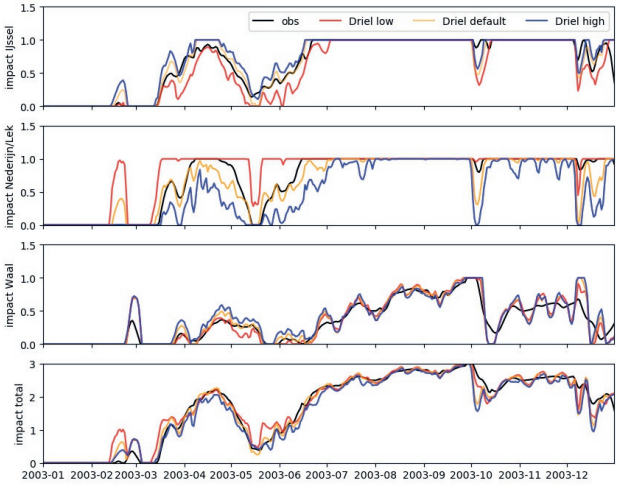


Figure B.6 Baseline impact calculation for 2003 based on simulation output. Shown are the highest, default and lowest scenario results for the three different river branches but also the total impact in the lowest panel.

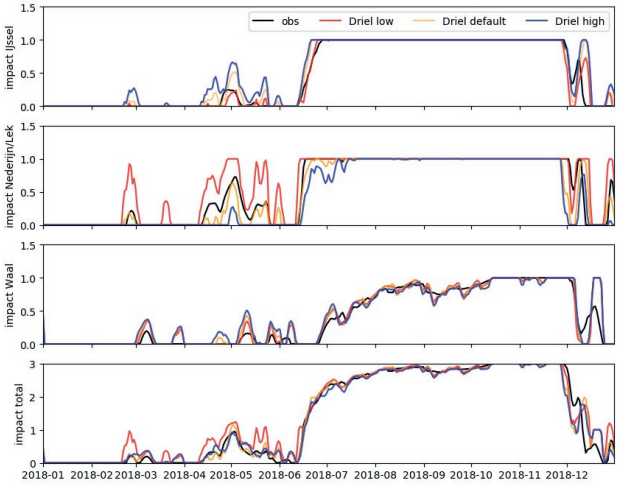


Figure B.7 Baseline impact calculation for 2018 based on simulation output. Shown are the highest, default and lowest scenario results for the three different river branches but also the total impact in the lowest panel.

B.2.4 Optimisation of water management to mitigate drought impacts

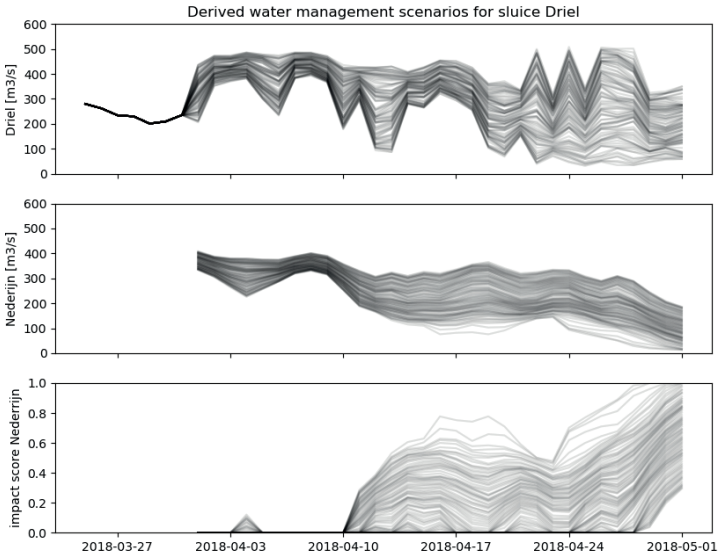


Figure B.8 Example of optimisation simulation starting in April 2018. First panel shows the input at Driel switching from observation to the scenarios. Second and third panel show the discharge simulations for the Nederrijn/Lek and the corresponding impacts for the different scenarios.

Appendix C | The suitability of a seasonal ensemble hybrid framework including data driven approaches for hydrological forecasting

C.1 Material and Methods

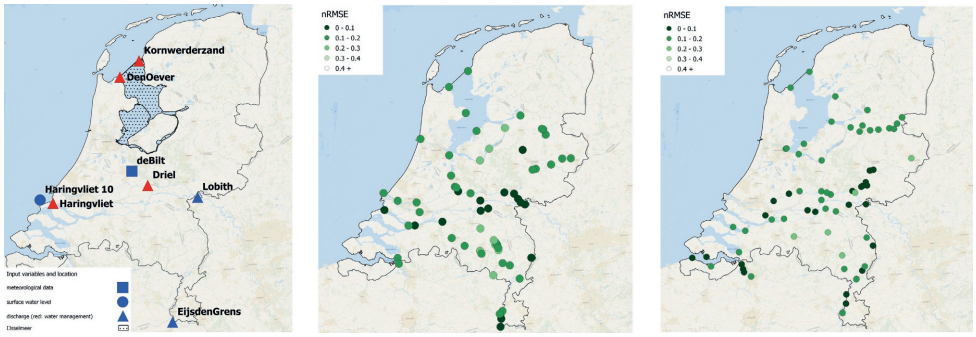


Figure C.1 Overview input stations (left) for ML model framework, as well as the locations of discharge (center) and surface water level (right) target stations used during model development, all figures directly taken from Hauswirth et al. (2021). Discharge and surface water level stations include the RMSE score achieved during evaluation of the modeling framework (more details can be found in Hauswirth et al. (2021)). Maps were created using QGIS (QGIS Development Team, 2022), HCMGIS plugin and basemap data from ESRI Ocean (Sources: Esri, GEBCO, NOAA, National Geographic, DeLorme, HERE, Geonames.org, and other contributors) and GADM database.

C.2 Results

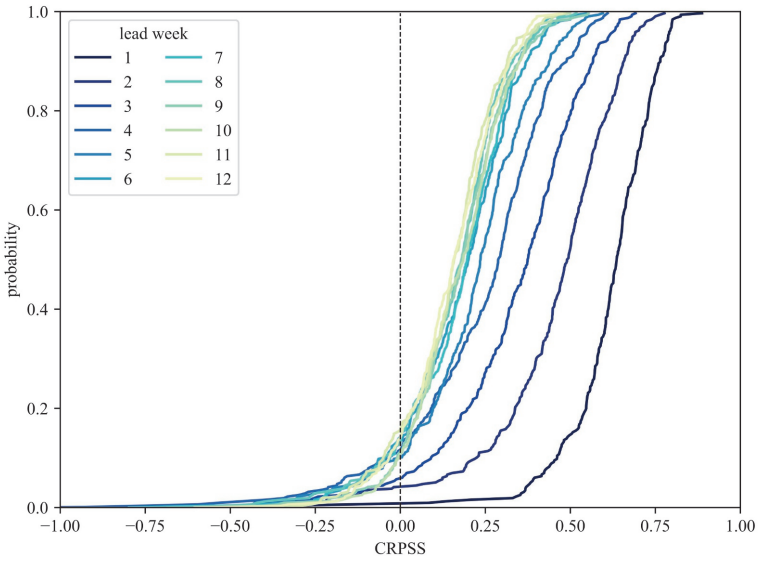


Figure C.2 CDFs of weekly CRPSS shown for different lead weeks with CRPSS being aggregated over all models and stations for fresh surface water level hindcasts (rivers, streams and lakes).

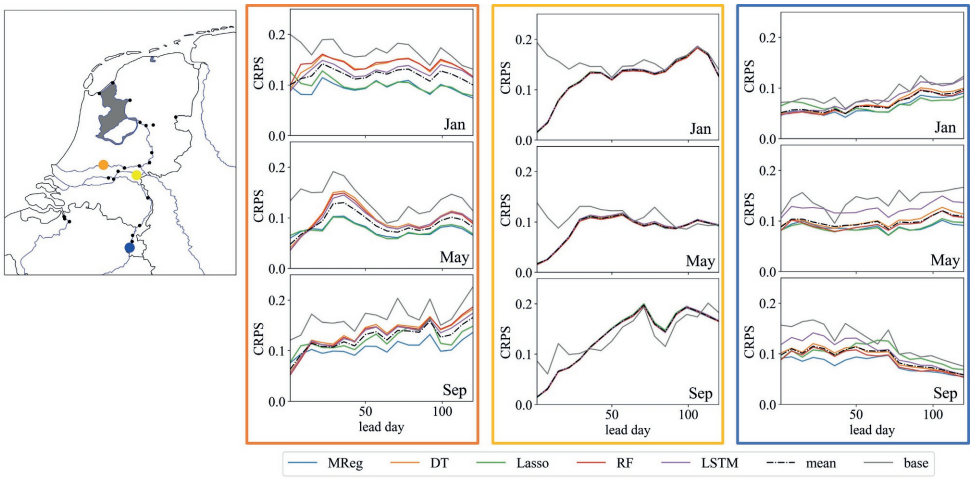


Figure C.3 Overview of weekly CRPS scores for month January, May and September. Different ML model scores, average of ML models score (dashed line) as well as climatological reference (grey) are shown for three surface water level stations along the main river networks (Hagenstein Boven (orange), Nijmegen Haven (yellow), Borgharen Julianakanaal (blue)). Maps were created using the python package Cartopy (Elson et al., 2022), which uses basemap data from Made with Natural Earth and © OpenStreetMap contributors 2022. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

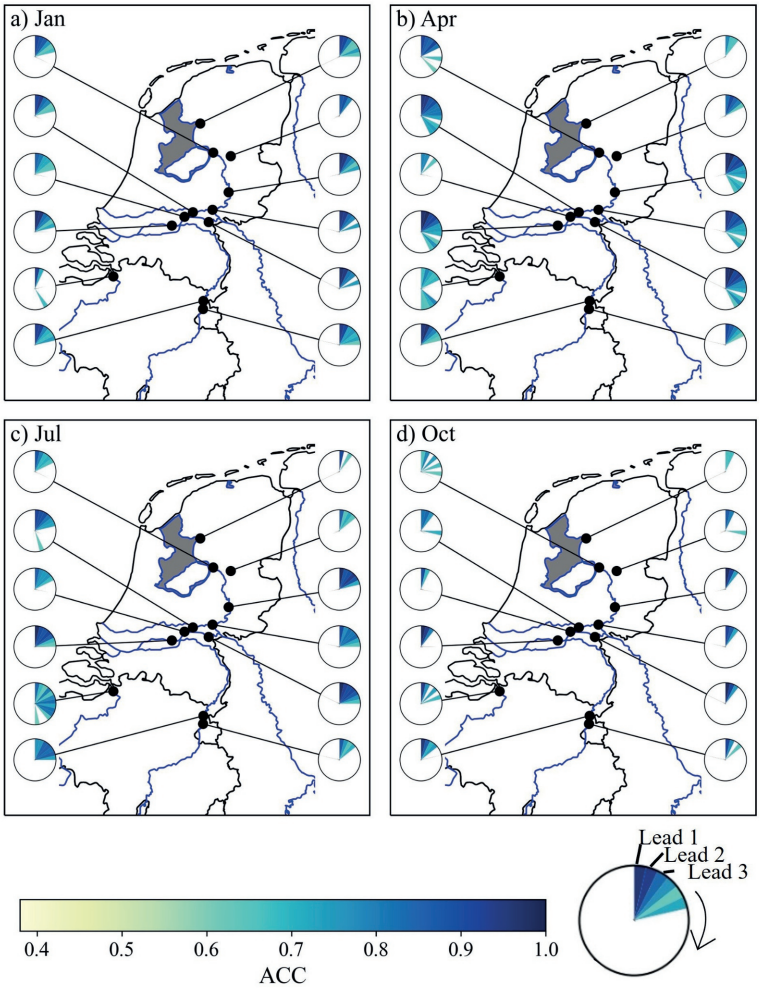


Figure C.4 Anomaly Correlation Coefficient (ACC, weekly) for a) January, b) April, c) July and d) October for surface water level hindcasts (fresh surface water levels such as rivers, streams and lakes). Simulation results are based on the RF model showing the results for different stations (limited selection for visual purpose) in the national monitoring network. Only major river network are shown, smaller streams or infrastructures are not highlighted (station along the coast is placed at a sluice along a stream). Maps were created using the python package Cartopy (Elson et al., 2022), which uses basemap data from Made with Natural Earth and © OpenStreetMap contributors 2022. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

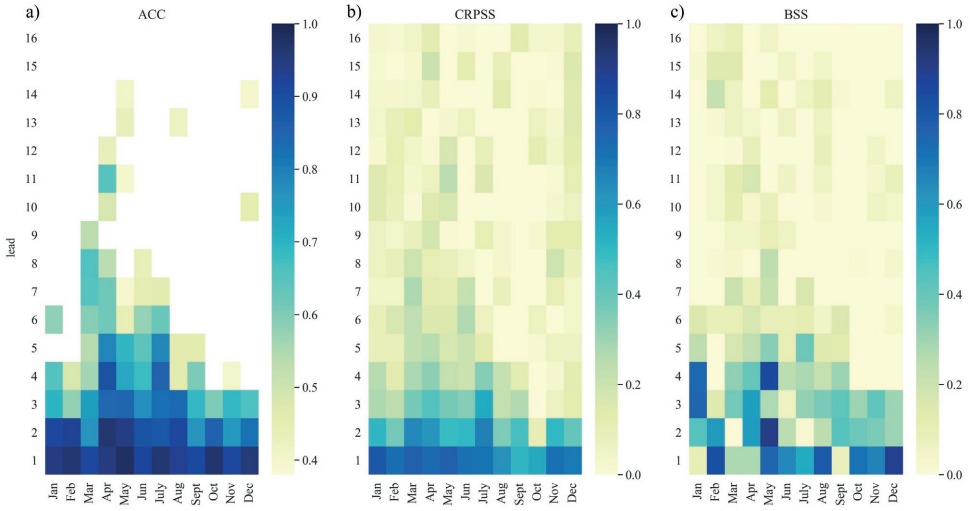
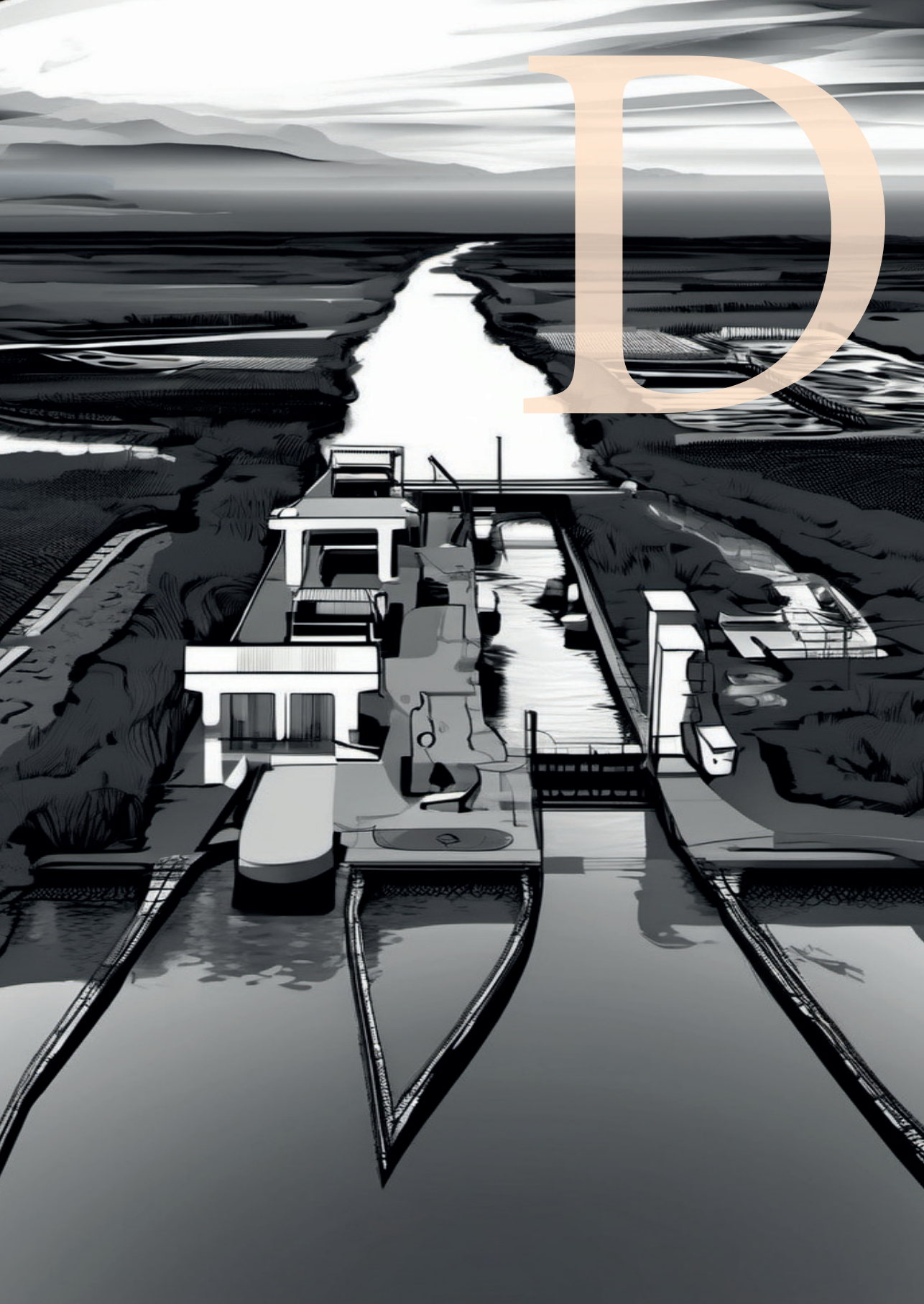


Figure C.5 Overview of weekly evaluation scores for fresh surface water level hindcasts for different initialization months. a) ACC , b) CRPSS and c) BSS heatmaps for example station Nijmegen.



Appendix D | Simulating hydrological extremes for different warming levels - combining large scale climate ensembles with local observation based machine learning models

D.1 Material and Methods

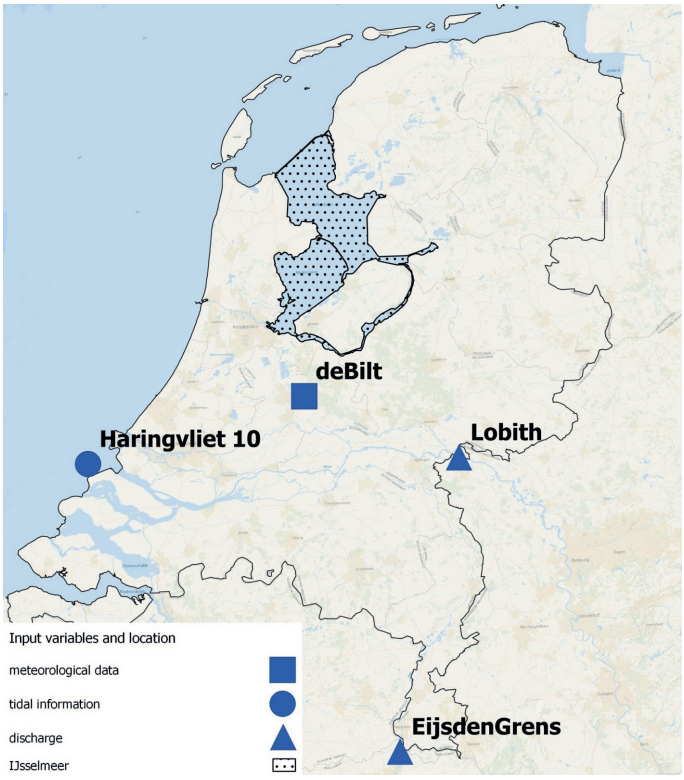


Figure D.1 Overview of locations used for the input data of the ML model framework. Input data includes discharge from stations Lobith and EijsdenGrens, precipitation and evaporation station deBilt as well as the tidal information from Haringvliet 10. Figure taken from Hauswirth et al. (2021), with slight modifications. Maps were created using QGIS (QGIS Development Team, 2022), HCMGIS plugin and basemap data from ESRI Ocean (Sources: Esri, GEBCO, NOAA, National Geographic, DeLorme, HERE, Geonames.org, and other contributors) and GADM database.

D.2 Results

D.2.1 Proof of concept

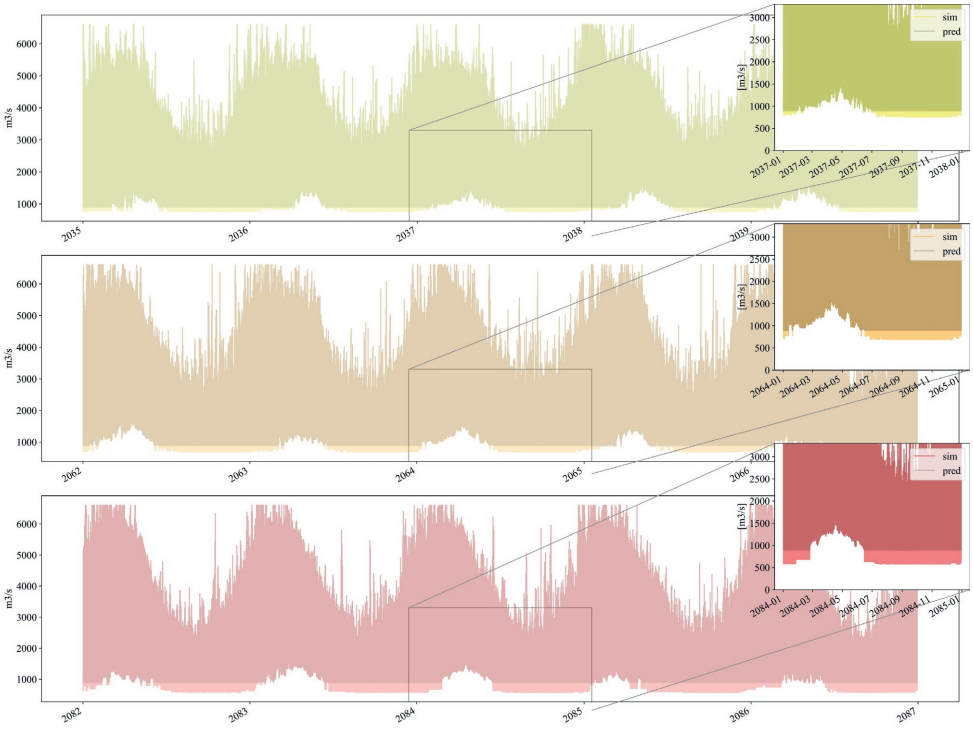


Figure D.2 Post-processing applied to LE discharge simulations for different scenarios (present day (PD, yellow), 2°C and 3°C warmer climate (2C, orange and 3C, red) for station Lobith. Simulations are based on MReg model for that location, while the raw prediction (pred, in grey) shows clear cutoffs for low flows, the simulation results (coloured) after post-processing indicate lower extremes.

D.2.2 Spatial characteristics - regional

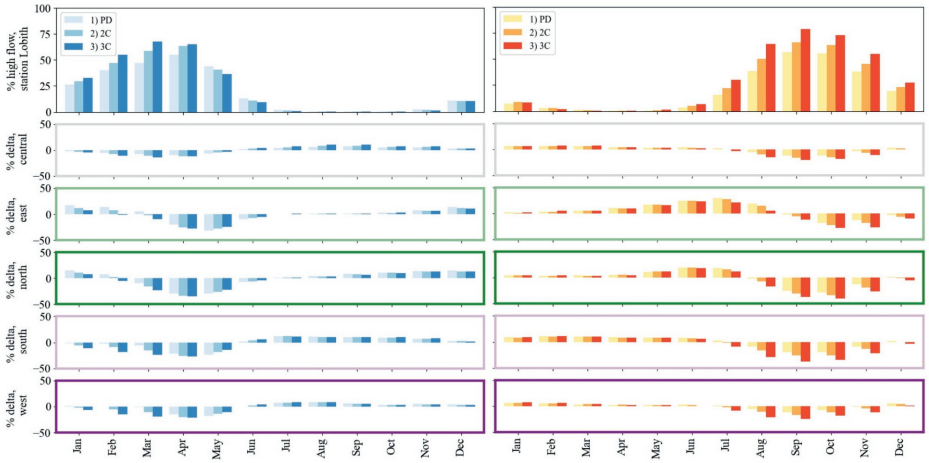


Figure D.3 Differences in seasonal cycle of percentage of high (left) and low (right) flows per scenario: present day (PD), 2°C and 3°C warmer climate (2C and 3C) compared to station Lobith. Results are averaged over different regions based on the LSTM model results.

D.2.3 Climate variability

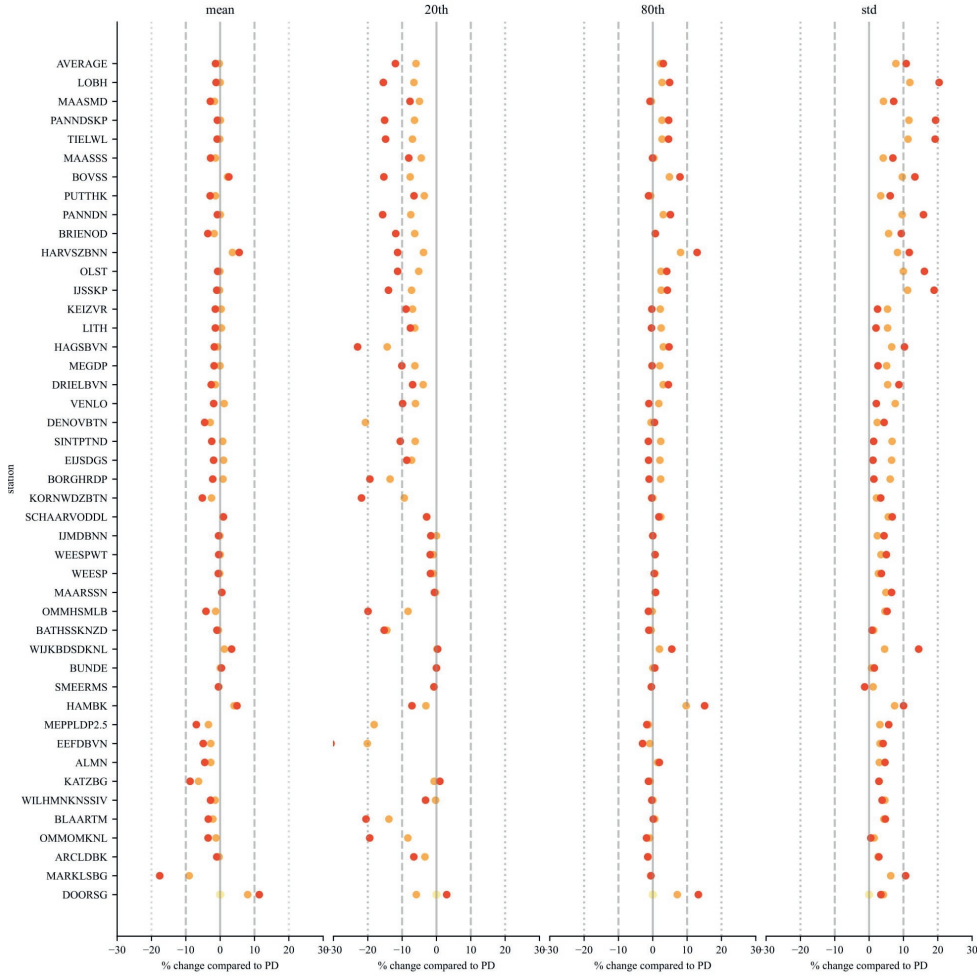


Figure D.4 Differences given in percentage difference compared to present-day scenario in mean discharge, 20^{th} and 80^{th} percentile as well as standard deviation for different scenarios and stations considered in analysis (stations sorted descending based on mean discharge). The differences observed are showing that the results are results based on a shift in climatology.

D.2.4 Frequency high and low flows

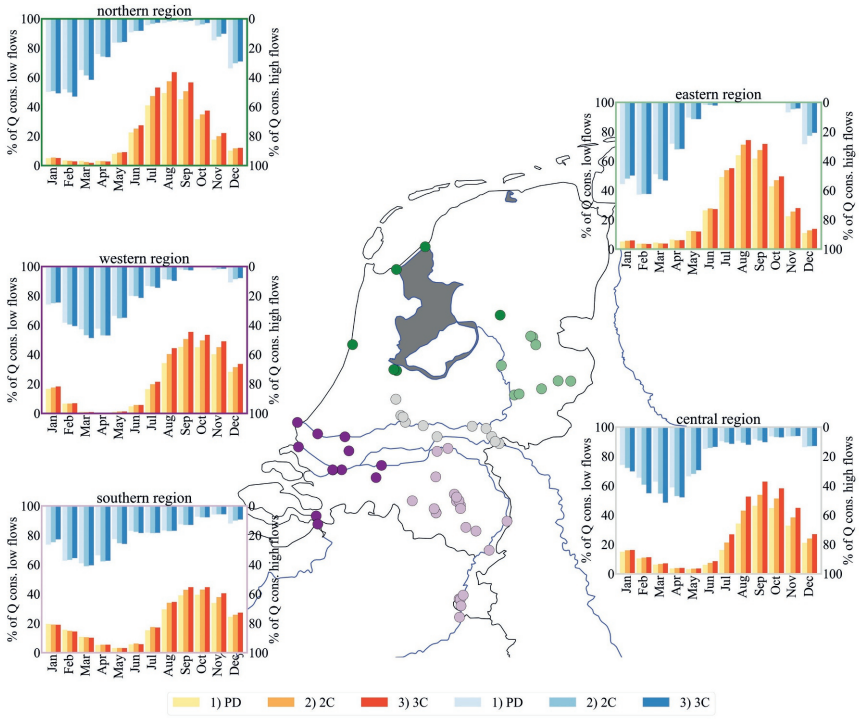


Figure D.5 Seasonal cycle of percentage of low and high flows under different scenarios including present day, 2°C and 3°C warmer climate (PD, 2C and 3C) averaged over different regions (indicated by the group of dots with same colour) based on the Multi-linear regression model results. Major rivers are presented by blue lines, while the dark grey area represents the IJsselmeer lake. Low flows are highlighted in yellow (PD), orange (2C) and red (3C), high flows in light blue (PD), blue (2C) and dark blue (3C).

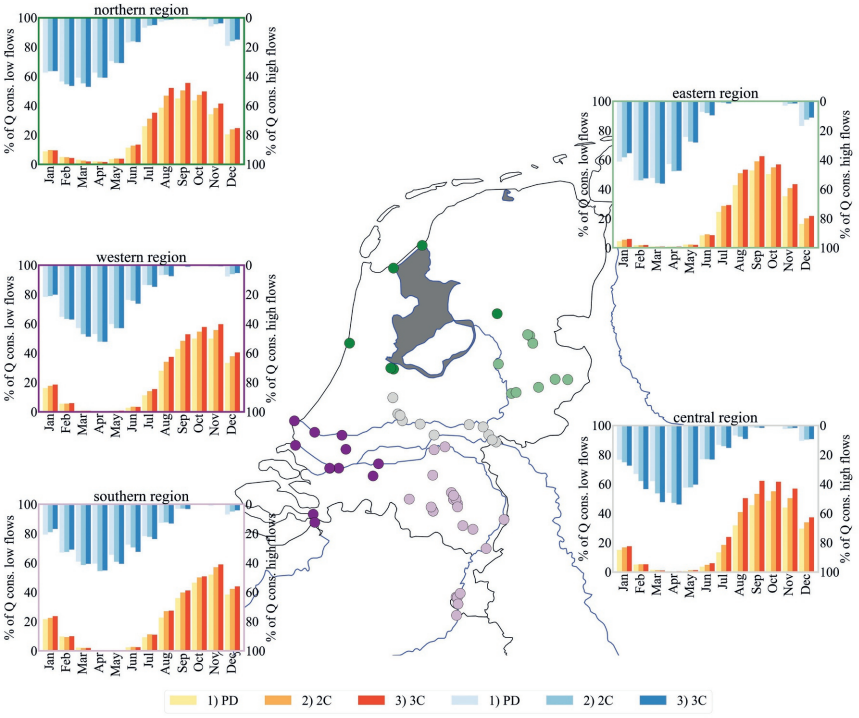


Figure D.6 Seasonal cycle of percentage of low and high flows under different scenarios including present day, 2°C and 3°C warmer climate (PD, 2C and 3C) averaged over different regions (indicated by the group of dots with same colour) based on the LASSO model results. Major rivers are presented by blue lines, while the dark grey area represents the IJsselmeer lake. Low flows are highlighted in yellow (PD), orange (2C) and red (3C), high flows in light blue (PD), blue (2C) and dark blue (3C).

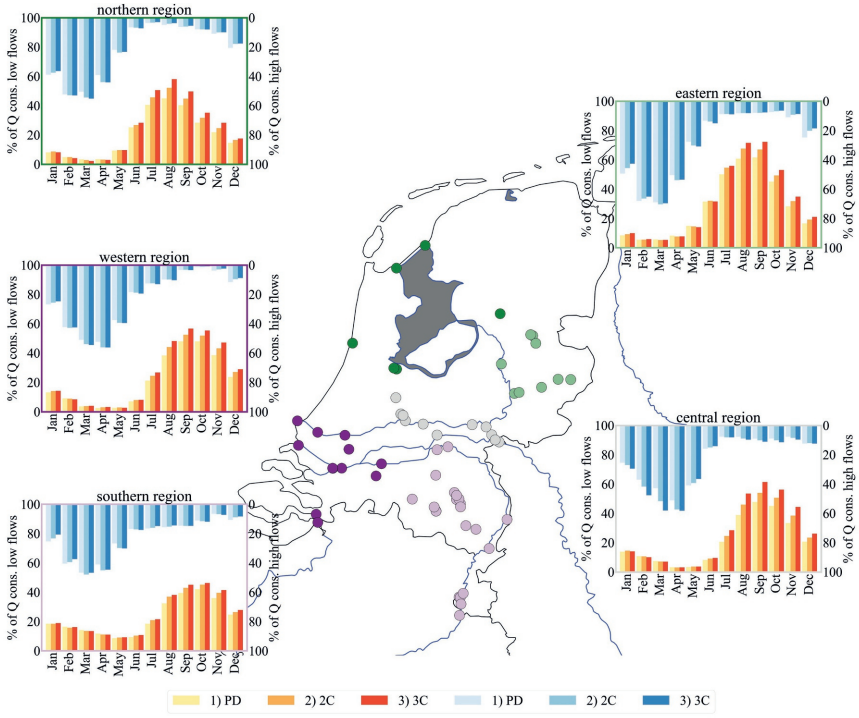


Figure D.7 Seasonal cycle of percentage of low and high flows under different scenarios including present day, 2°C and 3°C warmer climate (PD, 2C and 3C) averaged over different regions (indicated by the group of dots with same colour) based on the Decision Tree model results. Major rivers are presented by blue lines, while the dark grey area represents the IJsselmeer lake. Low flows are highlighted in yellow (PD), orange (2C) and red (3C), high flows in light blue (PD), blue (2C) and dark blue (3C).

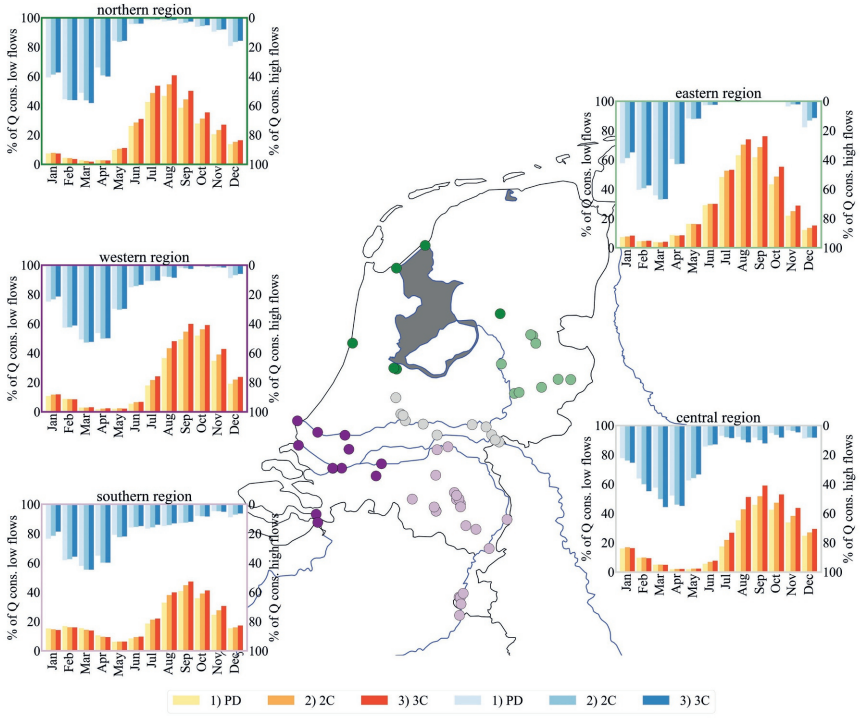


Figure D.8 Seasonal cycle of percentage of low and high flows under different scenarios including present day, 2°C and 3°C warmer climate (PD, 2C and 3C) averaged over different regions (indicated by the group of dots with same colour) based on the RF model results. Major rivers are presented by blue lines, while the dark grey area represents the IJsselmeer lake. Low flows are highlighted in yellow (PD), orange (2C) and red (3C), high flows in light blue (PD), blue (2C) and dark blue (3C).

References

AghaKouchak, A., Mirchi, A., Madani, K., Di Baldassarre, G., Nazemi, A., Alborzi, A., Anjileli, H., Azarderakhsh, M., Chiang, F., Hassanzadeh, E., Huning, L.S., Mallakpour, I., Martinez, A., Mazdiyasni, O., Moftakhari, H., Norouzi, H., Sadegh, M., Sadeqi, D., Van Loon, A.F. & Wanders, N. (2021). Anthropogenic Drought: Definition, Challenges, and Opportunities. en. *Reviews of Geophysics*, 59 (2). _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019RG000683, e2019RG000683. DOI:10.1029/2019RG000683.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J. & Pappenberger, F. (Mar. 2013). GloFAS - global ensemble streamflow forecasting and flood early warning. en. *Hydrology and Earth System Sciences*, 17 (3), pp. 1161–1175. DOI:10.5194/HESS-17-1161-2013.

Alpaydın, E. (2020). *Introduction to machine learning*. eng. Fourth edition. Adaptive computation and machine learning. The MIT Press: Cambridge, Massachusetts London.

Arnal, L., Cloke, H.L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B. & Pappenberger, F. (Apr. 2018). Skilful seasonal forecasts of streamflow over Europe? en. *Hydrology and Earth System Sciences*, 22 (4), pp. 2057–2072. DOI:10.5194/HESS-22-2057-2018.

Aydin, B.E., Tian, X., Delsman, J., Oude Essink, G.H., Rutten, M. & Abraham, E. (Feb. 2019). Optimal salinity and water level control of water courses using Model Predictive Control. en. *Environmental Modelling & Software*, 112, pp. 36–45. DOI:10.1016/J.ENVSOFT.2018.11.010.

Bachmair, S., Kohn, I. & Stahl, K. (June 2015). Exploring the link between drought indicators and impacts. en. *Natural Hazards and Earth System Sciences*, 15 (6), pp. 1381–1397. DOI:10.5194/NHESS-15-1381-2015.

Bachmair, S., Stahl, K., Collins, K., Hannaford, J., Acreman, M., Svoboda, M., Knutson, C., Smith, K.H., Wall, N., Fuchs, B., Crossman, N.D. & Overton, I.C. (July 2016). Drought indicators revisited: the need for a wider consideration of environment and society. en. *WIREs Water*, 3 (4), pp. 516–536. DOI:10.1002/WAT2.1154.

Bachmair, S., Svensson, C., Prosdocimi, I., Hannaford, J. & Stahl, K. (Nov. 2017). Developing drought impact functions for drought risk management. en. *Natural Hazards and Earth System Sciences*, 17 (11), pp. 1947–1960. DOI:10.5194/NHESS-17-1947-2017.

Bakke, S.J., Ionita, M. & Tallaksen, L.M. (Nov. 2020). The 2018 northern European hydrological drought and its drivers in a historical perspective. English. *Hydrology and Earth System Sciences*, 24 (11). Publisher: Copernicus GmbH, pp. 5621–5653. DOI:10.5194/HESS-24-5621-2020.

Bardsley, W., Vetrova, V. & Liu, S. (Oct. 2015). Toward creating simpler hydrological models: A LASSO subset selection approach. en. *Environmental Modelling & Software*, 72, pp. 33–43. DOI:10.1016/J.ENVSOFT.2015.06.008.

Bartholomeus, R.P., Van Der Wiel, K., Van Loon, A.F., Van Huijgevoort, M.H., Van Vliet, M.T.H., Mens, M., Muurling-van Geffen, S., Wanders, N. & Pot, W. (May 2023). Managing water across the flood-drought spectrum - experiences from and challenges for the Netherlands. en. *Cambridge Prisms: Water*, pp. 1–22. DOI:10.1017/WAT.2023.4.

Beillouin, D., Schauberger, B., Bastos, A., Ciais, P. & Makowski, D. (Sept. 2020). Impact of extreme weather conditions on European crop production in 2018. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375 (1810). Publisher: Royal Society, p. 20190510. DOI:10.1098/RSTB.2019.0510.

Bertsekas, D.P. (2011). Chapter 6 Approximate Dynamic Programming. en. In: *Dynamic Programming and Optimal Control 3rd Edition, Volume II*, p. 233.

Bierkens, M.F.P. (July 2015). Global hydrology 2015: State, trends, and directions. en. *Water Resources Research*, 51 (7), pp. 4923–4947. DOI:10.1002/2015WR017173.

Blauhut, V. (Nov. 2020). The triple complexity of drought risk analysis and its visualisation via mapping: a review across scales and sectors. *Earth-Science Reviews*, 210, p. 103345. DOI:10.1016/J.EARSCIREV.2020.103345.

Blauhut, V., Gudmundsson, L. & Stahl, K. (Jan. 2015). Towards pan-European drought risk maps: quantifying the link between drought indices and reported drought impacts. en. *Environmental Research Letters*, 10 (1), p. 014008. DOI:10.1088/1748-9326/10/1/014008.

Blauhut, V. et al. (June 2022). Lessons from the 2018-2019 European droughts: a collective need for unifying drought risk management. English. *Natural Hazards and Earth System Sciences*, 22 (6). Publisher: Copernicus GmbH, pp. 2201–2217. DOI:10.5194/NHESS-22-2201-2022.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45 (1), pp. 5–32. DOI:10.1023/A:1010933404324.

Brier, G.W. (Jan. 1950). VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. EN. *Monthly Weather Review*, 78 (1). Publisher: American Meteorological Society Section: Monthly Weather Review, pp. 1–3. DOI:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Camacho, E.F. & Bordons, C. (1999). *Model predictive control*. en. Advanced textbooks in control and signal processing. Springer: Berlin ; New York.

Cammalleri, C., Naumann, G., Mentaschi, L., Formetta, G., Forzieri, G., Gosling, S., Bisselink, B., De Roo, A. & Feyen, L. (2020). *Global warming and drought impacts in the EU: JRC PESETA IV project : Task 7*. en. Tech. rep. LU: Publications Office.

Candogan Yossef, N., van Beek, R., Weerts, A., Winsemius, H. & Bierkens, M.F.P. (Aug. 2017). Skill of a global forecasting system in seasonal ensemble streamflow prediction. en. *Hydrology and Earth System Sciences*, 21 (8), pp. 4103–4114. DOI:10.5194/HESS-21-4103-2017.

Clark, M.P., Bierkens, M.F.P., Samaniego, L., Woods, R.A., Uijlenhoet, R., Bennett, K.E., Pauwels, V.R.N., Cai, X., Wood, A.W. & Peters-Lidard, C.D. (July 2017). The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism. en. *Hydrology and Earth System Sciences*, 21 (7), pp. 3427–3440. DOI:10.5194/HESS-21-3427-2017.

Cook, B.I., Mankin, J.S., Marvel, K., Williams, A.P., Smerdon, J.E. & Anchukaitis, K.J. (2020). Twenty-First Century Drought Projections in the CMIP6 Forcing Scenarios. en. *Earth's Future*, 8 (6). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019EF001461, e2019EF001461. DOI:10.1029/2019EF001461.

De Bruin, H.A.R. & Lablans, W.N. (June 1998). Reference crop evapotranspiration determined with a modified Makkink equation. en. *Hydrological Processes*, 12 (7), pp. 1053–1062. DOI:10.1002/(SICI)1099-1085(19980615)12:7<1053::AID-HYP639>3.0.CO;2-E.

De Jong, J. (2019). *Eenvoudige scheepvaartrelatie voor vaarkosten Waal door afname vaardiepte*. nl. Deltares Memo.

De Lange, W.J., Prinsen, G.F., Hoogewoud, J.C., Veldhuizen, A.A., Verkaik, J., Oude Essink, G.H., van Walsum, P.E., Delsman, J.R., Hunink, J.C., Massop, H.T. & Kroon, T. (Sept. 2014). An operational, multi-scale, multi-model system for consensus-based, integrated water management and policy analysis: The Netherlands Hydrological Instrument. en. *Environmental Modelling & Software*, 59, pp. 98–108. DOI:10.1016/J.ENVSOFT.2014.05.009.

De Roo, A.P.J., Wesseling, C.G. & Van Deursen, W.P.A. (2000). Physically based river basin modelling within a GIS: the LISFLOOD model. en. *Hydrological Processes*, 14 (11-12). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1099-1085%2820000815/30%2914%3A11/12%3C1981%3A%3AAID-HYP49%3E3.0.CO%3B2-F, pp. 1981–1992. DOI:10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F.

Deser, C., Lehner, F., Rodgers, K.B., Ault, T., Delworth, T.L., DiNezio, P.N., Fiore, A., Frankignoul, C., Fyfe, J.C., Horton, D.E., Kay, J.E., Knutti, R., Lovenduski, N.S., Marotzke, J., McKinnon, K.A., Minobe, S., Randerson, J., Screen, J.A., Simpson, I.R. & Ting, M. (Apr. 2020). Insights from Earth system model initial-condition large ensembles and future prospects. en. *Nature Climate Change*, 10 (4). Number: 4 Publisher: Nature Publishing Group, pp. 277–286. DOI:10.1038/S41558-020-0731-2.

Di Baldassarre, G., Cloke, H., Lindersson, S., Mazzoleni, M., Mondino, E., Mård, J., Odongo, V., Raffetti, E., Ridolfi, E., Rusca, M., Savelli, E. & Tootoonchi, F. (Sept. 2021). Integrating Multiple Research Methods to Unravel the Complexity of Human-Water Systems. en. *AGU Advances*, 2 (3). DOI:10.1029/2021AV000473.

Elson, P. et al. (Jan. 2022). *SciTools/cartopy: v0.20.2*. Version v0.20.2. DOI:10.5281/ZENODO.5842769.

European Commission. Joint Research Centre. (2020). *Global warming and drought impacts in the EU: JRC PESETA IV project : Task 7*. en. Publications Office: LU.

European-Drought-Centre (2013). *European Drought Centre*.

Felsche, E. & Ludwig, R. (Dec. 2021). Applying machine learning for drought prediction in a perfect model framework using data from a large ensemble of climate simulations. English. *Natural Hazards and Earth System Sciences*, 21 (12). Publisher: Copernicus GmbH, pp. 3679–3691. DOI:10.5194/NHESS-21-3679-2021.

Feng, D., Fang, K. & Shen, C. (Sept. 2020). Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales. en. *Water Resources Research*, 56 (9). DOI:10.1029/2019WR026793.

Feng, D., Lawson, K. & Shen, C. (2021). Mitigating Prediction Error of Deep Learning Streamflow Models in Large Data-Sparse Regions With Ensemble Modeling and Soft Data. en. *Geophysical Research Letters*, 48 (14). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021GL092999, e2021GL092999. DOI:10.1029/2021GL092999.

Fowler, H.J., Blenkinsop, S. & Tebaldi, C. (2007). Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. en. *International Journal of Climatology*, 27, pp. 1547–1578. DOI:10.1002/JOC.1556.

Freeze, R.A. & Harlan, R.L. (Nov. 1969). Blueprint for a physically-based, digitally-simulated hydrologic response model. *Journal of Hydrology*, 9 (3), pp. 237–258. DOI:10.1016/0022-1694(69)90020-1.

Giorgi, F. (2019). Thirty Years of Regional Climate Modeling: Where Are We and Where Are We Going next? en. *Journal of Geophysical Research: Atmospheres*, 124 (11). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018JD030094, pp. 5696–5723. DOI:10.1029/2018JD030094.

Girons Lopez, M., Crochemore, L. & Pechlivanidis, I.G. (Mar. 2021). Benchmarking an operational hydrological model for providing seasonal forecasts in Sweden. en. *Hydrology and Earth System Sciences*, 25 (3), pp. 1189–1209. DOI:10.5194/HESS-25-1189-2021.

Goulart, H.M.D., van der Wiel, K., Folberth, C., Balkovic, J. & van den Hurk, B. (Dec. 2021). Storylines of weather-induced crop failure events under climate change. English. *Earth System Dynamics*, 12 (4). Publisher: Copernicus GmbH, pp. 1503–1527. DOI:10.5194/ESD-12-1503-2021.

Graham, L.P., Andréasson, J. & Carlsson, B. (May 2007). Assessing climate change impacts on hydrology from an ensemble of regional climate models, model scales and linking methods - a case study on the Lule River basin. en. *Climatic Change*, 81 (S1), pp. 293–307. DOI:10.1007/S10584-006-9215-2.

Gupta, H.V., Kling, H., Yilmaz, K.K. & Martinez, G.F. (Oct. 2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. en. *Journal of Hydrology*, 377 (1-2), pp. 80–91. DOI:10.1016/J.JHYDROL.2009.08.003.

Hari, V., Rakovec, O., Markonis, Y., Hanel, M. & Kumar, R. (Dec. 2020). Increased future occurrences of the exceptional 2018-2019 Central European drought under global warming. en. *Scientific Reports*, 10 (1), p. 12207. DOI:10.1038/S41598-020-68872-9.

Hattermann, F.F., Vetter, T., Breuer, L., Su, B., Daggupati, P., Donnelly, C., Fekete, B., Flörke, F., Gosling, S.N., Hoffmann, P., Liersch, S., Masaki, Y., Motovilov, Y., Müller, C., Samaniego, L., Stacke, T., Wada, Y., Yang, T. & Krysnaova, V. (Jan. 2018). Sources of uncertainty in hydrological climate impact assessment: a cross-scale study. en. *Environmental Research Letters*, 13 (1). Publisher: IOP Publishing, p. 015006. DOI:10.1088/1748-9326/AA9938.

Hauswirth, S.M., Bierkens, M.F.P., Beijk, V. & Wanders, N. (Mar. 2022). The suitability of a hybrid framework including data driven approaches for hydrological forecasting. English. *Hydrology and Earth System Sciences Discussions*. Publisher: Copernicus GmbH, pp. 1–20. DOI:10.5194/HESS-2022-89.

Hauswirth, S.M., Bierkens, M.F., Beijk, V. & Wanders, N. (Sept. 2021). The potential of data driven approaches for quantifying hydrological extremes. en. *Advances in Water Resources*, 155, p. 104017. DOI:10.1016/J.ADVWATRES.2021.104017.

Hazeleger, W., Wang, X., Severijns, C., Ştefănescu, S., Bintanja, R., Sterl, A., Wyser, K., Semmler, T., Yang, S., van den Hurk, B., van Noije, T., van der Linden, E. & van der Wiel, K. (Dec. 2012). EC-Earth V2.2: description and validation of a new seamless earth system prediction model. en. *Climate Dynamics*, 39 (11), pp. 2611–2629. DOI:10.1007/S00382-011-1228-5.

He, Q., Barajas-Solano, D., Tartakovsky, G. & Tartakovsky, A.M. (July 2020). Physics-informed neural networks for multiphysics data assimilation with application to subsurface transport. *Advances in Water Resources*, 141, p. 103610. DOI:10.1016/J.ADVWATRES.2020.103610.

Herrera-Estrada, J.E. & Diffenbaugh, N.S. (Sept. 2020). Landfalling Droughts: Global Tracking of Moisture Deficits From the Oceans Onto Land. en. *Water Resources Research*, 56 (9). DOI:10.1029/2019WR026877.

Hersbach, H. (Oct. 2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. en. *Weather and Forecasting*, 15 (5), pp. 559–570. DOI:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hochreiter, S. & Schmidhuber, J. (Nov. 1997). Long Short-Term Memory. en. *Neural Computation*, 9 (8), pp. 1735–1780. DOI:10.1162/NECO.1997.9.8.1735.

Hsu, K.-l., Gupta, H.V. & Sorooshian, S. (1995). Artificial Neural Network Modeling of the Rainfall-Runoff Process. en. *Water Resources Research*, 31 (10). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/95WR01955, pp. 2517–2530. DOI:10.1029/95WR01955.

Hunt, K.M.R., Matthews, G.R., Pappenberger, F. & Prudhomme, C. (Feb. 2022). Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. English. *Hydrology and Earth System Sciences Discussions*. Publisher: Copernicus GmbH, pp. 1–30. DOI:10.5194/HESS-2022-53.

Ionita, M., Tallaksen, L.M., Kingston, D.G., Stagge, J.H., Laaha, G., Van Lanen, H.A.J., Scholz, P., Chelcea, S.M. & Haslinger, K. (Mar. 2017). The European 2015 drought from a climatological perspective. en. *Hydrology and Earth System Sciences*, 21 (3), pp. 1397–1419. DOI:10.5194/HESS-21-1397-2017.

Johnson, S.J., Stockdale, T.N., Ferranti, L., Balmaseda, M.A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S.P.E., Mogensen, K., Zuo, H. & Monge-Sanz, B.M. (Mar. 2019). SEAS5: the new ECMWF seasonal forecast system. en. *Geoscientific Model Development*, 12 (3), pp. 1087–1117. DOI:10.5194/GMD-12-1087-2019.

Kelder, T., Wanders, N., Wiel, K.v.d., Marjoribanks, T.I., Slater, L.J., Wilby, R.l & Prudhomme, C. (Mar. 2022). Interpreting extreme climate impacts from large ensemble simulations—are they unseen or unrealistic? en. *Environmental Research Letters*, 17 (4). Publisher: IOP Publishing, p. 044052. DOI:10.1088/1748-9326/AC5CF4.

KNMI (2021). *KNMI Klimaatsignaal '21 - Hoe het klimaat in Nederland snel verandert*. nl. Tech. rep. De Bilt: KNMI, p. 72.

KNMI (2023). *KNMI'23 Klimaatscenario's voor Nederland*. nl. Tech. rep. De Bilt: KNMI.

Koch, J., Berger, H., Henriksen, H.J. & Sonnenborg, T.O. (Nov. 2019). Modelling of the shallow water table at high spatial resolution using random forests. en. *Hydrology and Earth System Sciences*, 23 (11), pp. 4603–4619. DOI:10.5194/HESS-23-4603-2019.

Koch, J., Gotfredsen, J., Schneider, R., Troldborg, L., Stisen, S. & Henriksen, H.J. (Sept. 2021). High Resolution Water Table Modeling of the Shallow Groundwater Using a Knowledge-Guided Gradient Boosting Decision Tree Model. en. *Frontiers in Water*, 3, p. 701726. DOI:10.3389/FRWA.2021.701726.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K. & Herrnegger, M. (Nov. 2018). Rainfall‐runoff modelling using Long Short-Term Memory (LSTM) networks. en. *Hydrology and Earth System Sciences*, 22 (11), pp. 6005–6022. DOI:10.5194/HESS-22-6005-2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S. & Nearing, G.S. (Dec. 2019a). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. en. *Water Resources Research*, 55 (12), pp. 11344–11354. DOI:10.1029/2019WR026065.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S. & Nearing, G. (Nov. 2019b). Towards Learning Universal, Regional, and Local Hydrological Behaviors via Machine-Learning Applied to Large-Sample Datasets. en. *arXiv:1907.08456 [cs, stat]*. arXiv: 1907.08456.

Li, W., Kiaghadi, A. & Dawson, C. (Feb. 2021). High temporal resolution rainfall‐runoff modeling using long-short-term-memory (LSTM) networks. en. *Neural Computing and Applications*, 33 (4), pp. 1261–1278. DOI:10.1007/S00521-020-05010-6.

Logar, I. & van den Bergh, J.C.J.M. (Apr. 2013). Methods to Assess Costs of Drought Damages and Policies for Drought Mitigation and Adaptation: Review and Recommendations. en. *Water Resources Management*, 27 (6), pp. 1707–1720. DOI:10.1007/S11269-012-0119-9.

Ma, K., Feng, D., Lawson, K., Tsai, W., Liang, C., Huang, X., Sharma, A. & Shen, C. (May 2021). Transferring Hydrologic Data Across Continents ‐Leveraging Data-Rich Regions to Improve Hydrologic Prediction in Data-Sparse Regions. en. *Water Resources Research*, 57 (5). DOI:10.1029/2020WR028600.

Magni, M., Sutanudjaja, E.H., Shen, Y. & Karssenberg, D. (Aug. 2023). Global streamflow modelling using process-informed machine learning. *Journal of Hydroinformatics*, 25 (5), pp. 1648–1666. DOI:10.2166/HYDRO.2023.217.

Maher, N., Milinski, S. & Ludwig, R. (Apr. 2021). Large ensemble climate model simulations: introduction, overview, and future prospects for utilising multiple types of large ensemble. English. *Earth System Dynamics*, 12 (2). Publisher: Copernicus GmbH, pp. 401–418. DOI:10.5194/ESD-12-401-2021.

Mai, J., Shen, H., Tolson, B.A., Gaborit, E., Arsenault, R., Craig, J.R., Fortin, V., Fry, L.M., Gauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D.G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N.K., Temgoua, A.G.T., Vionnet, V. & Waddell, J.W. (July 2022). The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL). English. *Hydrology and Earth System Sciences*, 26 (13). Publisher: Copernicus GmbH, pp. 3537–3572. DOI:10.5194/HESS-26-3537-2022.

Majumdar, S., Smith, R., Butler, J.J. & Lakshmi, V. (Nov. 2020). Groundwater Withdrawal Prediction Using Integrated Multitemporal Remote Sensing Data Sets and Machine Learning. en. *Water Resources Research*, 56 (11). DOI:10.1029/2020WR028059.

Marx, A., Kumar, R., Thober, S., Rakovec, O., Wanders, N., Zink, M., Wood, E.F., Pan, M., Sheffield, J. & Samaniego, L. (Feb. 2018). Climate change alters low flows in Europe under global warming of 1.5, 2, and 3 °C. en. *Hydrology and Earth System Sciences*, 22 (2), pp. 1017–1032. DOI:10.5194/HESS-22-1017-2018.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M.I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, R., Maycock, T.K., Waterfield, T., Yelekçi, Ö., Yu, R. & Zhou, B. (2021). *IPCC, 2021: Summary for Policymakers. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Tech. rep. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M.I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, R., Maycock, T.K., Waterfield, T., Yelekçi, Ö., Yu, R. & Zhou, B. (2022). *IPCC, 2021: Summary for Policymakers. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Tech. rep. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Mens, M., Minnema, B., Overmars, K. & van den Hurk, B. (Sept. 2021). Dilemmas in developing models for long-term drought risk management: The case of the National Water Model of the Netherlands. en. *Environmental Modelling & Software*, 143, p. 105100. DOI:10.1016/J.ENVSOFT.2021.105100.

Milly, P.C.D., Dunne, K.A. & Vecchia, A.V. (Nov. 2005). Global pattern of trends in streamflow and water availability in a changing climate. en. *Nature*, 438 (7066). Number: 7066 Publisher: Nature Publishing Group, pp. 347–350. DOI:10.1038/NATURE04312.

Miro, M.E., Groves, D., Tincher, B., Syme, J., Tanverakul, S. & Catt, D. (Jan. 2021). Adaptive water management in the face of uncertainty: Integrating machine learning, groundwater modeling and robust decision making. *Climate Risk Management*, 34, p. 100383. DOI:10.1016/J.CRM.2021.100383.

Nash, J.E. & Sutcliffe, J.V. (Apr. 1970). River flow forecasting through conceptual models part I —A discussion of principles. *Journal of Hydrology*, 10 (3), pp. 282–290. DOI:10.1016/0022-1694(70)90255-6.

Naumann, G., Cammalleri, C., Mentaschi, L. & Feyen, L. (June 2021). Increased economic drought impacts in Europe with anthropogenic warming. en. *Nature Climate Change*, 11 (6). Number: 6 Publisher: Nature Publishing Group, pp. 485–491. DOI:10.1038/s41558-021-01044-3.

O' Neill, B.C., Kriegler, E., Riahi, K., Ebi, K.L., Hallegatte, S., Carter, T.R., Mathur, R. & van Vuuren, D.P. (Feb. 2014). A new scenario framework for climate change research: the concept of shared socioeconomic pathways. en. *Climatic Change*, 122 (3), pp. 387–400. DOI:10.1007/s10584-013-0905-2.

Pachauri, R.K., Mayer, L. & on Climate Change, I.P., eds. (2015). *Climate change 2014: synthesis report*. eng. Intergovernmental Panel on Climate Change: Geneva, Switzerland.

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

Pechlivanidis, I.G., Crochemore, L., Rosberg, J. & Bosshard, T. (June 2020). What Are the Key Drivers Controlling the Quality of Seasonal Streamflow Forecasts? en. *Water Resources Research*, 56 (6). DOI:10.1029/2019WR026987.

Philip, S.Y., Kew, S.F., van der Wiel, K., Wanders, N. & Jan van Oldenborgh, G. (Sept. 2020). Regional differentiation in climate change induced drought trends in the Netherlands. en. *Environmental Research Letters*, 15 (9), p. 094081. DOI:10.1088/1748-9326/AB97CA.

QGIS Development Team (2022). *QGIS Geographic Information System*. QGIS Association.

Rakovec, O., Samaniego, L., Hari, V., Markonis, Y., Moravec, V., Thober, S., Hanel, M. & Kumar, R. (2022). The 2018-2020 Multi-Year Drought Sets a New Benchmark in Europe. en. *Earth's Future*, 10 (3). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021EF002394, e2021EF002394. DOI:10.1029/2021EF002394.

Rijkswaterstaat (Oct. 2011). *Factsheet Hoofdnetwerk, versie 1.0*. nl. Tech. rep.

Rijkswaterstaat & van Waterschappen, U. (Apr. 2019). *Watermanagement in Nederland*. nl. Tech. rep. the Netherlands.

Ruiter, M.C., Couasnon, A., Homberg, M.J.C., Daniell, J.E., Gill, J.C. & Ward, P.J. (Mar. 2020). Why We Can No Longer Ignore Consecutive Disasters. en. *Earth's Future*, 8 (3). DOI:10.1029/2019EF001425.

Rummukainen, M. (2010). State-of-the-art with regional climate models. en. *WIREs Climate Change*, 1 (1). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcc.8, pp. 82–96. DOI:10.1002/WCC.8.

Rösner, B., Benedict, I., Heerwaarden, C.C.v., Weerts, A.H., Hazeleger, W., Bissolli, P. & Trachte, K. (Sept. 2019). Sidebar 7.3: The long heat wave and drought in Europe in 2018. English. In: *State of the Climate in 2018*. American Meteorological Society, S222–S237.

Samaniego, L., Thober, S., Kumar, R., Wanders, N., Rakovec, O., Pan, M., Zink, M., Sheffield, J., Wood, E.F. & Marx, A. (May 2018). Anthropogenic warming exacerbates European soil moisture droughts. en. *Nature Climate Change*, 8 (5). Number: 5 Publisher: Nature Publishing Group, pp. 421–426. DOI:10.1038/s41558-018-0138-5.

Samaniego, L., Thober, S., Wanders, N., Pan, M., Rakovec, O., Sheffield, J., Wood, E.F., Prudhomme, C., Rees, G., Houghton-Carr, H., Fry, M., Smith, K., Watts, G., Hisdal, H., Estrela, T., Buontempo, C., Marx, A. & Kumar, R. (Dec. 2019). Hydrological Forecasts and Projections for Improved Decision-Making in the Water Sector in Europe. EN. *Bulletin of the American Meteorological Society*, 100 (12). Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, pp. 2451–2472. DOI:10.1175/BAMS-D-17-0274.1.

Schuldt, B. et al. (June 2020). A first assessment of the impact of the extreme 2018 summer drought on Central European forests. *Basic and Applied Ecology*, 45, pp. 86–103. DOI:10.1016/J.BAAE.2020.04.003.

Shen, C. (Nov. 2018). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. en. *Water Resources Research*, 54 (11), pp. 8558–8593. DOI:10.1029/2018WR022643.

Shen, C., Chen, X. & Laloy, E. (2021). Editorial: Broadening the Use of Machine Learning in Hydrology. *Frontiers in Water*, 3.

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., Ganguly, S., Hsu, K.-L., Kifer, D., Fang, Z., Fang, K., Li, D., Li, X. & Tsai, W.-P. (Nov. 2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. en. *Hydrology and Earth System Sciences*, 22 (11), pp. 5639–5656. DOI:10.5194/HESS-22-5639-2018.

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

Slater, L.J., Arnal, L., Boucher, M.-A., Chang, A.Y.-Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., Villarini, G., Wilby, R.L., Wood, A. & Zappa, M. (May 2023). Hybrid forecasting: blending climate predictions with AI models. English. *Hydrology and Earth System Sciences*, 27 (9). Publisher: Copernicus GmbH, pp. 1865–1889. DOI:10.5194/HESS-27-1865-2023.

Stahl, K., Kohn, I., Blauhut, V., Urquijo, J., De Stefano, L., Acácio, V., Dias, S., Stagge, J.H., Tallaksen, L.M., Kampragou, E., Van Loon, A.F., Barker, L.J., Melsen, L.A., Bifulco, C., Musolino, D., de Carli, A., Massarutto, A., Assimacopoulos, D. & Van Lanen, H.A.J. (Mar. 2016). Impacts of European drought events: insights from an international database of text-based reports. en. *Natural Hazards and Earth System Sciences*, 16 (3), pp. 801–819. DOI:10.5194/NHESS-16-801-2016.

Stahl, K., Weiler, M., van Tiel, M., Kohn, I., Hänsler, A., Freudiger, D., Seibert, J., Gerlinger, K. & Moretti, G. (2022). *Auswirkungen des Klimawandels auf die Abflussanteile aus Regen, Schnee und Gletscherschmelze im Rhein und seinen Zuflüssen*. de. Synthesebericht KHR Bericht Nr. I-28. Lelystad: Internationale Kommission für die Hydrologie des Rheingebietes (CHR/KHR), p. 30.

Sun, R., Pan, B. & Duan, Q. (Sept. 2023). A surrogate modeling method for distributed land surface hydrological models based on deep learning. *Journal of Hydrology*, 624, p. 129944. DOI:10.1016/J.JHYDROL.2023.129944.

Sun, Z., Long, D., Yang, W., Li, X. & Pan, Y. (Apr. 2020). Reconstruction of GRACE Data on Changes in Total Water Storage Over the Global Land Surface and 60 Basins. en. *Water Resources Research*, 56 (4). DOI:10.1029/2019WR026250.

Sutanto, S.J., van der Weert, M., Wanders, N., Blauhut, V. & Van Lanen, H.A.J. (Dec. 2019). Moving from drought hazard to impact forecasts. en. *Nature Communications*, 10 (1), p. 4945. DOI:10.1038/S41467-019-12840-Z.

Sutanudjaja, E.H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J.H.C., Drost, N., van der Ent, R.J., de Graaf, I.E.M., Hoch, J.M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M.W., Vannametee, E., Wisser, D. & Bierkens, M.F.P. (June 2018). PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. English. *Geoscientific Model Development*, 11 (6). Publisher: Copernicus GmbH, pp. 2429–2453. DOI:10.5194/GMD-11-2429-2018.

Tallaksen, L.M. & Lanen, H.A.J.v. (2004). *Hydrological Drought: Processes and Estimation Methods for Streamflow and Groundwater*. en. Google-Books-ID: eXLDwgxG0ikC. Elsevier.

Tartakovsky, A.M., Marrero, C.O., Perdikaris, P., Tartakovsky, G.D. & Barajas-Solano, D. (2020). Physics-Informed Deep Neural Networks for Learning Parameters and Constitutive Relationships in Subsurface Flow Problems. en. *Water Resources Research*, 56 (5). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019WR026731, e2019WR026731. DOI:10.1029/2019WR026731.

Thielen, J., Bartholmes, J., Ramos, M.-H. & de Roo, A. (Feb. 2009). The European Flood Alert System ‐Part 1: Concept and development. en. *Hydrology and Earth System Sciences*, 13 (2), pp. 125–140. DOI:10.5194/HESS-13-125-2009.

Thober, S., Kumar, R., Wanders, N., Marx, A., Pan, M., Rakovec, O., Samaniego, L., Sheffield, J., Wood, E.F. & Zink, M. (Jan. 2018). Multi-model ensemble projections of European river floods and high flows at 1.5, 2, and 3 degrees global warming. en. *Environmental Research Letters*, 13 (1). Publisher: IOP Publishing, p. 014003. DOI:10.1088/1748-9326/AA9E35.

Tibshirani, R. (Jan. 1996). Regression Shrinkage and Selection Via the Lasso. en. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58 (1), pp. 267–288. DOI:10.1111/J.2517-6161.1996.TB02080.X.

Tripathy, K.P. & Mishra, A.K. (2023). How Unusual Is the 2022 European Compound Drought and Heatwave Event? en. *Geophysical Research Letters*, 50 (15). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2023GL105453, e2023GL105453. DOI:10.1029/2023GL105453.

Tsai, W.-P., Fang, K., Ji, X., Lawson, K. & Shen, C. (2020). Revealing Causal Controls of Storage-Streamflow Relationships With a Data-Centric Bayesian Framework Combining Machine Learning and Process-Based Modeling. *Frontiers in Water*, 2.

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J. & Shen, C. (Dec. 2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. en. *Nature Communications*, 12 (1), p. 5988. DOI:10.1038/S41467-021-26107-Z.

Tyralis, H., Papacharalampous, G. & Langousis, A. (Apr. 2021). Super ensemble learning for daily streamflow forecasting: large-scale demonstration and comparison with multiple machine learning algorithms. en. *Neural Computing and Applications*, 33 (8), pp. 3053–3068. DOI:10.1007/S00521-020-05172-3.

Van Tiel, M., Weiler, M., Freudiger, D., Moretti, G., Kohn, I., Gerlinger, K. & Stahl, K. (2023). Melting Alpine Water Towers Aggravate Downstream Low Flows: A Stress-Test Storyline Approach. en. *Earth's Future*, 11 (3). _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022EF003408, e2022EF003408. DOI:10.1029/2022EF003408.

Van Vuuren, D.P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G.C., Kram, T., Krey, V., Lamarque, J.-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S.J. & Rose, S.K. (Aug. 2011). The representative concentration pathways: an overview. en. *Climatic Change*, 109 (1), p. 5. DOI:10.1007/S10584-011-0148-Z.

Van der Wiel, K., Stoop, L.P., van Zuijlen, B.R.H., Blackport, R., van den Broek, M.A. & Selten, F.M. (Sept. 2019a). Meteorological conditions leading to extreme low variable renewable energy production and extreme high

energy shortfall. en. *Renewable and Sustainable Energy Reviews*, 111, pp. 261–275. DOI:10.1016/J.RSER.2019.04.065.

Van der Wiel, K., Wanders, N., Selten, F.M. & Bierkens, M.F.P. (Feb. 2019b). Added Value of Large Ensemble Simulations for Assessing Extreme River Discharge in a 2 °C Warmer World. en. *Geophysical Research Letters*, 46 (4), pp. 2093–2102. DOI:10.1029/2019GL081967.

Van der Wiel, K., Batelaan, T.J. & Wanders, N. (Mar. 2023). Large increases of multi-year droughts in north-western Europe in a warmer climate. en. *Climate Dynamics*, 60 (5), pp. 1781–1800. DOI:10.1007/S00382-022-06373-3.

Van der Wiel, K. & Bintanja, R. (Jan. 2021). Contribution of climatic changes in mean and variability to monthly temperature and precipitation extremes. en. *Communications Earth & Environment*, 2 (1). Number: 1 Publisher: Nature Publishing Group, pp. 1–11. DOI:10.1038/S43247-020-00077-4.

Van der Wiel, K., Lenderink, G. & de Vries, H. (Sept. 2021). Physical storylines of future European drought events like 2018 based on ensemble climate modelling. en. *Weather and Climate Extremes*, 33, p. 100350. DOI:10.1016/J.WACE.2021.100350.

Van der Wiel, K., Selten, F.M., Bintanja, R., Blackport, R. & Screen, J.A. (Mar. 2020). Ensemble climate-impact modelling: extreme impacts from moderate meteorological conditions. en. *Environmental Research Letters*, 15 (3). Publisher: IOP Publishing, p. 034050. DOI:10.1088/1748-9326/AB7668.

Van Der Knijff, J.M., Younis, J. & De Roo, A.P.J. (Feb. 2010). LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, 24 (2). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13658810802549154, pp. 189–212. DOI:10.1080/13658810802549154.

Van Hateren, T.C., Sutanto, S.J. & Van Lanen, H.A. (Dec. 2019). Evaluating skill and robustness of seasonal meteorological and hydrological drought forecasts at the catchment scale - Case Catalonia (Spain). en. *Environment International*, 133, p. 105206. DOI:10.1016/J.ENVINT.2019.105206.

Van Lanen, H.A. et al. (Aug. 2016). Hydrology needed to manage droughts: the 2015 European case. en. *Hydrological Processes*, 30 (17), pp. 3097–3104. DOI:10.1002/HYP.10838.

Van Loon, A.F. & Laaha, G. (July 2015). Hydrological drought severity explained by climate and catchment characteristics. en. *Journal of Hydrology*. Drought processes, modeling, and mitigation, 526. XXXXX, pp. 3–14. DOI:10.1016/J.JHYDROL.2014.10.059.

Van Loon, A.F., Stahl, K., Di Baldassarre, G., Clark, J., Rangecroft, S., Wanders, N., Gleeson, T., Van Dijk, A.I.J.M., Tallaksen, L.M., Hannaford, J., Uijlenhoet, R., Teuling, A.J., Hannah, D.M., Sheffield, J., Svoboda, M., Verbeiren, B., Wagener, T. & Van Lanen, H.A.J. (Sept. 2016). Drought in a human-modified world: reframing drought definitions,understanding, and analysis approaches. en. *Hydrology and Earth System Sciences*, 20 (9), pp. 3631–3650. DOI:10.5194/HESS-20-3631-2016.

Vicente-Serrano, S.M., Beguería, S. & López-Moreno, J.I. (Apr. 2010). A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index. en. *Journal of Climate*, 23 (7), pp. 1696–1718. DOI:10.1175/2009JCLI2909.1.

Vörösmarty, C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R. & Davies, P.M. (Sept. 2010). Global threats to human water security and river biodiversity. en. *Nature*, 467 (7315). Number: 7315 Publisher: Nature Publishing Group, pp. 555–561. DOI:10.1038/NATURE09440.

Wanders, N., Karssenberg, D., de Roo, A., de Jong, S.M. & Bierkens, M.F.P. (June 2014). The suitability of remotely sensed soil moisture for improving operational flood forecasting. en. *Hydrology and Earth System Sciences*, 18 (6), pp. 2343–2357. DOI:10.5194/HESS-18-2343-2014.

Wanders, N. & Wada, Y. (July 2015a). Human and climate impacts on the 21st century hydrological drought. en. *Journal of Hydrology*, 526, pp. 208–220. DOI:10.1016/J.JHYDROL.2014.10.047.

Wanders, N., Thober, S., Kumar, R., Pan, M., Sheffield, J., Samaniego, L. & Wood, E.F. (Jan. 2019). Development and Evaluation of a Pan-European Multimodel Seasonal Hydrological Forecasting System. en. *Journal of Hydrometeorology*, 20 (1), pp. 99–115. DOI:10.1175/JHM-D-18-0040.1.

Wanders, N. & Wada, Y. (2015b). Decadal predictability of river discharge with climate oscillations over the 20th and early 21st century. en. *Geophysical Research Letters*, 42 (24). _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015GL066929, pp. 10,689–10,695. DOI:10.1002/2015GL066929.

Wanders, N. & Wood, E.F. (Sept. 2016). Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. en. *Environmental Research Letters*, 11 (9), p. 094007. DOI:10.1088/1748-9326/11/9/094007.

Wendt, D.E., Bloomfield, J.P., Van Loon, A.F., Garcia, M., Heudorfer, B., Larsen, J. & Hannah, D.M. (Oct. 2021). Evaluating integrated water management strategies to inform hydrological drought mitigation. en. *Natural Hazards and Earth System Sciences*, 21 (10), pp. 3113–3139. DOI:10.5194/NHESS-21-3113-2021.

Wilhite, D. (Jan. 2000). Chapter 1 Drought as a Natural Hazard: Concepts and Definitions. In: *Drought: A Global Assessment*. Vol. 1. London: Routledge, pp. 3–18.

Wilhite, D.A., Svoboda, M.D. & Hayes, M.J. (May 2007). Understanding the complex impacts of drought: A key to enhancing drought mitigation and preparedness. en. *Water Resources Management*, 21 (5), pp. 763–774. DOI:10.1007/s11269-006-9076-5.

Witte, J.-P., de Louw, P., van Ek, R. & Bartholomeus, R. (2020). AANPAK DROOGTE VRAAGT TRANSITIE WATERBEHEER. nl, p. 12.

Xu, T. & Liang, F. (2021). Machine learning for hydrologic sciences: An introductory overview. en. *WIREs Water*, 8 (5). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wat2.1533, e1533. DOI:10.1002/WAT2.1533.

Zhang, D., Lin, J., Peng, Q., Wang, D., Yang, T., Sorooshian, S., Liu, X. & Zhuang, J. (Oct. 2018). Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. en. *Journal of Hydrology*, 565, pp. 720–736. DOI:10.1016/J.JHYDROL.2018.08.050.

Zhu, Y., Zabaras, N., Koutsourelakis, P.-S. & Perdikaris, P. (Oct. 2019). Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *Journal of Computational Physics*, 394, pp. 56–81. DOI:10.1016/J.JCP.2019.05.024.

List of Publications

Journal articles as first author

Hauswirth, S. M., Bierkens, M. F., Beijk, V., and Wanders, N. (2021). The potential of data driven approaches for quantifying hydrological extremes. *Advances in Water Resources, 155, 104-017*, https://doi.org/10.1016/j.advwatres.2021.104017

Hauswirth, S. M., Bierkens, M. F., Beijk, V., and Wanders, N. (2022). The suitability of a seasonal ensemble hybrid framework including data-driven approaches for hydrological forecasting, *Hydrology and Earth System Sciences, pp. 501‑517*, https://doi.org/10.5194/hess-27-501-2023

Hauswirth, S. M., van der Wiel, K., Bierkens, M. F., Beijk, V., and Wanders, N. (2023). Simulating hydrological extremes for different warming levels‑combining large scale climate ensembles with local observation based machine learning models, *Frontiers in Water, Volume 5, 2023*, https://doi.org/10.3389/frwa.2023.1108108

Journal articles as co-author

van Jaarsveld, B., **Hauswirth, S.**, and Wanders, N.: Machine learning and Global Vegetation: Random Forests for Downscaling and Gapfilling, *Hydrol. Earth Syst. Sci. Discuss*. [preprint], https://doi.org/10.5194/hess-2022-430, in review, 2023.

Conference presentations, invited talks and posters

S.M. Hauswirth, Marc F. P. Bierkens, Vincent Beijk and Niko Wanders, *Exploring and Optimizing Water Management Strategies for Mitigating Local Drought Impacts in the Netherlands using a Multi-Target LSTM Model*, Poster, AGU23, San Francisco, USA, December 15, 2023

S.M. Hauswirth, *Simulating hydrological extremes using machine learning*, Presentation (invited), Workshop: Challenges in Defining and Modelling Hydrological Extremes in Regulated Catchments, Davos, Switzerland, October 24, 2023

S.M. Hauswirth, K. van der Wiel, Marc F. P. Bierkens, Vincent Beijk and Niko Wanders, *Simulating hydrological extremes for different warming levels‑combining large scale climate ensembles with local observation based machine learning models*, Presentation, AGU22, Chicago, USA, December 14, 2022

S.M. Hauswirth , Marc F. P. Bierkens, Vincent Beijk and Niko Wanders, *The suitability of a seasonal ensemble hybrid framework including data driven approaches for hydrological forecasting*, Presentation, EGU22, Vienna, Austria, May 26, 2022

S.M. Hauswirth , Marc F. P. Bierkens, Vincent Beijk and Niko Wanders, *The suitability of a seasonal ensemble hybrid framework including data driven approaches for hydrological forecasting*, Presentation, HydroML Symposium, Penn State University, State College, USA, May 18, 2022

S.M. Hauswirth , Marc F. P. Bierkens, Vincent Beijk and Niko Wanders, *The potential of data driven approaches for quantifying hydrological extremes*, vPICO presentation, vEGU2021, virtual, April 27, 2021

S.M. Hauswirth, Marc F. P. Bierkens, Vincent Beijk and Niko Wanders, *The potential of data driven approaches for quantifying hydrological extremes*, Boussinesq Lecture at Nederlands Aardwetenschappelijk Congres (NAC), virtual, April 8, 2021

Authorship contribution statement

This thesis was financially supported by the Cooperate Innovation Program and the Department of Water, Transport and Environment at the Dutch National Water Authority, Rijkswaterstaat. The general research direction and sub-projects were based on the project proposal, which was further formulated into a research plan for the period of the PhD project. Introduction and Synthesis (Chapter 1 and 6) were written by the PhD Candidate with minor suggestions from the promotors. Chapter 3 is based on the most recent work which is currently in preparation for submission to a peer-reviewed journal. Chapter 2, 4 and 5 are based on peer-reviewed articles in collaboration with the authors listed below:

SH: Sandra M. Hauswirth
NW: Niko Wanders
MB: Marc F.P. Bierkens
VB: Vincent Beijk
KvW: Karin van der Wiel

Chapter 2

Conceptualisation:	SH, NW, MB, VB
Data collection:	SH
Data analysis and interpretation:	SH in consultation with NW, MB, VB
Writing:	SH
Revision and approval:	SH, NW, MB, VB

Chapter 3

Conceptualisation:	SH, NW, MB
Data collection:	SH
Data analysis and interpretation:	SH in consultation with NW, MB, VB
Writing: SH	
Revision and approval:	SH, NW, MB,

Chapter 4

Conceptualisation:	NW, MB, VB
Data collection:	SH
Data analysis and interpretation:	SH in consultation with NW, MB, VB
Writing: SH	
Revision and approval:	SH, NW, MB, VB

Chapter 5

Conceptualisation:	SH, NW, MB, KvW, VB
Data collection:	Climate data acquisition was done by KvW. Hydrological data acquisition was done by SH
Data analysis and interpretation:	SH in consultation with NW, MB, KvW
Writing:	SH
Revision and approval:	SH, NW, MB, KvW, VB

Acknowledgements

Even prior to deciding on whether I wanted to do a PhD or not I reached out to a few people who were doing one at the time or had recently finished one. Just to see if I could get some nuggets of wisdom from them — and if the whole mystery of a PhD could be something for me. The answers were surprisingly (and annoyingly) vague but the reoccurring themes were:

a) it is going to be a journey — your *own* journey
b) know the *why* you want it, when times get rough this will be the reason to keep going
c) but most importantly: *enjoy* it while it lasts, it will be over way too soon

While I certainly cringed about some of these statements initially, looking back now myself, I have to admit, they were right. A journey it was, with many ups and downs, often feeling like the roller-coaster you wanted to do as a child — but then you realised you might have *slightly* underestimated it. Luckily, in this case, I am only left with a bittersweet feeling and the fact that I know I would not want to have missed any part of it. Especially not as this journey included so many lovely people, some rooting for me from the sideline and others from far away, and while words will never be able to fully describe how grateful I am to all of them, I will still try but also keep it short:

Niko and Marc, thank you for giving me this opportunity to embark on my *own* PhD journey. The appreciation for this is actually already going back to 2018, when you gave me the opportunity to do my MSc thesis in this setting. No doubt one of the key elements for my later journey in science.

Niko — thank you for everything (one of the few instances where words will never do it justice). Thank you for mentoring me all of these years, not just in my academic career and all aspects that come with it, but also outside as a friend. Thank you for showing me since day one that research can be not only interesting and exciting but most importantly fun, especially as a team.

Marc — your never ending energy and enthusiasm for science were showing in so many of our meetings and conferences and were a constant source of inspiration, thank you for sharing this enthusiasm with me since 2018.

Speaking of team — while I started this journey rather lonely and the COVID period did not help, seeing the team around me grow throughout the years made me realize how much I missed this part of the journey initially. I am beyond grateful to have been surrounded by such a bunch of great people, so thanks to all of you for the laughs and talks, coffees, teas, drinks and dinners, for being the best travel and conference companions and sticking around even through the more "grey" days of my PhD.

My paranymphs Ed and Emmy — the two people who have been there (and in the offices we shared together) since the beginning of the PhD journey and agreed to stick with me until the end of it. Thank you both for being there, navigating and exchanging the joys and frustrations (and everything in between) of a PhD, keeping up my spirits with endless meme conversations, walks, teas and drinks and the occasional sharing of support animals, the fluffiness of them always did wonders.

I also want to take a moment to thank all the other PhDs and colleagues, with whom I had shared many lovely moments, through more coffees, drinks and dinners that enriched my PhD journey also outside of work.

So far it sounds like my PhD only consisted of lots of coffees, teas and drinks, but I can guarantee there was work done as well, which can be found in the pages before in case you skipped that part...

Speaking of work: a special thank you also to my collaborators. Vincent — while the original plan to spend part of my time at Rijkswaterstaat got thrown out of the window due to COVID, the many meetings and opportunities to get a glimpse of your own work showed me how fulfilling it is to have a

project that is so closely linked to real life application. Karin — even though our project felt like it was flying by, I was able to take away a lot from it, not just scientific but also from the experiences you shared with me along the way.

Finally, my deepest thanks go to the people who have been with me the longest, and whom give me the feeling of being home, no matter where I am:

My parents, Margrit and Robert — thank you for your unconditional love and sacrifice, for opening up so many doors in my life and giving me the building blocks to form my own future, and for cheering me on every step along the way. Without your guidance, I would not have become the person I am today.

Adrian — thank you for being the best brother I could wish for, for being the person I can always look up to and ask for advice, and for making me believe that I can do much more than I would think myself.

My friends back home — thank you for sticking with me since all these years, cheering me on from afar and still putting up with me (even though my six month abroad turned into many years) and for making me feel like I never left whenever we meet up again.

And Maarten — we both know that this could potentially fill the whole page, so please remember what we once joked about what I could write for you and I will keep it short here. Thank you for being my rock, for your never ending support, love, and confidence in me. Thanks for being part of this adventure and to many more of them.

Sandra Margrit Hauswirth
February 2024

About the author

Sandra Margrit Hauswirth was born in Altdorf, Switzerland on the 28th of January 1994. After several moves during her childhood, she grew up in the small city of Spiez at the Thunersee.

The following years were often led by her internal motto of being open to trying more challenging journeys, with the option of always being able to turn back (but in the end never doing so).

After finishing school, she continued her education at the technical Gymnasium Thun Schadau (Thun, Switzerland), where she finished her bilingual 'Matura' with a major in Biology and Chemistry in 2012.

Following her interest in the earth, its problems and potential solutions she decided to start her student years at ETH Zurich (Zurich, Switzerland) in 2013, which turned out to be some of the most transformative years in her early life.

Throughout her time there she obtained a Bachelor's degree in Environmental Sciences (2016), with a major in Biogeochemistry, and a Master's degree in Environmental Engineering (2018), with a major in Resource Management. While spending most of these five years in Zurich, her MSc thesis was done abroad at Utrecht University (Utrecht, the Netherlands) and focused on the relative importance of climate and socio-economic changes on future global groundwater resources.

After her studies she continued working as a research assistant at Utrecht University before starting as a Junior Researcher/Consultant in the Groundwater Management Department at Deltares (Utrecht, the Netherlands).

Following her interest in water resources and working at the interface between research and application, she started her joint PhD project with Utrecht University and Rijkswaterstaat in 2020, the work that culminated in this thesis.

She currently works as a PostDoc on drought related topics at Utrecht University and while her initial idea of a *short* stay abroad back in 2018 has in theory "failed", she is looking forward to the new challenges as a PostDoc in a setting that she now calls a second home.

Utrecht University
Faculty of Geosciences
Department of Physical Geography

ISSN 2211-4335