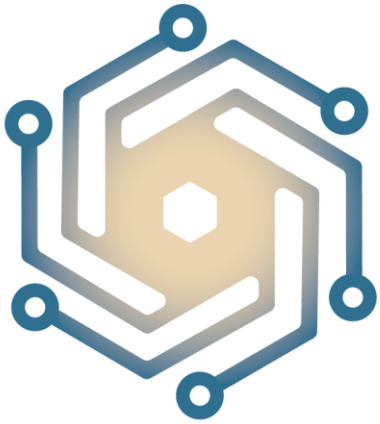

CLINT

CLIMATE INTELLIGENCE

D7.2
Preliminary AI-enhanced Climate Services for local decision-making

April, 2024

This project is part of the H2020 Programme supported by the European Union, having received funding from it under Grant Agreement No 101003876

Programme Call:	Building a low-carbon, climate resilient future: climate action in support of the Paris Agreement (H2020-LC-CLA-2018-2019-2020)
Grant agreement ID:	101003876
Project Title:	CLINT
Partners:	POLIMI (Project Coordinator), CMCC, HEREON, CSIC, SMHI, HKV, E3M, TCDF, DKRZ, IHE, ECMWF, UAH, JLU, OGC, UCM
Work-Package:	WP7
Deliverable #:	D7.2
Deliverable Type:	Document
Contractual Date of Delivery:	30 April 2024
Actual Date of Delivery:	30 April 2024
Title of Document:	D7.2 Preliminary AI-enhanced Climate Services for local decision-making
Responsible partner:	IHE
Author(s):	Andrea Ficchì, Celia Ramos Sanchez, Claudia Bertini, Matteo Giuliani, Paulina Kindermann, Michiel Pezij, Lucia De Stefano, Micha Werner, Andrea Castelletti, Schalk Jan van Andel
Content of this report:	This report describes the progress to-date of developing AI-enhanced climate services for the WP7 local scale case studies and presents the preliminary results for extreme event forecast performance and potential added value assessment for use case management practice.
Availability:	This report is public.

Document revisions		
Author	*Revision content*	*Date*
Schalk Jan van Andel	D7.2_draft structure	14 Dec 2023
Andrea Ficchì, Celia Ramos Sanchez, Claudia Bertini, Matteo Giuliani, Paulina Kindermann, Michiel Pezij, Lucia De Stefano, Micha Werner, Andrea Castelletti, Schalk Jan van Andel	D7.2_draft_v01	21 March 2024
Andrea Ficchì, Celia Ramos Sanchez, Claudia Bertini, Matteo Giuliani, Paulina Kindermann, Michiel Pezij, Lucia De Stefano, Micha Werner, Andrea Castelletti, Schalk Jan van Andel	D7.2_draft_v02, for internal revision	1 April 2024
Ronan McAdam, Elena Xoplaki	Internal revisions	25 April 2024
Andrea Ficchì, Celia Ramos Sanchez, Claudia Bertini, Matteo Giuliani, Paulina Kindermann, Michiel Pezij, Lucia De Stefano, Micha Werner, Andrea Castelletti, Schalk Jan van Andel	D7.2_draft_v03 after receiving internal reviewers' comments, for coordinator review	26 April 2024
Guido Ascenso, Andrea Castelletti	Final revisions	30 April 2024

Contents

LIST OF ACRONYMS

Abbreviations

AI: Artificial Intelligence

CS: Climate Service

DSS: Decision Support System

EAP: Early Action Protocol

EE: Extreme Event

EET: Extratropical Transition

GWL: Global Warming Level

ML: Machine Learning

RBD: River Basin District

S2S: Sub-seasonal to Seasonal

SWIO: Southwest Indian Ocean

TC: Tropical Cyclone

WP: Work Package

WEF: Water-Energy-Food

ZW: Zambezi Watercourse

EXECUTIVE SUMMARY

This report presents preliminary results of AI-enhanced Climate Services (CS) designed for decision-making in the local scale CLINT case studies across various climate change impact hotspots, aimed at managing and mitigating risks from different types of extreme events. For each of the case studies an analysis chain is presented in this deliverable, showing the steps to assess the potential added value of AI-enhanced climate services. These steps lead from comparison of AI-enhanced against benchmark predictions, to the use of an impact model, to the local decision-process as acquired from the user interactions (Deliverable 7.1). The results to-date are presented for each case study.

For the semi-arid Zambezi case study, the results of the AI-enhanced CS developed for droughts show that the simple data-driven forecasting model (NIPA) of seasonal streamflow outperforms global scale hydrological model-based seasonal forecasts. As a next step, further AI-enhanced NIPA inflow forecasts to Kariba dam, Zambezi, will be used to inform the lake water management and assess the added value of our enhanced forecasts with respect to the benchmark and no-forecast baseline. Concerning Tropical Cyclones (TC) and floods in the Zambezi case study, post-processing of benchmark TC rainfall forecasts, based on a variant of a state-of-the-art deep learning architecture (UNet), and a novel loss function was developed and tested. The AI enhancements showed a significant improvement with respect to the benchmark and the value for early warning has been demonstrated for extreme TC rainfall forecasts.

In the second semi-arid case study, focusing on Spain's Douro River basin, bias-correction of ECMWF seasonal precipitation forecasts using the BJP-SS approach showed improved accuracy in autumn, though the improvement was less pronounced in spring. The next steps for this case study will include an assessment of the quality of AI-enhanced reservoir inflow predictions up to 1-month lead time following the method developed in WP2 task 2.4. The added value of both bias corrected precipitation forecasts and AI-enhanced inflow forecasts will be assessed.

From the delta case studies in the Netherlands, for the Rijnland case study the potential added value of the AI-enhanced 1-month lead time precipitation predictions, using Extreme Learning Machines, has been quantified in terms of increased correct alerts. The number of correct drought alerts increased for two of the four thresholds used by the Rijnland water authority, while reducing false alerts, as compared to using ECMWF extended range ensemble mean precipitation forecasts. The benchmark ECMWF extended range forecasts outperform climatology and the AI-enhanced predictions for the lowest and highest precipitation deficit thresholds. For the next deliverable (D7.3), the assessment of potential added value for this case study will be extended to include AI-enhanced alerts for discharge of the Rhine river at Lobith dropping below the drought threshold (Deliverable 2.2).

In second delta case study in the Netherlands, Aa en Maas, a data-driven model has been developed to predict groundwater levels. The seasonal forecast of groundwater levels using this impact model with bias-corrected ECMWF SEAS5 meteorological forcing resulted in a predictive skill up to one to two months. As next steps for this case study the assessment of the potential added value of the

groundwater level predictions will be analysed in terms of hits and false alerts, and in consultation with the water authority end users.

Lastly for the delta case studies, the impact of extratropical transitions (ETT) on flood risk in the Netherlands under various climate change scenarios has been examined. The results imply that the impact of ETTs can be expected to be similar to typical winter storms in the North Sea, with respect to their statistical properties at specific locations along the Dutch coast. However, it is important to note that it is not known whether the analysed TC tracks would actually reach the Dutch coast and have any effect on the wind. ETTs occur most frequently in September, which is somewhat earlier than the regular stormy season along the Dutch coast. If the ETT intensity increases in the future, this could imply that the month September becomes unsuitable for maintenance of storm surge flood protection infrastructure.

For the snow hotspot case study of Lake Como, an AI-enhanced CS employing Reinforcement Learning was developed and tested. This approach aims to improve drought and flood management by effectively extracting the most valuable information from multi-timescale forecasts of lake inflows. The potential added value assessment reveals that using selected forecast information—namely, inflow forecasts with a 3-day aggregation period and lead time—provides the largest added value in terms of reduced water supply deficit for the multipurpose operation of Lake Como. For the development of an AI-enhanced CS for heatwaves and warm nights, the PRIM algorithm was used to discover the relationship between temperature extremes and crop failures to support farmers' adaptation to future climate conditions. The analysis of the projected heatwaves and warm nights shows that temperature extremes are expected to increase considerably over the coming years (all scenarios), suggesting the opportunity for agricultural sector users to cultivate more heat-tolerant crop varieties. The next step is to assess the impact of compound heatwave and drought events and evaluate the potential for responding to them through a dedicated AI-enhanced Climate Service. These findings will be reported in Deliverable 7.3.

The development of AI-enhanced CS and their performance compared to benchmark predictions suggest that the research for the local-scale WP7 case studies is progressing well. For the case studies of Zambezi, Rijnland, and Como, some of the AI-enhanced CS have been already assessed for their added value according to the user-defined impact indicators. We believe this deliverable provides a solid foundation for enhancing the developed CS and for completing the added value assessment compared to benchmark CS. This work leads into the next deliverable on final CS, D7.3, and the concluding benchmark analysis report, D7.4.

1 Introduction

The CLINT project seeks to improve the detection, causation, and attribution of Extreme Events (EE) through machine learning, with the ultimate goal of developing innovative and specialized AI-enhanced Climate Services (CS) to support adaptation, mitigation, and disaster risk management strategies. The role of Work Package 7 (WP7) is to integrate the AI-enhanced CS developed during the project into pilot services customized for specific local-scale case studies, and to assess their potential added value in terms of user-defined impact indicators.

In the first deliverable of WP7, the local-scale case studies, extreme events, and potential use cases of AI-enhanced predictions were described in detail, along with the current predictions used and requirements for improvement as kindly provided by stakeholder representatives. In this second WP7 deliverable, the case studies and EEs are only briefly listed in Section 1.1., and summarised at the beginning of each main case study section in the following chapters.

This Deliverable 7.2 marks the progress in developing AI-enhanced CS for the case studies. It includes the current prediction performance assessment results for the EE addressed, and first analysis of the results of added value assessment.

1.1 Climate change hotspots

The case studies in WP7 were chosen to cover a variety of extreme events, and to focus on regions especially vulnerable to climate change, including semi-arid areas, deltas, and snow-dependent regions—often referred to as climate change hotspots. CLINT WP7 is working on the following case studies:

Semi-arid areas
- Zambezi Watercourse - droughts, tropical cyclones and floods
- Douro basin - droughts

Deltas
- Rijnland - droughts
- Aa en Maas - droughts
- Main water system of the Netherlands - extratropical transition and flood risk

Snow-dependent areas
- Lake Como basin - droughts, floods, heatwaves and warm nights

1.2 Objectives of this deliverable

As CLINT nears the end of its third year, this deliverable aims to review the progress made in developing and testing AI-enhanced climate services for the WP7 case studies.

The specific objectives of this deliverable are to:
- describe the developed AI-enhanced pilot CS.

- describe how the AI-enhanced prediction information is used in impact models to derive CS tailored to each case study.
- Describe in detail for use case study the steps to assess the potential added value of the AI-enhanced CS: starting from comparing AI-enhanced predictions and benchmark predictions, via impact models, to arrive at impact indicators.
- present and analyse results of performance assessment of the AI-enhanced extreme event predictions for each case study, and compare the AI-enhanced performance with existing forecasting systems as a benchmark.
- present and discuss results of the potential added value of the AI-enhanced extreme event predictions in current event management and disaster risk reduction practices as described by the case study users.

1.3 Added value

What constitutes 'added value' can be defined in multiple ways. We start from a broad definition of this term in this section, and then the case study specific definition and method of assessment is given in each of the respective chapters discussing the pilot climate services.

General definitions of 'value' often relate the term to (a) something that has monetary worth or a fair return in money, services, or goods, (b) something useful, estimable, or important, and (c) a set of beliefs and concepts in individuals (McKeown and Summers, 2006). In the forecasting and climate services literature, value is usually related to the benefits for decision-making processes (Bruno Soares et al., 2018). The traditional approach for CS value estimation, also within the field of water resources management, is the comparison of expected or observed results derived from a CS-based decision with the results derived from a decision taken without CS, usually based on climatology. This approach assumes that decisions based on climatology will vary after the uptake of CS information. Whilst this is also the approach adopted in this project, it is worth noting that CS can still create value even if they do not change the course of decisions; for instance, through more qualitative factors such as the increasing user confidence throughout the decision-making process, avoidance conflicts or saving time.

1.4 Connection with other deliverables

This deliverable has been developed in conjunction with Deliverable 2.2 in which the ML methods selected, enhanced, and developed are described and tested. Here, we will briefly mention the main elements of an AI-enhanced CS tested, and refer the reader to D2.2 for the details of the ML algorithms and methods for use in EE prediction. Work package 2 closely works with the climate science WPs 3, 4, and 5, to support EE detection, causation, and attribution. Experience from WPs 3-5, feeds back to refining the AI-enhanced prediction methods of WP2, and where applicable feeds directly into deliverables of WP7 with test datasets of climate extreme predictions in either S2S forecasts or climate change projections. WP6 focuses on case studies at the pan-European scale. Since most of the local-scale case studies in WP7 are located in Europe, some of the pan-European predictions generated by WP6—such as streamflow forecasts—will be also tested for use in WP7.

For WP7, this deliverable follows D7.1, which provided a detailed description of the case study areas, the users involved, and the decision-making processes. This deliverable D7.2 is intended to review the progress made in developing AI-enhanced climate services, as well as the progress in evaluating the potential added value of these services in the local-scale use cases. As such, this deliverable leads up to D7.3 in which the final developed climate services will be presented, and D7.4 in which the final benchmark comparison of all services developed will be presented.

1.5 Structure of the document

In the three following chapters, grouped per climate change hotspot (semi-arid, delta, snow-dependent), each developed pilot CS for WP7 case studies is presented. At the start of each section, the analysis chain to assess the potential added value of each AI-enhanced CS is presented with a flowchart and explained in detail, describing datasets and benchmark predictions, impact models if applicable (e.g. hydrological model, reservoir model, crop-growth model, etc), and impact indicators that will shed light on the added value. Then, the next sub-section for each case study presents the preliminary test results obtained along the analysis chain. The final sub-section of each case study/extreme event use case, discusses the results and outlines the next steps. Chapter 5 concludes with a reflection of the status of the WP7 progress towards developing AI-enhanced CS and assessing their potential added value.

2 Semi-arid climate change hotspots

2.1 Zambezi Watercourse

2.1.1 Introduction

The Zambezi Watercourse (ZW) is the fourth largest basin in Africa, spanning 1.32 million km² across eight countries (Zambia, Zimbabwe, and Mozambique collectively sharing 70% of its area) and populated by about 40 million inhabitants. Water management in the ZW is key in sustaining irrigated agriculture and hydropower production, to ensure food and energy security in the region (Arnold et al., 2023; Spalding-Fecher et al., 2017). Several dams have been built since the 1970s with the main purpose of hydropower production which altered the ZW's natural flow, also impacting wetland ecosystems, especially in Mozambique's river delta. The total installed capacity of hydropower generation is about 5 GW (Stevanato et al., 2021), primarily concentrated in two major dams, Kariba and Cahora Bassa, located along the main ZW, between Zambia, Zimbabwe (Kariba) and Mozambique (Cahora Bassa). An increased frequency and severity of extreme events is expected in the ZW, as indicated by recent observations and climate projections for the region (IPCC, 2023). In particular, in CLINT we focus on the increasing risk from droughts and Tropical Cyclones (TCs), associated with heavy precipitation and flood risk, to enhance CS and better support adaptation actions, particularly for the sectors of water management (droughts) and early warning (TCs).

The main users and stakeholders of climate services identified for droughts, TCs, and floods in the ZW include: the Zambezi Watercourse Commission (ZAMCOM), dam operators and hydropower

companies (such as the Zambezi River Authority - ZRA and the Zambia Electricity Supply Corporation Limited - ZESCO), National Meteorological and Hydrological Services (NMHSs) of riparian countries, the Southern Africa Development Community (SADC), National Environmental Regulatory Agencies (like Zambia Environmental Management Agency - ZEMA and Zimbabwe's Environmental Management Agency - EMA), disaster management agencies (including the National Institute for Disaster Risk Management and Reduction of Mozambique - INGD), humanitarian agencies (such as the Mozambique Red Cross - Cruz Vermelha de Moçambique, CVM, and the World Food Programme - WFP), irrigation schemes' operators, national farmers unions, and water supply companies. As part of CLINT Deliverable D7.1 (Local CS), the outcomes of a survey and semi-structured interviews with some of these organisations, we highlighted the main challenges faced by the main CS users in the region. In particular, the information and answers provided by ZAMCOM and a National Environmental Agency (from one of the riparian countries) contributed to defining the CS needs for drought adaptation, while stakeholders from humanitarian agencies (Red Cross and WFP) guided us in defining the CS needs for TC and flood early warning. The main outcomes are summarised here below, focusing first on droughts (Section 2.1.1.1) and then on TCs and floods (Section 2.1.1.2).

Droughts

User definition of extreme event

ZAMCOM defines droughts as natural hazards caused by a lack of precipitation and increased evaporation, exacerbated by factors such as land use changes, increased pressure on resources, climate variability and change. Extreme droughts are of concern for ZAMCOM given the high negative socio-economic impacts on the hydropower and agriculture sectors. Another environmental regulatory agency provided a definition in line with ZAMCOM, emphasising the adverse impacts on the environment, including ecosystem degradation, land use change and disease outbreaks. The key variables used to define droughts are: rainfall deficit (mm), affecting agriculture, low water levels (m) in lakes and rivers, as well as streamflow (m3/s), impacting irrigation and hydropower. In the preliminary AI enhancements for the CLINT ZW drought use case (this deliverable), we focus on streamflow, which aim to enhance forecasts of inflows upstream of strategic regulated lakes (e.g., Kariba) to support dam operations and assess the value for hydropower and agriculture.

Decision process for preparedness, adaptation, and event or risk management for droughts

Using real-time monitoring systems and seasonal forecasting systems, ZAMCOM offers information to dam operators and irrigation scheme managers, aiding in local decision-making to optimise hydropower production and crop yields. Various data sources are integrated for monitoring, including in-situ hydro-meteorological observations from NMHS and SADC, remote sensing data and model-based (reanalysis) data in real-time or near-real-time ZAMCOM employs the Zambezi Water Resources Information System (ZAMWIS) for drought forecasting, which is continuously enhanced to meet evolving needs. ZAMWIS provides an operational web-based and software interface enabling users to access historical and real-time hydrological data and forecasts. Stakeholders, including government agencies of riparian countries, dam operators, hydropower companies, and irrigation schemes' operators utilise ZAMWIS to plan and manage the water resources in the basin, with a focus on droughts anticipation.

The ZAMWIS Flow Forecasting System (FFS) component is based on the MIKE river modelling software and generates river flow forecasts on the sub-seasonal to seasonal time scale (up to three months lead time) in parts of the basin. This forecasting system covers areas upstream of major dams like Kariba. Forecasting relies on combining meteorological precipitation forecasts (from NCEP's CFS) with rainfall gridded satellite observations, near-real-time river flow and reservoir water level data, and expected reservoir releases and operational guidelines. ZAMWIS employs streamflow thresholds at strategic river points, primarily distributed in the main watercourse upstream of the Kariba dam, to predict droughts effectively. The operator of Kariba dam uses forecasts issued by ZAMCOM via ZAMWIS to inform decisions regarding dam releases, enhancing water management strategies in the region.

User wishes and requirements for enhanced climate services for droughts

The main wish of ZAMCOM regarding the CS for droughts is to improve the skill and reliability of seasonal forecasts, thereby providing users with actionable information, with enhanced confidence for dam releases and water management. The objective of CLINT for the drought-related climate service in the ZW case study is to develop AI-enhanced seasonal (3-month) hydrological forecasts for extreme droughts with higher accuracy. These forecasts aim to guide the multipurpose operation of dams and improve system performance, particularly in hydropower production and irrigation supply.

Impact indicators for quantifying the value of AI-enhanced CS for droughts

The impact indicators for measuring the value of AI-enhanced climate services for droughts, as derived from a survey of ZW end-users (D7.1), fall into two categories:

- Hydropower production (to be maximised) or hydropower production deficit with respect to a target production (to be minimised); their formulation stems from the requirements and objectives of the dam operators and hydropower companies.
- Crop yield (to be maximised) or irrigation deficit with respect to the water demand (to be minimised); their formulation follows the requirements and objectives of irrigation schemes' operators and agricultural stakeholders in the basin.

In the preliminary AI enhancements for the CLINT ZW drought use case of this deliverable, we focus on assessing the improvements of seasonal hydrological forecast skill. In the next steps of the project, the enhanced seasonal drought predictions will be employed to improve dam management operations and optimise hydropower and agricultural (food) production, assessing and reporting their potential added value in Deliverable D7.3 (AI-enhanced Climate Services for local decision-making).

Tropical cyclones (TCs), floods

User definition of extreme event

TCs and associated flooding in the ZW basin are a frequent natural hazard with devastating impacts including loss of lives, injuries, significant damages to properties and infrastructures, displacement of people, disease outbreaks, agricultural losses (crops and livestock) and disruption to other economic activities. The key variables identified to define and characterise TCs and floods are TC wind speeds (km/hour), heavy rainfall (mm) and river levels (m). In the preliminary AI enhancements for the CLINT ZW TC-induced floods use case of this deliverable, we focus on heavy rainfall, and enhancement of medium-range forecasts of extreme TC total precipitation and thereby to support

early warning and early action for TCs that make landfall in Mozambique and impact the Zambezi region.

Decision process for preparedness, adaptation, and event or risk management for TCs

For TC forecasting and warnings, the National Meteorological Agency of Mozambique (INAM) relies on various forecast products. Primarily, TC forecasts are sourced from Météo-France La Réunion (MF), serving as the Regional Specialised Meteorological Centre (RSMC) with the mandate to monitor TCs in the SWIO region and to issue forecasts to national hydrometeorological services. Before TCs, MF issues daily updates on the meteorological situation and cyclogenesis potential. During a TC, MF provides technical bulletins and graphical warnings every 6 hours. The technical bulletins furnish detailed information on TC characteristics (location, size, intensity), aiming to support operational forecasters at NMHS. Graphical warnings illustrate the expected evolution of the TC track in maps, depicting the predicted track over a 5-day lead time, accompanied by a cone of uncertainty (indicating the potential track area) based on multi-model forecasts, as well as an indication of the expected TC intensity.

While the forecasts provided by the RMSC do not include information on rainfall or flooding, INAM's operational forecasters use multiple rainfall forecast products produced by global forecasting centres (like ECMWF) to prepare local rainfall forecasts based on their expert analysis (Emerton et al., 2020). The Red Cross in Mozambique (CVM) relies on INAM's meteorological forecasts (wind and rainfall) and on the hydrological bulletins from the National Hydrological Agency (DNGRH) for TC and flood early action plans (EAPs).

User wishes and requirements for enhanced climate services for TCs

In the survey of D7.1, the humanitarian stakeholders and CS users mentioned the need for better TC and flood triggers, requiring enhanced data and forecasts with increased quality and reliability, while taking into account also vulnerability and exposure layers. The objective of CLINT ZW's TC-induced floods use case is to generate an AI-enhanced early warning system for TCs to improve flood preparedness in Mozambique. In this deliverable, for the AI enhancements assessment, we focused on the improvements of medium-range (5-day) extreme TC rainfall forecasts and their operational value for early warnings and EAPs, following the requirements and current decision processes of humanitarian agencies, mainly based on the Red Cross Mozambique's EAP.

Impact indicators for quantifying the value of AI-enhanced CS for TCs

The impact indicators for quantifying the value of AI-enhanced CS for Tropical Cyclones have been identified following the exchanges with the ZW's end-users (see D7.1) and can be expressed as early warning/early action triggers' success indicators, expressed as:

- Hit rates (or other forms of the critical success index of the warnings), that need to be maximised to reach as much of the population at risk as possible, given a sufficient actionable lead time;
- False alarms, that need to be minimised to reduce adaptation costs and increase trust in the early warning systems.

Their formulation will follow the needs and current plans of the humanitarian organisations involved in flood early warning systems in the lower Zambezi (e.g. Red Cross and WFP in Mozambique). As

current levels of TC forecast skill in the region limit the possible actions to be implemented for TC and flood protection, a key objective here is to enhance medium-range flood and TC forecasts to be used to improve current EAP triggers thanks to more skilful forecasts and longer actionable lead times.

2.1.2 Analysis chain for AI-enhanced CS potential added value for droughts

The methodological workflow to assess the potential added value of AI-enhanced CS for droughts in the Zambezi River Basin is based on the comparison of the skill and value of the original (benchmark) seasonal forecasts against the AI-enhanced forecasts (see Figure 2.1).

Following the inputs from the CS users and stakeholders summarised above (see D7.1 for more details), we selected the seasonal inflow to Lake Kariba, as this is the major and most influential dam for drought control in the basin. In particular, in line with ZAMCOM's DSS for droughts (ZAMWIS), 3-month aggregated inflows are selected as the key variable of interest to inform the dam operation. The CS enhancement is performed in the forecast production and information selection step and is thus grounded in the real operation work of ZAMCOM, which provides 3-month streamflow forecasts of inflows to the dam managers of Kariba. In this use case, our CS enhancement model is used to forecast directly the variable of interest from which drought events can be detected (i.e., 3-month aggregated streamflow), following a purely data-driven model approach. Then, either the benchmark forecasts of inflows or the enhanced forecasts can feed an operational model, informing and optimising Kariba dam's releases (this will be part of Deliverable D7.3). Finally, the potential added value will be calculated as the extra hydropower production or the reduced irrigation deficit resulting from operations guided by AI-enhanced forecasts, compared to operations based on benchmark forecasts or a no-forecast baseline.

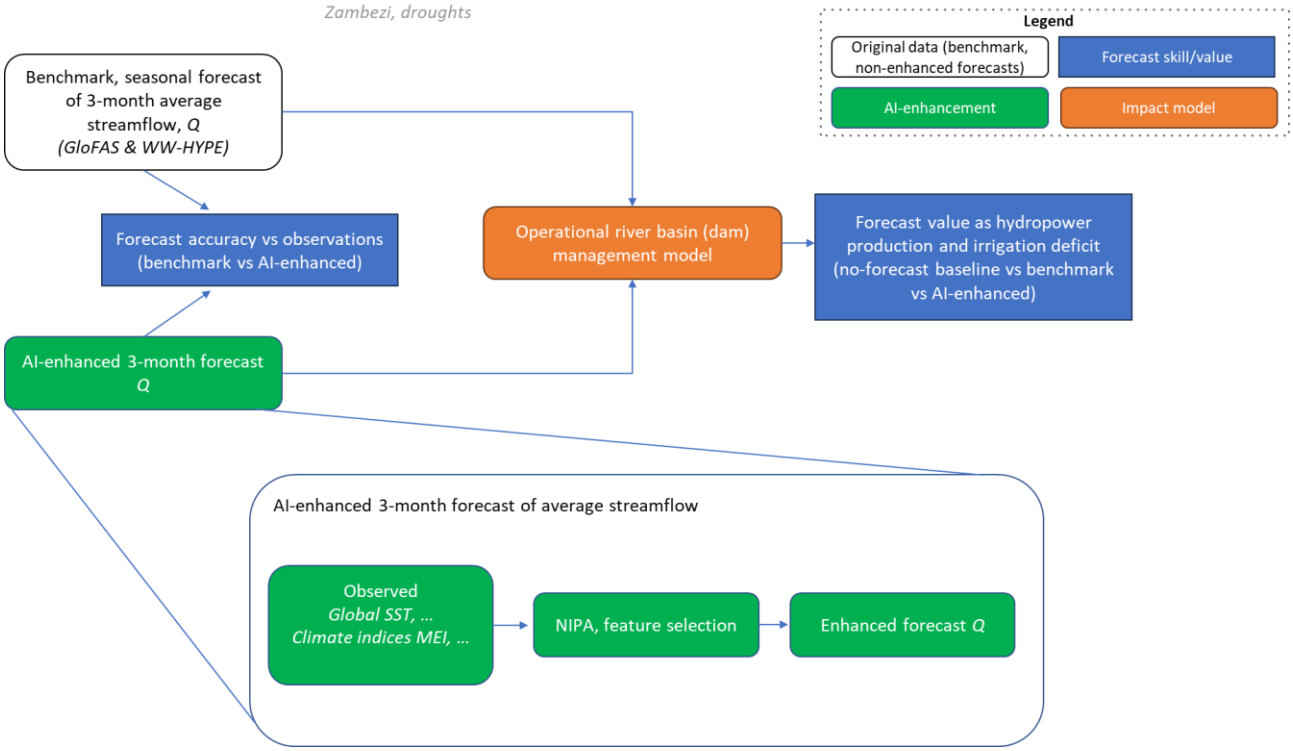

Figure 2.1. Flowchart for assessing potential added value of AI-enhanced CS for droughts in the Zambezi River Basin (see text for details).

Data and benchmark

To benchmark our AI-enhanced forecasts, we used seasonal forecasts from the Global Flood Awareness System GloFAS (version 3.1), produced by Copernicus Emergency Management Service (CEMS, run by ECMWF and EC-JRC), and seasonal forecasts from the Worldwide HYPE, WW-HYPE, system produced by SMHI, as these are operational and freely available systems and accessible to local users in the Zambezi River Basin. Operational (real-time) seasonal forecasts from these two forecasting systems have a maximum lead time of 7-months and a monthly frequency of forecast update. In particular, here we used re-forecasts, produced by running a consistent (i.e., the latest) version of the operational system over an historical record; these are available with a shorter lead time for GloFAS (4 months), but sufficiently long to cover the horizon of interest for the local users (3 months). Both systems are based on a spatial distributed hydrological model that is fed with operational seasonal ensemble forecasts produced by ECMWF (SEAS5); for WW-HYPE, these seasonal forecasts have been previously bias-adjusted through the Quantile Mapping technique. Both hydrological ensemble forecasts consist of 25 members. While streamflow forecasts in GloFAS are available as a gridded dataset (along the river network) at 0.05° grid resolution, WW-HYPE forecasts are provided at the outlet of sub-basins (average size 1000 km2). The calibration of the two hydrological models behind GloFAS and WW-HYPE is carried out on a global scale, using observations sourced from global datasets.

For the Zambezi Basin, the calibration process relies on a limited number of river stations available (from GRDC and shared by local agencies to the GloFAS/HYPE development teams), resulting in significant portions of the basin remaining largely uncalibrated, especially upstream of Kariba. For

both systems, streamflow forecast data were extracted at the grid cell / sub-basin outlet corresponding to the Kariba Lake inflow, identified based on coordinates and catchment area. As reference data, for the forecast assessment of this deliverable, daily historical observations of streamflow upstream of Kariba were available from the global Global Runoff Data Centre (GRDC, 2022) dataset. In particular, we used streamflow observations at the stream gauge station of Victoria Falls (Lat: -17.91°, Lon: 25.85°). We aggregated the daily observed time series at the monthly time step and computed the 3-month seasonal average flow over the calendar months (from the first day of each month over the available period), aligning observations with the seasonal forecasts. The common period of availability for observations and seasonal forecasts (GloFAS and WW-HYPE) is 1993-2015.

In Appendix A, we report a summary of a more detailed verification of the two seasonal hydrological forecast benchmarks that has been conducted for both systems across different locations. This additional analysis has been conducted at four selected locations in the upstream/middle part of the Zambezi Basin, corresponding to the point extracted here upstream of Kariba and other strategic locations to support other dams' operational decisions, with observed data available. In this deliverable, we focus on the AI-enhancement of forecasts of inflows to Lake Kariba only, as this is not only the most strategic location for drought control (see Section 2.1), but also where local observations are available over a much longer period (1924-2018) allowing to extend the period of analysis that is limited by (re-)forecast availability.

Impact model

The impact model for the ZW use case for droughts is an operational river basin model integrating a dam management simulation component and a multi-objective optimisation routine. The implementation of this model is ongoing and will be described in detail, alongside the final results of the value assessment, in Deliverable D7.3. The impact model for the ZW is similar to the one used for the Lake Como case study (presented in Section 2.5 of this Deliverable). As described in further details for Lake Como, the dam operations of Kariba are determined by a closed-loop operating policy that computes the release decision at each time step as a function of the time of the year, the water level in the lake and, in the forecast-informed scenarios, the upstream seasonal inflow forecast. A multi-objective optimal control problem (Castelletti et al., 2008) is solved to compute a set of Pareto-optimal solutions exploring different trade-offs between the operation objectives. For the ZW case study, the objectives considered are: (i) annual average hydropower production deficit (to be minimised), and (ii) irrigation deficit (to be minimised). The definition of the two objectives will be described in detail in D7.3.

AI enhancement

To generate our enhanced forecast of seasonal (3-month) streamflow, we used the Nino Index Phase Analysis (NIPA) framework proposed by Zimmerman et al. (2016). NIPA is a data-driven predictive tool that can be used for forecasting hydroclimatic variables on a seasonal time scale, leveraging on the well-established influence of the El Nino Southern Oscillation (ENSO) on hydro-climatic variables at the global scale. Global teleconnections exhibit a significant influence on hydrological variables over different regions, including on droughts in the Zambezi region (e.g., Cheon et al., 2021; Gaughan et al., 2016), and the phase of ENSO is known to affect the 'mean state' of the climate system and modulate the impact of global teleconnections (e.g., Taschetto et al.,

2020). Building on this knowledge, NIPA uses the state of ENSO to classify years into different phases (e.g., ENSO positive and negative years, if two phases are used). Then, based on the selected ENSO phases, NIPA extracts the correlation between lagged Sea Surface Temperature (SST) fields and the variable to predict (i.e, seasonal streamflow, in our case), identifying the most correlated SST fields. By splitting the data according to the different phases of ENSO, different global fields over each pre-season (prior to the target season to predict) can be selected as predictors, depending on the state of the climate system. These are then used in a seasonal forecast model, e.g., a linear regression model. As past SST observations are available at the global scale in near-real-time (i.e., SST data for previous months is available at the beginning of each month), this data-driven prediction approach can be used operationally.

In our implementation, following Giuliani et al. (2019), NIPA classifies years into two phases based on ENSO positive and negative years, using the Multivariate ENSO Index (MEI) from NOAA. As global SST data, we use the NOAA's Extended Reconstructed SST (ERSST, version 3b) dataset that provides monthly gridded data at spatial resolution of 2.5°. For selecting the predictors for seasonal streamflow, we use a lag time of 3 months, initialising each season at the beginning of each calendar month. We calibrate a distinct NIPA model separately over each month, using all years categorized in the two ENSO phases, thus building 24 models in total. For example, the model for the month of January of a year categorized as ENSO positive is calibrated using as target variables the aggregated average 3-m streamflow (JFM) of all ENSO-positive years, while the predictors are extracted from SST data over the previous 3 months (OND). As in Giuliani et al. (2019), we identify the SST predictors as the correlated regions at the 95% significance level, and we use a Principal Component Analysis (PCA) on the resulting SST fields to extract and use the first principal component (PC) as predictor in a linear forecast model, defined as:

$$\hat{y}_t = \beta * PC^1_{\tau-1} + \alpha$$

where: $\hat{y}_t$ is the predicted seasonal (3-month) streamflow, β is the regression coefficient and is the α intercept. Given the limited length of our data period (23-years) and the low interannual persistence of seasonal streamflow, to avoid overfitting, we used a leave-one-out cross-validation approach. The same approach, building on the NIPA framework, is also used in another CLINT case study, Rijnland, to generate AI enhanced predictions of precipitation (see Deliverable D2.2 for more details on the methodology).

2.1.3 Results towards potentially added value AI-enhanced CS

As our data-driven forecast model (NIPA) is deterministic, in this deliverable we focus on a comparison of the deterministic forecast performance between the ensemble mean of the two benchmark systems and the enhanced (NIPA) forecast. In Appendix A, we report a more complete probabilistic verification of the two seasonal hydrological forecast benchmarks, for further reference. As the forecast-based operation may be impacted by biases (lack of correlation or low accuracy of the relative variability between forecasts and observations), we focus on these different attributes of forecast quality to capture the forecast performance differences between the two benchmark systems and NIPA. In particular, six different metrics are considered to quantify the deterministic accuracy of the forecasts:

(i) the Mean Absolute Error, MAE;
(ii) the Mean Absolute Percentage Error, MAPE;
(iii) Kling-Gupta Efficiency, KGE;
(iv) variability ratio, alpha;
(v) bias ratio, beta; and
(vi) correlation, r.

The latter three metrics (alpha, beta and r) are the three components used to compute KGE. All metrics are computed on the target time series of 3-month average streamflow over the 23-year period (1993-2015).

The results show that our enhanced forecast model strongly outperforms both seasonal forecast benchmark systems, for all considered metrics (Table 2.1; Figure 2.2-2.5).

Table 2.1. Deterministic scores for mean 3-month flow forecasts of the benchmark (WW-HYPE and GloFAS) and AI-enhanced (NIPA) streamflow forecasts over 1993-2015 at Victoria Falls (WW-HYPE sub-basin ID: 207136). The ideal scores are: MAE=0, KGE=1, alpha=1, beta=1 and r=1.

Forecast product	MAE [m3/s]	KGE [-]	alpha	beta	r
GloFAS	3251.03	-3.106	3.745	4.037	0.68
WW-HYPE	936.73	-0.315	0.210	0.125	0.42
NIPA	230.2	0.878	0.908	0.972	0.93

The accuracy of the ensemble mean of GloFAS and WW-HYPE have different quality attributes and the performance ranking between the two benchmarks varies depending on the accuracy scores (Table 2.1). In particular, GloFAS forecasts have a larger mean absolute error (MAE) than WW-HYPE, and presents a large overestimation with respect to observations (as indicated by beta > 1), while WW-HYPE underestimates seasonal streamflow upstream of Kariba. The very large bias in GloFAS indicates that there is a systematic positive bias in the water balance of the hydrological model, as confirmed by the inspection of the hydrographs (see Figure 2.2). Overall, WW-HYPE seems to outperform GloFAS by looking at the aggregated performance metric (KGE) that combines bias, correlation and relative variability into a single metric. However, WW-HYPE also has a large margin of improvement, as it shows a large negative bias (Figure 2.2) and performs similarly to a simple mean flow benchmark (KGE just above -0.41; Knoben et al., 2019). The more refined diagnostics provided by the KGE components (alpha, beta and r) show that GloFAS seasonal forecasts outperform WW-HYPE in terms of correlation (r=0.68 vs. r=0.42), while being worse than WW-HYPE in terms of bias and relative variability of river flows with respect to observations (see Table 2.1 and Figures 2.2, 2.3 and 2.4).

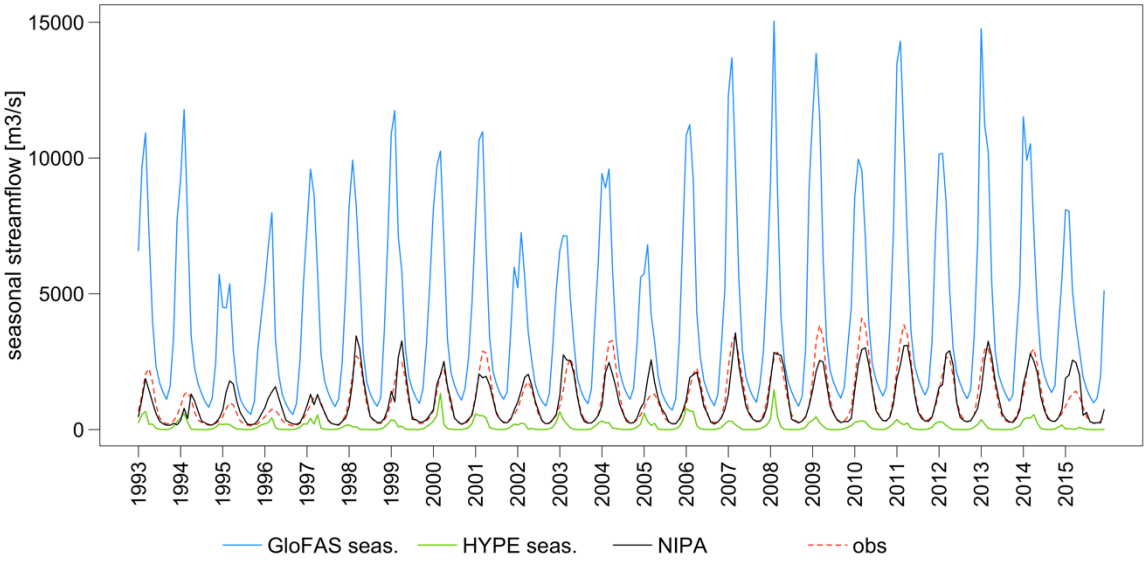

Figure 2.2. Average seasonal (3-month) streamflow (in m3/s) of GloFAS, HYPE and NIPA forecasts vs observations over the full period 1993-2015.

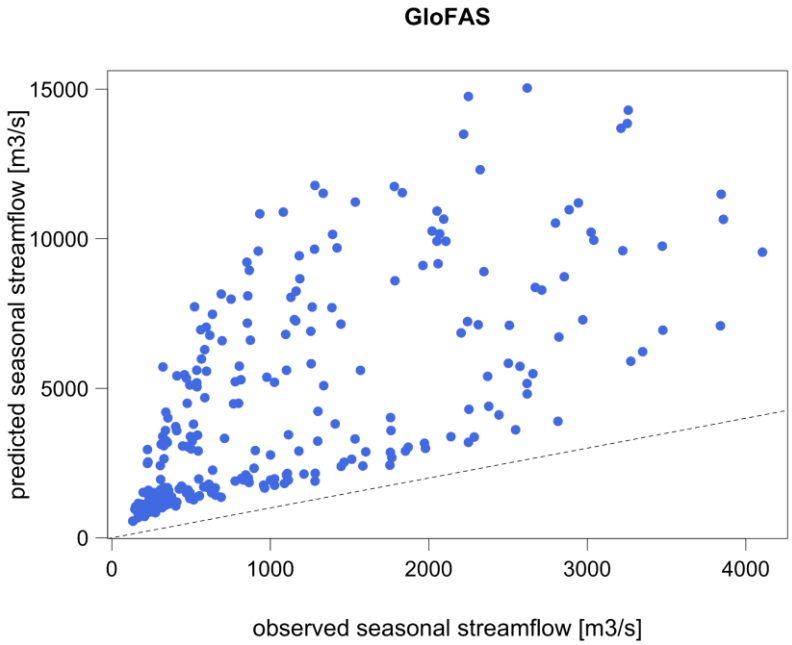

Figure 2.3. Scatterplot of average seasonal streamflow comparing GloFAS ensemble forecast mean and observed seasonal streamflow over the full period 1993-2015. The dashed line is the 1:1 line (for reference).

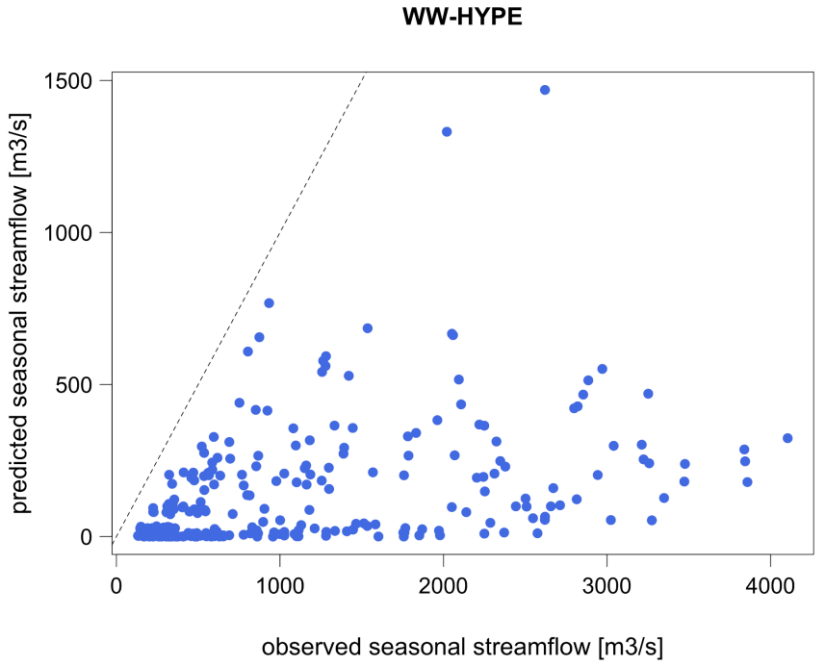

Figure 2.4 Scatterplot of average seasonal streamflow comparing WW-HYPE ensemble forecast mean and observed seasonal streamflow over the whole period (1993-2015). The dashed line is the 1:1 line (for reference).

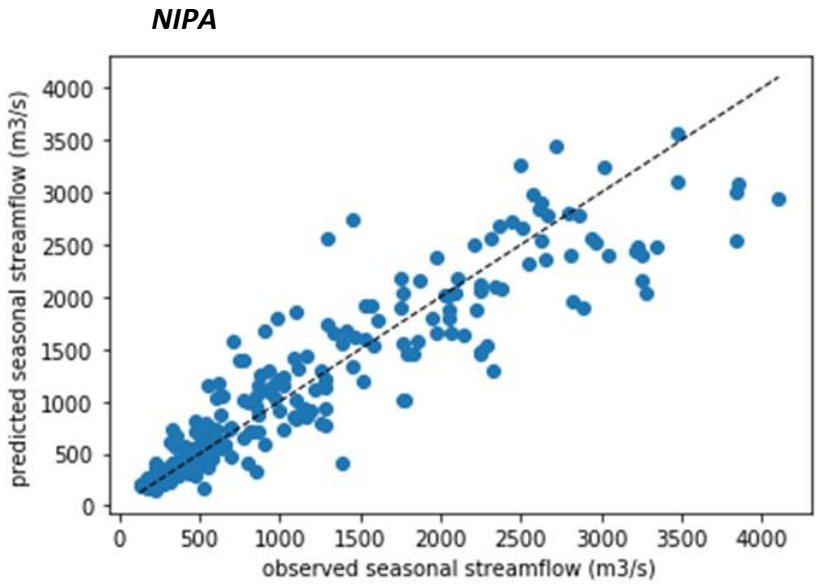

Figure 2.6 Scatterplot of average seasonal streamflow comparing NIPA forecast and observed seasonal streamflow over the whole period (1993-2015). The dashed line is the 1:1 line (for reference).

The results have also been broken down across seasons (Tables 2.2, 2.3 and 2.4) to assess any seasonal dependence of the quality of forecasts. In this case, we assess only the attributes of bias and relative variability, as the correlation across non-continuous time series would not give any sensible information. GloFAS and WW-HYPE present a large variability of biases across seasons, with strong changes along the year, especially when looking at relative biases with respect to observed streamflow (e.g., MAPE and beta). GloFAS shows the largest biases in relative terms (MAPE), with an overestimation of average flows (beta>1) and overestimation of the variability of flows (alpha>1) in DJF (Table 2.2). WW-HYPE has the largest relative biases, with the worst underestimation of observed flows (beta<1) and their variability (alpha<1) in SON (Table 2.3). On the other hand, the enhanced forecasts (Table 2.4) show a more stable relative variability (alpha) and bias, especially in relative terms with respect to the observed seasonal streamflow (MAPE and beta). Still, a dependence of the bias of NIPA forecasts along the calendar months can be observed, especially in absolute terms, with larger biases in the high-flow season, i.e. peaking in MAM, as it could be expected (Figure 2.6). Given our focus on droughts for this use case, the results of NIPA are particularly promising, as the largest improvement is found in the low-flow season (SON).

These results suggest that a data-driven model built upon observed preseason SST anomalies can lead to more skilful predictions of seasonal streamflow in the Zambezi than global scale hydrological model-based seasonal forecasts.

Table 2.2 Deterministic scores for DJF, MAM, JJA and SON average streamflow over the full period 1993-2015 for GloFAS forecasts against observations (MAE, MAPE, alpha and beta).

GloFAS performance	DJF	MAM	JJA	SON
MAE [m3/s]	5277.8	6193.1	952.6	734.6
MAPE [%]	880.5%	310.4%	122.0%	306.5%
alpha [-]	4.445	2.114	1.041	3.472
beta [-]	8.822	3.594	1.977	3.992

Table 2.3. Deterministic scores for DJF, MAM, JJA and SON average streamflow over the whole period (1993-2015) for WW-HYPE forecasts against observations (MAE, MAPE, alpha and beta).

WW-HYPE performance	DJF	MAM	JJA	SON
MAE [m3/s]	472.5	2067.6	960.7	242.4
MAPE [%]	66.0%	83.6%	98.5%	98.7%
alpha [-]	0.243	0.291	0.042	0.024
beta [-]	0.300	0.134	0.015	0.013

Table 2.4. Deterministic scores for DJF, MAM, JJA and SON average streamflow over the whole period (1993-2015) for NIPA forecasts against observations (MAE, MAPE, alpha and beta).

NIPA performance	DJF	MAM	JJA	SON
MAE [m3/s]	175.1	538.4	206.9	26.6
MAPE [%]	27.5%	29.7%	29%	12.6%
alpha [-]	0.740	0.796	0.767	0.799
beta [-]	0.964	0.988	0.995	0.992

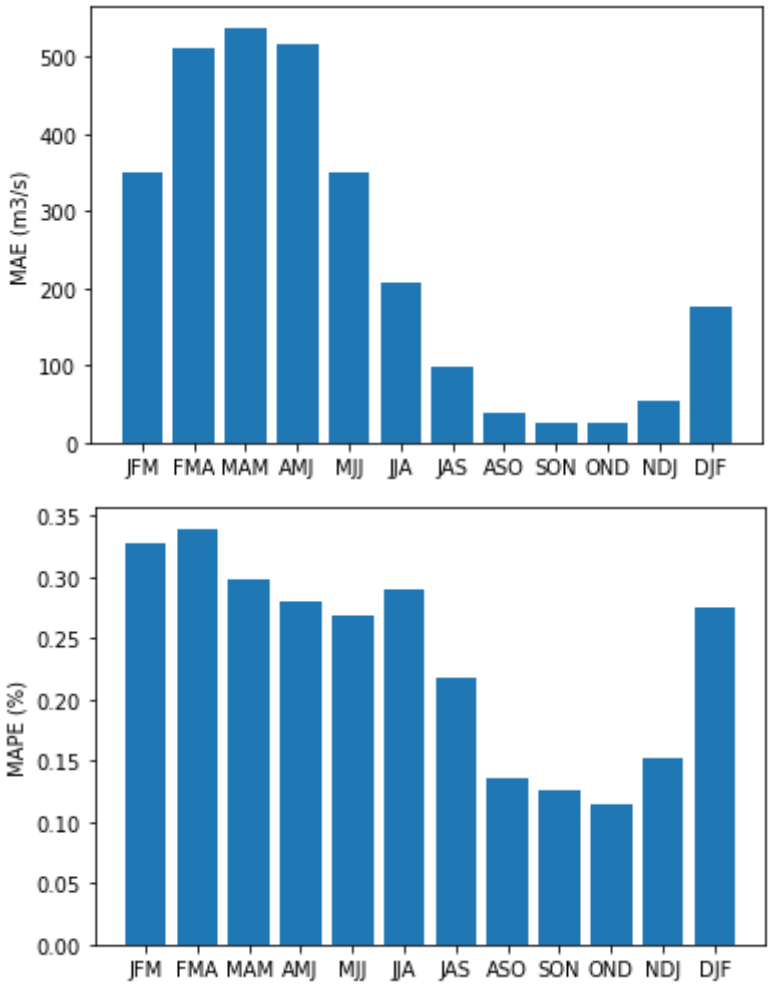

Figure 2.6. Bar chart of MAE and MAPE of NIPA forecast with respect to observed seasonal (3-month) streamflow by calendar month over the full period 1993-2015.

2.1.4 Analysis chain for AI-enhanced CS potential added value for Tropical Cyclones

The methodological workflow followed to assess the potential added value of AI-enhanced climate services for Tropical Cyclones (TCs) in the Zambezi River Basin is based on the comparison of the skill and value of the original (benchmark) medium-range forecasts against the AI-enhanced forecasts (Figure 2.7). We assessed the added value in terms of improvement of action-relevant scores designed based on the needs of humanitarian users and stakeholders involved in the early warning early action chain for TCs (see Deliverable D7.1), i.e. adjusted hit rates and false alarm ratios, that can lead to more effective warnings and actions, reducing costs and impacts. Following the inputs from D7.1 (as summarised also above; see Section 2.1), extreme rainfall has been selected as the key variable of interest to inform early warnings and action, being one of the variables indicated by the CS users and also used to feed flood prediction models and generate other variables (streamflow and inundation extent). Here, the AI-enhancement is performed in the forecast production and post-processing step, with a model that can be run to improve forecasts available in real-time within computation times compatible with operational needs (a few minutes). The

model can be quickly used, once it is trained over multi-year historical records off-line (only at a single point in time, before any TC events), which takes a few hours. Thus, the approach is grounded in the real operational work of the humanitarians and disaster managers who rely on available forecasts ahead of TCs impacts, of extreme rainfall and subsequent floods, based on which early warnings can be issued. In this use case, the AI enhancement is carried out via a deep learning model to post-process available operational rainfall forecasts, like our benchmark (ECMWF's HRES); thus, here we follow a hybrid AI model approach.

State-of-the art TC forecasts lack in correctly predicting TC tracks and rainfall peaks location with sufficiently long lead times; for example, 3 days ahead typical average track location errors are of ~200 km (e.g. Emerton et al., 2020). Given the needs and wishes expressed by the users (see Section 2.1) to improve early warnings for TCs in the region based on medium-range reliable forecasts, our goal here is to enhance forecast accuracy and value focusing on the medium-range (up to 5 days) which is a critical horizon for decision making. This allows us to start from a foundational level of forecast skill, even if often affected by large TC tracking errors, providing a base for enhancement efforts and mitigating the challenges related to missed TC detection and tracking at longer lead times, which would otherwise complicate efforts to improve rainfall forecasts. By concentrating on medium-range forecasts, we will address issues such as rainfall biases and lack of accurate prediction of extreme rainfall location. Thus, here we aim to investigate whether extreme rainfall forecasts can be improved, once a TC has been forecast to occur and possibly to move towards land or further inland, potentially leading to impacts to prepare for and respond to.

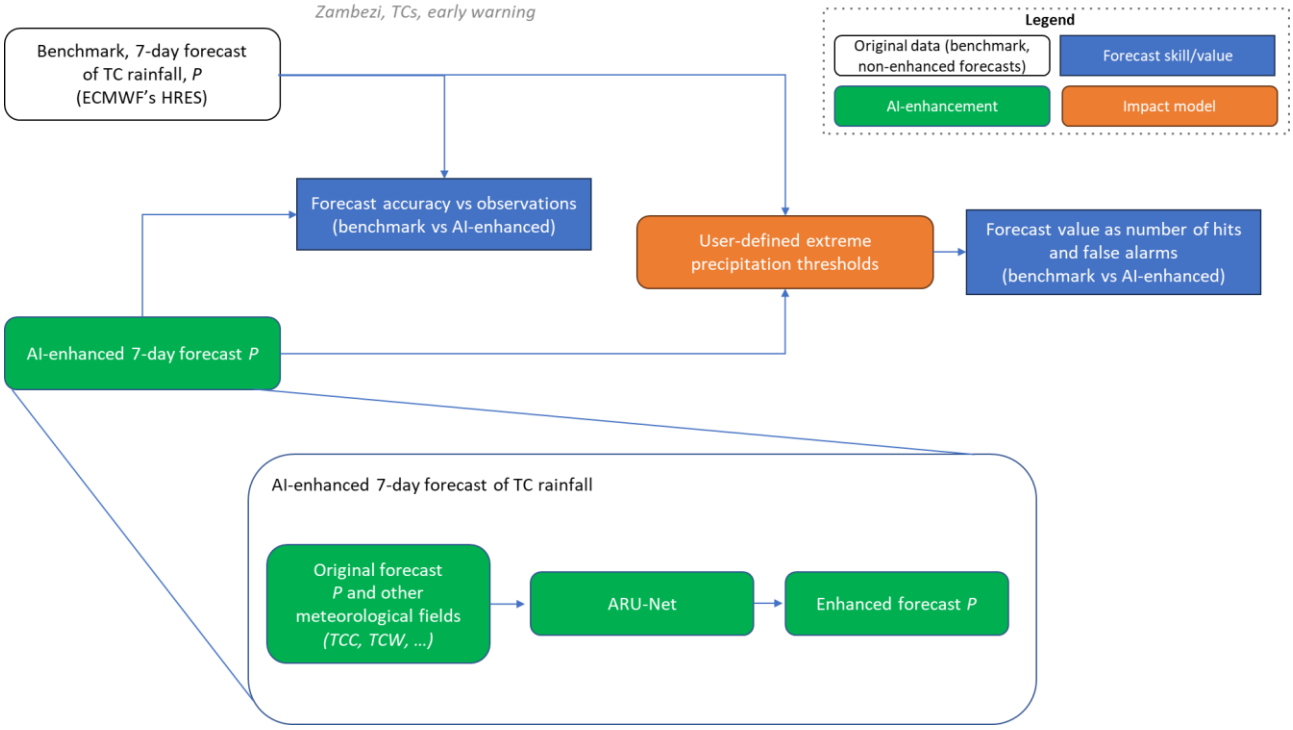

Figure 2.7 Flowchart for assessing the potential added value of AI-enhanced climate service for Tropical Cyclones for the Zambezi River Basin.

Data and benchmark

The ECMWF's High Resolution (HRES) forecasts were used as the benchmark to assess the potential added value of AI-enhanced forecasts for TCs early warning in the Zambezi River Basin. HRES forecasts are widely considered the top global deterministic operational forecasting system and often chosen as a benchmark in recent machine learning–based weather prediction studies (e.g., Lam et al., 2023; Rasp et al., 2020). HRES is a medium-range forecast system run twice daily (at 00 and 06 UTC) with a maximum lead time of 10 days, produced by the Integrated Forecasting System (IFS) model of ECMWF. Since March 2016 (IFS Cycle 41r2), HRES forecasts are run at the horizontal grid resolution of 9km (about 0.08 degrees) and stored at a temporal resolution between 1h to 6h, varying with lead time (1h up to 90h lead time, 3h up to 144h lead time, and 6h onwards). With its higher resolution compared to other global forecasting models, HRES offers the most precise single-run representation of large-scale weather patterns. Its use allows for a more detailed analysis of precipitation patterns than, for example, ECMWF's ensemble forecasts (ENS), especially in multi-year studies (Owens and Hewson, 2018; Magnusson et al., 2021). In June 2023, ENS was also upgraded to the same 9-km resolution as HRES, but a 9-km ENS reforecast is still under production and no multi-year ensemble reforecast run at 9 km is currently available. Given our need for a multi-year long data record for model training and validation (to include a large sample of TC events), we opted for the use of the operational HRES forecasts, which provide a record overlapping for 4 years (2016-2019) with the observational data available and used here (see below).

In terms of forecast horizons, we decided to focus on lead times up to 5 days, given the current levels of predictability and TC track errors, with TC position errors of HRES operational forecasts produced in 2020 of approximately 200 km and over 300 km at 3- and 5-day lead time, respectively (Magnusson et al., 2021). These large errors still limit forecast-based action for TCs over longer time scales than a few days (e.g., IFRC, 2024). For example, a 3-day lead time is used so far in the Red Cross' EAP for TCs in Mozambique (see Section 2.1), while 72-h and 30-h lead times (or slightly more) are used for TC early action pre-activation and activation in Bangladesh (IFRC and Bangladesh Red Crescent Society, 2021). We selected a 6-h target resolution, which is the current resolution of warnings issued by MF in the SWIO during a TC. Using a common resolution across the lead times considered makes the comparison of scores consistent, while avoiding increasing the resolution limits the correlation of the samples, which is better for training our model. Similarly, as HRES forecasts are issued at 00 and 12 UTC, but forecast maps issued 12 hours apart are expected to be highly correlated, only one daily forecast issue step was considered (00 UTC).

To help the deep learning model correct the biases of TC rainfall forecasts and improve spatial accuracy, we considered five different candidate inputs from HRES in addition to total precipitation:
(i) total column of water,
(ii) temperature at 850 hPa,
(iii) total cloud cover,
(iv) relative humidity at 850 hPa, and
(v) mean sea level pressure.

Their choice was based on first model development efforts on ERA5 (Ascenso et al., *under review*) and on previous studies (e.g., Sha et al., 2020; Hu et al., 2022; Ling et al., 2022). Finally, after testing

all possible multi-input combinations using HRES data, only the first two fields (i and ii) were selected as additional model inputs to total precipitation, as their use provided the best performance and reduced overfitting with respect to other two-input configurations. Including more inputs did not help or would keep the performance at the same level at the expense of increased computation.

As target of the AI enhancement and reference for the forecast skill assessment, we used the Multi-Source Weighted Ensemble Precipitation (MSWEP) dataset (Beck et al., 2019a). MSWEP provides global observational precipitation data derived from multiple sources, including ground-based observations, satellites, and reanalysis products, at a spatial resolution of 0.1 degrees (approximately 10 km) and a temporal resolution of 3 hours. The multi-source data integration of MSWEP enhances its performance and robustness with respect to other single source datasets (Beck et al., 2019b) and showed the highest accuracy in multi-datasets comparative studies (e.g., Sharifi et al., 2019; Beck et al., 2017), making it a suitable reference for validating the accuracy of AI-enhanced rainfall forecasts.

To define the domains for rainfall forecast evaluation and post-processing, we located TC centres using the International Best Track Archive for Climate Stewardship (IBTrACS) best-track data (version v04r003), which offers a global TC dataset at 3-hourly temporal resolution (Knapp et al., 2010). The IBTrACS reports instantaneous TC data every 3 hours starting at 00:00 UTC. Considering the target 6-h resolution of the two other products (HRES and MSWEP) and the forecast issue time (00 UTC), we sub-sample the IBTRrACS data to receive instantaneous data aligned with the centre of the MSWEP and HRES accumulation window. For example, TC data at 03:00 UTC are used to match rainfall maps accumulated between 00:00 and 06:00 UTC. Subsequently, for each time step in IBTrACS, we crop the HRES and MSWEP fields surrounding a 14-degree-side (i.e., about 1550 km) box centred on the TC location both temporally and spatially. This approach and box size allows us to encompass an area large enough to include all the grid cells with TC rainfall at each time step, as this distance (>1500 km) is larger than extreme TC sizes. The output is squared domains of 141-grid-cell side, with one channel for MSWEP and more channels for HRES, i.e. one for each selected input variable (total precipitation, total column of water, and temperature at 850 hPa). As MSWEP is a 3-hourly aggregated product and HRES rainfall forecasts at close hourly time steps are highly correlated, we perform a temporal aggregation at a common window of 6 hours for both products, i.e., to obtain 6-hourly accumulated values for HRES total precipitation and column of water and 3-hourly average for HRES temperature (and other instantaneous variables that were initially tested, like relative humidity).

In summary, to prepare the inputs for the deep learning model from the HRES forecast data (precipitation and other input variables), we followed these steps:
(i) HRES data were downloaded at the global scale from ECMWF's Meteorological Archival and Retrieval System (MARS);
(ii) 6-h step values were obtained from the original cumulated ones (e.g., de-accumulating values of two adjacent steps for rainfall or averaging close instantaneous values for temperature);
(iii) a spatial regridding was performed from the reduced gaussian grid system of IFS (octahedral from 2016) to the (close) regular 0.1/0.1 lon-lat grid resolution of MSWEP, applying the conservative interpolation method (using ECMWF's MetView Python library);

(iv) the regridded data is then crop over the regions of interest (14-degree-side boxes), defined based on the TC locations retrieved from IBTrACS.

Impact model

In this section, we present the valuation framework to assess forecasts and select triggers for humanitarian EAPs, that can be adapted to consider the parameters of specific anticipatory actions, similarly to what proposed by Coughlan de Perez et al. (2016). EAPs play a crucial role in the planning of forecast-based financing systems, facilitating the allocation of resources before a hazard occurs according to a predetermined forecast and trigger. The trigger analysis in EAPs acknowledges the possibility of actions to be taken "in vain" if a forecast hazard does not occur, with the goal of ensuring that the long-term benefits of preventive actions outweigh the costs of false alarms. The two key scores for this are False Alarm Ratios (FAR) and Hit Rates (HR):

$$FAR = \frac{false\ alarms}{hits + false\ alarms}$$

$$HR = \frac{hits}{hits + misses}$$

We propose a modified version of FAR and HR, to assess the capacity of a forecast in meeting the requirements for effective early actions, by redefining hits and false alarms in terms of spatial accuracy and acceptable margins of error. In particular, we considered the requirements of the Mozambique Red Cross Society (CVM), derived from information from their Early Action Protocol for TCs (IFRC, 2024). The activation of the Cyclone EAP in Mozambique is based on the TC forecast information distributed 72-hour ahead by INAM, as at this point the margin of error is approximately 240 km. Taking this information into account, we consider a more restrictive margin of error (<150 km), that we called ‘action scale’. In particular, for each grid cell with a forecast event (exceeding an extreme TC rainfall threshold) we take a box of half-side equal to 100 km and we define that forecast as either a false alarm if no observed event occurs in the box, or a hit if an event is observed in the box (Figure 2.8). In other words, we redefine hits and false alarms based on a specific action scale, i.e. how much farther the hazard can occur from a forecast location and it still counts as a ‘hit’.

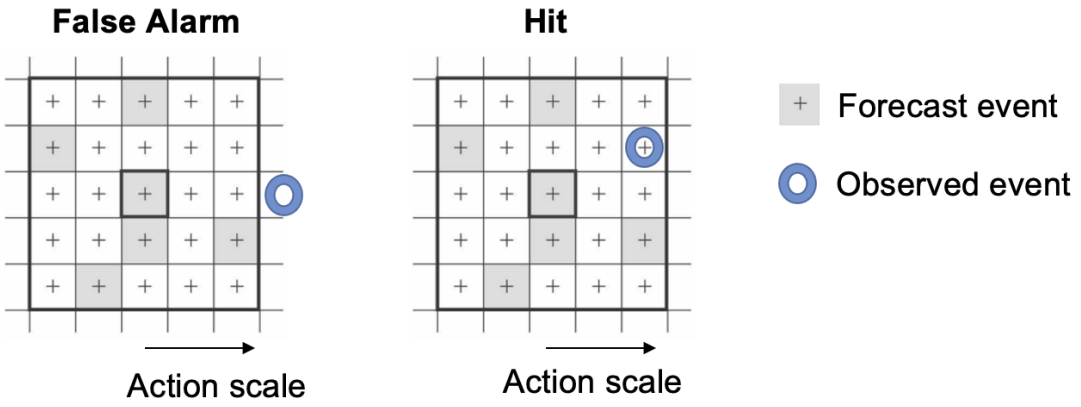

Figure 2.8. Scheme of the adjusted definition of False Alarms and Hits based on an action scale (maximum acceptable margin of error to inform early actions).

The trigger methodology that we adopted is based on the selection of a forecast (e.g., HRES) and thresholds (e.g., 99th percentile of TC rainfall peaks) for triggering early actions considering the users' willingness to act 'in vain' for anticipatory actions (e.g., FAR < 0.5) and to reach a high number of people at risk (e.g., HR > 0.5). In this deliverable, we focus on the value of reducing false alarms, as reduced FAR can be easily translated into a saved amount from a lower rate of actions in vain, for each specific action. Thus, our simple impact model consists of a binary-threshold classification to assess forecast value in terms of reduced costs. Here, as an example of application, we consider the planned early actions from the EAP for TCs in Mozambique, with the current activation lead time (3 days). We assume a different trigger (extreme precipitation forecast instead of wind speed) that is still considered of interest for users (see Section 2.1). The current EAP has a timeframe of 5 years and targets one activation for an event with a return period of 5 years, with a total allocation of CHF 195'962 for early actions, once the defined triggers are met, while the rest of the budget is for readiness and prepositioning actions. Thus, a target reduction of FAR of 10% would correspond in the long term to a saving of approximately CHF 20k/5yr.

AI enhancement

In this Section, we describe the AI enhancement approach that we implemented for the post-processing of TC rainfall forecasts, based on a variant of a state-of-the-art deep learning architecture, UNet, and a novel loss function (Figure 2.9). Given our evaluation framework for early action described above (Section 2.3.2), our primary objective is to improve the spatial discrimination of extreme events in the forecasts, refining the localization of rainfall peaks, and reducing errors within acceptable margins. To achieve this, we introduce a novel loss function, called the *compound loss*, including two components:
(i) the Mean Squared Error (MSE), to correct pixel-wise biases and
(ii) the Fractions Skill Score (FSS, Roberts and Lean (2008)),

to improve the accuracy of spatial patterns. Previous studies using deep learning for bias correction of rainfall maps predominantly relied on pixel-wise metrics, like the MSE, which overlook overall spatial accuracy and potentially lead to overly smoothed predictions. Recent studies have shown

that such local (pixel-wise) error metrics discourage models from making predictions with sharp gradients, often leading to "blurred out" predictions (e.g., Hess and Boers, 2022; Lagerquist and Ebert-Uphoff, 2022). By integrating the FSS into our compound function, we aim to mitigate such issues and enhance the accuracy of rainfall peak localization, reducing false alarms and increasing hit rates.

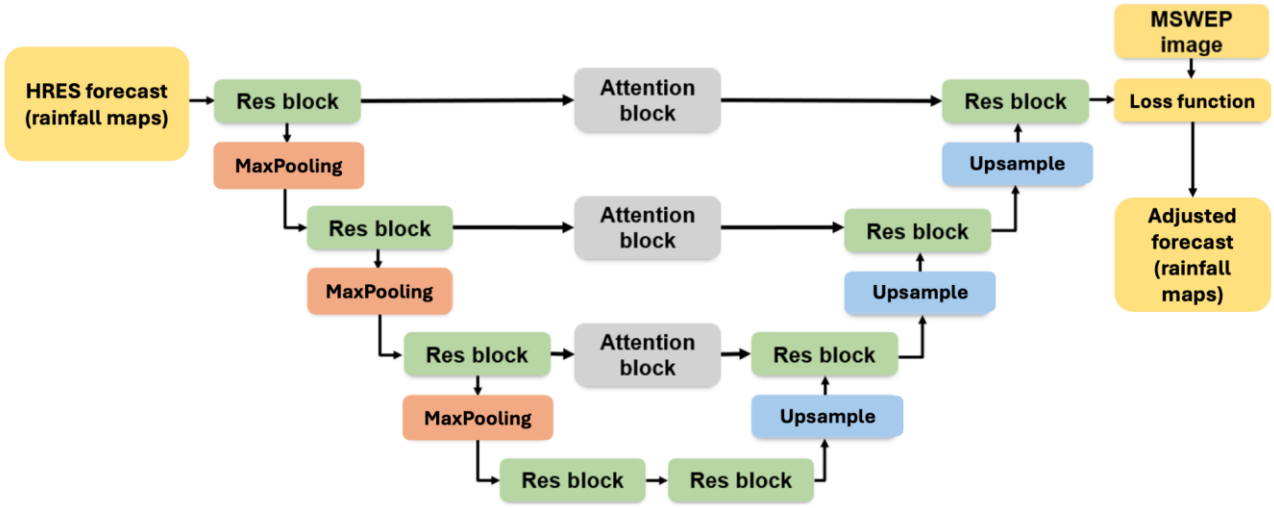

Figure 2.9. Flowchart of the AI-enhancement model for TC rainfall for the Zambezi River Basin based on RA-UNet.

The FSS is a popular spatial verification metric often used in meteorology to evaluate the resemblance between spatial patterns in two gridded datasets, typically comparing model predictions to observational data, yielding values between 0 (indicating no match) and 1 (perfect match). The FSS is computed through the following three steps:
(i) the rainfall maps (prediction and observation) are converted into binary maps, by applying a rainfall intensity threshold (Q) that can be either a fixed value or a percentile of rainfall intensity (calculated independently for each image);
(ii) fractional coverages of threshold exceedances are computed for various neighbourhood areas;
(iii) the fraction of positive pixels within patches of a specified size (N, number of grid cells) are then used to compute a skill score based on the mean squared difference of these fractions across all possible patches.
Thus, the FSS measures the average overlap between N-sized patches of the two binary grids.

In our implementation, the neighbourhood size N was set to 19 grid cells, following a grid search empirical testing (in the range [9, 21]). To allow the use of the FSS as a loss function for deep learning, we made some adjustments, mainly to ensure differentiability, by replacing the binary classification step based on hard threshold (Q) with an arctan function transformation. Our modified version of the FSS is referred to as FSS'. Also, we inverted the score to be used in the loss function (1-FSS'), so that the value of 0 indicates a perfect match.

Finally, our compound loss function (L_{cmpd}) consists of a weighted combination of FSS' values corresponding to different percentile thresholds (80th, 95th, and 99th percentiles) alongside the Mean Squared Error (MSE), as follows:

$$L_{cmpd} = 0.5\left(FSS'_{Q=80} + FSS'_{Q=95} + FSS'_{Q=99}\right) + 0.5\,MSE$$

The weights of the loss function components (0.5, 0.5) were determined empirically through systematic exploration of the weights space. By incorporating multiple percentile thresholds for FSS', our aim is to train the model to improve rainfall peak localization across varying intensities, thus enhancing spatial accuracy while mitigating pixel-level biases (represented by the MSE).

Regarding the deep learning model architecture, we selected a recent variant of the popular U-Net (Ronneberger et al., 2015) model, which has been shown to be effective to perform similar tasks and to outperform other deep learning networks for the prediction of precipitation extremes (Otero and Horton, 2023). As in the standard U-Net architecture, our model (Figure 2.9) is based on an encoder/decoder structure: first, information is encoded through a series of layers that reduce the resolution and perform an extraction of semantic information; second, the so-processed information is decoded via a series of layers that restore the spatial resolution to the original one (0.1°) as the input maps, and maintain the high-level extracted semantic information. Encoder and decoder blocks that are at the same depth in the network are linked via the so-called 'skip connections' that help transfer information across the network. We use the Residual Attention UNet (RA-UNet, or RA-U) variant, proposed by Jin et al. (2020), which further develops the standard UNet by replacing convolutional blocks with residual blocks and integrating attention modules along skip connections (Figure 2.9). Residual blocks facilitate gradient flow, mitigating the vanishing gradient issue, while attention modules augment the RA-U network's feature extraction capability, to emphasise salient features. We chose to omit batch normalisation and dropout, as we did not need these features to solve overfitting issues (see Results). For model training, we used the Adam optimizer with early stopping on an NVIDIA A100 GPU.

2.1.5 Results towards potentially added value AI-enhanced CS

First, we present the results at the global scale. The AI model (RA-UNet) has been trained and cross-validated over a large sample of 2872 TC time steps (6-hour resolution) from the global IBTrACS dataset. Based on our early warning/early action-oriented evaluation framework (with an action scale of 100 km), the AI-enhanced forecasts show a large improvement with respect to the HRES forecast benchmark (Figures 2.10 and 2.11). In particular, the largest improvement on False Alarm Ratios and Hit Rates is observed at 5-day lead time and for the higher rainfall thresholds, where more room for improvement in HRES is observed. Considering a threshold of 0.5 for maximum (minimum) acceptable FAR (HR), the use of our post-processing model is critical in making the forecast acceptable to support early warning and early actions for the highest rainfall thresholds.

Second, we extract the results for the Zambezi River Basin (Table 2.5-2.6). Over the study period considered (2016-2019), only four TCs are found to affect the Zambezi over a total of fifteen 6-h time steps, by selecting TC rainfall maps for which the footprint of rainfall exceedances above the 95th percentile intersects the river basin. For these TCs, the results are in line with the statistics

shown at the global scale. A large improvement is observed across all metrics, especially at 3- and 5-day lead time (Table 2.6 vs. Table 2.5). For the different components of our compound loss function, we see a substantial reduction of local biases (MSE) and improved spatial accuracy scores (FSS), especially for the highest rainfall intensities (FSS'_{q99}). At 5-day lead time, for the highest rainfall threshold (99-th percentile), False Alarm Ratios are reduced by approximately half, dropping from 67% to 29%, while Hit Rates increase from 33% to 71%. This large improvement in FAR suggests a potential monetary benefit in terms of reduced costs for early action in the long term. Considering this forecast and a 99th percentile as a possible trigger of the EAP for TCs at 3-day lead time (current lead time for early action in Mozambique), our model reduction in false alarms would correspond to a saving of approximately CHF 75k/5yr (see Section 2.3.2). Moreover, our results suggest a potential extension of the actionable lead times from 3 to 5 days, that could bring more benefits to early action planning and operations.

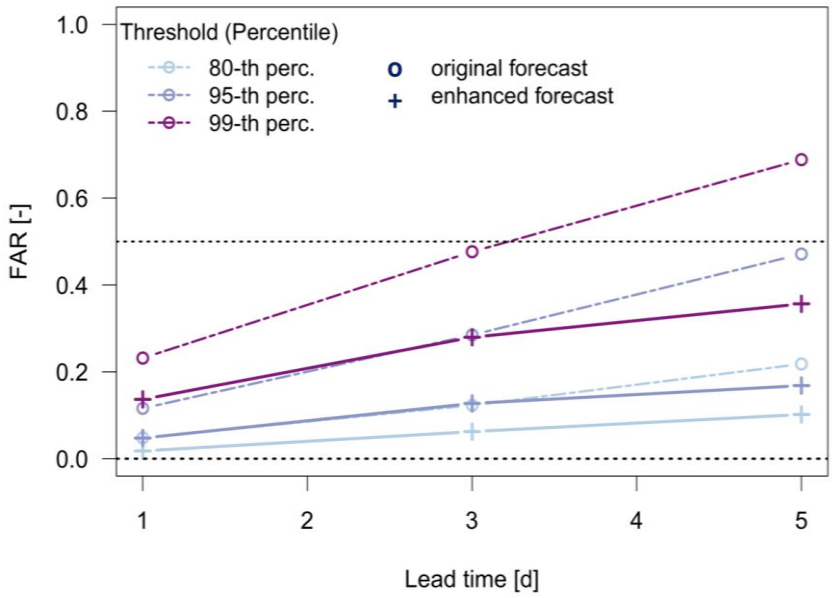

Figure 2.10 False Alarm Ratios of original HRES vs AI-enhanced TC rainfall forecasts at lead times from 1 to 5 days over the model test set of 2872 time steps (5-fold cross-validation sets) at the global scale.

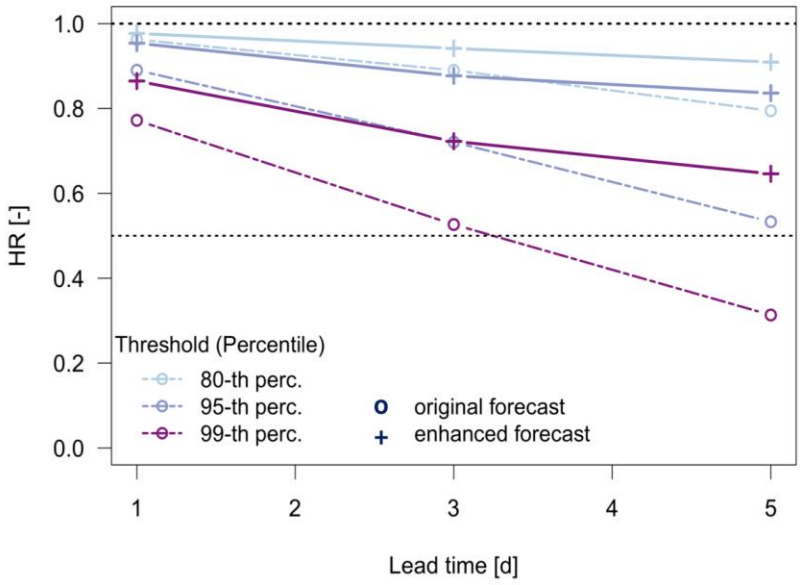

Figure 2.11 Hit Rates of original HRES vs AI-enhanced TC rainfall forecasts at lead times from 1 to 5 days over the model test set of 2872 time steps (5-fold cross-validation sets) at the global scale.

Table 2.5 Summary of scores for the original HRES rainfall forecasts by ECMWF at lead times from 1 to 5 days for four TCs impacting the Zambezi River Basin over fifteen 6-h time steps with TC rainfall above the 95th percentile over the basin. The ideal scores are: MSE=0, FSS'=1, FAR=0, HR=1. The percentile thresholds are computed over the rainfall map at each time step separately.

Original HRES forecast	**MSE [mm/6h]^2**	**FSS'$_{q95}$ [-]**	**FSS' $_{q99}$ [-]**	**FAR $_{q95}$ [-]**	**FAR $_{q99}$ [-]**	**HR $_{q95}$ [-]**	**HR $_{q99}$ [-]**
1 day	38.90	0.75	0.49	0.05	0.12	0.95	0.88
3 day	68.68	0.44	0.29	0.27	0.44	0.73	0.56
5 day	78.19	0.28	0.15	0.51	0.67	0.49	0.33

Table 2.6. Summary of scores for the AI-enhanced HRES rainfall forecasts at lead times from 1 to 5 days over the model test set (5-fold cross-validation sets) for four TCs impacting the Zambezi River Basin over fifteen 6-h time steps with TC rainfall above the 95th percentile over the basin. The ideal scores are: MSE=0, FSS'=1, FAR=0, HR=1. The percentile thresholds are computed over the rainfall map at each time step separately.

Enhanced HRES forecast	**MSE [mm/6h]^2**	**FSS'$_{q95}$ [-]**	**FSS' $_{q99}$ [-]**	**FAR $_{q95}$ [-]**	**FAR $_{q99}$ [-]**	**HR $_{q95}$ [-]**	**HR $_{q99}$ [-]**
1 day	28.32	0.76	0.76	0.01	0.08	0.99	0.92
3 day	36.86	0.59	0.70	0.11	0.12	0.90	0.88
5 day	38.78	0.52	0.64	0.13	0.29	0.87	0.71

2.1.6 Next steps

Droughts:

For the ZW AI-enhanced CS for droughts, the results of this deliverable show that our simple data-driven forecasting model (NIPA) of seasonal streamflow in the Zambezi outperforms global scale hydrological model-based seasonal forecasts, according to all forecast attributes considered (bias, correlation, and relative variability). As next steps, we will:

- test further improvements to the NIPA enhanced seasonal streamflow forecasts, considering a larger sample of predictors and more refined data-driven model options (than the current linear regression);
- use the enhanced seasonal forecasts of inflows to Kariba dam to inform the lake management and assess the added value of our enhanced forecasts with respect to the benchmark and no-forecast baseline.

The outputs of these next steps will be reported in Deliverable D7.3 (AI-enhanced CS for local decision-making).

TCs and floods:

For TC and floods, the AI enhancements presented in this deliverable showed a large improvement with respect to the benchmark and the value for early warning for extreme TC rainfall forecasts. Further work should study the improved rainfall forecasts as input of a hydro-dynamic model to produce enhanced TC-related flood (riverine and pluvial flood) forecasts. An improvement is to be expected given the known importance of accurate rainfall predictions as primary driver of flooding. However, this work goes beyond the planned outputs of the project and further results on this are not expected in D7.3.

2.2 Douro

2.2.1 Introduction

The Spanish part of the Douro River Basin (Douro RB) (Figure 2.12) constitutes one of the CLINT semi-arid climate change hotspots, and the extreme event of interest is droughts. The specific region within the Spanish Douro RB is the Orbigo System, which is one of the operational water resources management areas in the catchment. The prevailing water use in the Orbigo System is agriculture, which constitutes 90% of the total water demand (DRBA, 2021). Other water uses are domestic and industrial, as well as environmental water needs. The annual peak in demand occurs during the irrigation season, between May and August. Water demands are mostly satisfied with surface water resources regulated by annual reservoirs. The main reservoir in the region is Barrios de Luna, which represents 90% of the total storage capacity in the system. The institutions and users engaged with CLINT are the Douro River Basin Authority (Douro RBA), and Barrios de Luna Reservoir Union, an organization that gathers all the water user associations using water released from Barrios de Luna Reservoir.

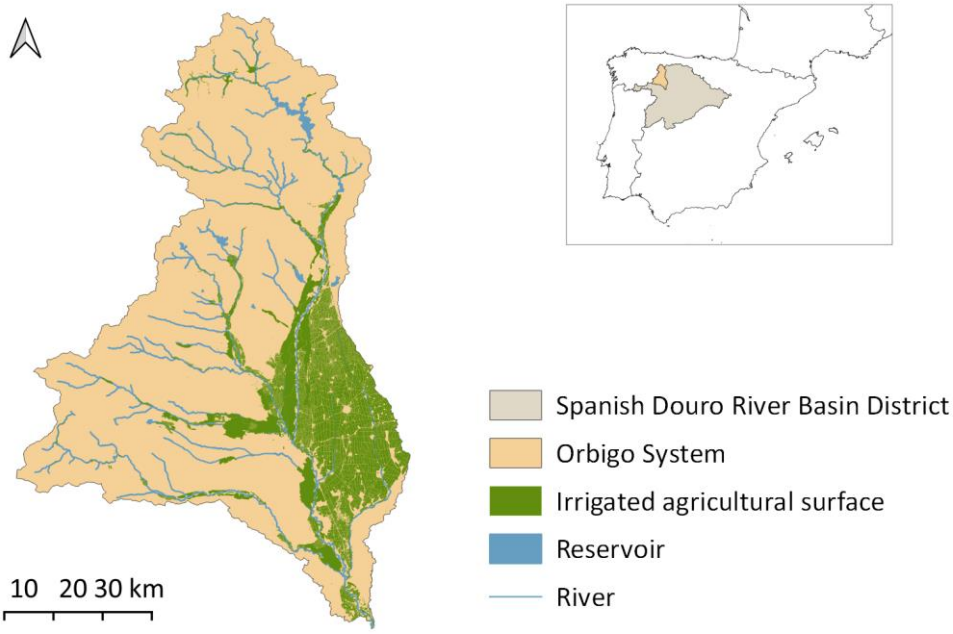

Figure 2.12 The Orbigo System in the Spanish Douro River Basin District in Spain.

The system is exposed and vulnerable to drought. During droughts, the irrigated agricultural sector is exposed to water shortages. Similarly, the minimum environmental flows are reduced during droughts and the degradation of aquatic ecosystem is temporarily tolerated. In addition, reservoir levels may incur extraordinary drawdowns, aggravating the risk of water shortage in the short and long term. Drought planning and management in the Douro RB, including the Orbigo System, are supported by the Douro RB Drought Management Plan (DRBA, 2023). The objectives of this plan are to avoid the degradation of aquatic ecosystems and minimizing the socioeconomic impacts of

droughts. This drought management plan defines a drought indicator system that monitors drought and defines some of the major measures that need to be taken. There are two types of indicators, the prolonged drought indicator and the water shortage indicator. Both are updated monthly. The prolonged drought indicator aims at detecting climate-induced drought episodes, integrating variables of accumulated precipitation and streamflow over nine and six months, respectively. It is associated to decisions related to freshwater ecosystems, such as the temporary tolerance to the degradation of the ecological status of water bodies and the decreased of minimum environmental flows. The water shortage indicator aims at detecting episodes of potential agricultural water curtailments, and it is based on reservoir levels.

In this context, decisions on reservoir management are guided by these two indicators, affecting water releases for environmental and agricultural purposes, as well as reservoir levels at the end of each month and at the end of the irrigation season (August). Moreover, the timing of the spring water meeting, where water allocations are set, is determined based on the water shortage indicator. Usually, this meeting occurs around early March, although it may be postponed to April in case of the water shortage scenario, because water managers and users prefer to delay decisions, hoping the drought situation will improve, rather than making early calls in March.

From the interviews (D7.1), users expressed interest in using S2S and seasonal forecasts to support drought management. Currently, agricultural water users and the River Basin Authority in the Douro RBA do not use any specific CS or seasonal forecast to support S2S to seasonal decision-making, even though several products are available to them on the national level (see Deliverable 7.1). Their reasons include:

(i) forecast quality is considered insufficient for users, particularly defined as uncertain; (ii) forecast information should be timely, this is to say, available at the moment they are making decisions (sporadic or irregular information cannot be integrated into the decision-making process); (iii) the information should ideally be integrated into their decision-support systems (i.e. impact models and drought indicators) and transformed into variables that are easily interpreted (e.g. reservoir levels).

Following these considerations, and the decision-making process summarized before, two objectives are established:

- Evaluating the improvement in forecast quality of the bias-corrected forecasts, and AI-enhanced forecasts, following a user-based verification approach. Several metrics are used to assess forecast reliability, accuracy, discrimination, and skill. These metrics are adapted to user needs in terms of temporal and spatial resolutions of interest, probability thresholds, variables included in the drought indicators, etc.
- Assess the potential added value for drought risk management of (AI)-enhanced forecast information. The added value is going to be measured according to the following criteria: (i) maintained minimum environmental flow during droughts; (ii) avoided unnecessary agricultural water curtailments during the irrigation season; (iii) prevented extraordinarily low reservoir levels at the end of the season; (iv) avoided postponed decisions for agricultural water allocations in March.

2.2.2 Analysis chain for AI-enhanced CS potential added value

General approach

The quality enhancement of the (AI)-enhanced forecast information is assessed through a user-based verification comparison of the AI-enhanced and the raw forecasts. Similarly, the assessment of the potential added value consists in comparing the outcomes of the water allocation strategy in the Orbigo System with and without forecast information and according to the criteria introduced in the second objective of the Introduction (Section 2.2.1). The workflow in Figure 2.13 depicts the main impact modelling steps (in orange), forecast enhancements methods (in green), and aspects to be evaluated (in blue) in this study. There is also information about the benchmark data. Below, the different elements of the workflow are described.

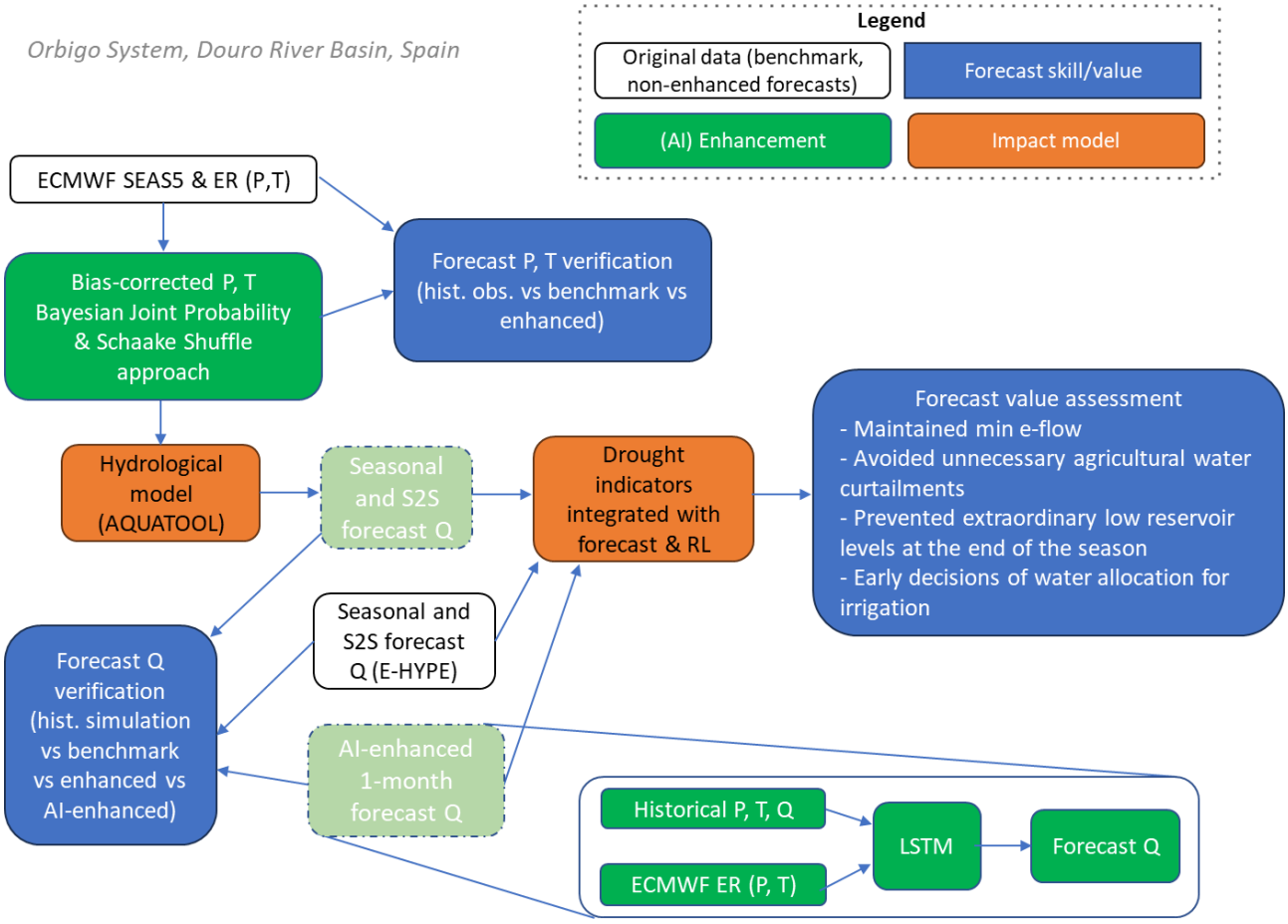

Figure 2.13 Flowchart for assessing the potential added value of AI-enhanced climate service Droughts Douro. (P = Precipitation; T = Temperature; RL = Reservoir Levels; Q = streamflow). (Transparent green refers to AI-enhancement still to be assessed)

Impact model

The two impact models are the EVALHID rainfall-runoff model (Paredes-Arquiola et al., 2012) and the SIMGES water management model (Andreu et al., 1992). Both software components are integrated within the AQUATOOL Decision Support System Shell (Andreu et al., 1996), which is the program and decision support system utilized by the Douro RBA for water resources planning and management. The EVALHID model is an aggregated semi distributed model. Each basin is divided into sub-basins that can be modelled according to different hydrological modelling approaches. The EVALHID hydrological model utilized in the Douro RBA is based on the Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Bergström, 1995) to represent the rainfall-runoff processes. The temporal time step of interest for the users is daily, and hence, it requires daily precipitation and temperature datasets as input. SIMGES is a water management model that simulates the functioning of the operational water resources scheme and allows the computation of the drought indicators (e.g. regulation and storage, intake, transport, consumption and use) (Andreu et al., 1992). Different operating rules can be defined to include priorities on water demands, including environmental flows.

Both the EVALHID and the SIMGES models used in this work has been provided by the Douro RBA, and they are being adapted to the requirements of this study. More particularly, these impact models will be simplified to represent only the elements in the Orbigo System that are relevant to assess the potential added value of the forecast, and they will be run in forecasting mode. In order to represent the measures, a decision-making process based on current management strategies (non-forecast informed) will be modelled in SIMGES, as well as a hypothetical decision model informed with forecasts that will be established based on the feedback received from the users during the following year of the project. The adaptation of the impact models and development of the conceptual decision-making processes are ongoing, and will be described in detail, alongside with the final results of the value assessment, in Deliverable D7.3.

Data and benchmark

Historical precipitation and temperature are derived from SPAIN02 (Herrera et. al., 2015), a 20 km daily rainfall analysis from the AEMET, the Spanish Meteorological Agency. Catchment precipitation and temperature are derived through a process of area-weighted averaging and are provided by the Douro RBA. The historical observations of streamflow are the simulated by the EVALHID model, and also the data series from the archives of the Ministry for Ecological Transition and Demographic Challenge (MITECO in Spanish).

Regarding forecast data, the 5th generation of the seasonal forecast systems (SEAS5) and the extended-range (ER) forecast from the European Centre for Medium Weather Forecast (ECMWF) are used as forcing data for both the E-HYPE model and the EVALHID model to generate the seasonal and sub-seasonal ensemble forecast of streamflow. In order to bias-correct the S2S and seasonal meteorological data for E-HYPE, the SEAS51 and ER datasets have been previously bias-adjusted through the Quantile Mapping technique using as reference the historical natural streamflow simulated by E-HYPE. To bias correct the forcing data for the EVALHID model, a Bayesian Joint Probability model and Schake Shuffle approach have been implemented. This part constitutes one

of the enhancements implemented in this case study, together with the generation of streamflow with an AI-based method.

Forecast enhancements

Two enhancements are implemented in the Douro case study to improve hydrometeorological forecast quality and value to support decision-making in the Orbigo System. The first one is the Bayesian Joint Probability (BJP) modelling and Schaake Shuffle (BJP-SS) approach, which post-process precipitation and temperature seasonal forecasts while conserving the temporal and spatial correlations across lead times and sub-catchments (Schepen et al., 2018). The BJP-SS approach has been selected for this case study because unlike other post-processing methods such as linear scaling and quantile mapping (e.g. Crochemore et al., 2016), it considers the correlation between forecasts and observations (Zhao et al., 2017), as well as the intrinsic skill of the seasonal climate model. In addition, it performs better at the S2S and seasonal time horizons, particularly for accumulated total precipitation (Schepen et al., 2018).

The second enhancement implemented in this study is the long short-term memory (LSTM) method, aimed at producing inflow forecast in Barrios de Luna Reservoir. The LSTM method is a type of neural network. It is trained with the historical precipitation, temperature and streamflow from the data sets described in the previous section, as well as with the ER sub-seasonal forecast for precipitation and temperature. The target lead time is one month. The LSTM method for streamflow generation in Barrios de Luna reservoir has been developed as part of the task on Machine Learning for Extreme Events forecasting and is described in detail in D2.2 Machine learning algorithms for extreme event forecasts and reconstruction. The generated streamflow will be directly integrated into the drought indicators as one-month accumulated streamflow, and will particularly contribute to the decisions in March (see verification framework in the next section). The results related to forecast quality and added value will be included in Deliverable 7.3.

Since the results presented in this deliverable are focused on the verification of the enhanced meteorological forecast, in the remainder part of the methodology a brief description of the BJP-SS approach and verification framework are provided.

All available precipitation hindcast from SEAS5 1 degree for the period 1981-2014 are post-processed. These re-forecasts start on the 1st of every month and have 25 ensemble members (Johnson et al., 2019). Unfortunately, it was not possible to apply the post-processing to the real time forecast because the historical precipitation data set is only available from 1950 to 2015. The BJP modelling creates a joint probability distribution to characterise the relationship between forecast ensemble means (predictors) and corresponding observations (predictands). The joint distribution is modelled as a bivariate normal distribution after transformation of the marginal distributions. Precipitation data is transformed using the log-sinh method. (Wang et al., 2012). The post-processing is applied at the daily time step, for each forecast initialization date and lead time. The specific steps followed in the post-processing procedure are detail below:

1. The BJP modelling is implemented at the catchment level and daily temporal scale. First seasonal daily precipitation ensemble means are transformed to catchment areas through a

process of area-weighted averaging. Before been transform to the normal space using the log-sinh transformation, both the predictand (observations) and predictor (ensemble mean) are pooled in an 11-day window. This is typically needed in dry climate, due to the predominance of days in the data sets with zero precipitation that prevents data inference. For more details about the 11-day window pooling, refer to Schepen et al. (2018).

2. After data pooling, both the predictand and the predictor data sets are rescaled within the range [0, 5], and zero values are treated as censored to allow the use of a continuous bivariate normal distribution.
3. The log-sinh transformation parameters (α and β) for the predictor and predictand are estimated using the maximum a-posteriori, and the transformation is applied to normalise both data sets according to Equation 2.4.1.

$$f(y) = \beta^{-1} \ln(\alpha + \beta y) \quad (2.4.1)$$

4. Consequently, a predictor (or predictand) x (or y) is transformed to g (or h), and the relationship between g and h is formulated by a bivariate normal distribution:

$$\begin{bmatrix} g \\ h \end{bmatrix} \sim N(\mu, \sum) \quad (2.4.2)$$

$$\mu = \begin{bmatrix} \mu_g \\ \mu_h \end{bmatrix} \quad (2.4.3)$$

$$\sum = \begin{bmatrix} \sigma_g^2 & \rho_{gh}\sigma_h\sigma_g \\ \rho_{gh}\sigma_h\sigma_g & \sigma_h^2 \end{bmatrix} \quad (2.4.4)$$

5. In forecasting mode, the predictor value is transformed using the predictor's log-sinh transformation parameters. The BJP model is conditioned on this new predictor value, and thus a new is sample for each parameter set following equation 2.4.5.

$$h_{new}|g_{new}, \theta \sim N\left(\mu_h + \rho_{gh}\frac{\sigma_h}{\sigma_h}(g_{new} - \mu_g), \sigma_g^2(1 - \rho_{gh}{}^2)\right) \quad (2.4.5)$$

6. Back-transform the ensemble members using the transformation for predictands and re-scale to the original space and set negative values to zero.
7. The forecast ensemble members after the BJP modelling are random sampled and thus, they do not conserve the temporal or spatial correlation. The Schaake Shuffle method is applied to reinstate the correlation, which is key for the successive hydrological modelling. For more detailed information about the Schaake Shuffle process, refer to Clark et al. (2004).

Assessment of the potential enhancement of forecast quality and value

The (AI)-enhanced forecast products are evaluated in terms of their quality and their value to support drought management decisions (blue boxes in the flow chart of Figure 2.13). The quality and skill of the enhanced forecasts are assessed for precipitation, temperature and streamflow,

while the value for decision making is evaluated based on the criteria introduced in the Introduction section.

The verification of forecast quality and skill according to user-oriented forecast variables constitutes one of the three major recent developments on forecast verification (Dorninger et al., 2020). It consists in verifying variables that are tailored to the user's needs, such as warnings, specific thresholds, or adapted to the spatial and temporal resolution relevant for decision making. Accordingly, this study investigates the enhancement in forecast quality and skill for variables that are relevant from the perspective of the indicators used in operational drought management, as well as to specific decisions which are key throughout the decision-making process. In this context, six specific sub-catchments within the Orbigo Catchment (Figure 2.14) are evaluated. The main catchment (306) corresponds to the entire Orbigo River Basin, while the sub-catchments (30601-30605) cover the upstream basin of the stream gauges that are integrated into the drought indicator system (Figure 2.14).

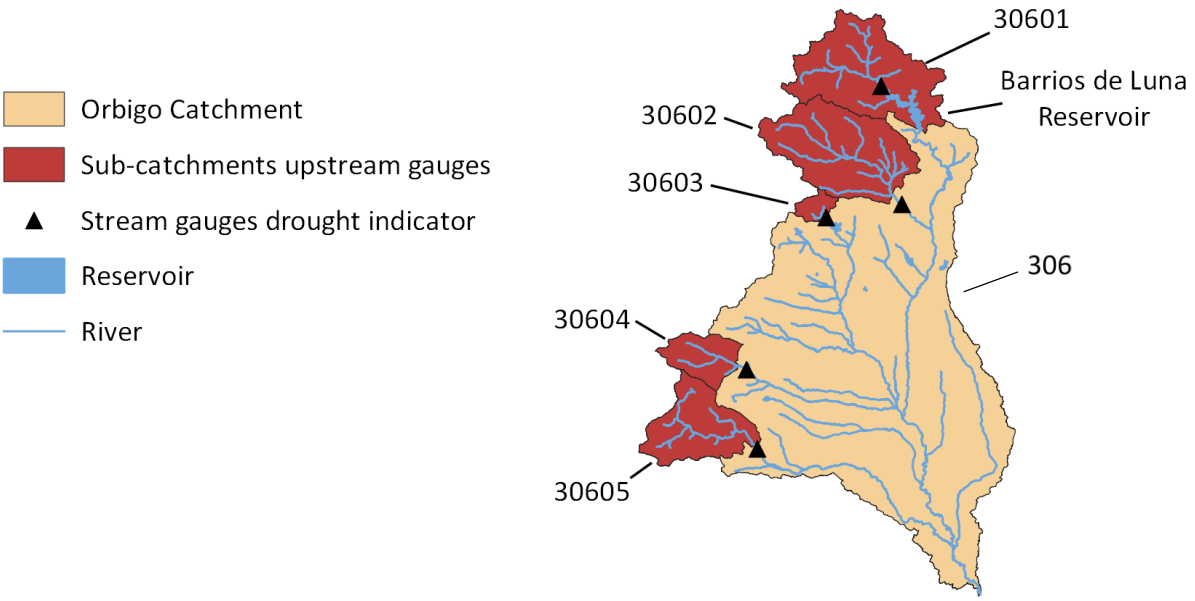

Figure 2.14 Map of the catchments utilized to carry out the bias correction of precipitation forecast in the Orbigo System. The red sub-catchments correspond to the basins upstream of the stream gauges that are integrated into the drought indicator system. The beige catchment corresponds to the Orbigo System catchment. Sub-catchment IDs are indicated.

Forecast quality and skill are investigated for accumulated rainfall as a function of lead time for three forecast initialization dates, namely October, March and September. To assess the enhancement of the bias-corrected forecasts compared to the raw (uncorrected) forecasts, three attributes of forecast quality are used, including overall performance, reliability and skill. To evaluate the overall performance, the Continuous Ranked Probability score (CRPS) is utilized, which is a measure of how good forecasts are in matching observed outcomes. The reliability, which is the statistical consistency between forecasts and observations (Schepen et al., 2018), is measured through the integral transform (PIT) diagram (Gneiting et al., 2007; Laio and Tamea, 2007). The PIT diagram is the cumulative distribution of the PIT values, which are defined by the values of the

predictive distribution function at the observations, computed at each time step. Skill is assessed based on the CRPS skill score, where the climatological forecast is utilized as reference forecast.

2.2.3 Results towards potentially added value AI-enhanced CS

Results show that forecast performance (CRPS), both for the raw and enhanced forecasts, improves at longer lead times when analyzed for accumulated rainfall totals. This is expected given the decrease effect in relative variability for aggregated times. Overall, the bias-corrected forecasts show greater improvement compared to the raw forecasts at longer lead times, beyond the sub-seasonal horizon. There are, however, differences between forecast initialization dates and catchments (Figure 2.15). Regarding spatial aggregation, catchment 306 seems to experience less enhancement than the sub-catchments considered for the drought indicators. For these drought-indicator sub-catchments, the major improvement is experienced in autumn, and summer at longer lead times. This can be observed in the October and August initializations, where the raw forecast has a higher CRPS compared to the bias-corrected forecast for lead times beyond the sub-seasonal. Results for the March initialization are less clear. For some catchments (e.g. 306), bias corrected forecasts perform better than raw forecasts in the sub-seasonal horizon, while for other sub-catchments (e.g. 30604) raw forecasts perform better. There is no improvement in performance for summer months.

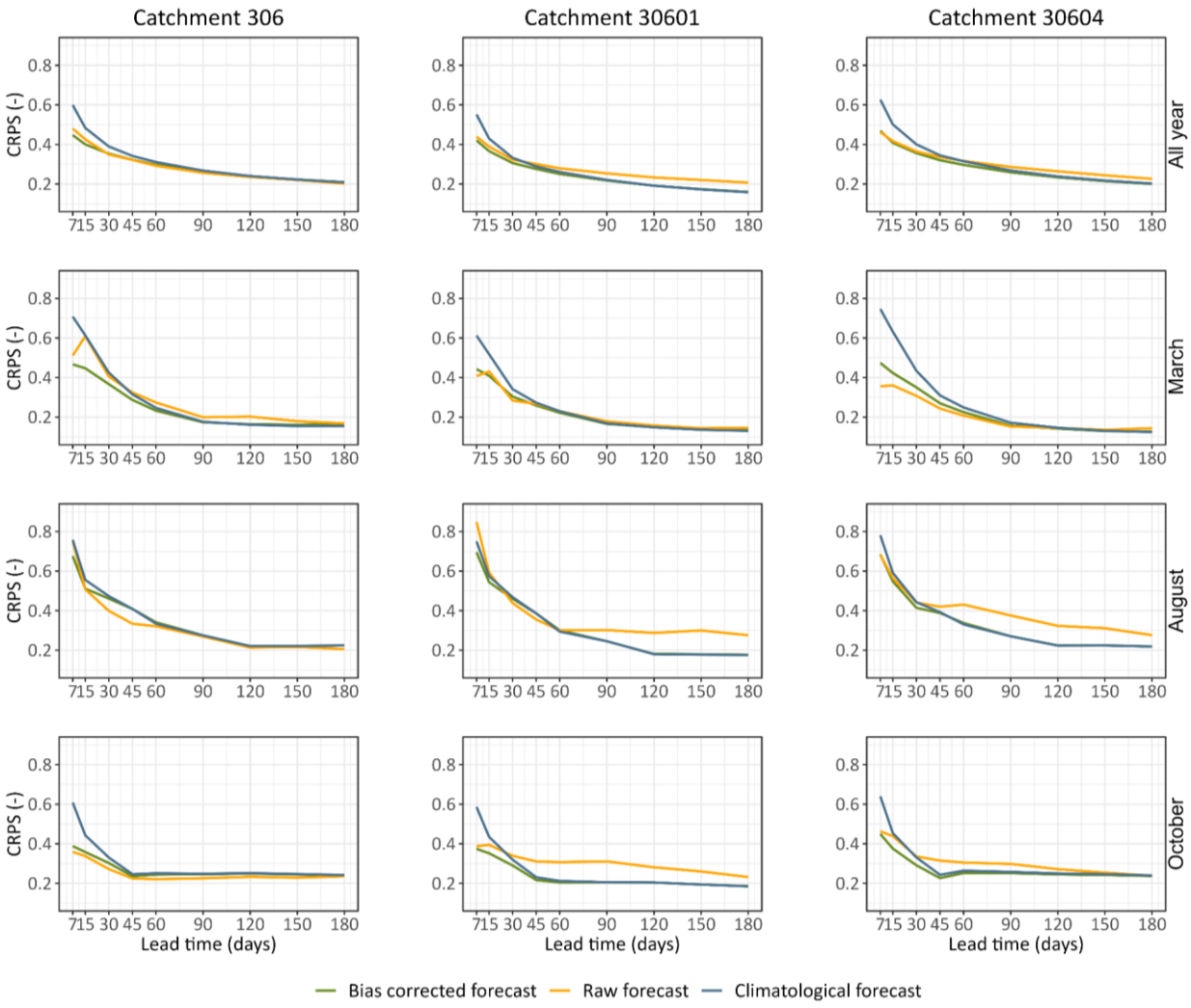

Figure 2.15. CRPS scores for forecast accumulated rainfall. Results for three catchments (columns) and for each initialization date (rows) are presented. CRPS scores for each initialization dates and lead times are standardized (i.e. divided by the observations mean) to eliminate the effect of magnitude differences in accumulated totals across lead times.

Figure 2.16 shows the PIT diagrams for all the catchments for the initialization dates of interest. For each forecast initialization date, lead times 30 and 120 days are shown. Overall, the bias corrected forecasts show a considerably better reliability than the raw forecasts for all initialization dates and lead times. This can be concluded from the shape of the bias-corrected PIT values, which are more parallel to the diagonal and lie closer. Despite the improvement in reliability, the bias-corrected forecasts tend to over-predict accumulated precipitation, especially for longer lead time (120 lead time row in Figure 2.16), where the PIT values lie further above the diagonal.

In addition, the bias corrected forecast is able to correct the jumps that can be observed in the raw forecast PIT curves a lead time of 30 days. This is an indication of narrowness and over prediction of the raw forecast, as well as a potential sign of difficulty of the system to forecast low precipitation (Crochemore et al., 2016). This tendency is usually lost at longer lead times, as can be seen in the

PIT diagrams for lead time 120 days. Therefore, the bias corrected forecast should be able to predict accumulated precipitation at shorter lead times with a higher reliability.

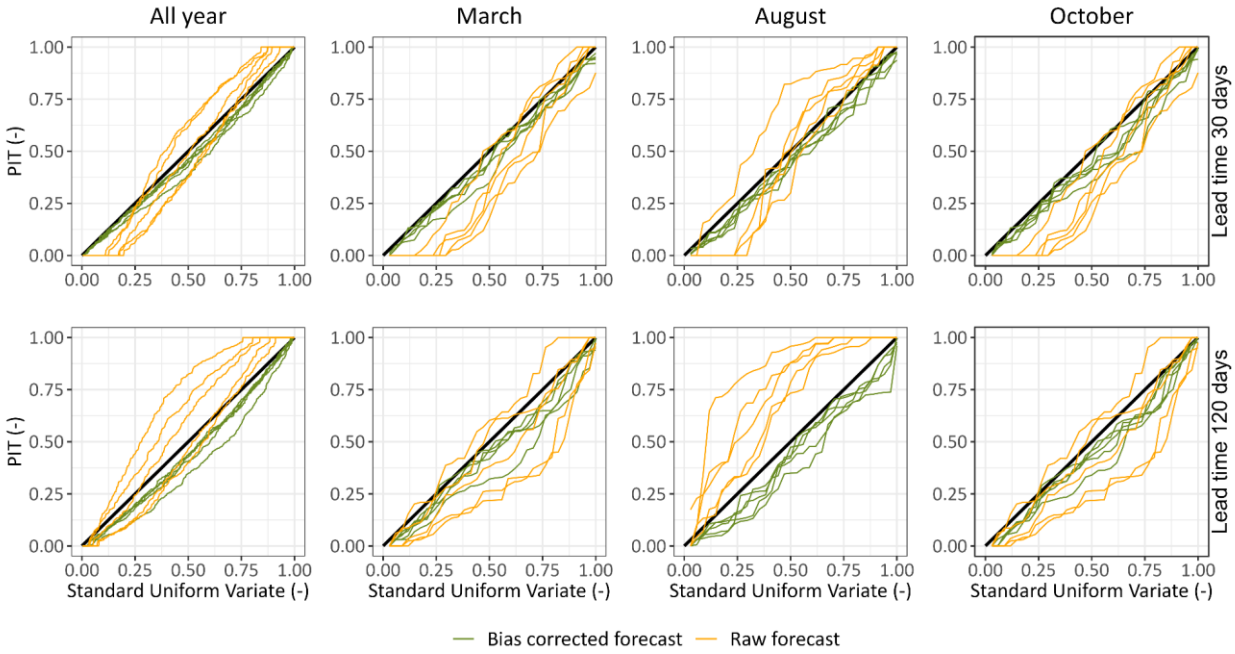

Figure 2.16. PIT diagrams for forecast accumulated rainfall for all catchments. Results for each initialization date (columns) and 30 and 120 lead times (rows) are presented.

Results from the Pearson correlation analysis (Figure 2.17) show that the raw forecast ensemble means present a stronger linear relationship with observations than the bias-corrected ensemble means in all the catchments and initialization dates. Ensemble mean, bias-corrected and raw forecasts correlate relatively well with observations at short lead times (correlation coefficient near or above 0.5). This means that the ensemble means vary similarly to observations in the short lead times. Nevertheless, the correlation is reduced beyond the sub-seasonal lead time horizon. While no distinctions are observed among catchments, there are seasonal differences. On the one hand, the linear correlation shows almost no differences in behaviour between the raw and the bias-corrected forecasts for all year forecast initialization dates and March forecast initialization dates. On the other hand, for August and October initialization dates, the correlation of the bias-corrected forecast decreases faster than the raw forecast with lead times, and becomes negative up to -0.5. The correlation of the climatological forecast present perfect, negative correlation for all seasons, while it starts negative and becomes zero when computed for all lead times. This is expected behaviour associated with the cross-validation process (Barnston and van den Dool, 1993).

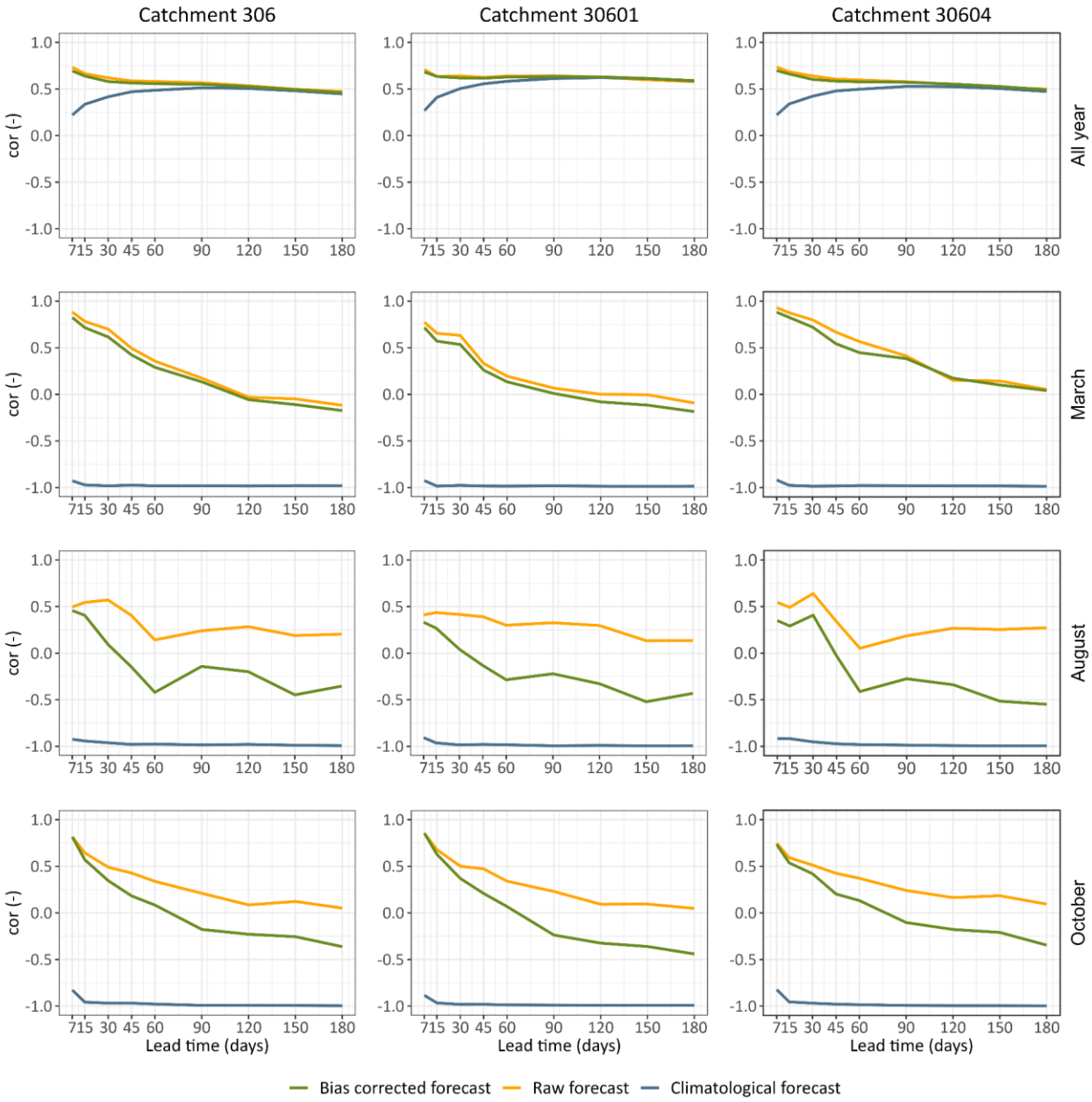

Figure 2.17 Pearson correlation values for accumulated rainfall. Results for three catchments (columns) and for each initialization date (rows) are presented.

Skill scores for accumulated total precipitation are presented in Figure 2.18. The bias-corrected forecast outperforms raw forecast for the majority of the cases, except for catchment 306 in August and October initialization dates, and sub-catchment 30604 for March initialization date and short lead times. Both bias corrected and raw forecast present skills above 30% in the sub-seasonal time horizon and for all initialization dates except summer. Beyond the sub-seasonal lead time, the raw forecast becomes worse than climatology, whereas the bias-corrected forecasts lose skill but never become worse than climatology. There is almost no skill in forecasts initialized in summer (August).

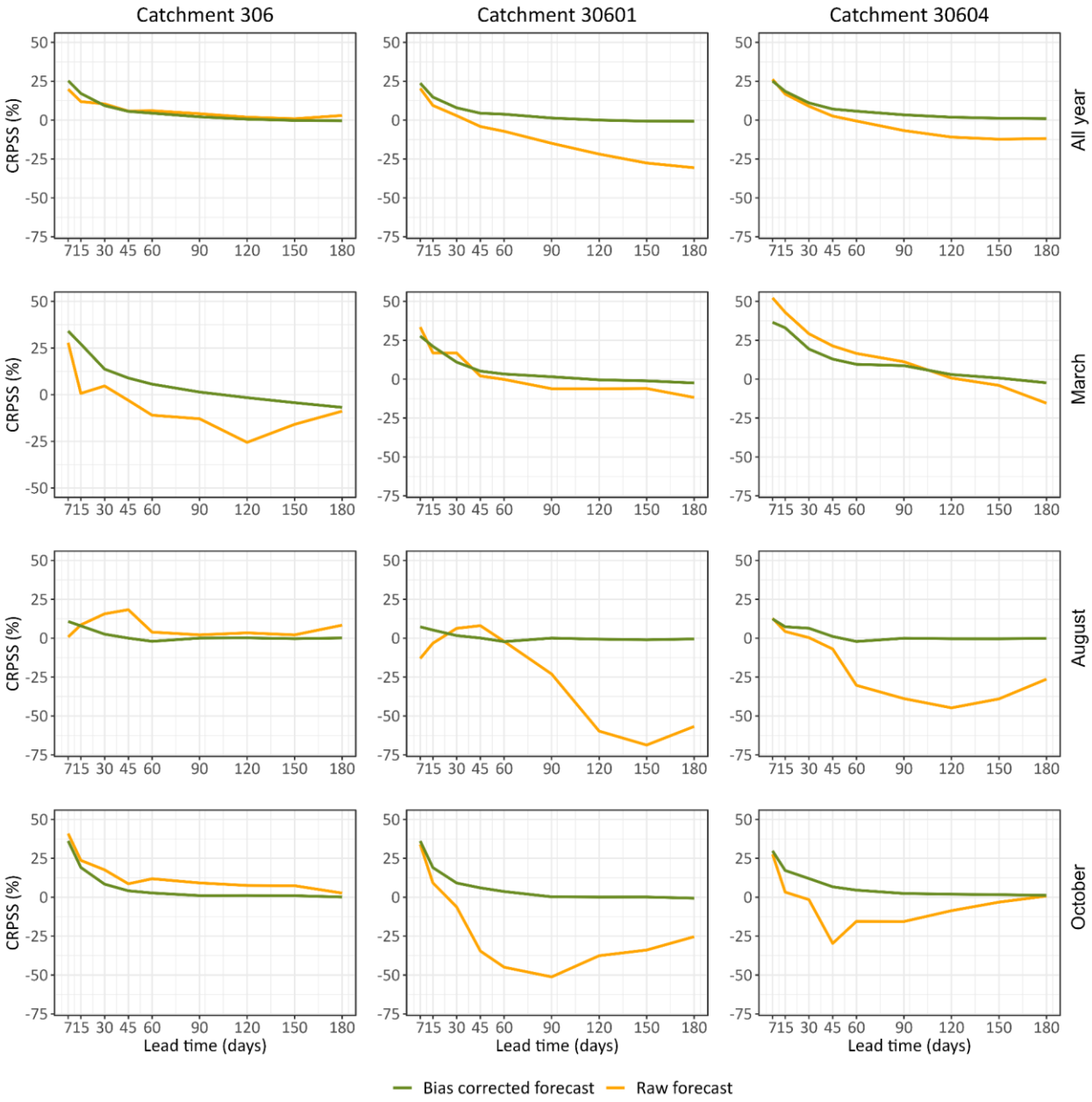

Figure 2.18 CRPS skill scores for accumulated total precipitation. Results for three catchments (columns) and for each initialization date (rows) are presented. Higher CRPS skills scores reflect better forecast performance. The skill of climatology is equal to zero.

2.2.4 Discussion

This work presents the results for the verification of the forecast quality of precipitation data after implementing the BJP-SS approach for bias-correction. Whereas results show minor differences in the bias-correction effectiveness regarding catchment size, there are seasonal differences between the bias and non-bias corrected forecasts. The bias-corrected forecast performs better than the raw forecast in autumn, while the enhancement in spring is less effective. Further research is needed to

understand how the number of ensemble members affects forecast quality, considering that different ensemble sizes might impact verification metrics (Ferro et al., 2008).

2.2.5 Next steps

As mentioned in the previous sections, the next steps will focus on assessing the quality of the forecast streamflow with the AI-enhancement and the assessment of the added value (to be reported in D7.3). Some other model configurations may also be tested to correct the rainfall (D2.2), in order to investigate whether the skill can be further improved in April.

3 Delta climate change hotspots

3.1 Rijnland

3.1.1 Introduction

The Rijnland command area is situated in the western part of the Netherlands, north of the mouth of the river Rhine, and is mostly consisting of low-lying land-reclamation areas below sea level, with a heavily controlled irrigation and drainage surface water system. The Rijnland regional water authority is responsible for water quantity and quality management of this system of inter-connected canals, river reaches, and lakes (D7.1).

This case study focusses on meteorological and hydrological droughts. When evapotranspiration from the area exceeds precipitation over a longer period of time (weeks to months), embankment's stability along the waterways may be at risk and fresh water demand for irrigation for agriculture (crops, flower bulbs, tree nurseries) will be high, increasing salinity pressure through seepage and from ship lock operation. This will require the water authority to intensify manual inspections of embankments and increase flushing of the water system with water from the River Rhine by operation of inlets and pumping stations in the South, and discharge the water through pumping stations in the north and west. Because of higher temperatures and longer day-light in the summer-half year in the NL, evapotranspiration to exceed precipitation occurs usually between April and September and is monitored accordingly as a cumulative precipitation deficit only in that period in the Netherlands.

When such period of precipitation deficit and high local fresh water demand coincides with a hydrological drought in the Rhine basin with low-flow of the Rhine when entering the Netherlands, the drought challenges increase for the Rijnland water authority. They can no longer let in water from the Rhine at their southern boundary (Gouda) because of too high salinity levels caused by sea water intrusion. Rijnland then has to negotiate and coordinate with neighbouring regional water authorities and the national water authority to start an alternative inlet route of Rhine water further upstream, called KWA. The capacity of that alternative fresh water inlet is however less compared to the normal route.

The water authority applies thresholds for increasing alert and warning level with increasing cumulative precipitation deficit, and a single alert threshold for too low Rhine discharge at Lobith

(where the Rhine enters the Netherlands) but varying per calendar month (D7.1). A drought monitoring report is issued monthly, bi-weekly or weekly, depending on the (near-)drought situation, and the maximum lead time of predictions in that report has been two weeks. As the measures of manual inspection of embankments and alternative water supply rout requires planning and coordination, the water authority expressed to CLINT its interest in sub-seasonal lead times up to a month. The potential added benefit thus would be in increased preparedness for an upcoming drought. In addition, the water authority and agricultural stakeholders, may also optimise operation of ship-locks, inlets at Gouda, and on-field storage basins to postpone the moment in which too high surface water salinity levels are reached during the drought.

The impact indicator and added value of AI-enhanced monthly drought predictions for Rijnland is, therefore, increased drought preparedness, which is to be quantified in terms of expected increase in correct drought alerts at lead times beyond two weeks.

3.1.2 Analysis chain for AI-enhanced CS potential added value

The flowchart of Figure 3.1 describes the steps applied for assessing the potential added value of the CLINT AI-enhanced predictions for early warning of droughts in the Rijnland case study. The analysis concerns prediction of the cumulative precipitation deficit from April to September, and then the alert decisions that would be taken based on the precipitation deficit thresholds that are used in the current operational drought event management practice of the Rijnland water authority.

Data and benchmark

Observed cumulative precipitation deficit for the Netherlands, from April to September, as reported by the Royal Netherlands Meteorological Institute (KNMI) from 2000 – 2019 is taken as the ground-truth data set (Source: https://www.knmi.nl/nederland-nu/klimatologie/geografische-overzichten/neerslagtekort_droogte. Last accessed March 2024). Rijnland water authority reports this country-average precipitation deficit as a first indicator in their bi-weekly drought monitor.

Observed precipitation and potential evapotranspiration (Makkink reference evapotranspiration), from the standard 13 monitoring stations used by KNMI for each variable to calculate the precipitation deficit, for the period 1980-1999, are used to create monthly climatology predictions as reference forecast, and for calendar-month specific correction of ECMWF extended range potential evaporation predictions to Makkink reference evapotranspiration. (Source: Royal Netherlands Meteorological Institute (KNMI). Last accessed March 2024)

As benchmark forecasts, re-forecasts of precipitation and potential evaporation from ECMWF extended range (S2S forecasts) issued with the operational forecasts from 13 June 2019 to 15 June 2020 is used (Source: https://www.ecmwf.int/en/forecasts/documentation-and-support/extended-range-forecasts/ecmwf-monthly-forecasting-system). Extended range forecasts are issued every 3 or 4 days, such that the issue date closest to the start of each month was selected and then the 30-day cumulative precipitation (lead time ~1 month) and potential evaporation was used as monthly.

CLINT AI- enhanced forecasts are Extreme Learning Machine calendar-month specific predictions of cumulative precipitation (1-month lead time), from 2000 – 2018. The machine learning method applied is summarised below and described in detail in deliverable D2.2.

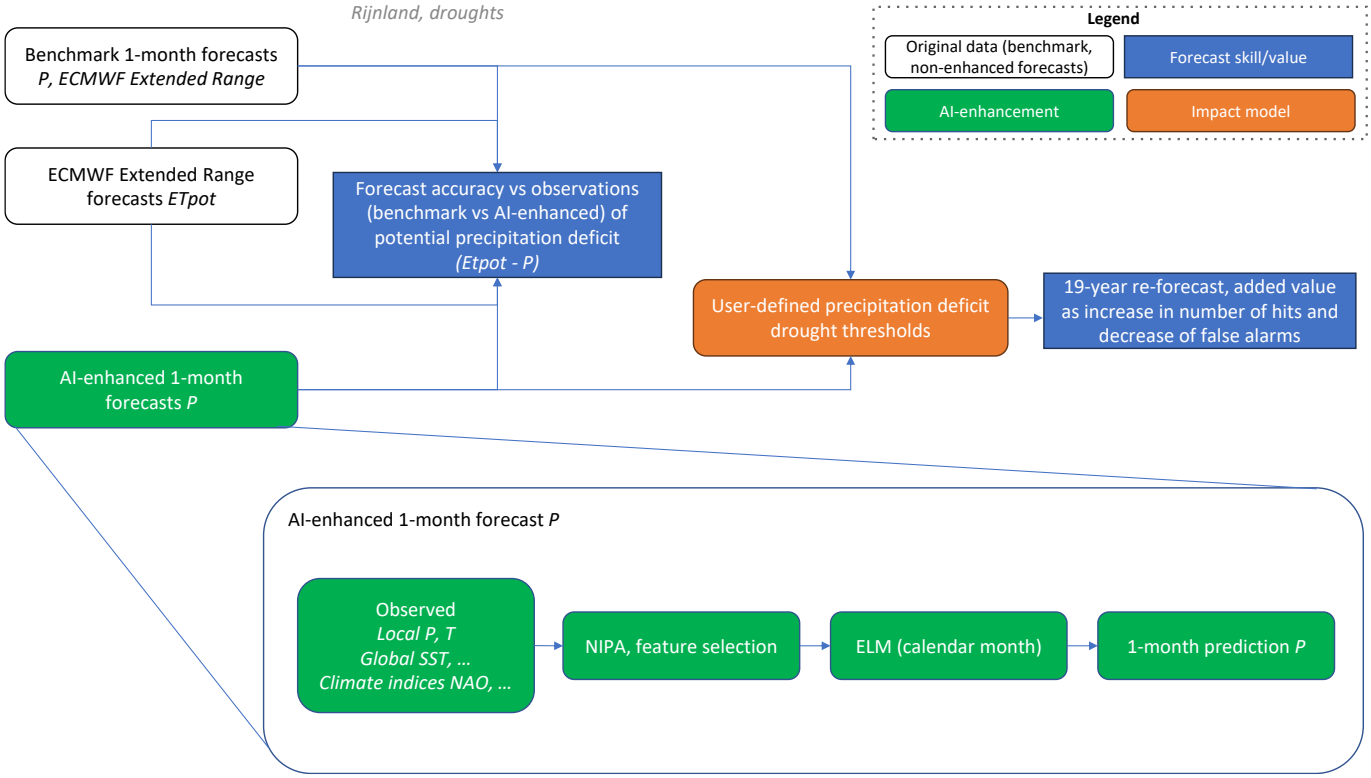

Figure 3.1 Flowchart for assessing potential added value of AI-enhanced climate service precipitation deficit drought alert Rijnland.

AI enhancement

Extreme Learning Machine (Explicit climate information approach)

To develop AI-enhanced forecasts of total precipitation of the upcoming month (green boxes, Figure 3.1), an approach that leverages on climate information and machine learning techniques was designed (Bosso, 2022). More specifically, this approach employs local atmospheric variables, global climate variables and teleconnection patterns, as sub-seasonal lead-times are short enough for the initial atmospheric conditions to still influence weather, but also long enough for the slow varying phenomena, such as ocean temperature variability and teleconnection patterns, to start to play a role in atmospheric circulation. Candidate variables were preselected among those that might influence precipitation processes in Rijnland (e.g. precipitation, specific humidity, wind speed, sea surface temperature, North Atlantic Oscillation, etc.).

The global climate variables (and the teleconnection patterns), then undergo a phase of dimensionality reduction and teleconnection detection, at the end of which only the first Principal Component (PC) of each variable and the relevant teleconnections are retained. This phase is a re-adaptation of the El Niño Phase Analysis (NIPA, Zimmerman et al., 2016) and of the Climate State Intelligence (CSI, Giuliani et al., 2019) framework, which is a statistical framework that was originally

developed to forecast seasonal precipitation based on prior state of atmospheric-oceanic variables. In CLINT, the method has been modified to predict monthly cumulative precipitation and to introduce multiple teleconnection patterns, such as North Atlantic Oscillation (NAO), El Niño Southern Oscillation (ENSO), the East Atlantic oscillation (EA), and the Scandinavian pattern (SCA) (see deliverable D2.2). The PCs in specific teleconnection phases retained were then used, together with local atmospheric variables, to train twelve different Extreme Learning Machine (ELM) models, one for each calendar month. ELM models were chosen because they are known to perform well with small sample-size like the one used in this work. Moreover, because of the small sample size (40 years in total, hence 40 data points per each calendar month-specific model), the maximum number of input features for the models was constrained to 5. All the different combinations of features were used to train each model independently, and the set that provided the best, i.e. lowest, Mean Squared Error (MSE) was selected. The training, validation and testing of the model follows a Leave One Out Cross Validation (LOOCV) approach, as the size of the sample dataset does not allow to adopt the canonical procedures. Further details on the model architecture and performance are given in the deliverable D2.2 *ML algorithms for EE forecast and reconstruction.*

In the next section, the resulting LOOCV monthly ELM precipitation predictions for the period 2000-2018 are compared to benchmark S2S precipitation predictions (blue box forecast accuracy AI-enhanced against benchmark, Figure 3.1). Their potential added value in drought alert decision making for Rijnland water authority is being assessed and discussed (blue box 19-year re-forecast, number of hits and false alarms, Figure 3.1).

Impact model

The impact model is relatively simple for this case study as it concerns the derivation of potential precipitation deficit. The method applied is as defined by the KNMI for the Netherlands, but then with a monthly time step instead of daily, as per equation:

$$Pdef(t) = \sum_{t=4}^{t=9}\left(ETpot(t) - P(t)\right)$$

Pdef = Cumulative potential precipitation deficit [mm]
ETpot = Cumulative Makkink potential evapotranspiration [mm]
P = precipitation [mm]
t = timestep [calendar month, where 4 is referring to April and 9 is referring to September]

The precipitation deficit starts at 0 each year 1st of April. The observed cumulative precipitation deficit is kept at 0 in case negative (precipitation surplus).

The impact model is used to emulate operational monthly prediction of potential precipitation deficit. At the beginning of each month the latest observed cumulative precipitation deficit is assimilated and the predicted cumulative ETpot for the coming month is added and the predicted P for the coming month is subtracted. This results in the 1-month lead time forecast of cumulative precipitation deficit, for each summer month from April to September. This forecast emulation is applied to the years 2000-2018 as test period.

3.1.3 Results towards potentially added value AI-enhanced CS

Deliverable D2.2 already presented the benchmark analysis comparing the AI-enhanced ELM monthly precipitation forecasts and ECMWF extended range forecasts against ERA5 re-analysis precipitation, showing consistently smaller prediction errors (MSE) from the AI-enhanced predictions. In this deliverable, however, analyses have progressed towards assessing the potential added value for the local scale use case of Rijnland. As this use case concerns droughts as indicated by precipitation deficit in the summer months, focus has to be on performance assessment for predicting low precipitation amounts in the months April to September. For example, Figure 3.2 shows the monthly prediction values of April and August for the years 2000-2018. It can be seen that the extreme learning machine predictions (ML-ELM) in most years are slightly better, also for low values. For some years, however, ECMWF extended range ensemble mean provided better predictions of precipitation, notably for the year 2003 of both April and August precipitation. Similar results have been found for the months May, June, July, and September.

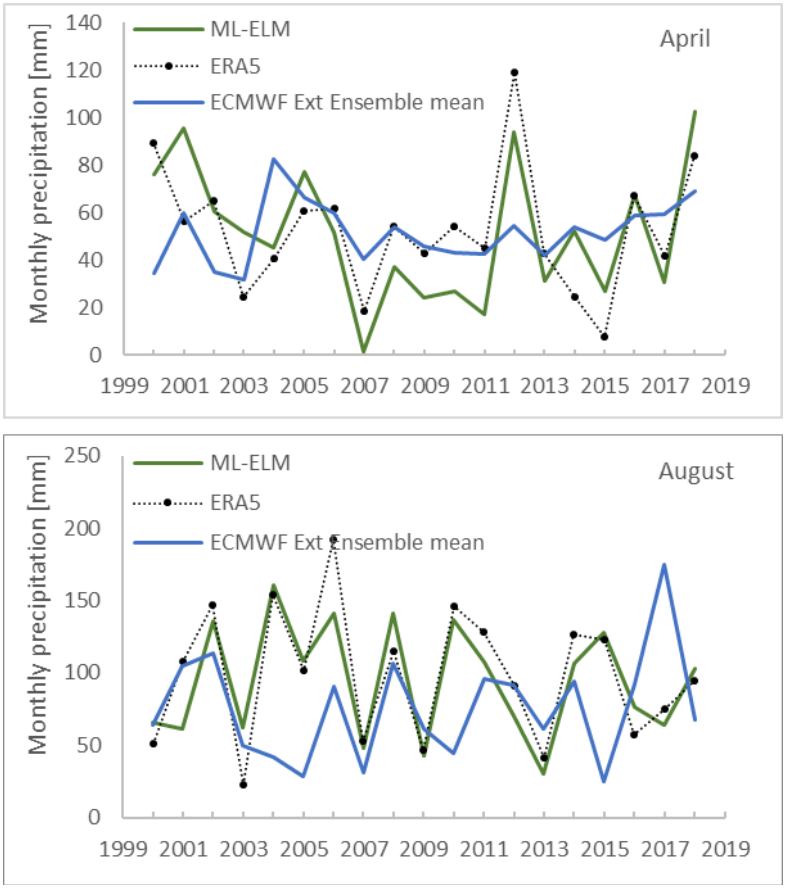

Figure 3.2 Comparison of predictions ECMWF S2S, ELM, and ERA5 cumulative 30-day precipitation for months July and August

The next step, following the flowchart (Figure 3.1), was to analyse the impact these AI-enhanced precipitation predictions would have on 1-month lead time drought alerts as per the operational

procedure of the Rijnland water authority of monitoring cumulative potential precipitation deficit. Figure 3.3 illustrates the comparison for ECMWF S2S ensemble mean, AI-enhanced ELM predictions, and observations for the years 2000, 2003, 2006, and 2018. The horizontal line indicates the pre-alert threshold applied by Rijnland for droughts (Table 3.1). It can be seen that the no-drought year of 2000 was well-predicted by both the ECMWF extended range-based and ELM-based precipitation deficit forecasts, with both staying well below the 125 mm threshold. For the drought years 2003, 2006, and 2018 a mix of results can be seen:

- the 2003 threshold was exceeded by ECMWF prediction but not by ELM;
- the 2006 threshold was exceeded by both the AI-enhanced ELM predictions and the ECMWF benchmark, with the ELM predictions showing a better match with the more extreme observed precipitation deficit that year;
- and for 2018 both predictions exceeded the threshold, this time with the ECMWF matching the maximum exceedance better, but both exceeding the pre-alert threshold of 125 mm a month too late (observed exceedance was beginning of July, while the predictions only exceed in August).

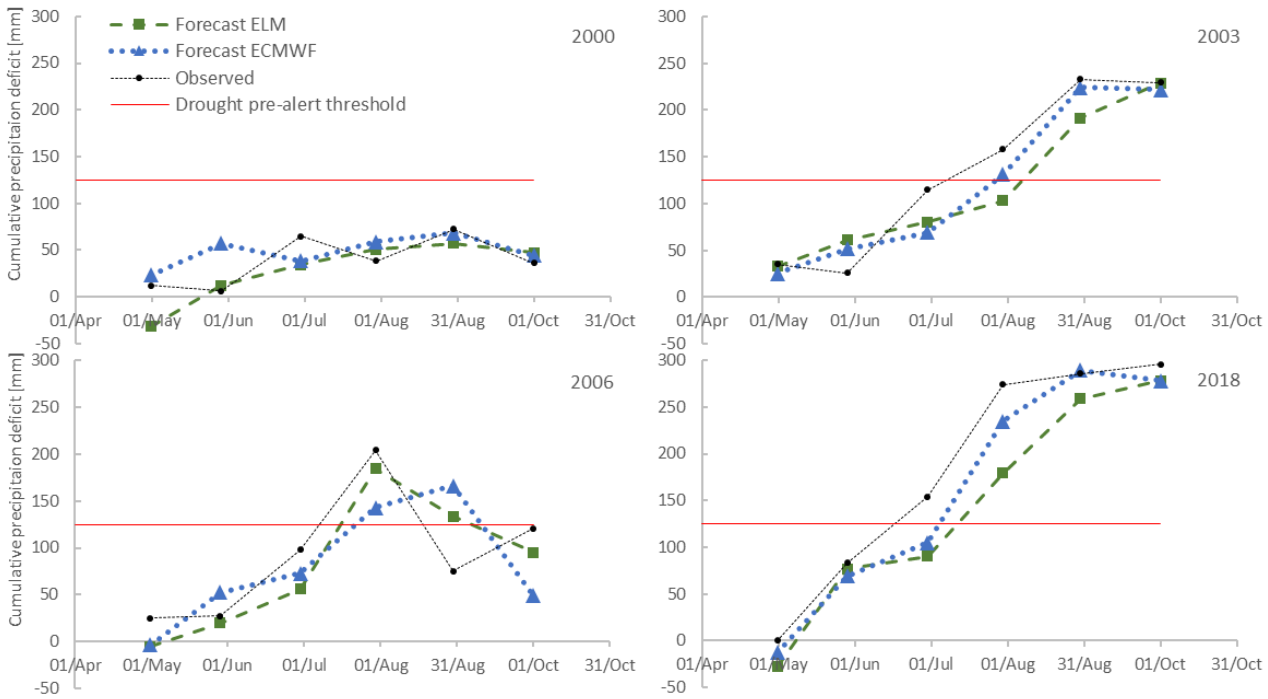

Figure 3.3 Comparison of 1-month lead time ECMWF S2S ensemble predictions and AI-enhanced ELM predictions with observed cumulative precipitation deficit for the Rijnland case study.

To come to a more comprehensive evaluation of the potential added value, threshold exceedances are assessed for all four thresholds operated by Rijnland, and for all years between 2000 and 2018. Rijnland operates four early drought alert thresholds for cumulative precipitation deficit as presented in Table 3.1.

Table 3.1 Cumulative potential precipitation deficit thresholds as used in operational practice for drought awareness and increasing alert-level by the Rijnland water authority

Alert level	**Cumulative potential precipitation deficit [mm]**
First alert level – pre-alert for drought awareness	125
Second alert level (potential need for stability check most drought-prone embankments)	150
Third alert level (potential need for stability check drought-prone embankments)	175
Fourth alert level (potential need for stability check all embankments)	200

By analysing for each year, false alarms can also be quantified, and verification metrics such as hit rate, false alarm ratio, and Critical Success Index (CSI) derived. This is analysed for climatology predictions as a reference, ECMWF extended range predictions as benchmark, and the ELM predictions as AI-enhanced. The results are presented in Table 3.2 for thresholds 125 and 150 mm as examples, and then for all thresholds for the key added value indicators of correct alerts, 'hits', and wrong alerts 'false alarms' in bar charts of Figure 3.4.

Table 3.2 shows that for the lower 125 mm alert threshold, which is assessed based on monthly predictions of precipitation deficit, the benchmark ECMWF predictions perform slightly better, with one more hit than the ELM predictions. However, neither of them demonstrates added value, as the simple climatology predictions score equally well, with a hit rate of 0.7, and all three have a critical success index of 0.6. For the 150 mm threshold, the AI-enhanced ELM predictions outperform the ECMWF ensemble mean predictions and the climatology predictions, predicting 14 out of the 20 months above the threshold, against 9 and 10 hits for the benchmark and reference forecasts respectively. The AI-enhanced predictions result in less false alarms as compared to the ECMWF and climatology predictions.

Table 3.2 Comparison of the performance of monthly alerts based on climatology, ECMWF extended range ensemble mean, and AI-enhanced ELM predictions of cumulative precipitation deficit for Rijnland 125 and 150 mm drought thresholds

Drought alert level:	**125**	**mm precipitation deficit**		
		1-month lead time forecasts		
Number of observed events	32	Climatology	ECMWF Extended Range	ML-ELM
Hits		22	22	21
Misses		10	10	11
False alerts		8	6	4
Correct no alerts		83	85	87
Hit rate		0.7	0.7	0.6
False alarm ratio		0.3	0.2	0.2
Critical succes index		0.6	0.6	0.6

Drought alert level:	**150**	**mm precipitation deficit**		
		1-month lead time forecasts		
Number of observed events	20	Climatology	ECMWF Extended Range	ML-ELM
Hits		10	9	14
Misses		10	11	6
False alerts		11	7	3
Correct no alerts		92	96	100
Hit rate		0.5	0.5	0.6
False alarm ratio		0.5	0.4	0.2
Critical succes index		0.3	0.3	0.6

It is, however, also important to consider the duration of events, especially when they sometimes exceed the forecast lead time. For droughts this is often the case. Scoring a ‘hit’ after the first alert of the same event has been given already is not relevant when the decision for action is already taken at the first alert. Even more so, when the drought duration is two months or longer, and the lead time is 1-month, in the standard verification approach a ‘hit’ may be scored when the event is already happening and will already be observed. The forecast alert is at such times no longer relevant for decision making. As the potential added value for the Rijnland case study is defined as ‘increase in number of correct drought alerts for drought preparedness’, the contingency table type analysis above is now repeated with hits and missed events conditional to the event not yet taking place. The results for alert thresholds 125 and 150 mm are presented in Table 3.3.

Table 3.3 Comparison of the performance of alerts for the onset of drought based on climatology, ECMWF extended range ensemble mean, and AI-enhanced ELM predictions of cumulative precipitation deficit for Rijnland 125 and 150 mm drought thresholds

Drought alert level: 125 mm precipitation deficit				
Only one 'hit' per drought event		**1-month lead time forecasts**		
Number of observed events	13	Climatology	ECMWF Extended Range	ML-ELM
Hits		4	6	4
Misses		9	7	9
False alerts		8	6	4
Correct no alerts		83	85	87
Hit rate		0.3	0.5	0.2
False alarm ratio		0.7	0.5	0.5
Critical succes index		0.2	0.3	0.2

Drought alert level: 150 mm precipitation deficit				
Only one 'hit' per drought event		**1-month lead time forecasts**		
Number of observed events	10	Climatology	ECMWF Extended Range	ML-ELM
Hits		1	0	4
Misses		9	10	6
False alerts		11	7	3
Correct no alerts		92	96	100
Hit rate		0.1	0.0	0.3
False alarm ratio		0.9	1.0	0.4
Critical succes index		0.0	0.0	0.3

The results show an added value of the benchmark ECMWF forecasts with 2 more hits than the climatology and ELM predictions for the 125 mm, and a clear added value of the ELM predictions for the 150 mm threshold of 4 hits against 0 and 1 hits of the ECMWF and climatology predictions respectively. It also indicates that, out of a total of 10 events, the hit rate for predicting drought onset is not high (0.4).

Figure 3.4, which summarizes correct alerts (hits) and false alarms for all four drought alert thresholds, shows that the benchmark ECMWF forecasts perform better for the 125 mm and 200 mm thresholds, while the AI-enhanced ELM prediction provides more accurate alerts for the 150 mm and 175 mm thresholds. The lower panel of Figure 3.4 also shows that the AI-enhanced ELM predictions result in a reduction of false alarms consistently for all drought alert threshold.

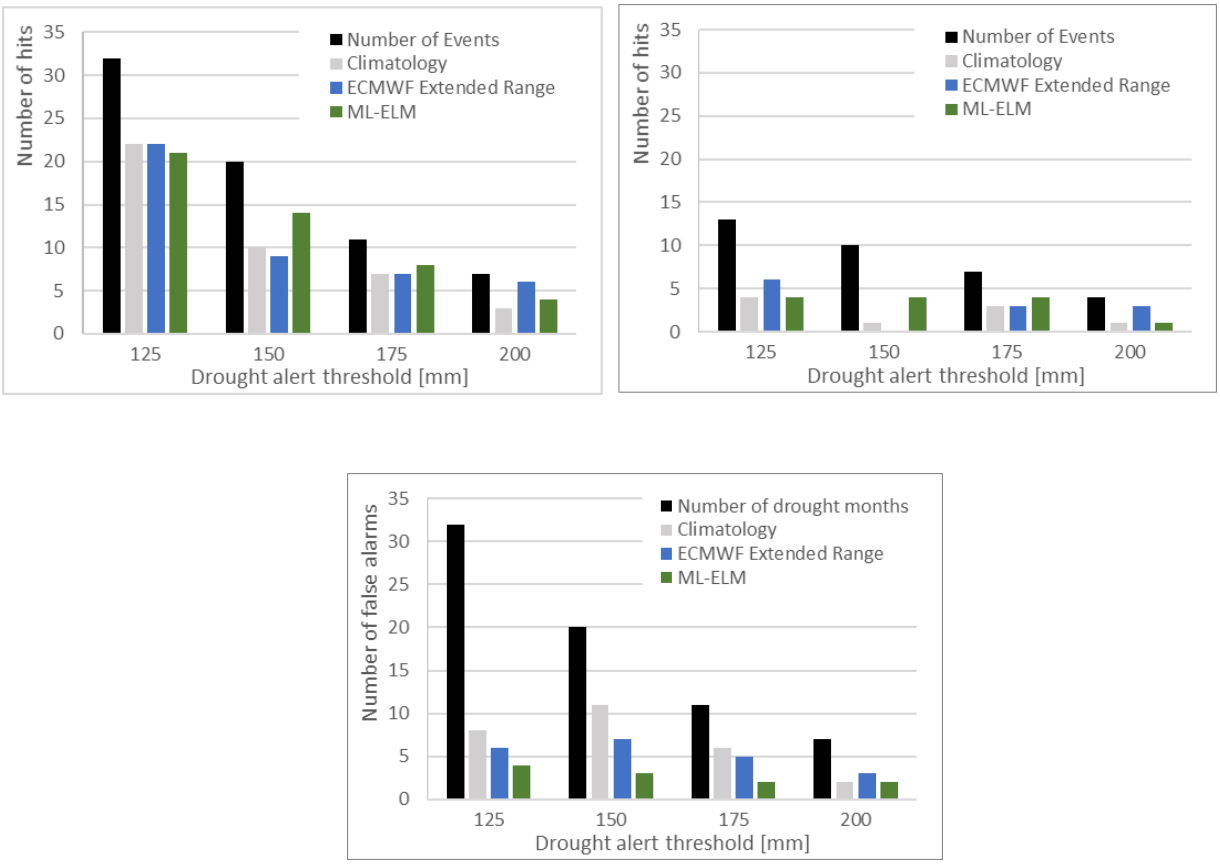

Figure 3.4 Potential added value assessment in terms of correct alerts (monthly hits top-left, and only drought onset hit top-right) and false alarms for the four drought alert levels as used in monitoring operation by Rijnland water authority for cumulative precipitation deficit from April to September, comparing 1-month lead time climatology predictions as reference, with ECMWF ensemble mean predictions as benchmark, and CLINT ELM predictions as AI-enhanced, for the reforecast period from 2000 to 2018.

3.1.4 Discussion

From these decision-based 1-month lead time prediction verification analyses, there are several interesting findings to be highlighted. For the lower alert threshold of a 125 mm precipitation deficit, two-thirds of the months exceeding this threshold, can be predicted a month in advance using basic climatology. This alone could be a reason for the water authority to start preparing for and monitoring such predictions, as they currently lack outlooks beyond two weeks. This holds, however, mainly for assessing whether a current drought will endure for another month or not: predicting the first month at which the lower threshold will be exceeded is not that successful, with a maximum hit rate of 0.4 for ECMWF extended range. On the other hand, in 13 out of the 19 years analysed the precipitation deficit being over the threshold of 125mm, it could also be discussed whether this threshold is perhaps too precautionary, even as a pre-alert for awareness. The AI-enhanced 1-month predictions of precipitation, using key climatic drivers as input to calendar-month specific extreme learning machines, do have potentially added value over the climatology and benchmark ECMWF extended range predictions for the 150 and 175 mm thresholds, ranging

from 1 to 5 more correct drought alert decisions over the reforecast period from 2000-2018 (19 years). The benchmark ECMWF extended range outperforms climatology and the AI-enhanced predictions for the lowest and highest precipitation deficit threshold.

The AI-enhanced predictions consistently perform better, for all thresholds, in terms of reduced number of false alerts. This will, however, often be of minor added value in the current operational practice of the Rijnland water authority because they have indicated that once drought measures have been activated, they are preferably kept active to the end of the summer season, September, to avoid switching on-off measures multiple times. It would need to be analysed how many of the false alarms occur before the event actually starts, or are given in a year that no drought occurs, because such false alerts would preferably be minimised.

3.1.5 Next steps

Overall, the first potential added value of the AI-enhanced 1-month lead time ELM precipitation predictions has been assessed with increased number of correct drought alerts for two of the four thresholds applied by Rijnland, and at the same time consistently resulting in fewer false alerts, as compared to the ECMWF extended range ensemble mean precipitation predictions.

As a next step the ECMWF extended range predictions and the AI-enhanced ELM predictions will be analysed in ensemble mode. The trade-off between hits and false alarms needs to be analysed. For preparedness (planning and preparation of mitigation measures) false alarms may have only limited adverse effects, while if the warnings will also lead to control measures, such as limiting surface water supply and ship-lock operation, false alarms are more damaging.

For the next deliverable (D7.3), also other AI-enhanced prediction methods will be assessed, which will extend the assessment of potential added value for the Rijnland case study over more lead times (weekly up to a month, and monthly up to 6 months), and to the alert thresholds for too low discharge of the Rhine river at Lobith (Deliverable 2.2).

3.2 Aa en Maas

3.2.1 Introduction

The Aa en Maas local case study focuses on a sub-catchment of the Dutch Rhine-Meuse delta. The sub-catchment covers the river Aa and its tributaries and has a size of approximately 1,600 km^2. Benninga et al. (2019) describe the study area in detail. The land use in the area consists of agricultural fields used for animal and crop farming. The regional water authority Aa en Maas manages the sub-catchment. Their daily operations consist of operational management and real-time control of the surface water system, operation of wastewater treatment plants and the maintenance of embankments. The water authority operates a system of weirs and pumping stations to minimize water shortages during dry periods. The water authority uses various information sources for their operational water management, for example up-to-date observations and forecasts of meteorological and hydrological variables, such as precipitation, evapotranspiration, and discharge. In addition, the regional water authority cooperates with the

national water authority Rijkswaterstaat to align the management of the regional water system and the management of the Meuse river.

User definition of extreme event

The climate of the sub-catchment is temperate oceanic and experiences precipitation spread evenly throughout the year with an average amount of 767 mm (at KNMI station Volkel). Temperatures typically vary between 3 C˚ in January and 19 C˚ in July. Therefore, evapotranspiration is strongly seasonal and exceeds rainfall rates in summer periods, leading to a precipitation deficit. In addition, climate change projects show that extreme events will become more pronounced in the area, as emphasized by the recently published KNMI'23 climate projections. Thus, future summers will become drier and future winters will become wetter.

Meteorological droughts (expressed as precipitation deficit) lead to water supply shortages in both the surface water and groundwater systems. Various users in the catchment depend on the availability of sufficient water in summer periods. Examples of users are the agricultural sector, nature conservation agencies, the transport sector, industry, and domestic water use (e.g., drinking water). Droughts due to precipitation deficits have large impacts on these users. For example, farmers irrigate their fields from both groundwater reservoirs and surface water due to the precipitation deficit in summer periods. The water authority can impose irrigation bans to limit the decrease in groundwater levels based on thresholds for groundwater levels. Water shortages thus lead to decreased productivity and financial losses for the users. Also, nature conservations are impacted as droughts are potentially devastating for susceptible nature ecosystems such as the peaty area De Peel.

Decision process

The water authority is constantly monitoring groundwater levels in its management area. They calculate the current drought status by comparing current groundwater observations with long-term average conditions. The information is used for decision-making on both strategic and operational levels, for example:

- Optimize the system of weirs to optimally distribute surface water in the area;
- Determine whether irrigation bans should be imposed.

In addition, the water authority is communicating the drought status of the groundwater system to the end users using an online dashboard, see Figure 3.5.

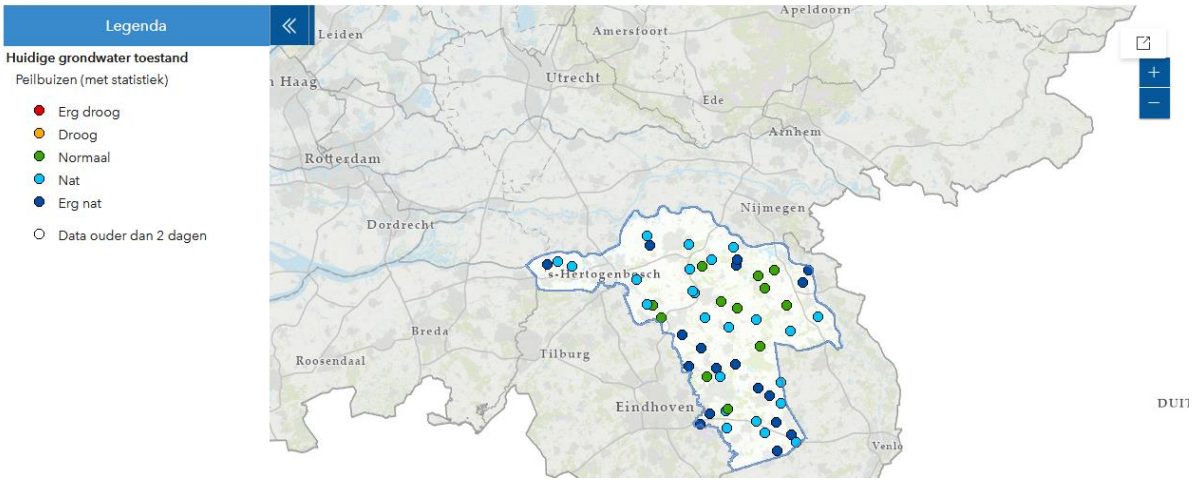

Figure 3.5 Dashboard of regional water authority Aa en Maas to communicate the current drought status of groundwater system using observations. Screenshot from https://www.aaenmaas.nl/actueel/. (accessed 25 March 2024)

User wishes

The water authority has a good understanding of the current drought status of the groundwater system. A valuable addition would be information on middle to long term evolution of the groundwater droughts. Therefore, the goal of this case study the development of an impact indicator that informs the water authority on the expected development of groundwater droughts on sub-seasonal to seasonal time horizons.

Impact indicators for quantifying the value of AI-enhanced CS

The water authority currently does not have a climate service that provides (sub-) seasonal groundwater forecasts. Thus, a first step is to develop a climate service that provide (sub-) seasonal groundwater forecasts using meteorological (sub-) seasonal forecasts and assess its skill. Insight in the (sub-) seasonal development of groundwater drought will be of added value for the water authority. In addition, we will use CLINT-derived AI-enhanced climate information in the bias-correction procedure to enhance the skill of the (sub-) seasonal groundwater forecast.

3.2.2 Analysis chain for AI-enhanced CS potential added value

Figure 3.6 shows the analysis chain for this case study. We will elaborate on the steps:

1. The ECMWF seasonal forecast SEAS5 (P and ET_{ref}) is the meteorological input we use in the impact model (Johnson et al., 2019). A bias correction procedure is needed to ensure the climatology of SEAS5 is consistent with the climatology of local KNMI observations that are used as validation of the impact model.
    a. We perform the bias correction in this stage using local P and ET_{ref} observations from KNMI (Source: Royal Netherlands Meteorological Institute (KNMI). Last accessed March 2024);
    b. We want to enhance the bias correction using CLINT-derived AI-enhanced climate data in the coming months.

2. We develop a data-driven impact model for groundwater level forecasting:
    a. The model is trained using both local groundwater observations as well as local meteorological observations of P and ET_{ref} retrieved from KNMI (Source: Royal Netherlands Meteorological Institute (KNMI). Last accessed March 2024);
    b. The data-driven model uses transfer functions to simulate groundwater dynamics. We perform a feature analysis to select specific transfer functions that are important for explaining groundwater dynamics.
3. Next, we use the bias-corrected seasonal meteorological forecasts as input to the trained impact model to retrieve a six-month forecast of groundwater levels;
4. Finally, we validate the forecasts in a hindcasting procedure using historical groundwater observations to determine the forecast skill and the added value of the climate service. We use the full ensemble of SEAS5, so we can calculate terciles (dry, regular, wet conditions).
5. We will repeat steps 3 and 4 in the coming months using the CLINT-derived AI-enhanced bias-corrected seasonal meteorological forecasts.

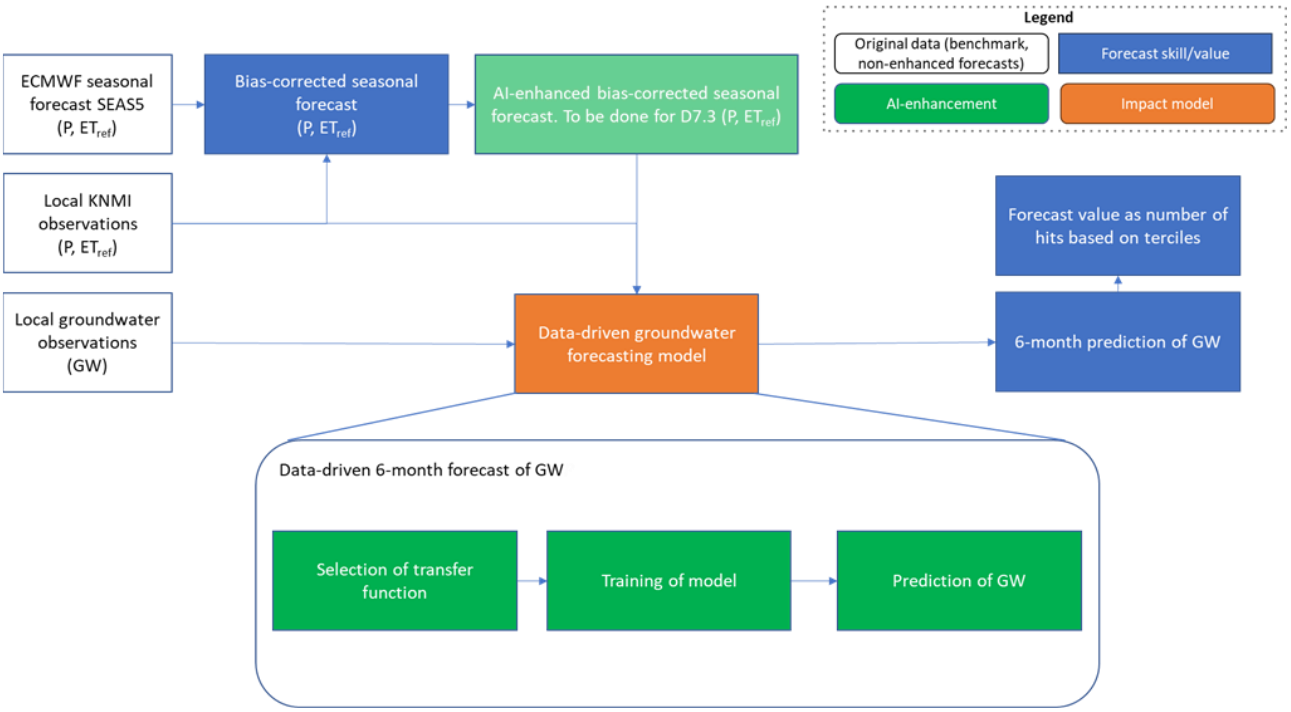

Figure 3.6 Flowchart for assessing potential added value of AI-enhanced climate service for Aa en Maas. (Transparent green refers to AI-enhancement still to be assessed)

3.2.3 Results towards potentially added value AI-enhanced CS

We discuss the results of three aspects in this deliverable: training of the impact model, the result of the bias correction, and the groundwater level forecast skill. The climate service is developed for various groundwater measuring locations. We show the results for one location in this report.

Training of impact model

Figure 3.7 shows the results of training the impact model. The impact model is trained using local groundwater observations which are indicated in the figure by the black dots. The seasonal variation in groundwater level is clearly visible: high groundwater levels during winters which decrease during summer. The results of the trained impact model are visualised using the blue line in the figure. The impact model simulates groundwater levels using a daily time step. The impact model is able to simulate groundwater level dynamics quite well on a daily scale. We resampled the model results to a weekly time step for the remainder of this study. As a general rule of thumb for these kinds of simulations, the model is assumed to be accurate when the model validation in terms of the coefficient of variation exceeds well over 70% (Collenteur et al, 2019; Pezij et al., 2020). The coefficient of variation of this particular model is 78.6%, which means we assume that the model can be used for forecasting.

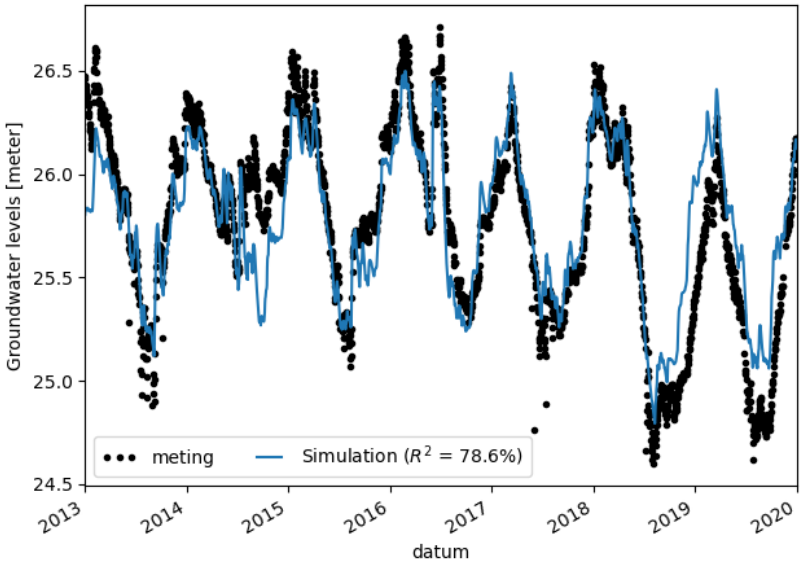

Figure 3.7 Observed groundwater time series (meting) and trained impact model.

Results of bias correction

Figure 3.8 shows the results of the bias correction. The figure shows the empirical cumulative density function of the seasonal forecasts and the reference (local) datasets. The climatology of the precipitation seasonal forecast already matches the climatology of the KNMI reference data quite well. Thus, the bias correction does not have a considerable impact on the precipitation seasonal forecast. The climatology of the evapotranspiration seasonal forecast does not match the climatology of the KNMI reference data well. Therefore, the seasonal forecast is corrected considerably. The result (SEAS5 bias corrected) still shows a deviation to the empirical cumulative density function of the reference dataset. In particular, the higher evaporation values, which can be expected during dry summer periods, are underestimated by the bias corrected seasonal forecast.

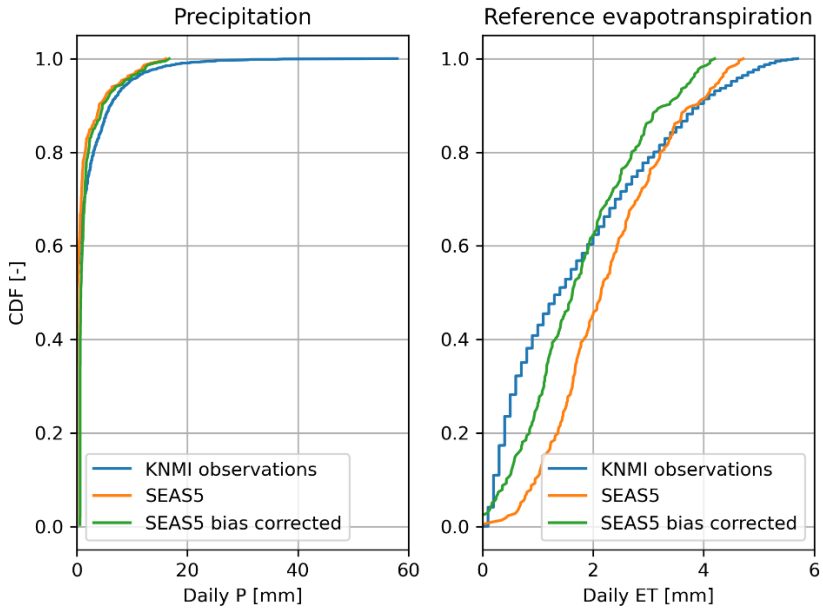

Figure 3.8 Result of the linear scaling bias correction approach on precipitation and reference evapotranspiration forecasts.

Groundwater level forecasts: skill and added value

We used a hindcast approach to assess the forecast skill of the impact model. Figure 3.9 shows the forecast of the model as well as the skill assessment in terms of hit ratio. We use all ensemble members of the bias corrected seasonal meteorological forecast to generate 50 groundwater level seasonal forecasts. Next, we determine a tercile group per ensemble member: whether the conditions per simulated time step represent average conditions or will become more dry or more wet in comparison with long-term climatological conditions.

The top panel of Figure 3.9 visualizes this analysis. Each ensemble member is colour coded. In addition, we count the number of occurrences in each tercile group. The lower panel visualises the observed groundwater conditions during a time step as a colour code referent to one of the three tercile groups. The percentage indicates the hit ratio: the number of occurrences of the forecast in the observed tercile group. In other words, the forecast indicates for 15 June 2020 that 27% of the ensemble simulations indicate a development towards dry groundwater conditions on that date.

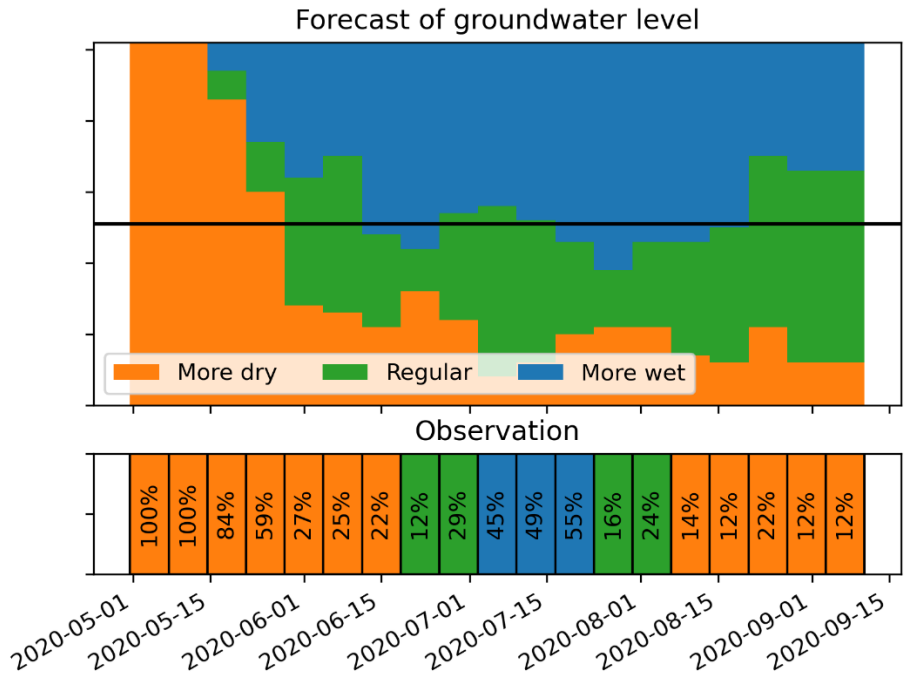

Figure 3.9 Groundwater level forecast in top panel and skill assessment in lower panel in terms of hit ratio. The orange colour indicates more dry conditions in comparison with long-term average climatological conditions, the blue colour indicates more wet conditions, and the green colour represents average conditions.

3.2.4 Discussion

Perspective on the forecast skill

In general, the seasonal forecast of groundwater levels has a forecast skill up to one-two months. We combine two aspects achieve this skill:

- Groundwater systems have a long-term memory in comparison with meteorological variables. The impact model utilizes this aspect by defining the correct initial conditions for the forecast runs;
- The SEAS5 seasonal meteorological forecasts are available for a period up to six months. However, the meteorological forecasting ability in Europe is currently limited and has a skill for quite smaller periods. Therefore, we are considering continuing this project by using the ECMWF Extended Range forecasts, which provide forecasts up to 42 days. This forecast horizon is consistent with the found forecast horizon with skill for the groundwater modelling exercise.

Model

The impact model shows potential in terms of describing the local groundwater system. We have identified the following possible enhancements to improve the results of this study:

- The long-term climatological conditions are now estimated for a small period (8-10 years) due to the lack of long-term groundwater observations. We want to extend the availability by combining long-term historically meteorological observations with the impact model to

generate long-term synthetic groundwater observations, so we can improve our understanding of long-term groundwater conditions.
- The bias correction procedure can be improved for the reference evapotranspiration dataset. CLINT partner SMHI developed an AI-based bias correction method including a European reference dataset. We want to utilize this new innovative method in the coming months.

3.2.5 Next steps

So far, we used a simple bias correction method. The next step is to improve the bias correction using CLINT-generated AI-enhanced climate data We can then repeat steps 3 and 4 of the analysis chain to assess the potential added value of using the AI-enhanced climate data for this climate service. In addition, we want to work together with operational water managers of regional water authority in the summer of 2024 to assess the added value of the CS for their benefits. This will be reported in D7.3.

3.3 Main water system of the Netherlands

3.3.1 Introduction

In this case study, we consider the impact of extratropical transitions (ETT) on flood risk in the Netherlands, in current and future climate. The flood risk along the Dutch coast is primarily influenced by storm surges. Currently, storm surges are mainly caused by storm depressions, whereas extratropical induced storm events may become relevant for flood risk management if the occurrence of extratropical transitions increases in future climate. Potentially, the probability of occurrence of extreme wind speeds and corresponding wind directions changes in future climate. Moreover, the season in which extreme storms occur may change. The wind statistics play an important role in the derivation of design loads for flood defences along the Dutch coast. On the other hand, the storm season determines the optimal timing for maintenance of the flood defences. Within this case study, we investigate the contribution of extratropical-induced storm surges to wind statistics and to the optimal timing of maintenance in the Netherlands in future climate.

The user organisation for this case study is the technical directive for water management of the Ministry of Infrastructure and Water, Rijkswaterstaat (RWS). The resulting insights from this case study can support future flood risk management for the Dutch coast, concerning the design of flood defence measures or changes in maintenance policy, which is of keen interest to RWS.

User-definition of extreme event

An extreme event is defined as an event that poses a significant flood risk for primary flood defences. Each defence has its own safety standard, depending on the area it is protecting. The safety standard is expressed as an allowable probability of flooding, which is related to the probability of occurrence of an extreme event and the probability that a flood defence fails during this extreme event. During this case study, we focus on the first. For the coast, extreme events concern high sea levels and waves, which are mainly caused by wind: on the relatively shallow North Sea, wind storms can cause set-up (storm surge) of the water, especially from the NW direction.

Besides the direction of the storm, also the duration and the timing of the storm with respect to high tide play an important role as so-called driving mechanisms for extreme sea levels.

Decision process for preparedness, adaptation, and event or risk management

The Dutch approach to flood risk adaptation is a so-called 'multi-layer safety' approach, in which three layers are distinguished:

1. Reducing the probability of flooding, for example by constructing flood defences
2. Reducing the consequences of flooding. This concerns spatial planning, for example avoiding buildings in vulnerable areas.
3. Improving disaster management, for example by informing inhabitants about evacuation procedures.

This case study focuses on the first step by providing information that supports a correct derivation of the design load for coastal flood defences in the future, which are based on storm statistics. Moreover, the results will support decision-making processes for maintenance, which is also an important aspect in reducing the probability of flooding.

User wishes and requirements for enhanced climate services

The effect of ETTs in future climate is not incorporated in the current statistical models for the hydraulic loads at the Dutch coast. RWS is interested in methods that provide insights into the contribution of ETTs to future design loads and maintenance planning. It is important that the resulting climate service is consistent and compatible with the current models, in order to ensure convenient applicability within the Dutch design context.

Impact indicators for quantifying the value of AI-enhanced CS

When considering flood risk research in the Netherlands, climate services for coastal flood risk are primarily used for two applications: (1) early warning systems (forecasting of wind and sea levels) to ensure a timely closure of the storm surge barriers; and (2) extreme value analysis of causal factors for floods that are used for design and assessment of flood defences. This case study focuses on the second application, but the insights may be of interest for the first application as well, as new indicators for forecasting. An indicator for the value of AI-enhanced CS is therefore a better insight in causal factors of these extreme events. The problem with extreme value statistics is that it is difficult to determine whether the estimate is right. That is why insight into the causal factors is the most valuable result to obtain.

3.3.2 Analysis chain for AI-enhanced CS potential added value

Currently, the design hydraulic loads for the coast are based on historic data of measurements, which is a period of about 100 years. However, for design of flood defences in the Netherlands, the sea level with a return period of 1,000 years or more is relevant (Kok et al., 2016). It is very uncertain to estimate this extreme sea level from a limited dataset. Therefore, extensive research is being conducted in recent years to use large simulation datasets, such as the wind from SEAS5, for the estimation of extreme events (de Valk and van den Brink, 2023). The different ensemble members can be combined to create a synthetic time series of 9,000 years. This wind data has been used as

input to create an equally long time series of simulated sea levels, using the hydrodynamic model WAQUA-DCSMv5.

The same dataset, consisting of SEAS5 wind and WAQUA sea levels, is used for this case study. Within this dataset, we want to compare ordinary winter storms and storms that were caused by ETTs. The method consists of two research directions: one is focused on comparing the statistics of wind speed and wind direction, and the second is focused on comparing the so-called driving mechanisms for extreme sea levels. The analysis chain is illustrated in Figure 3.10 and the different components are explained below.

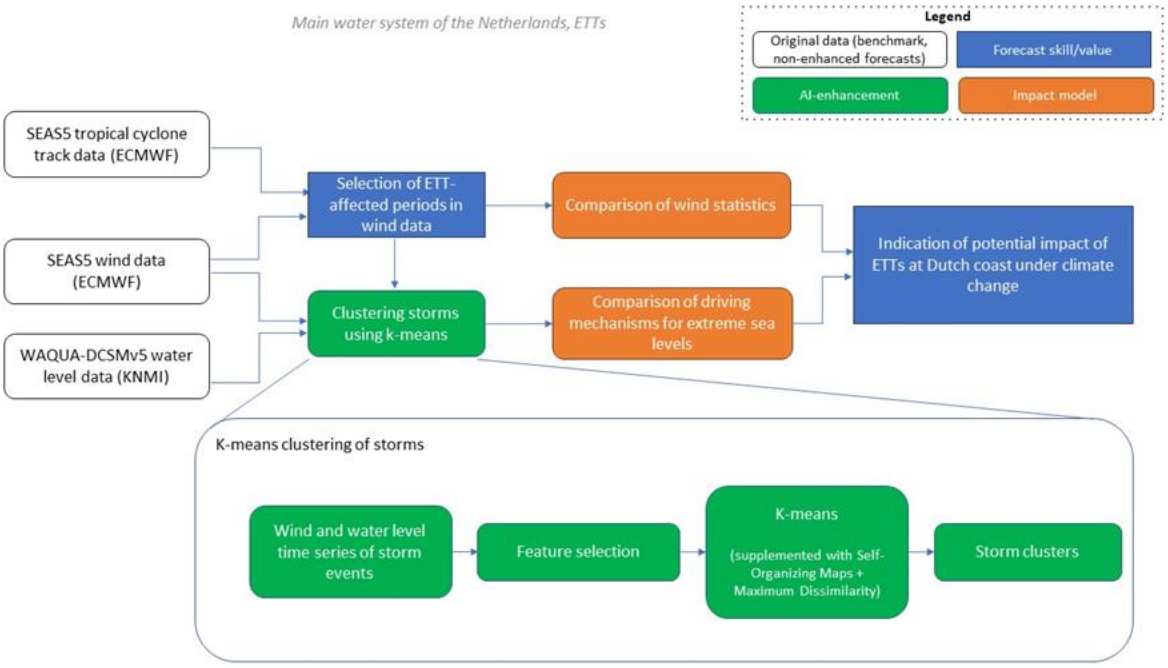

Figure 3.10 Flowchart for assessing the potential impact of ETTs for flood risk management of the coastal water system in the Netherlands.

SEAS5 tropical cyclone track data

For the identification of ETTs within the SEAS5 dataset, we link the SEAS5 time series of wind speed of each ensemble member to the corresponding tropical cyclone (TC) tracks. We consider the maximum sustained wind speed as the measure for TC intensity, because it is more consistent with the observed climatological distribution. As can be seen in Figure 3.11, the TC tracks are only available for the Atlantic Ocean. Only the TCs that reach the area shown in the right map are relevant for the Dutch coast. For these TCs (depicted by blue dots) we derive the time step *t* within the corresponding SEAS5 ensemble at which the TC reaches its shortest distance to the Netherlands. Since the TC track area is restricted to east of the UK, there is a time lag between *t* and the moment in time that the TC reaches the Dutch coast (depicted in red). To account for this time lag, an assumed period of 6 days after *t* is selected from the corresponding SEAS5 ensemble member as the period in which the ETT could have an effect on the wind at the Dutch coast. Several sensitivity

analyses will be performed concerning the selection of the ETT-affected period, e.g. for the 6 days period and for the area in which TCs are selected.

Comparison of wind statistics

The ETT-affected time periods will be compared to reference situations with no ETT storm occurring, in terms of wind statistics. For the reference situation, the wind data within the same time period is selected from a different member of the ensemble for which it is known that no ETT has occurred in a period of 12 days before and after *t*. This is illustrated in Figure 3.12. Then, the selected time periods with and without ETT occurrence can be compared. For this, both the mean and standard deviation of the wind speed are of interest. Besides, we compare the wind direction during maximum wind speed of a storm and the timing of storm occurrence within a year.

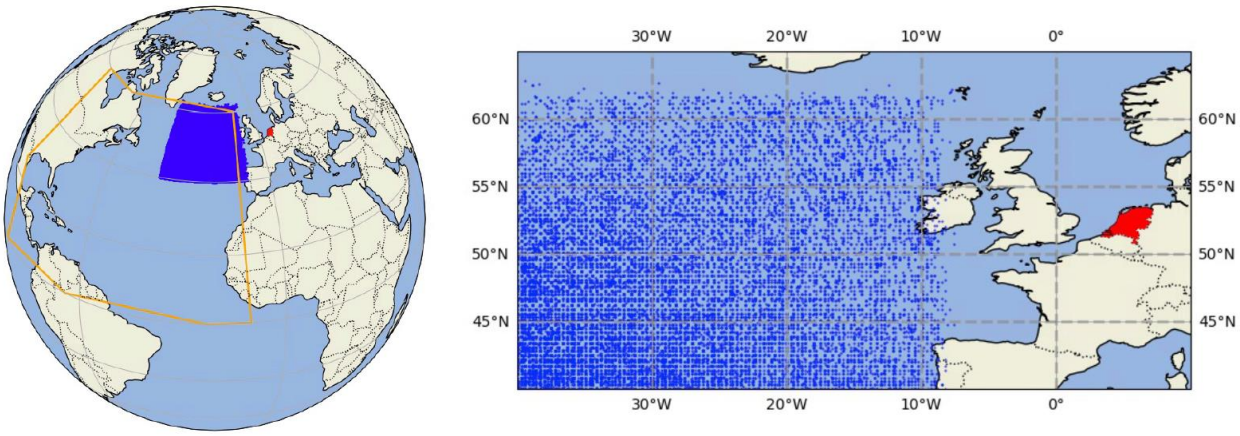

Figure 3.11 Selection area for TC tracks (blue dots) in SEAS5 hindcast data. The red area indicates the Netherlands.

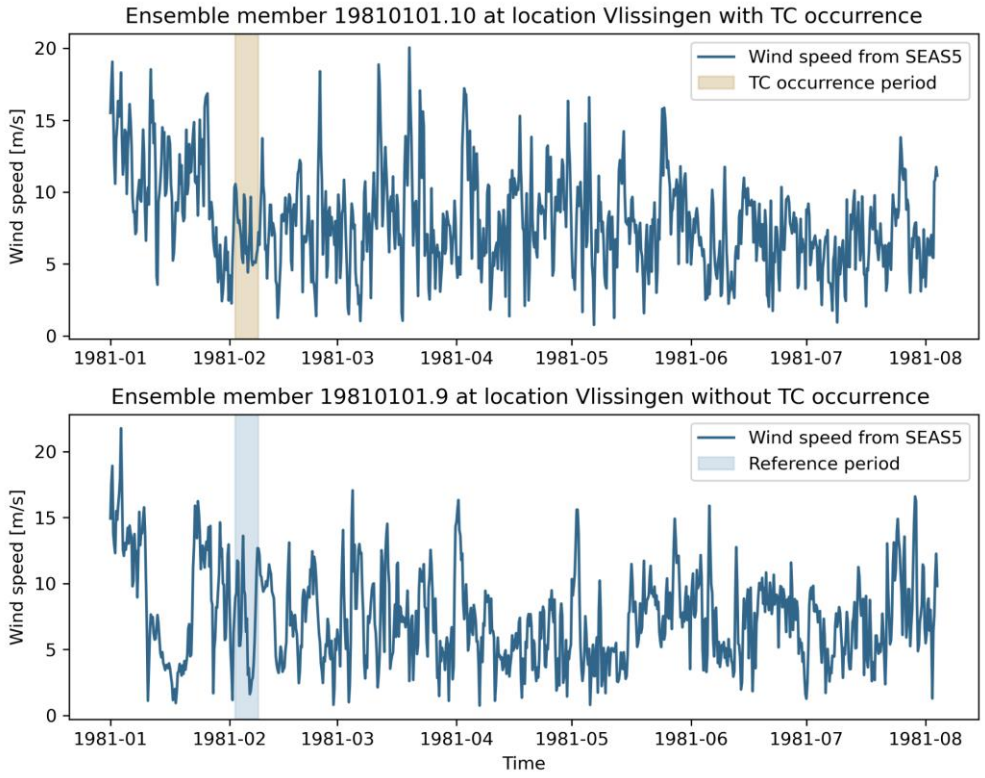

Figure 3.12 Illustrative example of the selection of wind speed time series with (upper graph) and without TC occurrence (lower graph), for location Vlissingen, for one SEAS5 ensemble member (19810101.10). The data has a time step of 6 hours.

Clustering storms using k-means

Beside the statistics of wind speed and wind direction, also other properties of the storm play a role as driving mechanisms of extreme sea levels. More specifically, the complete set of driving mechanisms consist of the maximum wind speed, the wind direction, the maximum surge height, the astronomical tide level, the timing with respect to the next high tide, the duration and the rotation of the storm. These storm properties are used as the features for clustering storms. First, a set of storms is selected from the SEAS5 data, using Peak-Over-Threshold. For the storms within the set, the features are defined. Based on these features, the storms will be clustered using k-means. Self-organizing maps (SOMs) and a maximum dissimilarity algorithm (DMA) will be used to support the k-means approach, in order to obtain representative clusters and gain understanding.

Comparison of driving mechanisms of extreme sea levels

The next step is to identify storms related to ETTs within the different clusters. For this, the same approach as described for the statistics is applied. Consequently, it is possible to indicate the amount of ETTs in each cluster and to investigate their characteristic features, in comparison to the other clusters. If the ETTs are concentrated within one or a few clusters, this could indicate that ETTs show a specific behaviour regarding the driving mechanisms of extreme sea levels. If not, it could indicate that ETTS behave very similar to 'regular' winter storms at the Dutch coast.

Indication of the potential impact of ETTs at the Dutch coast in future climate

The last step will be to describe how ETTs may impact future flood risk management. On the one hand, this consists of conclusions concerning the differences between ETTs and regular storms concerning their impact on extreme sea levels. On the other hand, it is necessary to give an indication on how the occurrence of ETTs at the Dutch coast may change in the future, for this the results from TC tracking algorithms (TRACK or TempestExtremes) from global climate models (GCMs) will be used.

3.3.3 Results towards potentially added value

Regarding the comparison of wind statistics, the results are presented below. Figure 3.13 shows the resulting histograms of some statistical properties of wind speed and wind direction during the selected 6-day periods (as shown in the figure above). In total, the dataset consists of 6.820 periods of TC occurrence (and the same amount for the reference periods). It can be observed from the histograms that the statistical properties of wind time series with and without TC occurrence are almost identical. Next, we compare the timing of TCs and regular winter storms at the North Sea, which is illustrated in Figure 3.14. It can be seen that the peak period for ETTs is in September, which is one month earlier than the regular stormy season.

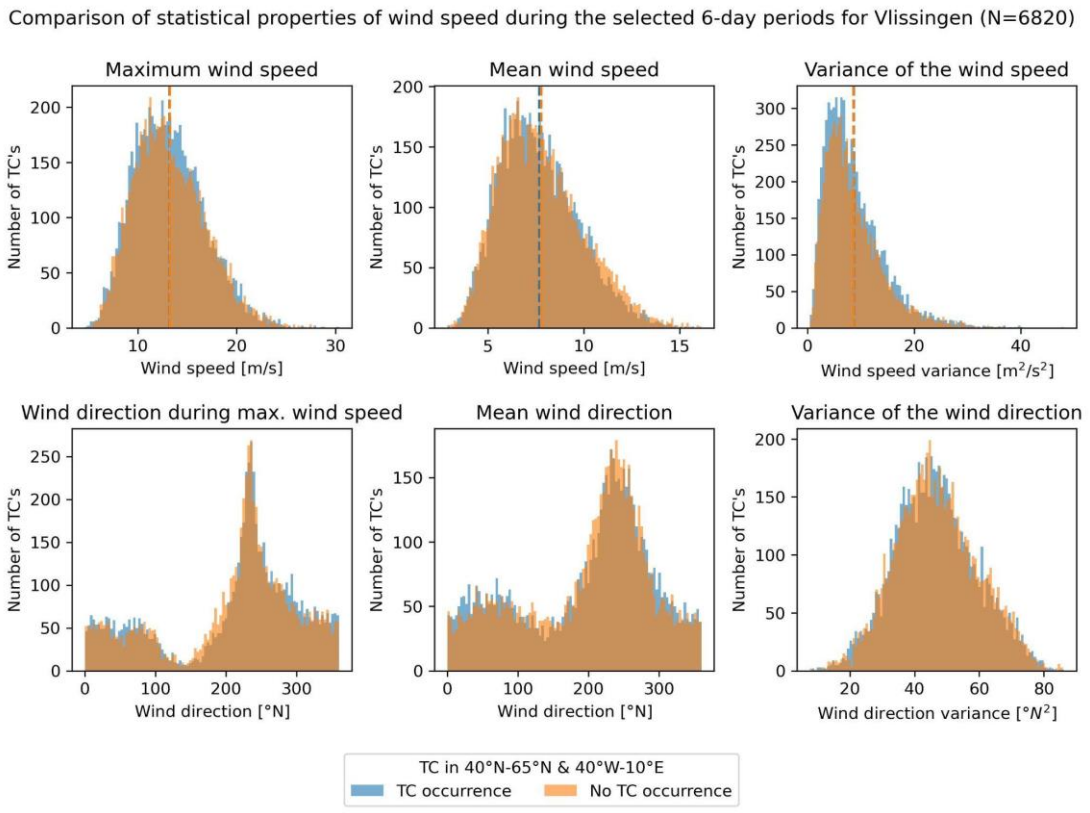

Figure 3.13 Comparison of the statistical properties of wind during periods of ETT occurrence and no ETT occurrence

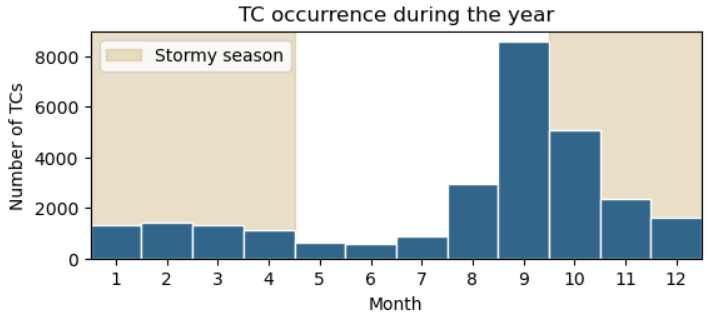

Figure 3.14 Comparison of the months with the highest ETT occurrence (dark blue bins) and the months of the regular stormy season for the North Sea

3.3.4 Discussion

The results imply that the impact of ETTs can be expected to be similar to regular winter storms in the North Sea, with respect to the statistical properties at specific locations along the Dutch coast. However, it is important to notice that it is unknown whether the TC tracks that are shown in Figure 3.11 actually reach the Dutch coast and cause any effect on the wind. Therefore, the results may be distorted in the sense that we falsely include 6-hourly periods to the TC occurrence data set, while these periods are not affected by the occurrence of an ETT. To partly overcome this problem, a next step is to look into a smaller area, e.g. the upper right corner of the rectangle in Figure 3.11. It may be that differences are more significant if we focus on the area with a larger probability that the storms reach the Dutch coast.

The timing of ETTs is somewhat different than the regular stormy season at the Dutch coast. ETTs occur most frequently in September, which is somewhat earlier than the regular stormy season. If the ETT intensity increases in the future, this could imply that the month September could become unsuitable for maintenance.

3.3.5 Next steps

Concerning the statistics, the next steps are to focus on a smaller, more relevant selection of ETTs and to look at the large-scale storm patterns. For now, we only focused on local wind speed and wind direction, while more distinct differences between ETTs and regular storms may be visible in larger scale patterns. Concerning the clustering of storms, the features have been selected and the k-means clustering algorithm (KMA) is in development. The next steps will be to further optimize the KMA in order to result in representative storm clusters.

Subsequently, the combined results from both research directions will be used to formulate conclusions regarding the relevance of ETTs for flood risk management in the Netherlands, and estimations regarding the situation in future climate will be made.

4 Snow climate change hotspots

4.1 Lake Como basin

Lake Como, Italy, is a large, regulated lake with an active storage capacity of more than 200 Mm^3. The hydrological regime in the area is that of sub-alpine regions: inflows are high during spring and autumn due to snow melt and precipitation, respectively, while low inflows are observed during winter and summer. Summer irrigation demand generally exceeds natural water availability, making it essential to regulate the lake to increase the summer release by accumulating water in the lake during the spring season. On the other hand, storing more water in the lake increases the lake level and flood risk, as it limits the buffer capacity of the lake to control floods. The portion of the Adda river downstream of the lake feeds a dense network of irrigation canals, which supply four irrigation districts with a total surface of about 1,400 km^2. The releases from the lake also feed seventeen run-of-river hydroelectric plants and are used as cooling water in two thermal power stations. The total concession along the downstream stretch of the Adda river oscillates between a minimum of 88 m^3/s in March to a maximum of 226 m^3/s between June and July, adding up to 4,418 Mm^3 each year. In the irrigated area served by the lake releases, the most common crop is maize (57.48%), followed by cereals (5.76%), rice (2.41%), melon (0.68%) and soy (0.44%). (Giuliani et al., 2016)

Drought events have increased in recent decades, challenging the reliability of the irrigation supply. For example, two droughts in 2003 and in 2005 led to severe crop failures and exacerbated the conflicts between agriculture and other sectors (Anghileri et al., 2013). Droughts also negatively impact lake navigation along with other recreational and touristic interests that suffer during periods of low lake levels. The occurrence of floods along the lake shores is also a recurrent problem, especially in autumn, when floods are driven by intense rainfall events, but some flooding events may occur in late spring due to intense snow-melt peaks (Denaro et al., 2017). Lastly, extreme temperature is another factor of risk for agriculture in the area, since early or in-season heatwaves and summer persistent anomalously warm nights may jeopardize the yield.

According to the survey's responses reported in Deliverable D7.1, drought and flood risks in Lake Como are primarily managed by Consorzio dell'Adda through the regulation of the lake, which is informed by short-term (3-days) hydrological forecasts that are primarily used for flood preparedness. These forecasts are however not formally integrated into any Decision Support System. No climate services are implemented for predicting or projecting the risk associated with temperature extremes. Further interviews with 460 farmers in the region carried out by Ricart et al. (2024) revealed the most common adaptation measures implemented, which include the reduction in the use of fertilizers, or improving their efficiency in combination with crop diversification and rotation. Moreover, 40% of the interviewed farmers also take measures related to the crop planting process: changing dates, planting earlier, employing different varieties or drought-tolerant crops.

Among the users' requirements for enhanced climate services identified by our survey (see again Deliverable D7.1), in this report we focus on investigating the value of sub-seasonal to seasonal forecasts for informing the lake operations in managing flood and drought risk (Section 4.2), and on the value of climate projections of heatwaves and warm nights for informing farmers adaptation in terms of changing cropping patterns (Section 4.3).

4.1.1 Analysis chain for AI-enhanced CS potential added value for droughts and floods

The methodological workflow to assess the potential added value of AI-enhanced Climate Services for droughts and floods in the Lake Como Basin is based on the assessment of the skill and value of multi-timescale forecasts of the lake inflows combined with a Reinforcement Learning method that jointly learns how to extract the most useful forecast information (e.g., which product to use, which aggregation time) and how to use it for informing the lake operation (Figure 4.1).

The AI-enhancement is therefore associated with the automatic extraction of the most useful information from three diverse forecast products, including the short-term forecasts currently used by the lake operator and the sub-seasonal and seasonal reforecasts produced by Copernicus EFAS, thus grounding our work in the real operational context of the Consorzio dell'Adda. The novelty of our RL method is the idea of making a selection not based on forecast skill, but rather on forecast value. This allows, for example, the potential selection of a less skilful forecast over a longer lead time if this provides more valuable information for drought management than a short-term product characterized by a higher accuracy.

We are also working towards a second AI-enhanced CS that will focus on the production of sub-seasonal to seasonal hydrologic forecasts of the lake inflows using ML models. The results describing the added value of this CS will be described in Deliverable D7.3.

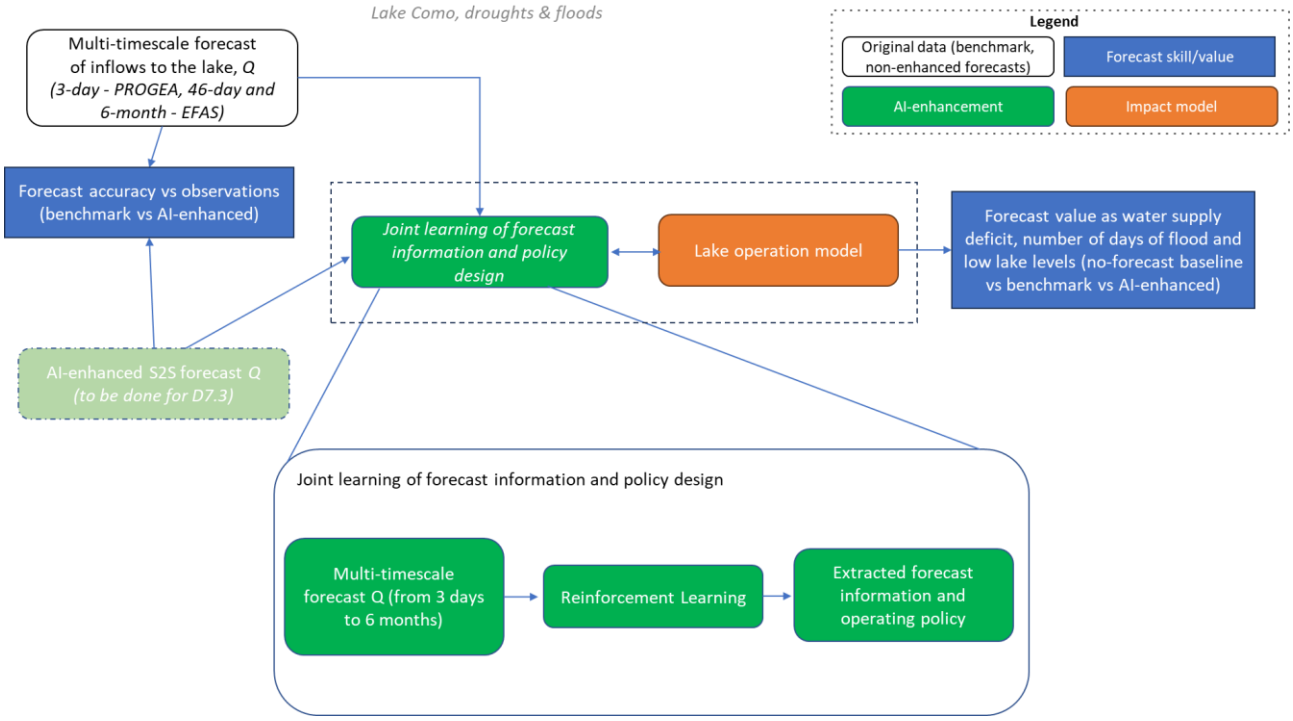

Figure 4.1 Flowchart for assessing the potential added value of AI-enhanced CS Lake Como floods and droughts. (Transparent green refers to AI-enhancement still to be assessed)

Data and benchmark

Daily time series of observed lake levels and releases are available from 1946 (the start of the lake regulation after the dam construction) to 2022. From these observations, the net inflow to the lake is estimated by inverting the mass balance equation of the lake storage. This is used as input for the simulation of the impact model (see next section) and as a reference for the assessment of the forecast accuracy.

The multi-timescale forecasts available to inform the Lake Como operations are the following:

- Short-term deterministic forecasts (PRO), provided by the local company PROGEA and obtained by feeding a locally-calibrated hydrological model with short-term weather forecasts from COSMO[1]. They are single trajectories with an hourly time step and update frequency, a lead time of up to 60 hours, and initially available between 2014 and 2022.
- Sub-seasonal probabilistic re-forecasts (EFRF) produced by Copernicus European Flood Awareness System (EFAS) over the whole European domain by forcing the LISFLOOD (Knijff et al., 2010) hydrological model (uncalibrated for the Lake Como basin) with extended-range ensemble forecasts. These ensemble forecasts comprise 11 members with a 6-hour time step, a twice-weekly update frequency, a 46-day lead time, and availability over 1999-2018 (Barnard et al., 2020). In Appendix B, we report a summary of verification of the different hydrological forecasts that has been conducted focusing on different aspects of probabilistic performance for the EFAS ensemble forecasts.
- Seasonal probabilistic re-forecasts (EFSR) are produced by EFAS, too. Similarly to the sub-seasonal product, these ensemble forecasts are obtained by LISFLOOD but forced here with seasonal meteorological forecasts from the SEAS5 model. Their characteristics are 25 ensemble members, daily time step, issued on the first day of each month, up to 6 months lead time, and availability over 1999-2019 (Wetterhall et al., 2020).

Given the workflow in Figure 4.1, these forecasts are not used as real benchmarks for assessing a new ML-based forecast product but they are rather processed in a hybrid setting by a RL algorithm to extract the most valuable information to advance the lake operation.

Impact Model

The operational model of the lake is focused on reproducing the controlled dynamics of the lake, which is described by the mass balance equation of the lake storage assuming a modeling and decision-making time-step of 24 hours, i.e.

$$s_{t+1} = s_t + q_{t+1} - r_{t+1}$$

where s_t is the lake storage m^3, while q_{t+1} and r_{t+1} are the net inflow (i.e., inflow minus evaporation losses) and the outflow volumes in the time interval [t, t+1), respectively. The release volume is determined by a nonlinear, stochastic function that depends on the release decision (Soncini-Sessa et al., 2007). The actual release might not be equal to the decision due to existing legal and physical constraints on the reservoir level and release, including spills when the reservoir level exceeds the maximum capacity. The lake regulation is determined by a closed-loop operating policy p that

[1] https://www.cosmo-model.org

computes the release decision at each time step as a function of the day of the year, the lake level and, possibly, inflow forecasts.

Historically, two primary competing objectives have driven the lake regulation: (i) flood control to avoid flooding that affects Como and other populated areas on its shoreline, and (ii) water supply to satisfy the demand of downstream agricultural districts and run-of-the-river hydropower plants. Recently, a new objective has also been taken into account; this is related to preventing extremely low lake levels that are detrimental to several users, including navigation, tourism, and the environment. According to previous studies and interactions with the local stakeholders, these objectives are formulated as follows:

- **Flood days**: the average annual number of days when the lake level (h_t) is above the threshold h^{flo}=1.1 m:

$$J^{flo} = \frac{1}{H/T}\sum_{t=0}^{H-1} g_{t+1}^{flo}; \qquad g_{t+1}^{flo} = \begin{cases} 1 \ \ if \ h_{t+1} > h^{flo} \\ 0 \ \ otherwise \end{cases}$$

where H is the simulation horizon (days), and T is the annual period of the year (days).

- **Water supply deficit**: the daily mean deficit considering the water released from the lake (r_{t+1}) and the water demand of the downstream users (w_t):

$$J^{def} = \frac{1}{H}\sum_{t=0}^{H-1} g_{t+1}^{def}; \qquad g_{t+1}^{def} = [\max(w_t - (r_{t+1} - q^{MEF}), 0)]^{\beta_t}$$

where q^{MEF} = 22 m^3/s is the Minimum Environmental Flow constraint ensuring adequate environmental conditions in the Adda River, and β_t is a time-varying exponent that penalizes with different importance the deficit during summer and winter. This parameter was tuned to mimic the decision-making preferences of the operator, with the deficit squared during the summer (1 April to 10 October), while the unitary value is taken during winter.

- **Low lake levels days**: the average annual number of days when the lake level (h_t) is below the threshold h^{low}=-0.2 m:

$$J^{low} = \frac{1}{H/T}\sum_{t=0}^{H-1} g_{t+1}^{low}; \qquad g_{t+1}^{low} = \begin{cases} 1 \ \ if \ h_{t+1} < h^{low} \\ 0 \ \ otherwise \end{cases}$$

The Pareto optimal operating policies are computed by solving a multi-objective optimal control problem (Castelletti et al., 2008) formulated as follows:

$$p^* = arg\min_p \boldsymbol{J}(p) = |J^{flo}, J^{def}, J^{low}|$$

Note that the resolution of this problem does not yield a unique optimal solution but a set of optimal solutions exploring different trade-offs between the three competing objectives. A solution is defined as Pareto optimal (or nondominated) if no other solution gives a better value for one objective without degrading the performance in at least one other objective. The image in the objective space of the Pareto-optimal solutions is the Pareto front.

AI enhancement

In this section, we illustrate the proposed RL approach (Figure 4.2) for the design of the optimal operations of a multipurpose reservoir leveraging the most valuable information (I_t) to be extracted from a set of candidate forecast available at time t ($\hat{Q}_t$). Specifically, we introduce a generic parametric function representing the extraction of information from available forecasts:

$$I_t = F_\zeta(\hat{Q}_t)$$

This function can include the following operations:

- selection of the best forecast product ($\gamma \in \Gamma$)
- selection of the best forecast lead time ($\lambda \in LT^\gamma$), here also called Aggregation Time (AT) as forecasts are aggregated over it;
- selection of the best temporal aggregation operator of the forecasts over the selected lead time (ψ_λ);
- selection of the best operator to deal with the forecast uncertainty (ψ_{n_e}, where n_e is the dimension of the forecast ensemble).

Moreover, an implicit operation is always performed to use only the most recent forecast between those available at time t.

The formulation of such a parametric information extraction function is then coupled with a Direct Policy Search formulation of the operating policy design problem. DPS is based on the parameterization of the operating policy (p_θ) within a given family of functions and the exploration of the parameter space ($\theta \in \Theta$) to find a parameterized policy that is optimal with respect to the operating objectives (Ruckstiess et al., 2010). Given the presence of multiple competing objectives, we used the Evolutionary Multi-Objective Direct Policy Search method (Giuliani et al., 2016) that allows an efficient search of the optimal parameters with respect to a multidimensional objective space.

Combining these two formulations, the daily release decision of Lake Como is now determined as

$$u_t = p_\theta(d_t, s_t, F_\zeta(\hat{Q}_t))$$

The multi-objective optimal control problem introduced in the previous section can be then reformulated as finding the best parameters of an Extended Operating Policy (EOP) that will specify both forecast information extraction (ζ^*) and reservoir operation (θ^*):

$$[\zeta^*, \theta^*] = \arg\min_{[\zeta,\theta]} \boldsymbol{J} \qquad s.t. \zeta \in \mathrm{Z}, \theta \in \Theta$$

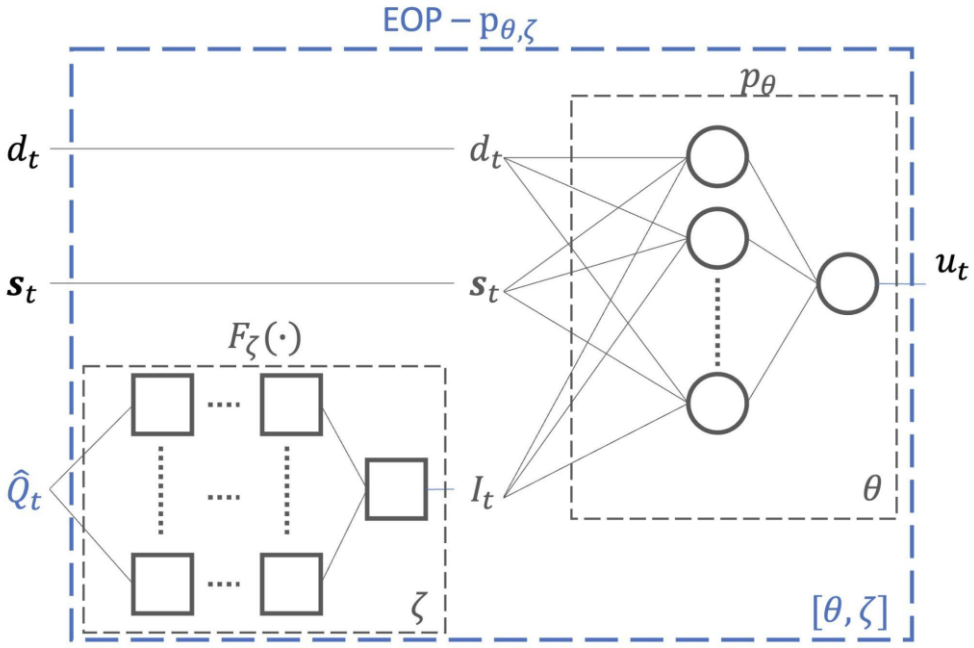

Figure 4.2 Internal structure of the Extended Operating Policy optimized by the joint learning of forecast information and operating policy. In the scheme, the circles represent the activation functions in the non-linear approximating network used to parameterize the operating policy, and the squares represent the operations that the EOP can perform on the forecasts (e.g., selection, temporal aggregation, or post-processing).

4.1.2 Results towards potentially added value AI-enhanced CS

Assessment of Forecast Skill

A preliminary analysis is performed to quantify the accuracy of the available forecasts. As more than one forecast product is available, comparing their performance may be beneficial to understand the results and steer the policy design in the right direction.

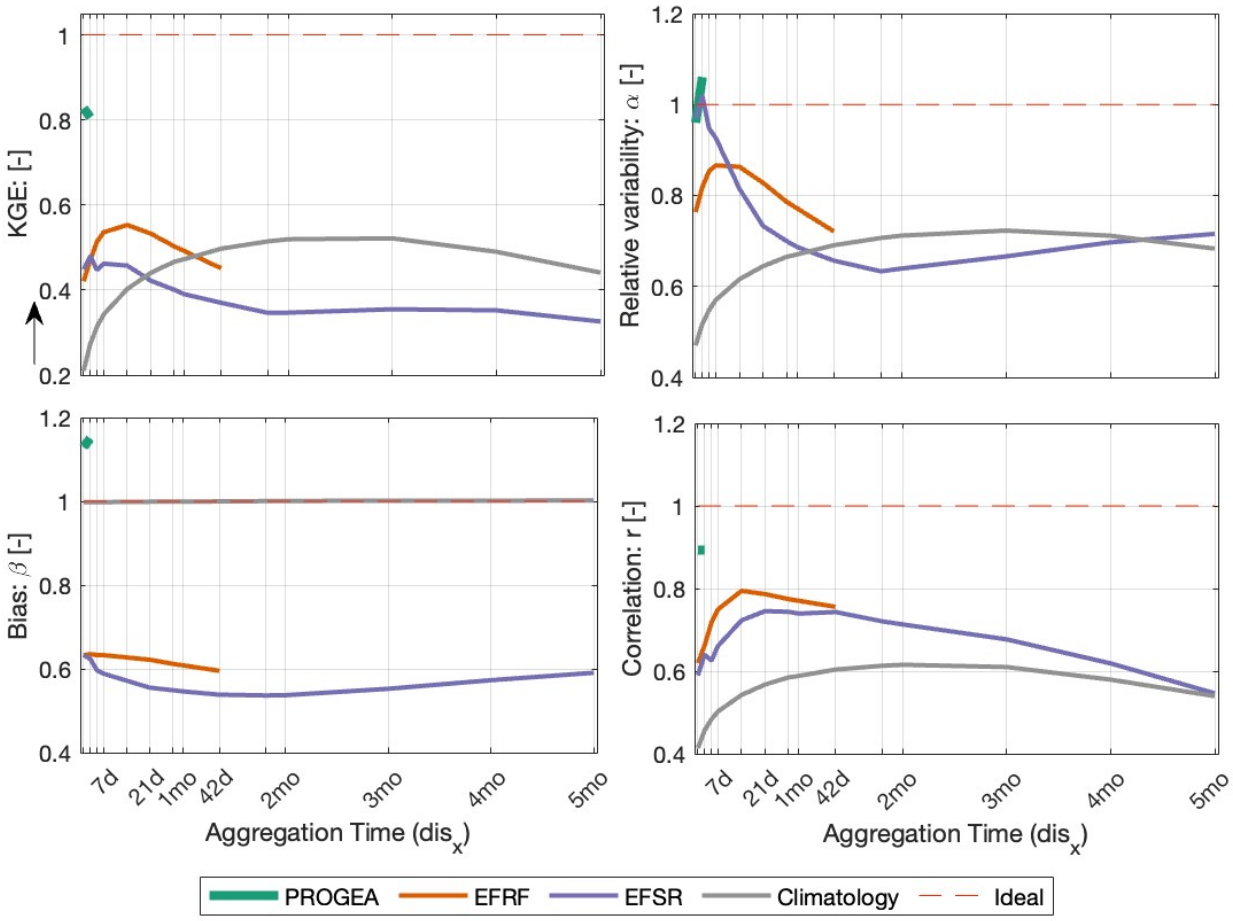

Figure 4.3 Kling-Gupta Efficiency score (KGE) and its three components as a function of the Aggregation Time (AT) for all the available forecast products and a climatology benchmark. The dashed red line indicates the ideal value for the score and each component.

Figure 4.3 shows the KGE as a function of the Aggregation Time (i.e., the period over which forecasts are averaged). The KGE score is based on the ensemble mean for the probabilistic products (EFRF and EFSR) because, in this work, we focus on using deterministic forecast according to the PROGEA product currently used by the lake operator. Although forecast accuracy is often analyzed with respect to the lead time at the original resolution (e.g., 6-hourly and daily values for EFRF and EFSR, respectively), the relationship between KGE and AT is more interesting here, because the lake regulation is expected to benefit from information on the cumulative inflow (e.g., the average discharge over the next month), and so the period of aggregation of the inflow over a future horizon is the key informative parameter here, encompassing more than one forecast lead time. Moreover, the strategy used to inform the lake operations is by seamless integration, i.e., by always using the most recent forecasts; hence certain lead times may be merged or never used separately.

The forecasts are evaluated using as a benchmark the climatology, i.e., the cyclostationary mean of the observed net inflows. PROGEA forecasts outperform the EFAS products at their short AT (i.e., up to 3 days). Similarly, the sub-seasonal EFAS forecasts (EFRF) outperform their seasonal

counterpart (EFSR) up to their maximum AT of 42 days. This result is in agreement with previous studies (Wetterhall and Di Giuseppe, 2018) and is motivated by the more frequent update of the initial condition of the sub-seasonal forecasting model, providing more accurate initial conditions. However, the climatology benchmark outperforms the forecasts when considering AT longer than one month. This is mainly determined by the large bias of EFAS to the local observations from which the climatology is derived. This problem is an expected consequence of the lack of calibration of the LISFLOOD hydrological model in the region. However, the correlation component of the KGE suggests that EFAS forecasts are more correlated with the observations than the climatology benchmark at all ATs, and their performance peaks around an AT equal to 14 days.

Assessment of forecast value

Building on the forecast skill assessment reported in the previous Section, we perform a first experiment to verify the learning of the best Aggregation Time of a single fictitious seamless product (i.e., $\zeta = \lambda$). This is called the best-skill product and combines the three available forecast products by selecting the forecast with the best KGE score for each AT. This means using PROGEA for the first 3 days, EFAS EFRF between 4 and 42 days, and EFAS EFSR from 43 days onwards. The performance of the EOPs is benchmarked against a set of Basic Operating Policies (BOPs) not informed by any forecast and a set of upper-bound solutions obtained by solving a deterministic problem, which we called Perfect Operating Policies (POPs).

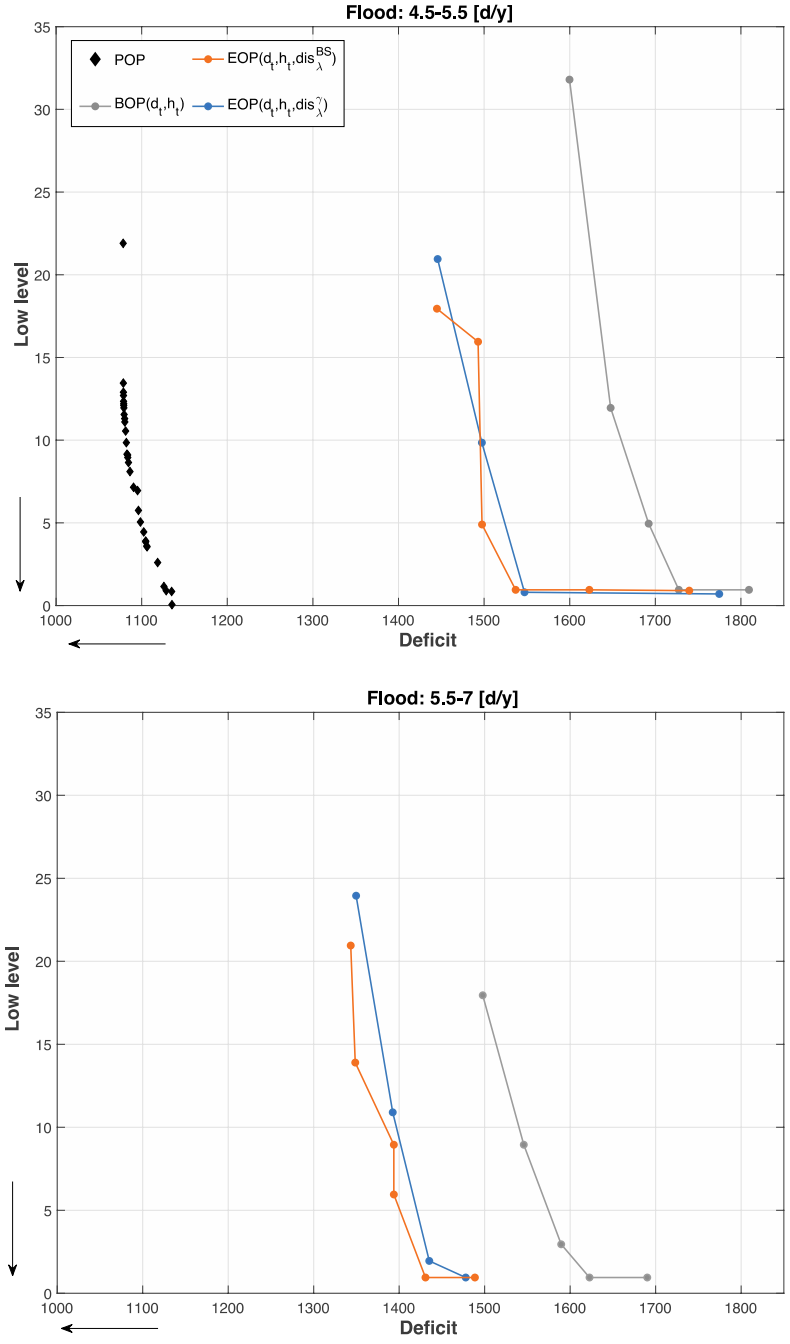

Figure 4.4 Performance of Extended Operating Policies informed by the best-skill product (orange) and all products processed in one input (blue). Solutions are grouped for different levels of flood days considering only the ones with less than 7 flood days per year. Arrows indicate the direction of preference, with the preferred solutions in the bottom-left corner of the figures. There are no POP solutions in the bottom panel as the perfect knowledge of future inflows allows attaining a performance in flood control that is always lower than 5.5 flood days per year.

Results in Figure 4.4 show that the EOPs improve the performance of the BOPs, especially for solutions with less than 5.5 flood days per year (top panel). The policy design consistently selects 3 days as AT, corresponding to using the PROGEA forecasts. This choice can be explained by the

substantially higher accuracy of the PROGEA short-term forecasts with respect to the sub-seasonal and seasonal EFAS products (Figure 4.3). Using more skillful, shorter-term forecasts results in policies that outperform those using lower-skill, longer-term products. However, it is interesting to observe that the EOPs select the PROGEA forecasts at their maximum AT, although their KGE is lower than the one of PROGEA forecasts with 1 day as AT (0.806 vs 0.828).

Figure 4.4 also shows that it is possible to further improve the Water Deficit and Low-Level objectives by accepting more flood events (e.g., in the bottom panel, where the flood performance is worse than the top one, the three curves shift towards the left thus attaining a better performance in the deficit objective). However, the forecast value decreases when moving to solutions with higher numbers of flood days because, in this case, the knowledge of future inflows is less critical for the lake operation, which can store water in favor of the other objectives without being limited by the increasing flood risk.

Given these promising results in learning the best AT, we run a second experiment in which the policy design simultaneously learns the best AT and the best forecast product (i.e., $\zeta = [\lambda, \gamma]$). The performance of the resulting EOPs (Figure 4.4) is very similar to the solutions informed by the best-skill product. This result is not surprising, as the EOP design selects again the PROGEA forecast at 3 days AT to inform the lake operation.

The use of simulated trajectories of lake level and release under different policies (Figure 4.5) can help us better understand the various contributions made by selected forecast information. Thanks to the additional information about the predicted inflows, the EOP keeps a higher average lake level than the BOP, thus saving more water to meet the summer demand. This is particularly evident at the beginning of the irrigation period (April), which represents the crucial timing to have water stored for being able to face potential drought periods over the coming summer. Then, all policies show similar release trajectories during summer, with only relatively small deviations observed mainly at the end of the irrigation season (September).

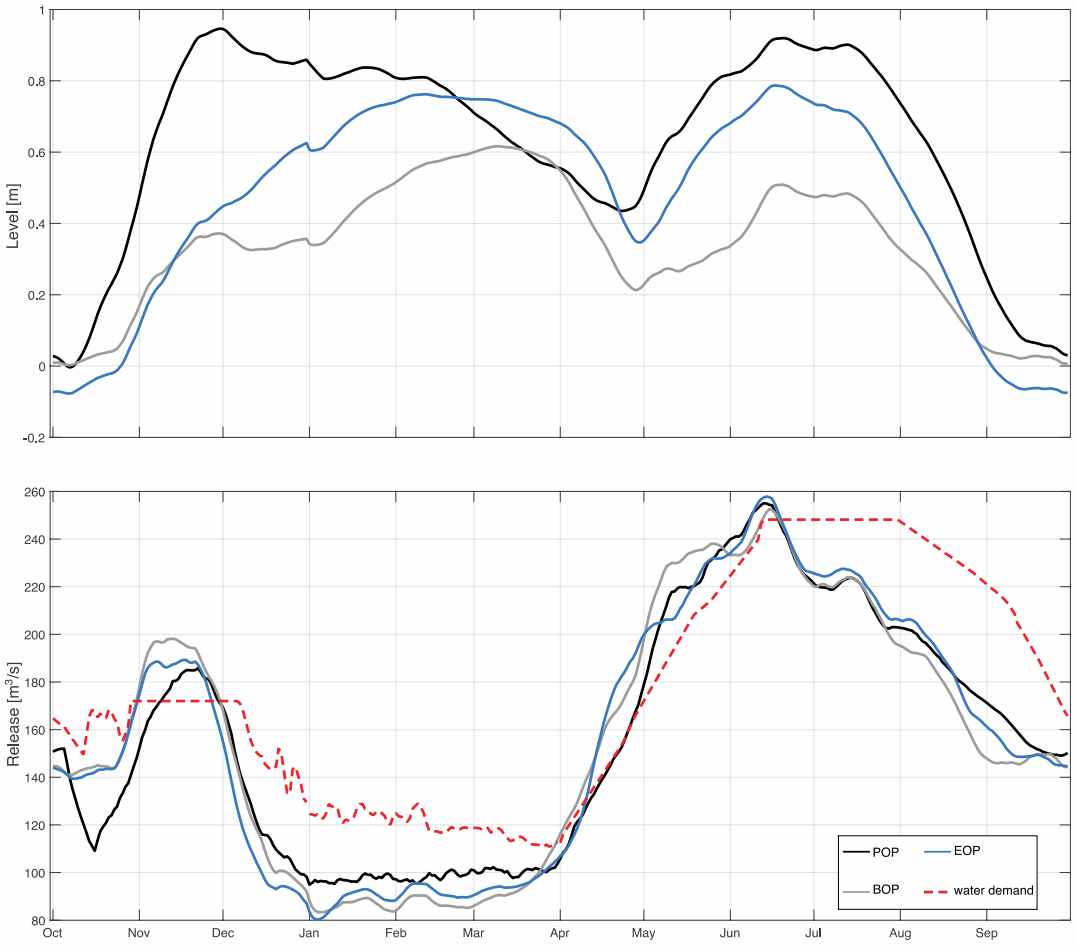

Figure 4.5 Lake Como level (top) and release (bottom) cyclostationary average trajectories (with 5-d moving window) under the Basic Operating Policy, the Perfect Operating Policy, and the Extended Operating Policy informed by PROGEA forecasts.

4.1.3 Analysis chain for AI-enhanced CS potential added value for heatwaves and warm nights

The methodological workflow to assess the potential added value of AI-enhanced Climate Services for heatwaves and warm nights in the Lake Como Basin is based on the detection of the relationship between heatwaves (and warm nights) indices with the crop yield in the area (see Figure 4.6). The AI-enhancement is therefore related to the identification of the most critical indices associated with the crop failure. The subsequent analysis of the projected values of these selected indices over different climate change scenarios provides value to the farmers for planning adaptation strategies in terms of changing cropping patterns favouring the cultivation of heat-tolerant varieties.

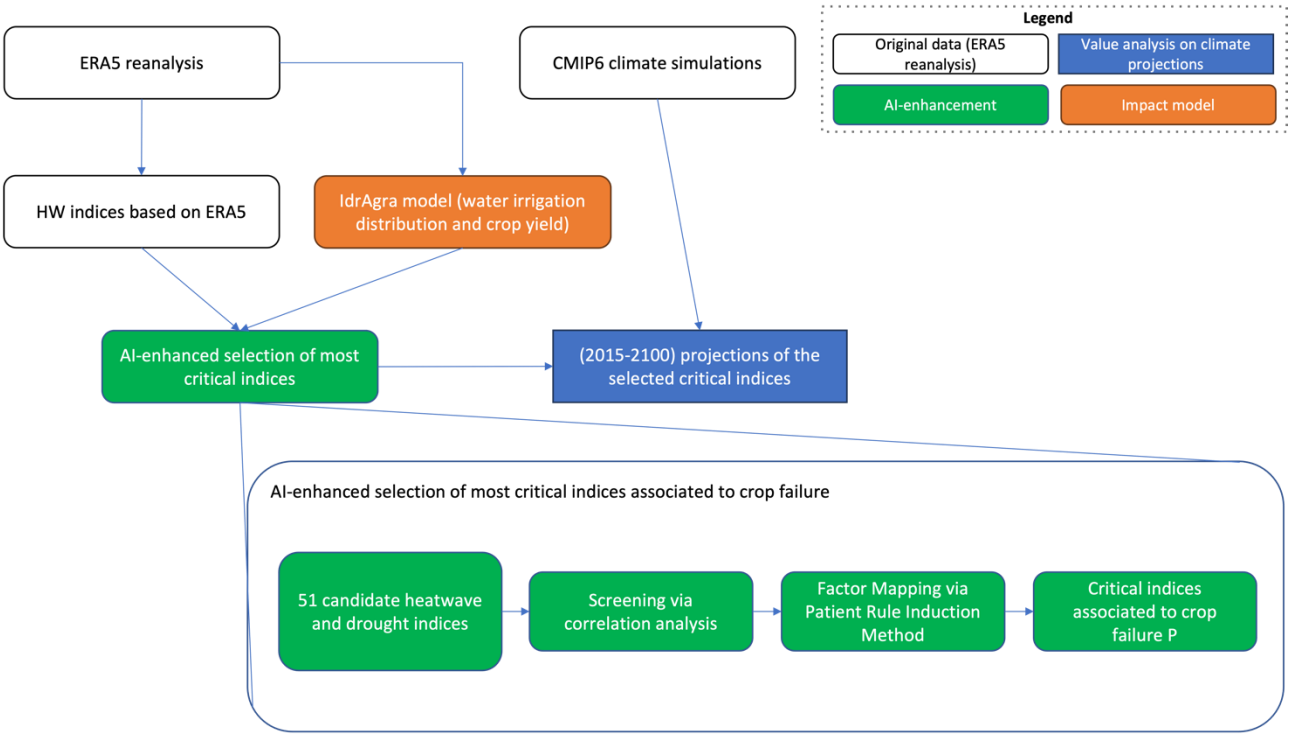

Figure 4.6 Flowchart for assessing the potential added value of AI-enhanced climate service Lake Como for heatwaves.

Data

Since the local meteorological data collected by the stations operated by the Environmental Protection Agency are available over a relatively short time period (i.e., less than 20 years), we used data from the ERA5 hourly reanalysis extracted for the box 46.5° North, 10.9° East, 44.5° South, 8.65° West. Specifically, the impact model described in the next section requires the following inputs:

- daily minimum and maximum temperature
- total precipitation
- daily minimum and maximum relative humidity
- daily average wind speed (derived from the V and U components)
- daily total solar radiation.

Daily time series of observed levels, releases, and net inflows of Lake Como are available from 1946 (the start of the lake regulation after the dam construction) to 2022. The release data are here used to simulate the irrigation supply to the agricultural area considered in the analysis. Moreover, the inflow timeseries is used for the detection of agricultural drought events by using the Standardized Streamflow Index in line with the information collected in Deliverable D7.1.

The projections of the heatwave indices are based on the climate simulations of the EC- Earth model (member 6), part of the Coupled Model Intercomparison Project Phase 6 (CMIP6) of the World Climate Research Programme. The scenarios considered are the ones used in the IPCC's Sixth Assessment Report (IPCC, 2021), namely SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5.

Impact Model

The impact model simulating the dynamic processes in the irrigation districts served by the Lake Como releases is the IdrAgra model (Figure 4.7). The model is composed of three distinct modules devoted to specific tasks:
(i) a distributed-parameter water balance module that simulates water sources, conveyance, distribution, and soil–crop water balance, including the application of irrigation (Facchi et al., 2004);
(ii) a crop phenology module that computes the sequence of growth stages as a function of the temperature according to the Heat Units theory (Neitsch et al., 2011); and
(iii) a crop yield module that estimates the optimal and actual yields, accounting for the effects of stresses due to insufficient water supply that may have occurred during the agricultural season (Steduto et al., 2009).

The water balance module partitions the irrigation district with a regular mesh of cells with a side length of 250 m (i.e., each cell covers an area of 6.25 hectares), which allows for the representation of the space variability of crops, soil types, meteorological inputs, and irrigation distribution. The study area consists of 32,820 grid cells, for a total cultivated area that amounts to 205,125 hectares. In addition, the water flows in the different irrigation canals that convey water to the cells of the domain were calculated using a set of Support Vector Machine models, one for each canal, that were identified using the B-AMA (Basic dAta-driven Models for All) protocol (Amaranto and Mazzoleni, 2023). These models use as input the flow in the Adda River released by Lake Como and the flow in the Oglio River, which is estimated through a polynomial model that uses as input the discharge in the Adda River.

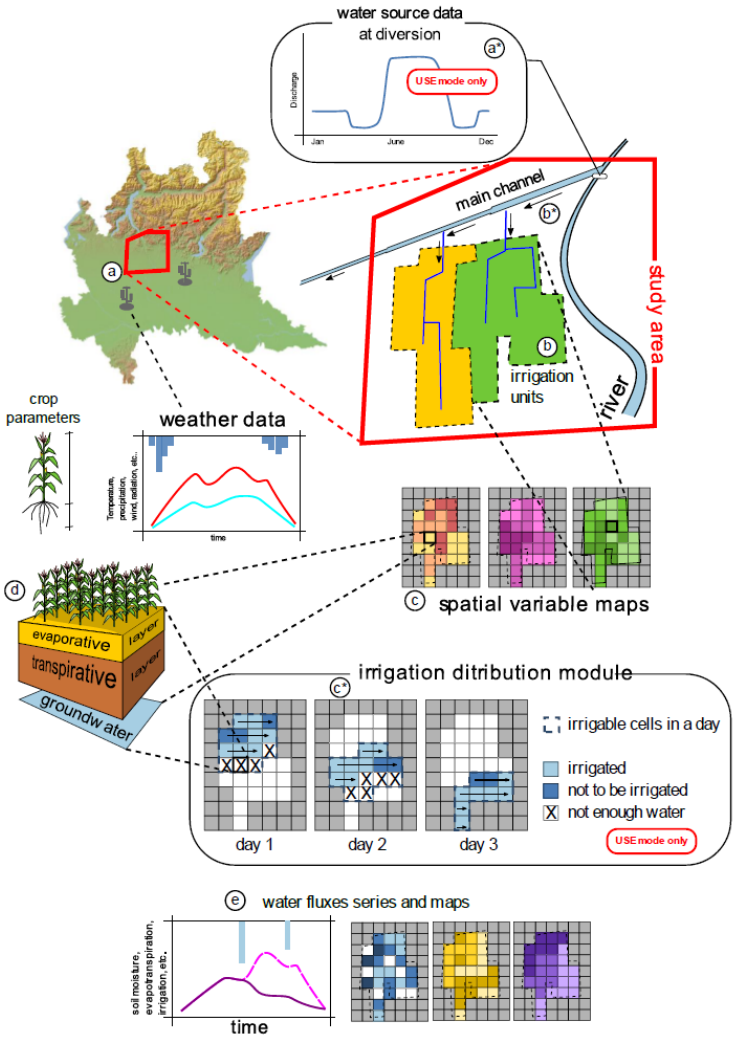

Figure 4.7 Overview of the IdrAgra impact model: a) identification of the study area, meteorological stations and crop types; a) identification of the water sources; b) identification of the Irrigation Units (optional); b*) definition of the conveyance path from each source to IU or cells; c) space discretization with a regular mesh; c*) irrigation distribution module; d) computation of the daily water balance for each cell; e) simulation results in each cell.*

AI enhancement

The Patient Rule Induction Method (PRIM) was used as an aid in the analysis of the relationship between crop production and heatwave-drought indices. PRIM is a statistical clustering method originally introduced by Friedman and Fisher (1999). It belongs to a group of algorithms called "bump-hunting" algorithms, which are used to find regions, called *scenario boxes*, in the input variable space that are associated with the highest or lowest mean value for the outcome (Nannings et al., 2008). Boxes correspond to a simple square in a two-dimensional input variable space and to a hypercube in a multi- dimensional space. This is unlike regression models, which seek to model the whole population by optimizing a likelihood function. In this work, the input variable space is composed of the heatwave and drought indices, whereas the outcome is the yearly yield. More specifically, what is of interest is the identification of the indices associated with the low yields, which are labelled as crop failures. Therefore, a threshold (defined in the next section) has to be set

on the outcome variable so that the algorithm can distinguish between failures and non-failures when building the scenario boxes.

The scenario boxes are constructed by optimizing different competing metrics, namely coverage - how many failure scenarios are captured within a box - and density - how many of the captured scenarios in each box belong to the failure set. Ideally, a scenario box should have both a high coverage and density. This guarantees that the inputs used to build a box are able to explain the highest number of failure points possible and that the noise generated by uninteresting points is minimum, which happens when the density is high.

4.1.4 Results towards potentially added value AI-enhanced CS

Heatwaves impact on crop yield

Figure 4.8 shows the total yearly yield simulated by the IdrAgra model over the period 1946-2022. The average yield over the 77 years considered is of 2.45×10^6 tons, with a standard deviation of 2.16×10^5 tons. The maximum yield is equal to 2.99×10^6 tons and it was recorded in 1972, while the minimum is equal to 1.95×10^6 tons and it was recorded in 2006.

Overall, there has been a slow decline of the average yield in the years 1992-2022 (when the heatwave indices show their greatest growth). In fact, the average in the last 30 years is 5% lower compared to the average of the period from 1946 to 1991. Despite this, there are still strong fluctuations from year to year, and the peaks reached in the last decades are comparable to some of the values recorded before 1992.

To identify the potential connections between heatwaves and droughts with crop failures, we chose as the threshold the 25^{th} percentile of the empirical distribution of the simulated yields. It is interesting to notice that the majority of the most significant failure years have happened in the last 30 years.

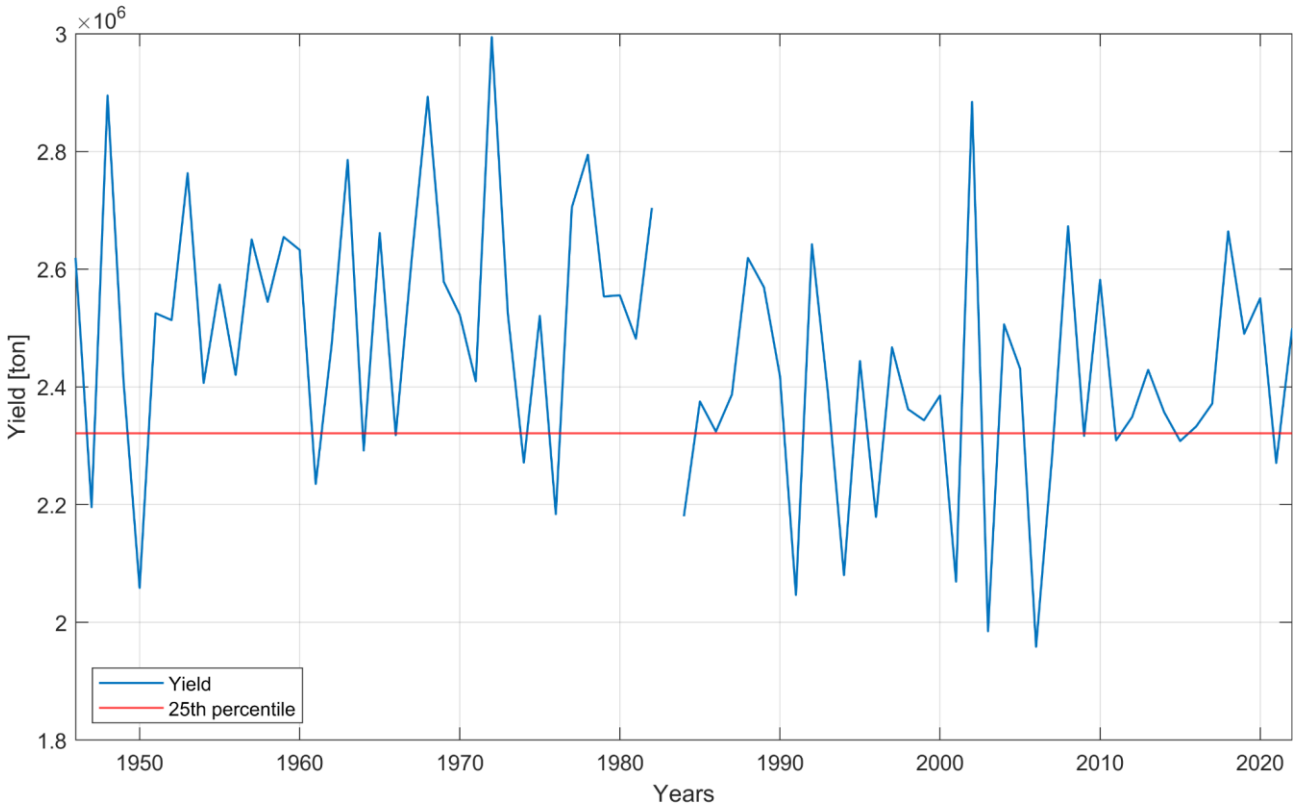

Figure 4.8 Simulated yield (1946-2022). The red line represents the 25th percentile of the empirical distribution used to identify crop failures. The 1983 yield is missing because of corrupted data.

To understand the link between yearly yield and extreme temperatures, and potentially droughts, we considered 51 indices, which fall into the following categories:

- HWMI calculated with maximum (tmax) and minimum temperature (tmin);
- NDQ90 (tmax and tmin);
- Number of heatwave occurrences over the agricultural season (April to September) (tmax and tmin);
- Sum of heatwave intensity over the season (tmax and tmin);
- Number of heatwave occurrences in the individual months over the season (tmax and tmin);
- Sum of heatwave intensity in the individual months over the season (tmax and tmin);
- NDQ90 in the individual months over the season (tmax and tmin);
- Number of drought months each year;
- Average SSI during the year;
- Average drought intensity each year;
- September SSI, aggregated over 3, 6, 9 and 12 months.

In particular, the monthly heatwave indices were considered to assess whether the yield is more sensitive to temperature extremes in specific moments of the season. The September SSI was chosen because, since it is calculated at the end of the agricultural season, it can give information on water scarcity during the crops' growth period. The standard aggregation is 6 months, but different aggregation periods were tested to understand whether considering hydrological

anomalies further back in time can be informative for capturing crop stress that negatively affects the yield in the Adda River basin. However, numerical results show that correlations with SSI aggregated over 12 and 9 months turned out to be statistically insignificant, meaning that anomalies that happen before April do not influence crop production.

After an initial filtering based on a correlation analysis, only 27 out of the 51 indices proved to be statistically significantly correlated with the simulated yield. All heatwave indices are negatively correlated with the yield, while the SSI values show a positive correlation. Since it was impossible to rank the indices that showed a statistically significant correlation based only on the values of the correlation coefficients and p-values, we used PRIM as a supporting tool for completing the factor mapping task, i.e. the identification of the most relevant drivers (indices) of the crop failures.

Specifically, we explored different solutions generated by PRIM that rely on an increasing number of indices and differently balance coverage and diversity:

- NDQ90 in June calculated with maximum temperature;
- NDQ90 in June calculated with maximum temperature and HWMI calculated with minimum temperature
- NDQ90 in June calculated with maximum temperature, HWMI calculated with minimum temperature, NDQ90 in August calculated with maximum temperature and the SSI aggregated over 3 months and calculated in September
- NDQ90 in June calculated with maximum temperature, HWMI calculated with minimum temperature, NDQ90 in August calculated with maximum temperature, the SSI aggregated over 3 months and calculated in September, number of heatwave days in June calculated with maximum temperature and number of yearly heatwave days calculated with maximum temperature

The PRIM results show that the most important drivers selected as responsible of crop failures in the Adda River basin are the NDQ90 in June (tmax) and the HWMI (tmin). The former, by definition, also includes days that are not necessarily part of a heatwave event, meaning that extreme temperatures in general are detrimental to crop yield. The latter is particularly interesting because it shows the significant role that night-time temperature extremes play in the agricultural sector (minimum temperatures are normally reached during the night). A drought index also appears among the heatwave indices which signifies that water stress contributes to crop failure but not as much as temperature stress since it is selected only starting from the 4-dimensional solution.
After analyzing the trade-offs between coverage and density for the four scenario boxes, we chose to consider only the solution with two drivers (coverage 89.5% and density 44.7%) for our subsequent analyses on climate projections. Although we are aware that the density is not extremely high, we concluded that moving to the solution with four drivers allowed for an improvement of this metric that did not compensate for the reduction in coverage.

Projections of future heatwaves

The two indices selected by PRIM, namely HWMI calculated with minimum temperature and NDQ90 June calculated with maximum temperature, were projected into the future (2015-2100) using the EC-Earth model and four different scenarios. The projected trajectories of HWMI in SSP1-2.6 and

SSP2-4.5 do not exhibit any evident trend, but spikes are present throughout the entire century (Figure 4.9). For SSP2-4.5, in particular, the values from 2060 onwards are for the most part higher than the highest value recorded for the historical period. SSP3-7.0 and SSP5-8.5, on the other hand, show a clear rise in HWMI values starting from the late mid-century. In SSP5-8.5, values from 2060 onwards are constantly greater than the historical maximum, while this is true for SSP3-7.0 from circa the mid 2070s onwards. The values projected for the last decade of the 21st century in SSP5-8.5 are at least 5 times higher than the historical maximum.

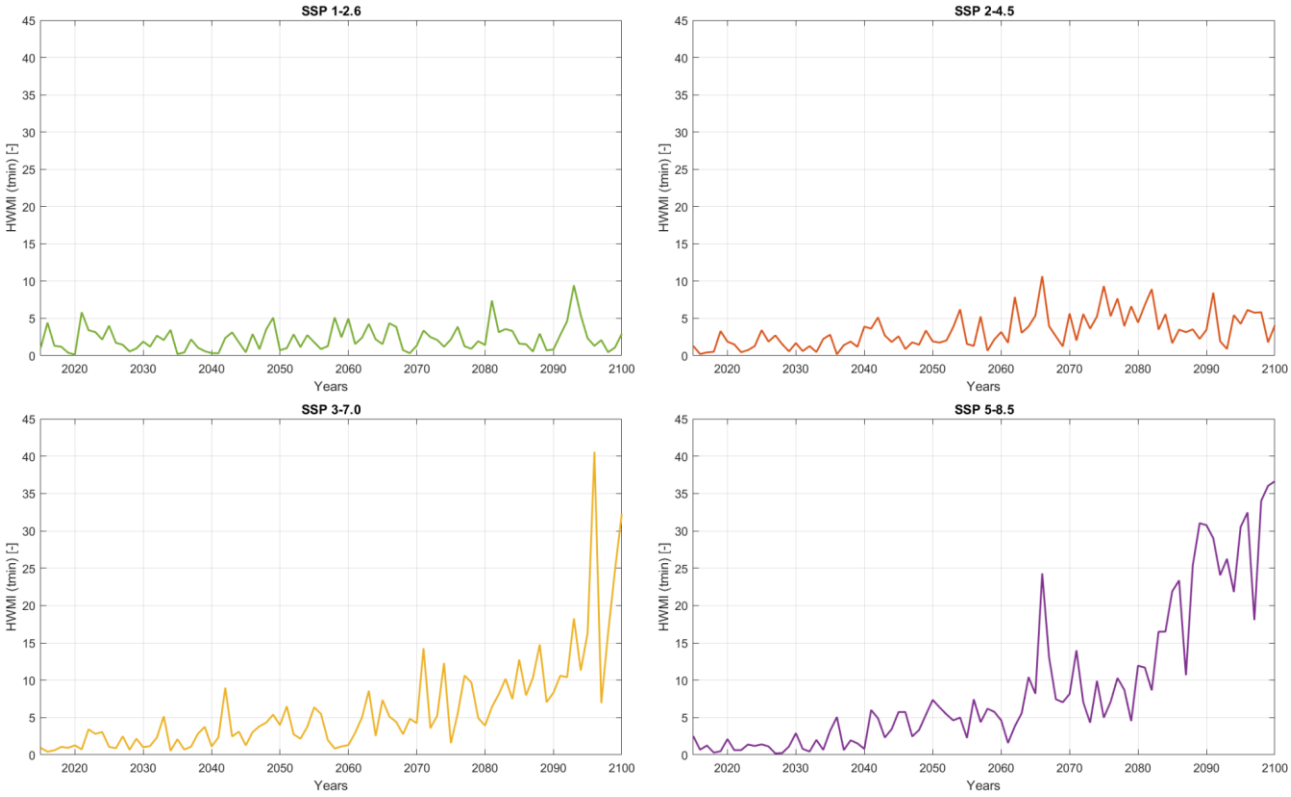

Figure 4.9 Projected trajectories (2015-2100) of HWMI (tmin) for the four scenarios considered.

The projections of the NDQ90 June index (Figure 4.10) show a clear growth trend in SSP5-8.5 only. SSP3-7.0 also shows an increasing trend from 2060 onwards, with values of the index never reaching zero, except for one occasion. Despite not showing any apparent trend, SPP1-2.6 and SSP3-7.0 have values of the index that are on average greater than the historical average (2.82), and SSP3-7.0 often has peaks that surpass the historical maximum (20). What is interesting in the instance of the NDQ90 June index is that the range of values is rather similar for all four scenarios, contrary to what happens for the HWMI.

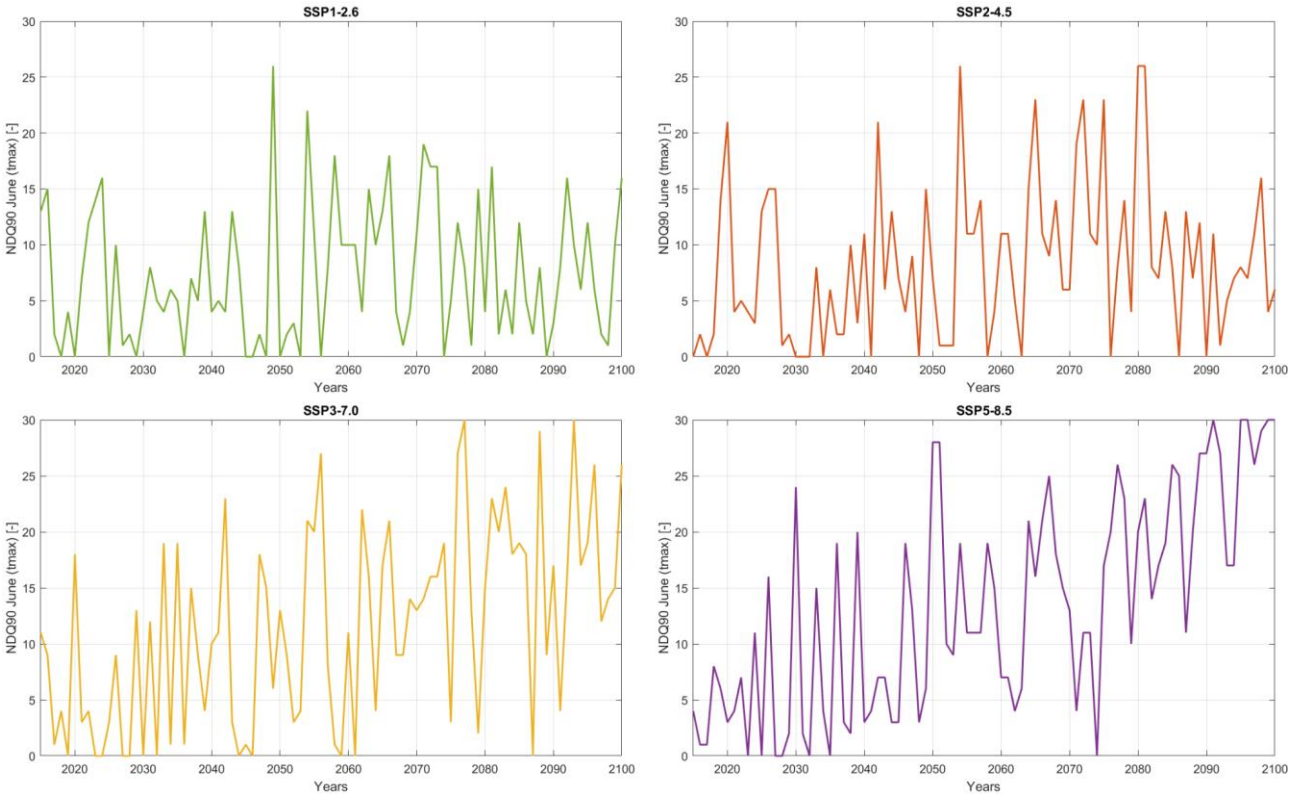

Figure 4.10 Projected trajectories (2015-2100) of NDQ90 June (tmax) for the four scenarios considered.

The comparison of the HWMI (tmin) projections in the middle (2036-2065) and end (2071-2100) of the century with the historical period (1993-2022) shows that there is high variability in the distributions of the future index values across scenarios (Figure 4.11, top panels). Especially at the end of the century, we can observe how the maximum values of SSP1-2.6 and SSP2-4.5 are expected to be lower than the median of the other two scenarios. More specifically, the median for SSP3-7.0 is two times bigger than the 90th percentile in SSP1-2.6 and two units higher than the 90th percentile in SSP2-4.5, while the median for SSP5-8.5 is more than three times bigger than the one for SSP1-2.6 and twice the median for SSP2-4.5.

The distributions of NDQ90 June (tmax) look more similar than those of HWMI (tmin) in the middle of the century (Figure 4.11, bottom panels). Towards the end of the century, the distributions in SSP3-7.0 and SSP5-8.5 move decisively upwards. In fact, the median for SSP5-8.5 is projected to be three times bigger than in SSP1-2.6 and more than twice the median in SSP2-4.5, whereas the median for SSP3-7.0 results more than two times bigger than the ones in the lower emissions scenarios.

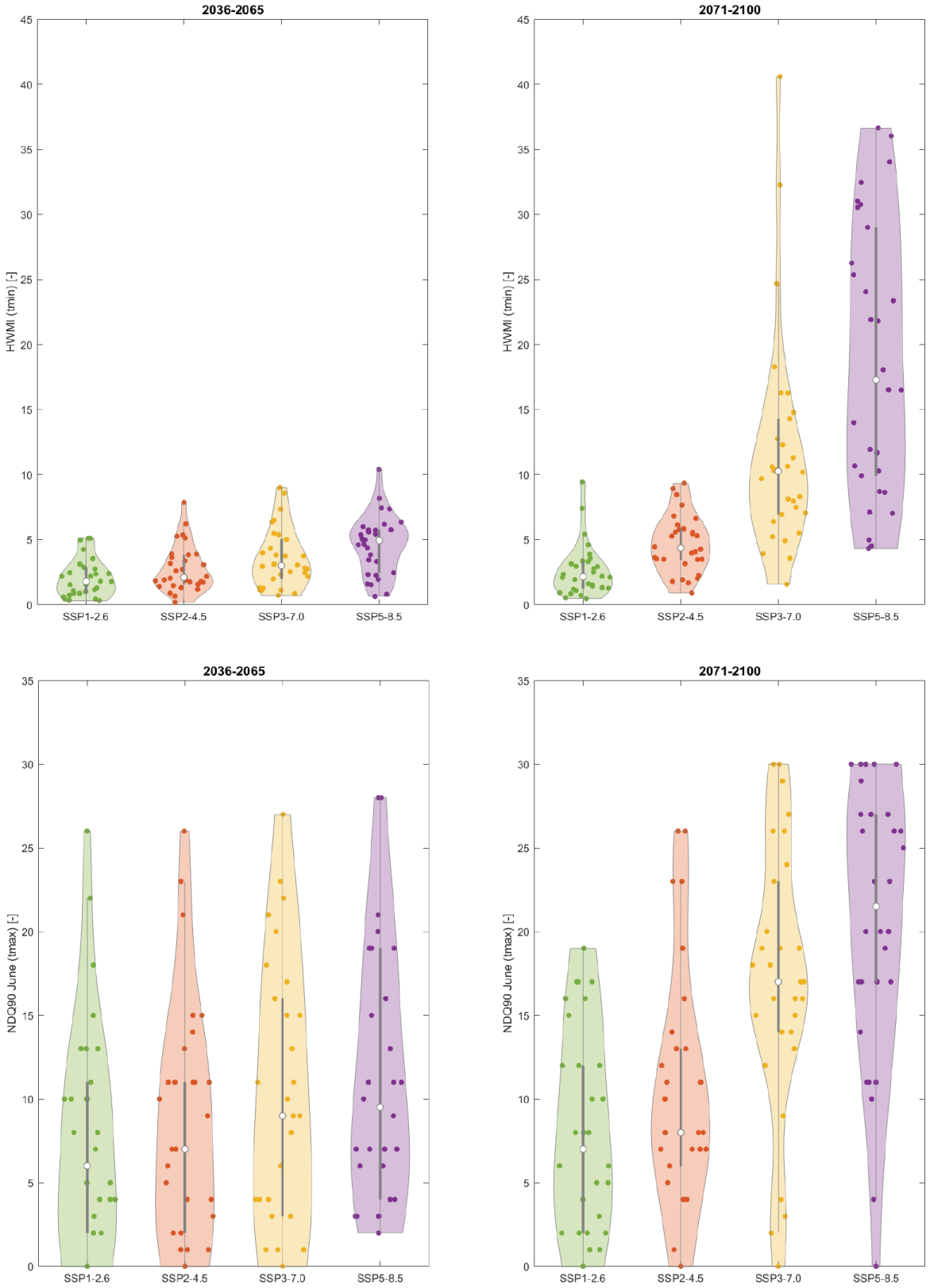

Figure 4.11 Comparison of HWMI (tmin) – top panels – and NDQ90 June (tmax) – bottom panels - across scenarios in the middle (left) and at the end (right) of the 21st century.

Finally, we summarize the analysis of these climate projections using the references provided by Global Warming Levels. The violin plots in Figure 4.12 contrast the projected distributions with the one of the indices computed on ERA-5 data, which corresponds to a GWL equal to 0.63°C. Results clearly show substantially higher HWMI (tmin) values projected in a world that is 4.0°C warmer than pre-industrial times compared to the other two lower degrees of warming. On average, the index is equal to 9.78 at GWL 4.0°C while it is 1.23 at GWL 1.5°C and 0.89 at GWL 0.63°C. Besides, the 25th percentile at the highest warming level is almost two times bigger than the 90th percentile at GWL 1.5°C and more than double the 90th percentile at GWL 0.63°C.

The NDQ90 June (tmax) distributions are not blatantly diverse at GWL 0.63°C and GWL 1.5°C: the average value of the index is, respectively, 3.77 and 5.25, while the median of the distribution is in both cases 3.00 and the 90th percentile is 10.50 and 13.50, respectively. On the other hand, the projected average value of NDQ90 June (tmax) at GWL 4.0°C is 15.88 (four times higher than the average at GWL 0.63°C and three times the mean at GWL 1.5°C). The 90th percentile of the distribution is more than twice the 90th percentile at the other two GWLs.

These results can provide valuable information for the farmers in the Lake Como basin as they support a better understanding about crops' vulnerability to projected temperature extremes. In fact, the correlation analysis between the critical heatwave indices with the yields of individual crops, instead of with the total yield, shows that maize and rice are negatively correlated with both indices, whereas melon is correlated only with NDQ90 June (tmax). The correlations on cereals and soy are instead not statistically significant, suggesting these crops are particularly resistant to extreme temperatures.

In light of the analysis done on the projected heatwave indices HWMI (tmin) and NDQ90 June (tmax), it emerges that the HWMI (tmin) is the index that is expected to increase the most - the average at GWL 4.0°C is 10 times higher for HWMI (tmin) and three times higher for NDQ90 June (tmax). This means that those crops that exhibit a stronger negative correlation with this index are the ones that would be more at risk in the future because it is to be expected that, for higher values of the index, the values of yield will be increasingly lower. For example, rice shows the greatest negative Pearson correlation coefficient with HWMI (tmin) (and also with NDQ90 June (tmax)); this probably means that rice will not be a suitable crop in the area, particularly not in the worst-case scenarios or at the highest levels of global warming.

Maize shows the second-greatest negative correlation coefficient with HWMI (tmin), therefore it is also a crop at risk. It should be noted that maize represents the main crop in the area. A partial switch from maize to a more resistant crop type will probably be necessary to avoid frequent crop failures in the years to come. If a crop switch were not to be feasible because of reasons dictated by market demand or other socio-economic reasons, then investments in heat-tolerant varieties of maize would be of the essence, considering the high correlation that this crop also has with NDQ90 June (tmax).

Melon is not significantly correlated with HWMI (tmin) and exhibits a weaker negative correlation with NDQ90 June (tmax) compared to the previous crops. Therefore, it might still be a crop worth cultivating in the area in the near future in the case of the lowest emission scenarios, for example. It might also partially (and momentarily) substitute maize and rice, because even though it is not the optimal crop for the area, it, at least, seems to perform better and would not need farmers to adapt to a completely new crop variety since it is already present on the territory.

Lastly, neither soy nor cereals showed any significant correlation with the two considered indices. Our results suggest these are, therefore, the most promising crops that should be considered in the Adda River basin to face the projected increase in temperature. It must be noted, though, that, even though there are no apparent correlations with the historical indices' series, there is no guarantee that with extremely higher values of the indices soy or cereals will not start to be impacted as well. Therefore, constant monitoring and analyses are crucial.

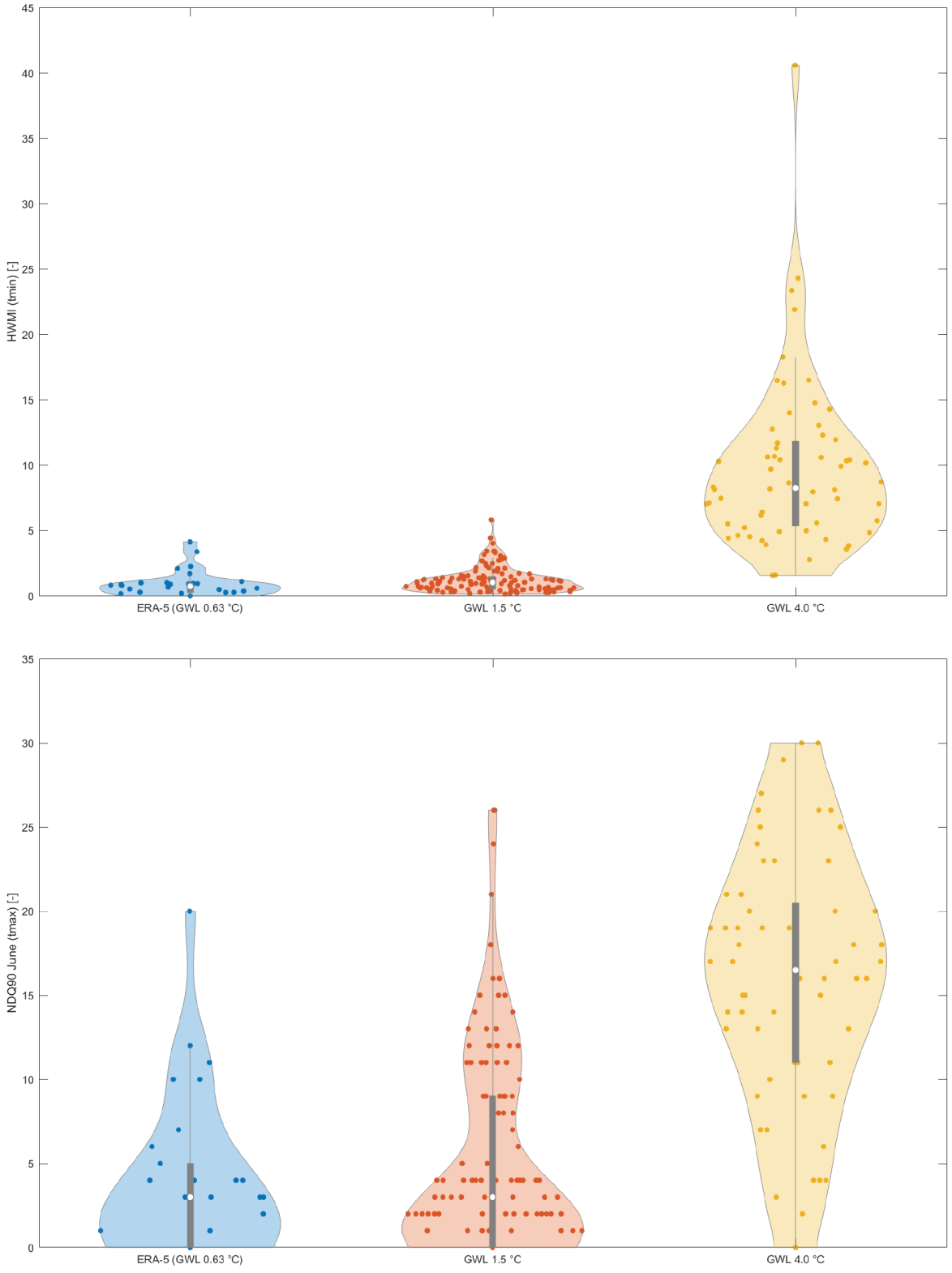

Figure 4.12 Values of the two heatwave indices at different warming levels. The number of points in each plot is different because the ERA-5 data is composed of yearly values over 30 years, at GWL 1.5°C there are 120 points (30 for each of the four scenarios), and at GWL 4.0°C there are 60 points (30 each for SSP3-7.0 and SSP5-8.5 only).

4.1.5 Next steps

This chapter reports the results of Task T7.4 focused on the development and demonstration of AI-enhanced Climate Services for the snow Climate Change Hotspot of Lake Como. Specifically, we report a preliminary AI-enhanced CS that leverages Reinforcement Learning algorithms to improve drought and flood management by extracting the most valuable information from multi-timescale forecasts of the lake inflows. Moreover, we describe a second preliminary AI-enhanced CS that uses the PRIM algorithm to discover the relationship between temperature extremes (both heatwaves and warm nights) and crop failures to support farmers' adaptation to future climate scenarios.

Our results show that the use of selected forecast information, i.e. inflow forecasts with a 3 days aggregation time, produce the largest added value for the multipurpose operation of Lake Como, especially in terms of reducing the water supply deficit. The analysis of the projected heatwaves and warm nights shows that temperature extremes are expected to increase considerably over the coming years in all scenarios. These trends suggest the opportunity to replace some of the crops currently cultivated in the area, primarily maize, in favour of heat-tolerant varieties, such as soy or cereals, in order to ensure more reliable productions in the coming years.

These preliminary results motivate ongoing research efforts aimed to further improve these CS, which will be reported in Deliverable D7.3. These new CS will focus on the development of AI-enhanced sub-seasonal to seasonal hydrologic forecasts by adapting and evolving the approach described in Deliverable D2.2 for the meteorological drought forecasting with ML and climate data. Moreover, we will investigate the impact of compound heatwave and drought events and the possibility of addressing them through a dedicated AI-enhanced CS.

5 Conclusions and outlook

From the preliminary results of the AI-enhanced climate services as developed to-date for the local scale CLINT case studies, it can be concluded that for several extreme events improved forecast skill over benchmark predictions has been found already, and for several use cases also the potential added value following the user-defined impact indicators has been shown.

For the Zambezi (droughts and tropical cyclone rainfall), Rijnland (droughts), Aa en Maas (droughts), and Lake Como (droughts and floods, and heatwaves and warm nights) case studies, AI-enhancement of CSs has been shown. For Douro (droughts) and Main water system of the Netherlands (flood risk), work on AI-enhancement is ongoing.

For Zambezi (tropical cyclone rainfall), Rijnland (droughts), and Lake Como (droughts and floods, and heatwaves and warm nights), the potential added value in local decision making context for extreme event impact mitigation has been shown. For the other case studies, analysis of potential added value of the AI-enhanced services is ongoing.

In our opinion, the analysis chain developed for each case study as presented in this deliverable, outlining the complete steps up to the added value assessment within the operational context of each case study's extreme event management practices, provides a solid foundation for advancing the AI-enhanced climate services and their analysis in the next WP7 activities, to be reported in D7.3.

The results and discussion also showed that increase in forecast skill or added value in decision making is sometimes only limited and may be sensitive to the benchmark forecast systems chosen. For each of the case studies, therefore, next steps have been identified, both for further AI-enhancement and for strengthening the robustness of the added value analyses.

The outcomes of these next steps will be reported in the final deliverables of this work package, D7.3 on CS developed and their demonstrated added value, and D7.4 on benchmark analysis.

References

Amaranto A., and Mazzoleni M. (2023). B-ama: A python-coded protocol to enhance the application of data-driven models in hydrology. *Environmental Modelling & Software*, 160:105609.

Andreu J., Capilla J., Sanchis E., and Tormo P. (1992). AQUATOOL: Sistema Soporte de Decisión para la Planificación de Recursos Hídricos. Manual del Usuario. Departamento de Ingeniería Hidráulica y Medio ambiente. Universidad Politécnica de Valencia, Valencia.

Andreu J., Capilla J., and Sanchís E. (1996). AQUATOOL, a generalized decision-support system for water-resources planning and operational management, *J. Hydrol.* 177 (1996) 269–291, https://doi.org/10.1016/0022-1694(95)02963-X.

Anghileri D., Castelletti A., Pianosi F., Soncini-Sessa R., and Weber E. (2013). Optimizing watershed management by coordinated operation of storing facilities. *Journal of Water Resources Planning and Management*, 139, 492–500. doi: https://doi.org/10.1061/(ASCE)WR.1943-5452.0000313.

Arnold W., Salazar J.Z., Carlino A., Giuliani M., and Castelletti A. (2023). Operations Eclipse Sequencing in Multipurpose Dam Planning. *Earth's Future*, 11(4), e2022EF003186. https://doi.org/10.1029/2022EF003186.

Ascenso G., Ficchì A., Giuliani M., Scoccimarro E., and Castelletti A. (under review). Downscaling, Bias Correction, and Spatial Adjustment of Extreme Tropical Cyclone Rainfall in ERA5 Using Deep Learning. *Submitted to Weather and Climate Extremes*.

Barnard C., Krzeminski B., Mazzetti C., Decremer D., Carton de Wiart C., Harrigan S., Blick M., Ferrario I., Wetterhall F., Thiemig V., Salamon P., and Prudhomme C. (2020). Reforecasts of river discharge and related data by the European Flood Awareness System, version 4.0, Copernicus Climate Change Service (C3S) Climate Data Store (CDS). doi: https://doi.org/10.24381/cds.c83f560f.

Barnston A. G., and van den Dool H.M. (1993). A degeneracy in cross-validated skill in regression-based forecasts. *Journal of Climate*, 6(5), 963-977.

Beck H.E., Vergopolan N., Pan M., Levizzani V., Van Dijk A.I., Weedon G.P., Brocca L., Pappenberger F., Huffman G. J., and Wood E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, 21(12):6201–6217.

Beck H.E., Wood E.F., Pan M., Fisher C.K., Miralles D.G., Van Dijk A.I., McVicar T.R., and Adler R.F. (2019a). MSWEP v2 global 3-hourly 0.1 precipitation: methodology and quantitative assessment.

Beck H.E., Pan M., Roy T., Weedon G.P., Pappenberger F., Van Dijk A.I., Huffman G.J., Adler R.F., and Wood E.F. (2019b). Daily evaluation of 26 precipitation datasets using stage-iv gauge-radar data for the conus. *Hydrology and Earth System Sciences*, 23(1):207–224.

Benninga H.-J.F., Carranza C.D.U., Pezij M., van Santen P., van der Ploeg M.J., Augustijn D.C.M., and van der Velde R. (2018). The Raam regional soil moisture monitoring network in the Netherlands, *Earth Syst. Sci. Data*, 10, 61–79, https://doi.org/10.5194/essd-10-61-2018.

Bergström S. (1995). The HBV model. *Computer models of watershed hydrology*., 443-476.

Bosso F. (2022). Improving sub-seasonal drought forecasting via machine learning to leverage climate data at different spatial scales, MSc thesis POLITECNICO DI MILANO, Geoinformatics Engineering Master of Science School of Civil Environmental and Land Management Engineering.

Caspers J.J., and Kindermann P.E. (2023). Hydraulic load model for the Dutch shore. HKV, Lelystad, Final report, PR4529.20.

Castelletti A., Pianosi F., and Soncini-Sessa R. (2008). Water reservoir control under economic, social and environmental constraints. *Automatica,* 44(6), 1595–1607. doi: https://doi.org/10.1016/j.automatica.2008.03.003.

Cheon S.H., Hamlington B.D., Reager J.T., and Chandanpurkar H.A. (2021). Identifying ENSO-related interannual and decadal variability on terrestrial water storage. *Sci Rep* 11, 13595. https://doi.org/10.1038/s41598-021-92729-4.

Clark M., Gangopadhyay S., Hay L., Rajagopalan B., and Wilby R. (2004). The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5(1), 243-262.

Collenteur R.A., Bakker M., Caljé R., Klop S.A., and Schaars F. (2019), Pastas: Open Source Software for the Analysis of Groundwater Time Series. *Groundwater*, 57: 877-885. https://doi.org/10.1111/gwat.12925.

Coughlan de Perez E., van den Hurk B., van Aalst M.K., Amuron I., Bamanya D., Hauser T., Jongma B., Lopez A., Mason S., Mendler de Suarez J., Pappenberger F., Rueth A., Stephens E., Suarez P., Wagemaker J., and Zsoter E. (2016). Action-based flood forecasting for triggering humanitarian action, *Hydrol. Earth Syst. Sci.*, 20, 3549–3560, https://doi.org/10.5194/hess-20-3549-2016.

Crochemore L., Ramos M.H., and Pappenberger F. (2016). Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, 20(9), 3601-3618.

Denaro S., Anghileri D., Giuliani M., and Castelletti A. (2017). Informing the operations of water reservoirs over multiple temporal scales by direct use of hydro-meteorological data. *Advances in Water Resources*, 103, 51-63. doi: https://doi.org/10.1016/j.advwatres.2017.02.012.

De Valk C.F., and van den Brink H.W. (2023). Update van de statistiek van extreme zeewaterstand en wind op basis van meetgegevens en model simulaties. KNMI, De Bilt, Technical report TR-406.

Dorninger M., Ghelli A., and Lerch S. (2020). Recent developments and application examples on forecast verification. *Meteorological Applications*, 27(4), e1934.

DRBA (2021). Plan Hidrológico de la parte española de la Demarcación Hidrográfica del Duero Revisión de tercer ciclo (2022-2027). Anejo2. Inventario de recursos hídricos naturales. Duero River Basin Authority, Valladolid (Spain) (Draft version).

DRBA (2023). Plan Especial de Sequía de la parte española de la Demarcación Hidrográfica del Duero. Valladolid (Spain).

Emerton R., Cloke H., Ficchì A., Hawker L., de Wit S., Speight L., Prudhomme C., Rundell P., West R., Neal J., Cuna J., Harrigan S., Titley H., Magnusson L., Pappenberger F., Klingaman N., and Stephens E. (2020). Emergency flood bulletins for Cyclones Idai and Kenneth: A critical evaluation of the use of global flood forecasts for international humanitarian preparedness and response. *International Journal of Disaster Risk Reduction*, 50, 101811. https://doi.org/10.1016/j.ijdrr.2020.101811.

Facchi A., Ortuani B., Maggi D., and Gandolfi C. (2004). Coupled SVAT–groundwater model for water resources simulation in irrigated alluvial plains, *Environ. Modell. Software*, 19(11), 1053–1063.

Ferro C.A., Richardson D.S., and Weigel A.P. (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, 15(1), 19-24.

Friedman J.H., Fisher N.I. (1999). Bump hunting in high-dimensional data. *Statistics and computing*, 9(2):123–143.

Gaughan A.E., Staub C.G., Hoell A., Weaver A., and Waylen P.R. (2016). Inter-and Intra-annual precipitation variability and associated relationships to ENSO and the IOD in southern Africa. *Int. J. Climatol.* 36, 1643–1656.

Giuliani M., Castelletti A., Pianosi F., Mason E., and Reed P.M. (2016). Curses, tradeoffs, and scalable management: Advancing evolutionary multi-objective direct policy search to improve water reservoir operations. *Journal of Water Resources Planning and Management*, 142(2), 04015050. doi: https://doi.org/10.1061/(ASCE)WR.1943-5452.0000570.

Giuliani M., Li Y., Castelletti A., and Gandolfi C. (2016). A coupled human-natural systems analysis of irrigated agriculture under changing climate, *Water Resources Research*, 52, 6928–6947.

Giuliani M., Zaniolo M., Castelletti A., Davoliq G., and Block P. (2019). Detecting the state of the climate system via artificial intelligence to improve seasonal forecasts and inform reservoir operations. *Water Resources Research*, 55. https://doi.org/10.1029/2019WR025035.

Gneiting T., Balabdaoui F., and Raftery A.E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2), 243-268.

Herrera S., Fernández J., and Gutiérrez J.M. (2015). Update of the Spain02 Gridded Observational Dataset for Euro-CORDEX evaluation: Assessing the Effect of the Interpolation Methodology. *International Journal of Climatology*. DOI: 10.1002/joc.4391.

Hess P., and Boers N. (2022). Deep learning for improving numerical weather prediction of heavy rainfall. *Journal of Advances in Modeling Earth Systems*, 14(3):e2021MS002765.

Hu Y.-F., Yin F.-K., Zhang W.-M., and Deng K.-F. (2022). A hybrid fusion precipitation bias correction approach for Yin-he global spectral model. *Meteorological Applications*, 29(5):e2097.

IFRC (2024). Mozambique, Cyclone Early Action Protocol Summary (EAP Number: EAP2023MZ03). Retrieved from https://reliefweb.int/report/mozambique/mozambique-cyclone-early-action-protocol-summary-25-january-2024-eap2023mz03.

IFRC, and Bangladesh Red Crescent Society (2021). Bangladesh, Cyclone Early Action Protocol Summary (EAP Number: EAP2021BD04). Retrieved from https://reliefweb.int/report/bangladesh/bangladesh-cyclone-early-action-protocol-summary-eap-number-eap2021bd04.

IPCC (2021). Climate Change 2021: the physical science basis. https://www.ipcc.ch/report/sixth-assessment-report-working-group-i/.

IPCC (2023). Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland, 184 pp., doi: 10.59327/IPCC/AR6-9789291691647.

Jin Q., Meng Z., Sun C., Cui H., and Su R. (2020). Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans. *Frontiers in Bioengineering and Biotechnology*, 8:605132.

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., and Monge-Sanz B.M. (2019). SEAS5: the new ECMWF seasonal forecast system, *Geosci. Model Dev*., 12, 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019.

Knapp K.R., Kruk M.C., Levinson D.H., Diamond H.J., and Neumann C.J. (2010). The international best track archive for climate stewardship (IBTrACS) unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91(3):363–376.

Knijff J.M.V.D., Younis J., and Roo A.P.J.D. (2010). LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, 24(2), 189-212. doi: https://doi.org/10.1080/13658810802549154.

Knoben W.J.M., Freer J.E., and Woods R.A. (2019). Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. *Hydrol Earth Syst Sci* 23, 4323–4331.

Kok M., Jongejan R., Nieuwjaar M., and Tánczos I. (2016). Fundamentals of Flood Risk. On behalf of ENW. Breda: NPN Drukkers.

Lagerquist R., and Ebert-Uphoff I. (2022). Can we integrate spatial verification methods into neural network loss functions for atmospheric science? *Artificial Intelligence for the Earth Systems*, 1(4):e220021.

Laio F., and Tamea S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11(4), 1267-1277.

Lam R., et al. (2023). Learning skillful medium-range global weather forecasting. *Science*, 382,1416-1421. DOI:10.1126/science.adi2336.

Ling F., Li Y., Luo J.-J., Zhong X., and Wang Z. (2022). Two deep learning based bias-correction pathways improve summer precipitation prediction over China. *Environmental Research Letters*, 17(12):124025.

Magnusson L., Majumdar S., Emerton R., Richardson D., Alonso-Balmaseda M., Baugh C., Bechtold P., Bidlot J.-R., Bonanni A., Bonavita M., Bormann N., Brown A., Browne P., Carr H., Dahoui M., De Chiaraw G., Diamantakis M., Duncan D., English S., Forbes R., Geer A., Haiden T., Healy S., Hewson T., Ingleby B., Janousek M., Kuehnlein C., Lang S., Lock S.-J., McNally T., Mogensen K., Pappenberger F., Polichtchouk I., Prates F., Prudhomme C., Rabier F., de Rosnay P., Quintino T., and Rennie M. (2021). Tropical Cyclone Activities at ECMWF. ECMWF Technical Memorandum, No. 888. DOI:10.21957/zzxzzygwv.

Nannings B., Abu-Hanna A., and De Jonge E. (2008). Applying prim (patient rule induction method) and logistic regression for selecting high-risk subgroups in very elderly icu patients. *International Journal of Medical Informatics*, 77(4):272–279.

Neitsch S., Arnold J., Kiniry J., and Williams J. (2011). Soil and water assessment tool theoretical documentation Version 2009, Tech. Rep. 406, Grassland, Soil and Water Res. Lab., Agric. Res. Serv. Blackland Res. Cent., Tex. AgriLife Res., College Station.

Otero N., and Horton P. (2023). Intercomparison of deep learning architectures for the prediction of precipitation fields with a focus on extremes. *Water Resources Research*, 59, e2023WR035088. https://doi.org/10.1029/2023WR035088.

Owens R.G., and Hewson T.D. (2018). ECMWF Forecast User Guide. Reading: ECMWF. DOI: 10.21957/m1cs7h.

Paredes-Arquiola J., Solera A., Andreu J., and Lerma N. (2012). Manual técnico de la herramienta EVALHID para la evaluación de recursos hídricos.

Pezij M., Augustijn D.C.M., Hendriks D.M.D., and Hulscher S.J.M.H. (2020). Applying transfer function-noise modelling to characterize soil moisture dynamics: a data-driven approach using

remote sensing data. *Environmental Modelling & Software*, 131: 104756. https://doi.org/10.1016/j.envsoft.2020.104756.

Rasp S., Dueben P.D., Scher S., Weyn J.A., Mouatadid S., and Thuerey N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12, e2020MS002203. https://doi.org/10.1029/2020MS002203.

Ricart S., Gandolfi C., and Castelletti A. (2024). How do irrigation district managers deal with climate change risks? Considering experiences, tipping points, and risk normalization in northern Italy. *Climate Risk Management*, 44, 100598.

Roberts N.M., and Lean H.W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136(1):78–97.

Ronneberger O., Fischer P., and Brox T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer.

Rückstiess T., Sehnke F., Schaul T., Wierstra D., Sun Y., and Schmidhuber J. (2010). Exploring parameter space in reinforcement learning. Paladyn, *Journal of Behavioral Robotics*, 1(1), 14–24. doi: https://doi.org/10.2478/ s13230-010-0002-4.

Schepen A., Zhao T., Wang Q.J., and Robertson D.E. (2018). A Bayesian modelling method for post-processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments. *Hydrology and Earth System Sciences*, 22(2), 1615-1628.

Sha Y., Gagne II D.J., West G., and Stull R. (2020). Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part ii: Daily precipitation. *Journal of Applied Meteorology and Climatology,* 59(12):2075–2092.

Sharifi E., Eitzinger J., and Dorigo W. (2019). Performance of the state-of-the-art gridded precipitation products over mountainous terrain: A regional study over Austria. *Remote Sensing*, 11(17):2018.

Soncini-Sessa R., Castelletti A., and Weber E. (2007). *Integrated and participatory water resources management: Theory.* Elsevier.

Spalding-Fecher R., Joyce B., and Winkler H. (2017). Climate change and hydropower in the Southern African Power Pool and Zambezi River Basin: System-wide impacts and policy implications. *Energy Policy*, 103, 84–97. https://doi.org/10.1016/j.enpol.2016.12.009.

Taschetto A.S., Ummenhofer C.C., Stuecker M.F., Dommenget D., Ashok K., Rodrigues R.R., and Yeh S.-W. (2020). ENSO Atmospheric Teleconnections. In El Niño Southern Oscillation in a Changing

Climate (eds M.J. McPhaden, A. Santoso and W. Cai). https://doi.org/10.1002/9781119548164.ch14.

Steduto P., Hsiao T., Raes D., and Fereres E. (2009). AquaCrop—The FAO crop model to simulate yield response to water: I. Concepts and underlying principles, *Agron. J.*, 101(3), 426–437.

Stevanato N., Rocco M.V., Giuliani M., Castelletti A., and Colombo E. (2021). Advancing the representation of reservoir hydropower in energy systems modelling: The case of Zambesi River Basin. *PLoS ONE* 16(12): e0259876. https://doi.org/10.1371/journal.pone.0259876.

Wang Q. J., Shrestha D.L., Robertson D.E., and Pokhrel P. (2012). A log-sinh transformation for data normalization and variance stabilization. *Water Resources Research*, 48(5).

Wetterhall F., Arnal. L., Barnard C., Krzeminski B., Ferrario I., Mazzetti C., and Prudhomme C. (2020). Seasonal forecasts of river discharge and related data by the European Flood Awareness System, v4.0, Copernicus Climate Change Service (C3S) Climate Data Store (CDS). doi: https://doi.org/10.24381/cds.768eefc2.

Wetterhall F., and Di Giuseppe F. (2018). The benefit of seamless forecasts for hydrological predictions over Europe. *Hydrology and Earth System Sciences*, 22(6), 3409–3420. doi: https://doi.org/10.5194/hess-22-3409-2018.

Zhao T., Bennett J.C., Wang Q.J., Schepen A., Wood A.W., Robertson D.E., and Ramos M.H. (2017). How suitable is quantile mapping for postprocessing GCM precipitation forecasts? *Journal of Climate*, 30(9), 3185-3196.

Zimmerman B.G., Vimont D.J., and Block P.J. (2016). Utilizing the state of ENSO as a means for season-ahead predictor selection. *Water Resources Research*, 52, 3761–3774. https://doi.org/10.1002/2015WR017644.

Appendix A - Zambezi Watercourse: Benchmark CS for droughts

Here we provide a more complete probabilistic evaluation and skill assessment of the two benchmark systems considered for the ZW, i.e. GloFAS (v. 3.1) and WW-HYPE seasonal re-forecasts, across four different strategic locations in the basin, upstream of the four main dams (Kariba, Cahora Bassa, Ithezi-thezi and Kafue Gorge). Monthly historical observations of streamflow upstream of the dams are used as reference datasets to assess the seasonal forecast products. This monthly historical streamflow dataset is retrieved from ZAMCOM and includes observed inflows at four stations: (1) Kafue Hook Bridge along the Kafue River; (2) Luangwa Great East Road Bridge along the Luangwa River; (3) Victoria Falls along the Zambezi; (4) Mangochi along the Shire River.

The probabilistic forecast performance is assessed via the Brier Scores (BS) and the Continuous Ranked Probability Score (CRPS), considering the 25 ensemble members of the two systems. Skill is assessed with respect to climatology using the Brier Skill Scores (BSS) and the Continuous Ranked Probability Skill Score (CRPSS).

First, the evolution of the Brier Scores with lead time is analysed for two different low-flow thresholds (15-th and 25-th percentiles) for the seasonal forecasts (see Figure A.1). Results show a very small variability of the score with lead time, suggesting that the low-flow detection is similarly good (or poor) at 1-4 month lead times for GloFAS and 1-7 months for HYPE. The low dependency on lead times may be due to the strong seasonality of these low flow events in the basin and a good long-range predictability of the dry season start. The lower performance of WW-HYPE for low flow detection with respect to GloFAS is probably due to the presence of periods of several months with zero-flow in HYPE during the dry seasons that suggest a lack of baseflow in the hydrological model and is likely to jeopardise the low-flow detection capability of the system.

Second, the CRPS indicates that GloFAS seasonal forecasts have a much larger bias than WW-HYPE (Figure A.2), in line with previous results on the MAE and the bias component of KGE (see Table 2.1). WW-HYPE seasonal forecasts are better than GloFAS in terms of biases, but still similar to or worse than climatology at almost all stations, except for Kafue Gorge for which WW-HYPE just reaches the performance of the benchmark and does not add much information in terms of the overall distribution of probabilities of events with respect to a simple mean observed flow benchmark.

In terms of skill assessment, the Brier Skill Scores show the skill of the two forecast systems removing the bias component effect, as ‘unbiased’ (forecast-based) thresholds are used, where thresholds are computed based on forecasts for defining forecasted events (i.e., thresholds exceedances). The unbiased definition of the BS is in contrast with the normal procedure for calculating the BS that would be extracting the thresholds from the historically observed data to define both observed and forecasted events. As expected from previous results on the better accuracy of low-flow event probabilities of GloFAS with respect to HYPE (Figure A.1), the skill of GloFAS for drought prediction at the seasonal scale is relatively good (Figure A.3). The skill is then further assessed by using the Continuous Ranked Probability Skill Score (CRPSS) that is again clearly affected by the forecast bias (Figure A.4), which is large in GloFAS. The score is close to zero for WW-HYPE, showing that despite a better agreement of the water balance with observations, HYPE does not add much information on the probability distribution of streamflow events of all classes with respect to a simple

climatology benchmark. This is in line with the results of other scores and shows that a poorer correlation of WW-HYPE with respect to observation (Table 2.1) and a poor capacity of detecting low-flow events (Figure A.1) brings down the general skill of this system in the Zambezi Basin.

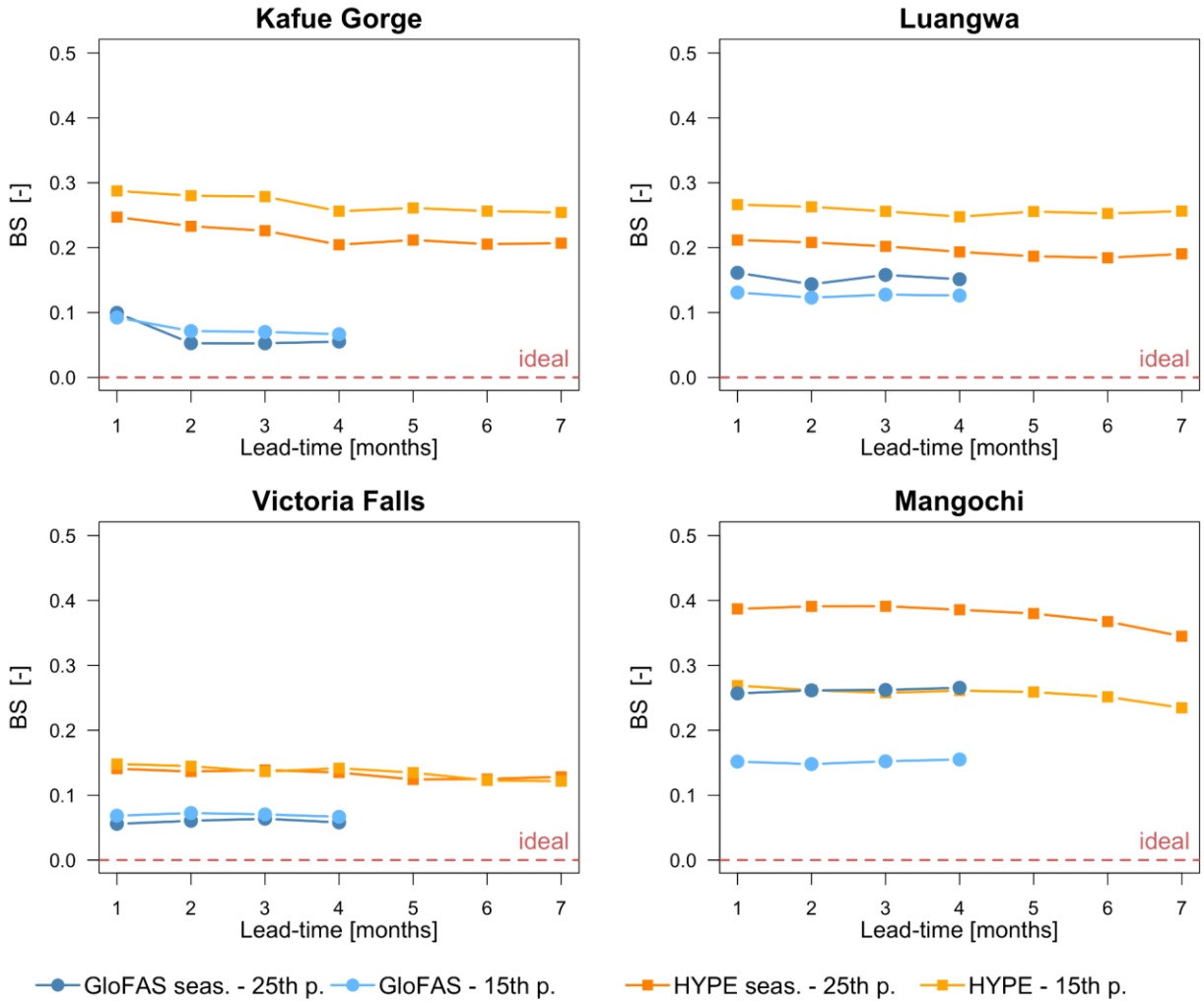

Figure A.1. Brier Score (BS) obtained for GloFAS and WW-HYPE seasonal ensemble forecast using two low-flow percentile thresholds (15th and 25th monthly flows) at the four selected river gauge stations in the Zambezi River Basin, as a function of lead time (up to the maximum available lead time for each product). The optimal value is 0.

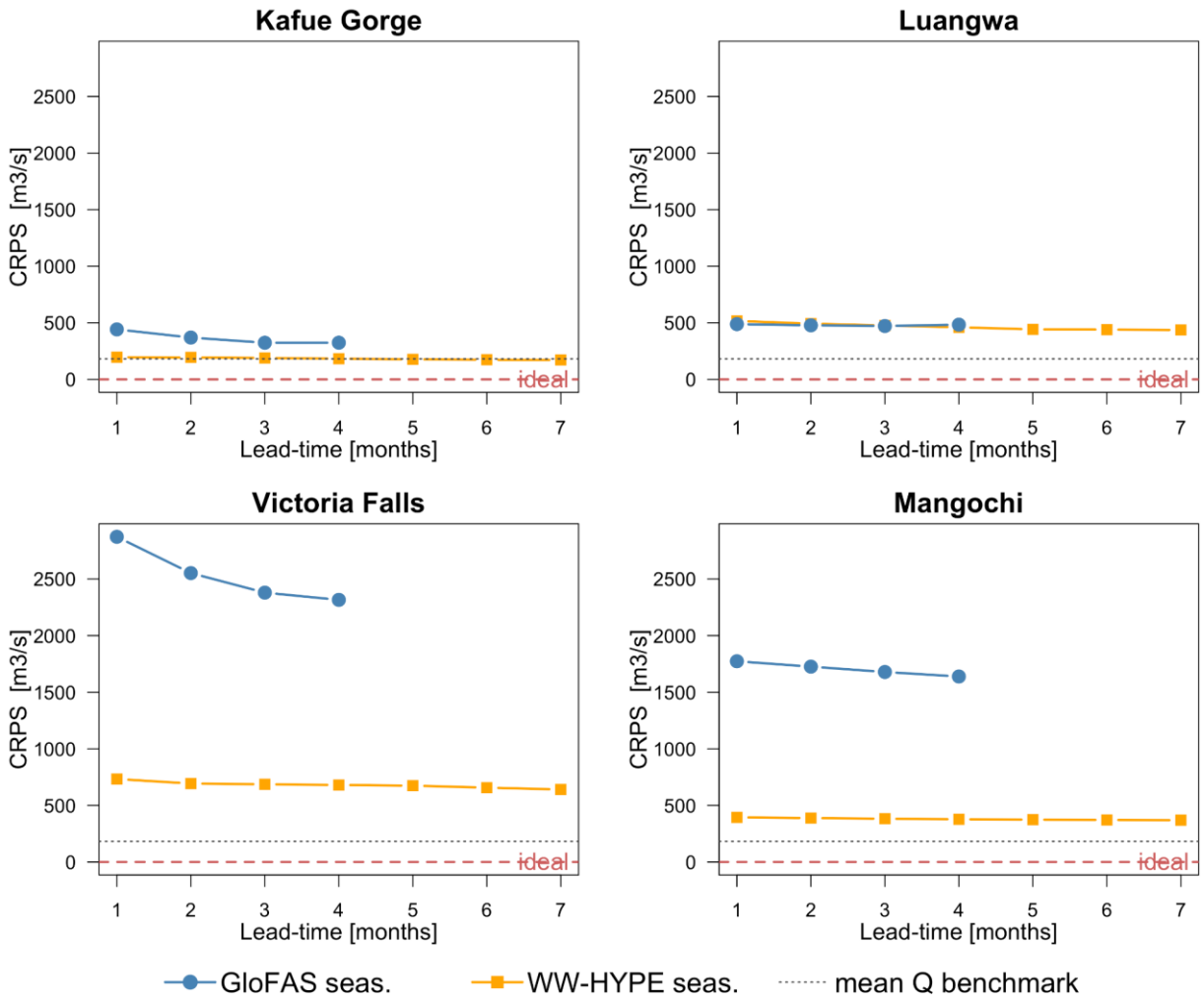

Figure A.2. Continuous Ranked Probability Score (CRPS) obtained for GloFAS and WW-HYPE seasonal ensemble forecast at the four selected river gauge stations in the Zambezi River Basin, as a function of lead time (up to the maximum available lead time for each product). The optimal value is 0.

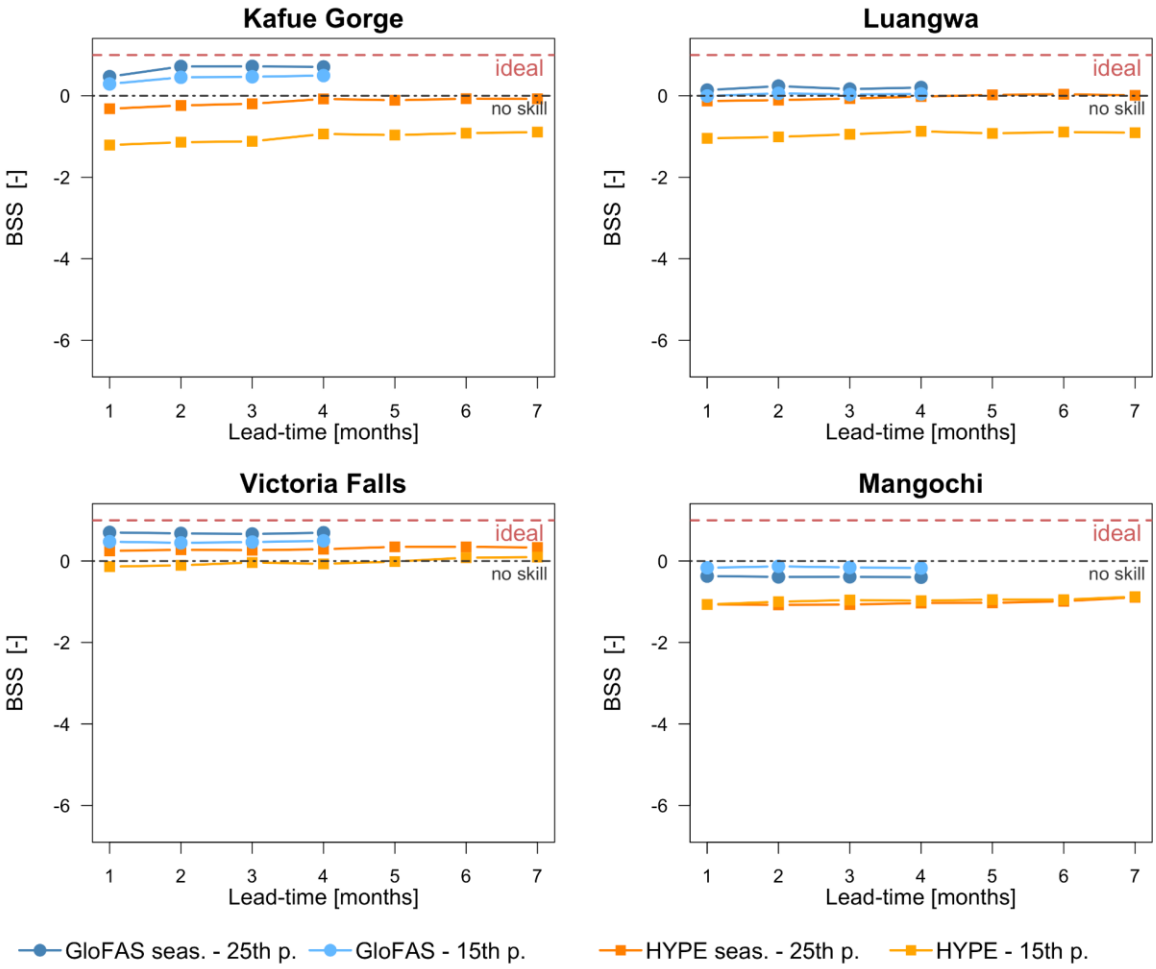

Figure A.3. Brier Skill Score (BSS) obtained for GloFAS and WW-HYPE seasonal ensemble forecasts at the four selected river gauge stations in the Zambezi River Basin, as a function of lead time (up to the maximum available lead time for each product). The optimal value is 1. The reference benchmark is climatology.

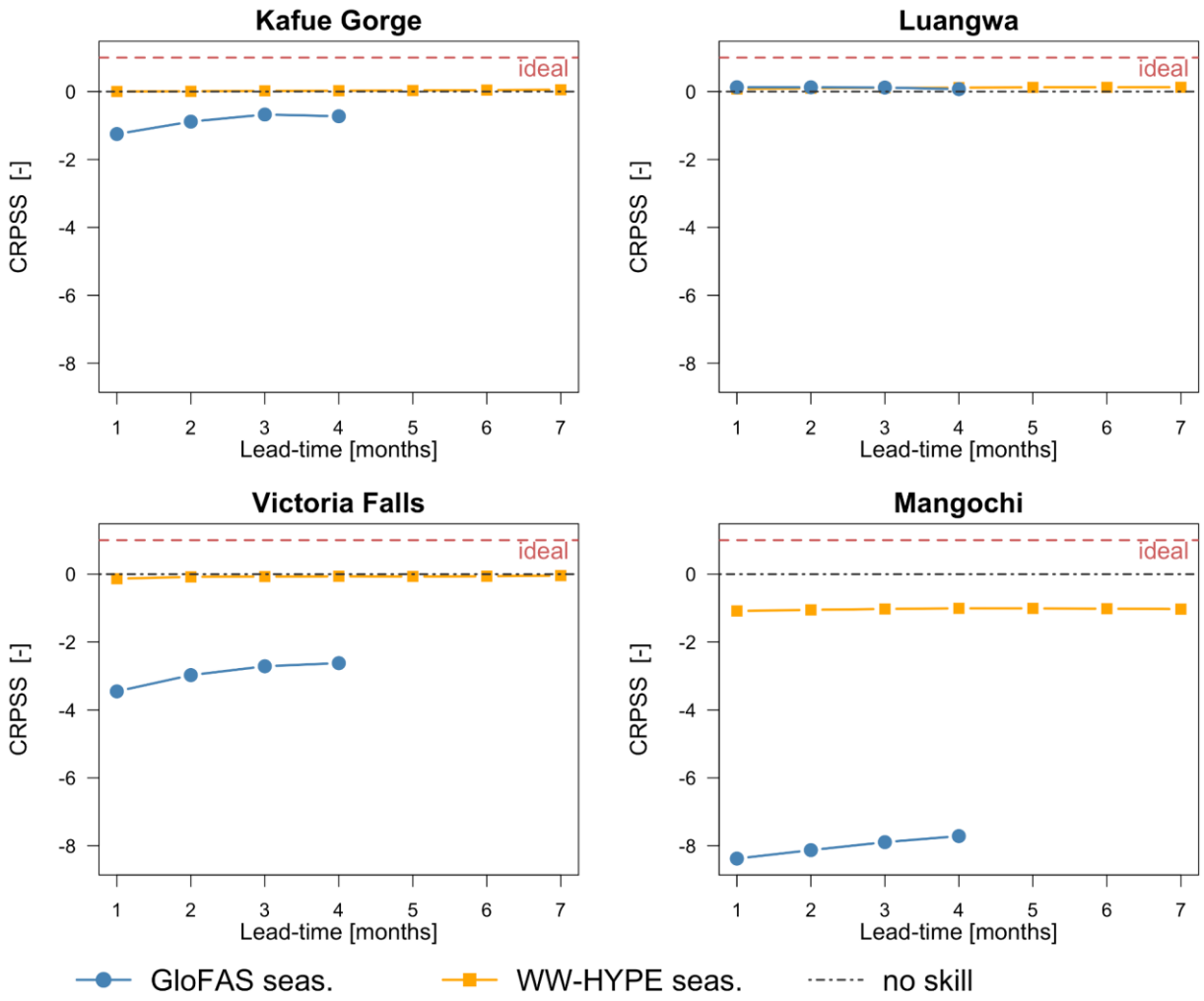

Figure A.4 Continuous Ranked Probability Skill Score (CRPSS) obtained for GloFAS and WW-HYPE seasonal ensemble forecast at the four selected river gauge stations in the Zambezi River Basin, as a function of lead time (up to the maximum available lead time for each product). The optimal value is 1. The reference benchmark is climatology.

Appendix B - Lake Como Basin: Benchmark CS

Here we provide a more complete probabilistic evaluation and skill assessment of the probabilistic benchmark systems considered for the Lake Como Basin, i.e. EFAS sub-seasonal (EFRF) and seasonal (EFSR) ensemble re-forecasts. The probabilistic overall performance is assessed via the Continuous Ranked Probability Score (CRPS) and the Brier Scores (BS), considering all the ensemble members of the EFAS systems (11 members for EFRF and 25 for EFSR). As the CRPS can be easily interpreted also for deterministic forecasts, corresponding to the Mean Absolute Error (MAE), we compare the overall performance of the two EFAS products also with respect to the deterministic forecasts of PROGEA based on the CRPS. Forecast skill is assessed with respect to climatology using the Continuous Ranked Probability Skill Score (CRPSS) and the Brier Skill Scores (BSS).

First, the evolution of the CRPS with lead time (LT) is analysed (see Figure B.1). Results show a small variability of the score with lead time for EFAS forecasts, that are again outperformed by PROGEA's forecasts at short lead times, as expected, given that the CRPS is affected by the forecast bias which is much larger in EFAS (see Figure 4.3). For the seasonal forecasts, despite the erratic behaviour of the CRPS with lead time, it is interesting to note that there is a slight improvement in performance on average for longer lead times (>35 days). The same behaviour can be noted even more clearly for the CRPS as function of the aggregation time (AT; see Figure B.2), where the average error of the probabilistic seasonal forecasts decreases from about 70 m^3/s when the prediction is aggregated at a weekly-scale (or shorter), to about 55 m^3/s with an aggregation over 5 months.

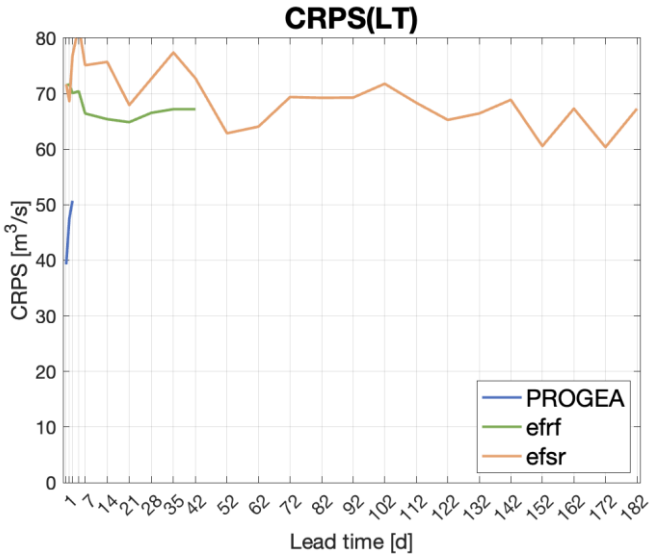

Figure B.1. Continuous Ranked Probability Score (CRPS) obtained for the inflows forecasts from PROGEA's short-term, EFAS sub-seasonal (efrf) and seasonal (efsr) ensemble, as a function of the lead time (LT), up to the maximum available LT for each product. For the CRPS, the optimal value is 0.

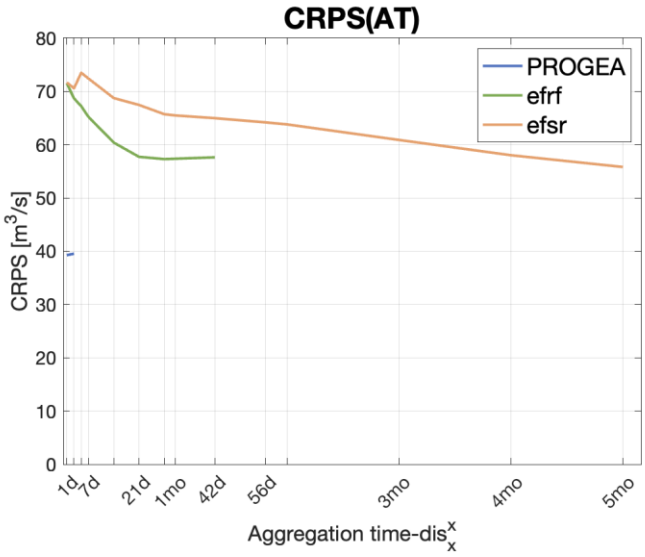

Figure B.2. Continuous Ranked Probability Score (CRPS) obtained for the inflows forecasts from PROGEA's short-term, EFAS sub-seasonal (efrf) and seasonal (efsr) ensemble, as a function of the aggregation time (AT), up to the maximum available LT for each product. For the CRPS, the optimal value is 0.

For the Brier Score, a comparison is carried out between the BS for EFAS sub-seasonal forecasts using thresholds computed with two different methods (to define forecasted events): (i) thresholds computed as percentiles from observations, that we refer to as 'biased', given the known bias in EFAS forecasts with respect to observations; and (ii) thresholds computed as percentiles from forecasts, i.e. 'unbiased' (Figure B.3*Figure*). The difference between the BS from the two methodologies shows that the unbiased version of the score outperforms the biased one. This confirms that biased forecasts can be more informative in terms of EE detection when events are defined based on thresholds that are derived from the biased forecast itself. Thus, the unbiased thresholds are then used to assess the quality of EFAS for drought and flood event prediction at both sub-seasonal and seasonal scales (Figure B.4 and Figure B.5). Results show that the forecast performance does not vary much (and not steadily) with the lead time, with a slightly erratic behaviour of the seasonal forecast scores that can be explained by the sampling of EE that changes with lead-time. EFAS sub-seasonal forecasts perform slightly worse than the seasonal one, in the short-range and up to the Extended Range (ER) for droughts (Figure B.4), but also for floods (to a lower extent) in the short- to medium-range (Figure B.5) which is a surprising finding. This can be explained again by the different sampling of EE of the two reforecast datasets, given the interplay of the lead time with the different forecast update frequency (monthly for EFAS seasonal, twice per week for EFAS sub-seasonal reforecasts) and given the different resulting unbiased thresholds (higher for the sub-seasonal than the seasonal system).

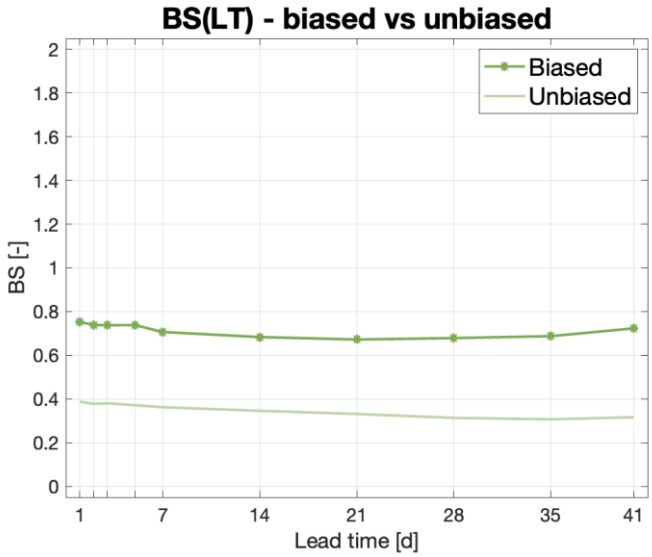

Figure B.3. Brier Score (BS) obtained for the forecast inflows from EFAS sub-seasonal (efrf) ensemble, with the observed (biased) and forecast-based (unbiased) threshold definition, as a function of the lead time from 1 day to 41 days. For the BS, the optimal value is 0. The selected threshold is the 20th percentile of daily flows over the whole period.

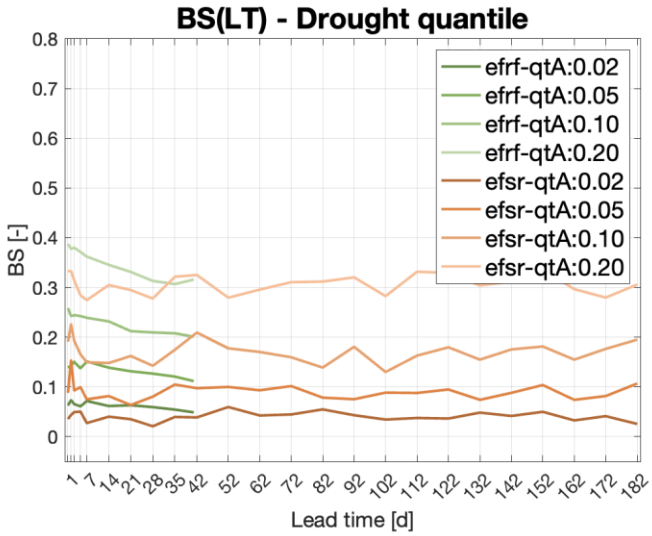

Figure B.4. Brier Score (BS) obtained for the forecast inflows from EFAS sub-seasonal (efrf) and seasonal (efsr) ensemble, for different low-flow thresholds with the unbiased definition, as a function of the lead time (up to the maximum available LT for each product). For the BS, the optimal value is 0. The thresholds are the 2nd, 5th, 10th and 20th-percentile of daily flows over the whole period (as indicated by the quantile number, after 'qtA', in the legend).

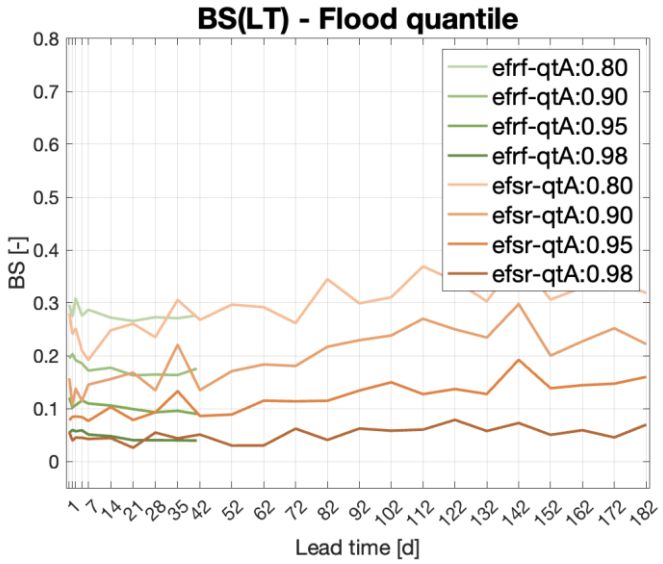

Figure B.5. Brier Score (BS) obtained for the forecast inflows from EFAS sub-seasonal (efrf) and seasonal (efsr) ensemble, for different high-flow thresholds with the unbiased definition, as a function of lead time (up to the maximum available LT for each product). For the BS, the optimal value is 0. The thresholds are the 80th, 90th, 95th and 98th-percentile of daily flows over the whole period (as indicated by the quantile number, after 'qtA', in the legend).

The forecast skill is first assessed by using the Continuous Ranked Probability Skill Score (CRPSS), where the benchmark is the cyclostationary mean (climatology). Again, this score is clearly affected by the forecast bias, which is large in EFAS. Despite this problem, the score is close to zero for both EFAS products (Figure B.6Figure Figure) across lead times, meaning that they are performing similarly to the cyclostationary mean. However, given the known large bias in EFAS, we expect a better correlation and ability in predicting events of the forecast over the climatology benchmark, as the larger bias compensates for this.

The longer the time scale over which we aggregate the observed/forecast inflow, the more predictable it becomes by a simple climatology benchmark. This is reflected by the fact that the cyclostationary mean, which is used as a benchmark in the CRPSS, is gaining more and more power, leading to an apparent decrease in the skill of the forecast (Figure B.7).

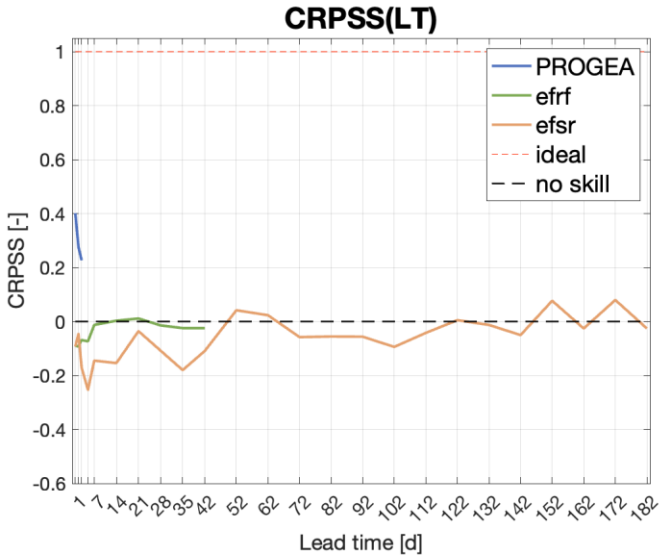

Figure B.6. Continuous Ranked Probability Skill Score (CRPSS) obtained for the inflows forecasts from PROGEA's short-term, EFAS sub-seasonal (efrf) and seasonal (efsr) ensemble, as a function of the lead time (up to the maximum available LT for each product). For the CRPSS, the optimal value is 1.

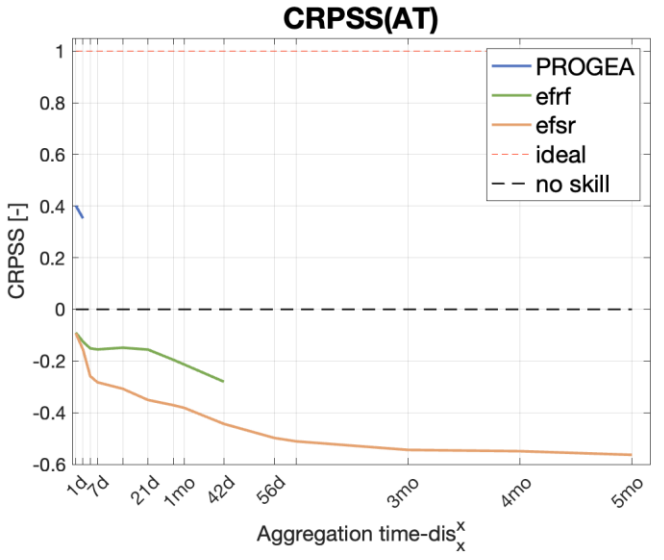

Figure B.7. Continuous Ranked Probability Skill Score (CRPSS) obtained for the inflows forecasts from PROGEA's short-term, EFAS sub-seasonal (efrf) and seasonal (efsr) ensemble, as a function of the aggregation time (up to the maximum available AT for each product). For the CRPSS, the optimal value is 1.

Finally, the Brier Skill Score (BSS) are analysed to evaluate the skill of the forecasts with respect to climatology in detecting drought and flood events, defined by using different inflow thresholds. Figure B.8 shows that for drought detection there is an improvement with respect to climatology for the prediction of droughts of medium to high severity (between 20^{th} to 5^{th} percentiles) for extended range to seasonal aggregation time scales (> 3 weeks). On the other hand, there is no skill

for the most extreme events (2nd percentile of daily flows) for both sub-seasonal and seasonal forecasts, as the skill improves for both systems only for less extreme events (5th percentile or above), especially as the aggregation time increases. Focusing on the class of low-flow events of interest for the lake operator (10th percentile) there is skill with respect to climatology for both sub-seasonal and seasonal forecasts for aggregation time scale of two weeks or longer. A similar improvement with respect to climatology is observed for the prediction of floods of severity between 80th to 98th percentiles (Figure B.9) for time scales longer than two weeks, on average.

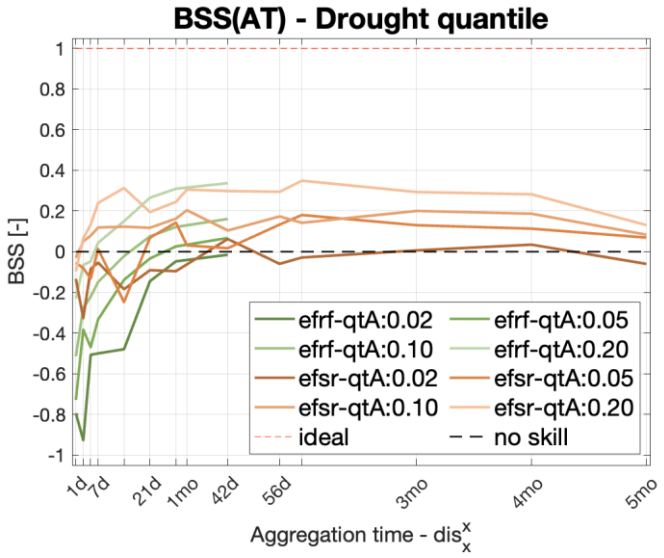

Figure B.8. Brier Skill Score (BSS) obtained for the forecast inflows from EFAS sub-seasonal (efrf) and seasonal (efsr) ensemble, for different low-flow thresholds with the unbiased definition, as a function of the aggregation time (up to the maximum available AT for each product). For the BSS, the optimal value is 1. The thresholds are the 2nd, 5th, 10th and 20th-percentile of daily flows over the whole period (as indicated by the quantile number, after 'qtA', in the legend).

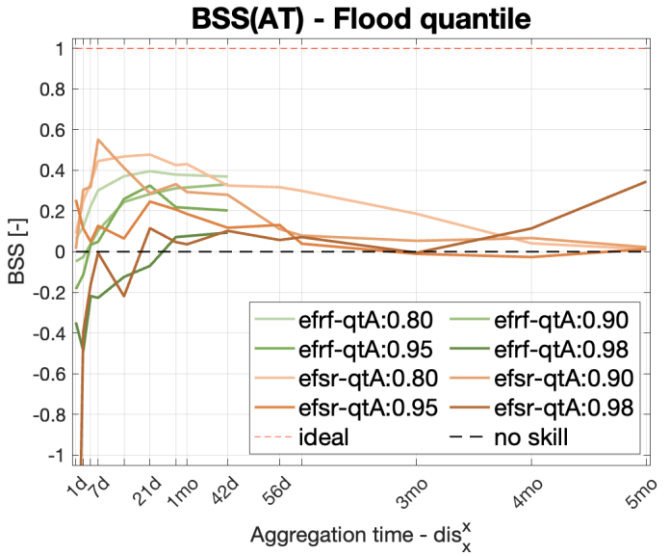

Figure 1 Brier Skill Score (BSS) obtained for the forecast inflows from EFAS sub-seasonal (efrf) and seasonal (efsr) ensemble, for different high-flow thresholds with the unbiased definition, as a function of the aggregation time (up to the maximum available AT for each product). For the BSS, the optimal value is 1. The thresholds are the 80th, 90th, 95th and 98th-percentile of daily flows over the whole period (as indicated by the quantile number, after 'qtA', in the legend).

CLINT

CLIMATE INTELLIGENCE

This project is part of the H2020 Programme supported by the European Union, having received funding from it under Grant Agreement No 101003876