



**I-CISK**  
HUMAN CENTRED CLIMATE SERVICES

## Deliverable D3.2

# **Skill assessment and comparison of state-of-the-art methods for forecasts and projections of extremes**

April 2024





# I-CISK

Innovating Climate services through Integrating Scientific and local Knowledge

Deliverable Title:	DL3.2 Skill assessment and comparison of state-of-the-art methods for forecasts and projections of extremes
Author(s):	Ilias Pechlivanidis (SMHI), Yiheng Du (SMHI), Ilaria Clemenzi (SMHI).
Date	April 2024
Suggested citation:	Pechlivanidis I., Du Y., Clemenzi I. (2024) Skill assessment and comparison of state-of-the-art methods for forecasts and projections of extremes
Availability:	<input checked="" type="checkbox"/> PU: This report is public [Please select] <input type="checkbox"/> CO: Confidential, only for members of the consortium (including the Commission Services)

## Document Revisions:

Author	Revision	Date
Ilias Pechlivanidis, Yiheng Du, Ilaria Clemenzi	First draft	March 2024
Micha Werner, Ilyas Masih	Review	April 2024
Ilias Pechlivanidis, Yiheng Du, Ilaria Clemenzi	Final version	April 2024



## Executive Summary

Climate Services (CSs) have a crucial role in empowering citizens, stakeholders and decision-makers in taking climate-smart decisions that are resilient to climate change and compatible with achieving climate neutrality. The results supported by a scientific evidence base contribute towards a sustainable economy, lifestyle, environmental protection and resource use. CSs aim to transform climate-related data and information into customised products, among others projections, forecasts, information, trends, economic analysis etc., in order to further support adaptation, mitigation and disaster risk management. To achieve this, advanced scientific knowledge, monitoring and modelling of climate change and the impacts of climate extremes are needed. A key barrier that impedes the current generation of CSs achieving the full opportunity of their value-proposition relates to the failure to incorporate the social and behavioural factors and the local knowledge and customs of their users. Additional challenges are in: (i) the understanding of the multi-temporal and multi-scalar dimension of climate-related impacts and actions; (ii) the translation of CS-provided data into actionable information; (iii) the consideration of reinforcing or balancing feedback loops associated to users' decisions based on CSs; (iv) the lack of transdisciplinary approaches across the full CS value chain; and (v) need to deliver tailor-made and robust services at the scale relevant to users.

The I-CISK project aims to seize these untaken opportunities by developing next-generation CSs that follow a social and behaviourally informed approach for co-producing CSs that meet the climate information needs of citizens, decision makers and stakeholders at the spatial and temporal scale relevant to them. In seven geographically diverse living labs (LL), each with different relevant sectors, I-CISK showcases its human-centred co-design, co-creation, co-implementation, and co-evaluation approach across key sectors vulnerable to climate change in Europe and beyond.

This document builds on the previously delivered report of the I-CISK project; D3.1 "Preliminary report on the skill assessment and comparison of state-of-the-art methods for forecasts and projections of extremes" delivered in October 2022. The current report presents new skill assessments from hydrological modelling developments and climate change impact assessments at the I-CISK living lab (LL) scale. The results presented here set new benchmarks for the user-centric climate services within I-CISK.

The document presents two methodologies that aim to better simulate the hydrological conditions and drive long-term decision-making, accounting for impact indicators of hydrological extremes. The report focuses on five European LLs, excluding Lesotho and Georgia, due to geographical domain covered by the hydrological impact model used. In particular, the first methodology applied in three LLs due to data availability aims to enhance the scientific knowledge through artificial intelligence and hybrid hydrological modelling driven by local hydrological data, while the second methodology applied in five LLs accounts for long-term mean changes of extremes (linked to floods and droughts) in different future periods and emission scenarios. These results are of important to the I-CISK climate services, which aim to improve decision-making at different time horizons (from sub-seasonal to centennial).

## Keywords

Climate Services; Artificial Intelligence; Post-processing; Machine Learning; Climate change impacts; Climate Impact Indicators; Extreme events; Local data

## About I-CISK

I-CISK's ambition is to innovate how climate information is used, interpreted and acted on through a next-generation of Climate Services that follow a human centred, social and behaviourally informed approach; integrating the knowledge, needs and perceptions of citizens, decision makers and stakeholders with climate information at spatial and temporal scale relevant to them.

Climate Services (CSs) are crucial to empowering citizens, stakeholders and decision-makers in taking climate-smart decisions that are informed by a solid scientific evidence base, that contribute towards a sustainable European economy, lifestyle, environmental protection and resource use, and that are resilient to climate change and compatible with achieving climate neutrality. European and international collaborative research efforts, including Copernicus and GEOSS have established a solid scientific foundation for an effective CS value chain, including advanced scientific knowledge, monitoring and modelling of climate change and the impacts of climate extremes. However, several barriers challenge the current generation of CSs in achieving the full opportunity of their value-proposition. These challenges include the failure to incorporate the social and behavioural factors and the local knowledge and customs of climate services users. Additionally, the effectiveness of climate services is challenged by; the still poorly developed understanding of the multi-temporal and multi-scalar dimension of climate-related impacts and actions; the translation of CS-provided data into actionable information; consideration of reinforcing or balancing feedback loops associated to users' decisions; and the lack of trans-disciplinary approaches across the full CS value chain.

I-CISK aims to seize these untaken opportunities through a human-centred framework for co-production of next generation CSs that spans the full CS value chain taking the downstream part of the value chain as a starting point. The I-CISK framework realises the full potential of information provided through CSs by empowering actors to take the impacts of extreme climatic events and climate change into account in their decisions.

## Disclaimer

Use of any knowledge, information or data contained in this document shall be at the user's sole risk. Neither the I-CISK consortium nor any of its members, their officers, employees or agents shall be liable or responsible, in negligence or otherwise, for any loss, damage or expense whatever sustained by any person as a result of the use, in any manner or form, of any knowledge, information or data contained in this document, or due to any inaccuracy, omission or error therein contained.

The European Commission shall not in any way be liable or responsible for the use of any such knowledge, information or data, or of the consequences thereof.

This document does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of it.

## Table of Contents

1	Introduction.....	1
1.1	Purpose of this document.....	1
1.2	Structure of this document.....	2
2	Living labs, models and data.....	3
2.1	Summary of CSs for the LLs.....	3
2.2	Description of the hydrological model.....	4
2.3	Historical meteorological forcing.....	5
2.4	Local streamflow observations.....	6
2.5	Simulated centennial projections.....	8
3	Methodology.....	10
3.1	Improving process understanding through hybrid hydrological modelling.....	10
3.1.1	Statistical and ML-based post-processing of hydrological extremes.....	10
3.1.2	Process understanding through machine learning.....	13
3.1.3	Evaluation framework.....	15
3.2	Understanding changes of hydro-climatic extremes under future conditions.....	16
3.2.1	Experimental setup.....	16
3.2.2	Evaluation framework.....	16
4	Results - Improving process understanding at the local scale through hybrid hydrological modelling.....	17
4.1	Improving historical model performance through post-processing.....	17
4.2	Spatial patterns of the post-processing results.....	19
4.3	Performance attribution to basin descriptors.....	22
4.4	Summary of the results.....	25
5	Results - Understanding changes of hydro-climatic extremes under future conditions.....	26
5.1	The Rijnland Living Lab - The Netherlands.....	26
5.2	The Crete Living Lab - Greece.....	28
5.3	The Emilia Romagna Living Lab - Italy.....	31
5.4	The Guadalquivir Living Lab - Spain.....	34
5.5	The Budapest Living Lab - Hungary.....	37
6	Towards the future evolution of the I-CISK climate services.....	42
6.1	Conclusions from state-of-the-art investigations for understanding and predicting extremes.....	42
6.2	Moving beyond the state-of-the-art operational climate services.....	43
	References.....	44

## List of Figures

<b>Figure 2.1:</b> Locations of the I-CISK Living Labs. ....	4
<b>Figure 2.2:</b> (Left) Schematic representation of the processes described in the HYPE model structure. (Right) Domain of the pan-European E-HYPE hydrological model.....	5
<b>Figure 2.3:</b> The HydroGFD system that creates meteorological data for hydrological modelling, tracking both current and historical weather conditions. ....	6
<b>Figure 2.4:</b> Location of the ~2000 streamflow gauges including the length of the observation data: (left) across the pan-European domain, and (right) the stations with available observation data in the Living Labs (the Netherlands, Spain, Italy, Greece and Hungary).....	7
<b>Figure 3.1:</b> Workflow for post-processing of hydrological model outputs. ....	11
<b>Figure 3.2:</b> Conceptualised illustration of the statistical and ML-based post-processing methods: Generalised Linear Model (GLM), Quantile Mapping (QM), Random Forest (RF), and Long Short-Term Memory (LSTM). 12	
<b>Figure 4.1:</b> An example of post-processing results in the station from Netherlands Living Lab. ....	18
<b>Figure 4.2:</b> Scatterplot for the performance achieved after post-processing versus the raw E-HYPE performance: (a) performance for each method at the available streamflow stations in the Italian (IT), the Netherlands (NL) and Spanish (SP) Living Labs; (b) zoomed-in plots for the same metrics in (a) with four post-processing methods in the same plot. Green dotted lines denote the reference performance of a perfect prediction, 0 for SMAE and 1 for NSE and logNSE.....	19
<b>Figure 4.3:</b> Spatial distribution of the raw E-HYPE model performance (SMAE, NSE and log NSE) and the skills achieved after post-processing with the four different methods.....	21
<b>Figure 4.4:</b> Improvements achieved from post processing methods (represented by skills) for regulated/unregulated river systems in the Living Labs. ....	22
<b>Figure 4.5:</b> Key drivers identified for the raw E-HYPE model and post-processing performance based on the SMAE metric. ....	23
<b>Figure 4.6:</b> Key drivers identified for the raw E-HYPE model and post-processing performance based on the NSE and logNSE metrics.....	24
<b>Figure 5.1:</b> Boxplots representing the mean model ensemble duration, number and surplus/deficit volume of the flood (a,b,c) and drought (d,e,f) events in the historical, early, mid and late century for the RCP 2.6, 4.5, 8.5 emission scenarios in the Rijnland, The Netherlands. The boxplots are generated from the values of all the sub-basins in the Living Lab.....	26
<b>Figure 5.2:</b> Change in surplus volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Rijnland, The Netherlands.....	27
<b>Figure 5.3:</b> Change in deficit volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Rijnland, The Netherlands.....	28
<b>Figure 5.4:</b> Boxplot representing the mean model ensemble duration, number and surplus/deficit volume of the flood (a,b,c) and drought (d,e,f) events in the historical, early, mid and late century for the RCP 2.6, 4.5 and 8.5 emission scenarios in the Crete basins. The boxplots are generated from the values of all the sub-basins in the Living Lab. ....	29
<b>Figure 5.5:</b> Change in surplus volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Crete, Greece. ....	30
<b>Figure 5.6:</b> Change in deficit volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Crete, Greece. Basin with no drought events are in dark grey colour. ....	31
<b>Figure 5.7:</b> Boxplot representing the mean model ensemble duration, number and surplus/deficit volume of the flood (a,b,c) and drought (d,e,f) events in the historical, early, mid and late century for the RCP 2.6, 4.5,	

8.5 emission scenarios in Emilia Romagna, Italy. The boxplots are generated from the values of all the sub-basins in the Living Lab..... 32

**Figure 5.8:** Change in surplus volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Emilia Romagna, Italy..... 33

**Figure 5.9:** Change in deficit volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Emilia Romagna, Italy..... 34

**Figure 5.10:** Boxplot representing the mean model ensemble duration, number and surplus/deficit volume of the flood (a,b,c) and drought (d,e,f) events in the historical, early, mid and late century for the RCP 2.6, 4.5, 8.5 emission scenarios in the Upper and Middle Guadalquivir, Spain. The boxplots are generated from the values of all the sub-basins in the Living Lab..... 35

**Figure 5.11:** Change in surplus volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in the Upper and Middle Guadalquivir, Spain. .... 36

**Figure 5.12:** Change in deficit volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in the Upper and Middle Guadalquivir, Spain. .... 37

**Figure 5.13:** Points representing the mean model ensemble duration, number and surplus/deficit volume of the flood (a,b,c) and drought (d,e,f) events in the historical, early, mid and late century for the RCP 2.6, 4.5, 8.5 emission scenarios in Budapest, Hungary. Note that this living lab is covered by the E-HYPE model setup with a single sub-basin. .... 38

**Figure 5.14:** Change in surplus volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Budapest, Hungary. .... 39

**Figure 5.15:** Change in deficit volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Budapest, Hungary. .... 40

## List of Tables

<b>Table 2.1:</b> Information from the available streamflow gauges in each Living Lab. Stations with data length over 10 years are selected for further analysis, where stations with less than 10 years of data are denoted in grey colour.....	7
<b>Table 2.2:</b> The ensemble of Euro-CORDEX projections used to produce hydrological impacts. All members are available for the RCP2.6, 4.5 and 8.5 emission scenario. Note that RCMs have a 0.11 degree (about 12.5 Km) horizontal spatial resolution. The GCM model MPI-ESM-LR has two realisations, r1i1p1 (r1) and r2i1p1 (r2), indicating different initial Earth system states of the GCM scenarios. ....	8
<b>Table 3.1:</b> Sample weights for the LSTM algorithm. ....	13
<b>Table 3.2:</b> Potential drivers considered in this study, including topography, climate, human impact and hydrological regimes. The column “Selected / Replaced by” denotes if the corresponding variable is selected and kept after removing the interdependency (✓), or replaced by other highly correlated variables. ....	14
<b>Table 3.3:</b> The evaluation metrics used to quantify the potential improvements for different characteristics of the streamflow time series.....	15
<b>Table 5.1:</b> Summary of climate change impacts on the different indicators for the five living labs, 4 periods and 3 RCPs (RCP2.6/4.5/8.5). ....	40

# 1 Introduction

## 1.1 Purpose of this document

In the Description of Work of the I-CISK project it is stated that efforts will be targeted towards innovation and enhancement of existing climate services (CSs) and downstream impact-based products, and consequently on the support of decisions and policies in multiple sectors accounting for their local trade-offs. In Work Package (WP) 3, one of the aims is to address the local needs and sectoral gaps of existing CSs and therefore various state-of-the-art methods will be used together with tools/methods to integrate local state-of-the-art observations and local knowledge. Both continental/global and local-scale process-based impact models, e.g. for the water and agriculture sectors, will be used to assess sub-seasonal, seasonal and centennial changes and impacts at the Living Lab (LL) scale. Therefore, a continuous dialogue with various WPs, e.g. WP1, WP2 and WP4, has been established to ensure a continuous exchange and feedback of information required to translate datasets into tailored information and indicators for local use.

The objectives of WP3 are to:

- To advance local impact predictions and projections of climate change and future extremes through developing modelling chains that efficiently integrate existing CSs while also combining local data and knowledge for local tailoring.
- To explore different scientific state-of-the-art methods to bridge data and services that are currently separated on temporal and spatial scales (from forecasts to projections) and increase the trust in local predictions.
- To evaluate the usefulness of the integrated impact predictions and assessments for local operations and decision-making from both a scientific and a user perspective.
- To unlock the benefits of transformation of data to information for and within the climate-sensitive LL regions and sectors by improving the confidence information of indicators while enhancing their usability.
- To develop user-driven visualisation tools that assure robust and seamless transfer of produced information from CSs, and communicate predictions, explicitly including uncertainty, for guided decision-making.
- To provide recommendations for product adaptations, extensions and CS improvements, and deliver fit for purpose tools, methods and products for user-tailored real-time operational services

To achieve some of the objectives listed above, this document presents the recent state-of-the-art scientific efforts with continental hydrological models and reports on the progress in WP3, while it addresses a series of specific objectives that include:

- Enhance the quality of streamflow simulations derived from the continental hydrological model through post-processing at the regional and local scale of the living labs, thereby enhancing the applicability of the impact model, and hence service, for local decision-making processes.
- Identify the drivers that are linked to the quality of the statistical and machine-learning based post-processors, and diagnose underlying similarities between living lab applications.
- Improve understanding of the local impact of climate change under different emission scenarios and future periods.
- Derive and quantify impact indicators for hydrological extremes, and assess their changes, accounting for uncertainty coming from the climate models.

## 1.2 Structure of this document

The deliverable is structured in six chapters:

- **Chapter 1** (current) is the introduction to the document presenting the scope.
- **Chapter 2** describes the living labs, the impact models and available data for historical assessments and future projections.
- **Chapter 3** presents the methodology for hybrid hydrological modelling and quantification of future changes of hydro-climatic extremes.
- **Chapter 4** presents the results of the hybrid hydrological modelling and shows the drivers that are linked with the quality of the post-processing at local conditions.
- **Chapter 5** presents the changes of hydro-climatic extremes under future conditions at the living lab scale.
- **Chapter 6** concludes the main body of the document and summarises future work towards operationalisation.

## 2 Living labs, models and data

### 2.1 Summary of CSs for the LLs

Here we aim to provide a better understanding of the role and efficacy of living labs in integrating and utilising climate services to meet user requirements. This inquiry is rooted in the research activities of Work Package (WP) 1 and 2 of the project, which also carries significance for WP3 as it delves into diverse scientific approaches for tailoring to user needs at the scale of the living labs. As described in Deliverable D3.1 (a preliminary first version of this report), the process of delineating the project's scope, through discussions, the formation of I-CISK Living Labs (LLs), and the subsequent targeted surveys and interviews, has yielded an initial assessment of CS utilisation and needs within the LLs. This encompasses an understanding of decision-making processes, challenges in leveraging existing CS, and the demand for enhanced CS solutions (as documented in Deliverable D2.1). Spanning the Netherlands, Spain, Italy, Greece, Hungary, Georgia, and Lesotho, the seven LLs represent a spectrum of climate-related challenges. Within WP2, a comprehensive collection of information was facilitated on decision-making practices, CS utilisation, and user needs, despite the variability in the development phases of the LLs. A pivotal aspect of I-CISK involves the iterative co-exploration with stakeholders/users regarding the value of CSs, climatic data, and insights, which is crucial for designing and advancing CSs. This exploration aims to deepen our understanding of the CS end-users' decision-making contexts, identify barriers to existing CS utilisation, and uncover ways to overcome these barriers in crafting next-generation CSs that are not only useful but also user-friendly and effectively meet user requirements.

Feedback from WP2's interviews and questionnaires has highlighted several motivators for enhancing CSs and their application. These include the significance of fostering preparedness and adaptation strategies, mitigating the risks of climatic hazards and extreme events, preventing conflicts over water resources, reducing the impact of such events across sectors, simplifying access to existing information, supporting policy-making, and encouraging stakeholders to adopt a proactive stance in decision-making through the direct benefits of CSs (Moschini, Emerton et al., 2022). Current challenges in CSs adoption involve issues with resolution (both temporal and spatial), accessibility (such as difficulties in data acquisition and distribution to targeted audiences), and a lack of impact variables critical for informed decision-making. The demands from LLs vary, encompassing the enhancement of CSs availability across different timescales, provision of spatial resolutions pertinent to decision-making needs, and the development of sector-specific CSs offering impact-based forecasts and additional essential variables.

Deliverable D2.1 (Tables 2 to 7; Moschini, Emerton et al., 2022) sheds light on the CSs available and utilised within the seven LLs. It remains uncertain if all identified CSs are actively employed in decision-making processes, and there might be other CSs that have not been integrated into these processes or captured through the conducted surveys/interviews. Most CSs identified depend on national and regional/local sources, such as hydro-meteorological services or environmental agencies. The availability of large-scale or international CSs, including seasonal outlooks from regional authorities, forecasts from ECMWF, and insights from the Copernicus Land Monitoring Service, prompts further exploration into the potential for incorporating global CSs alongside local sources to address the highlighted challenges and to tailor CSs to meet local needs more effectively. Hence there is potential to explore how such large-scale CS can be adapted to the local scales.

Deliverable D3.1 presents an evaluation of seasonal hydro-meteorological forecasts (up to 6 months ahead) tailored to each LL's scale, including their overall accuracy and the predictability of extreme events. These evaluations serve as benchmarks for ongoing scientific efforts to customise state-of-the-art CSs to local user needs. The current deliverable is instead focusing on 2 pillars:

- Application of post-processing methods rooted in machine learning and statistical approaches to enhance the quality of the modelled streamflow at the local conditions using local data.

- Quantification of climate change impacts on hydrological extremes using state-of-the-art climate projections, accounting for different emission scenarios and future periods.

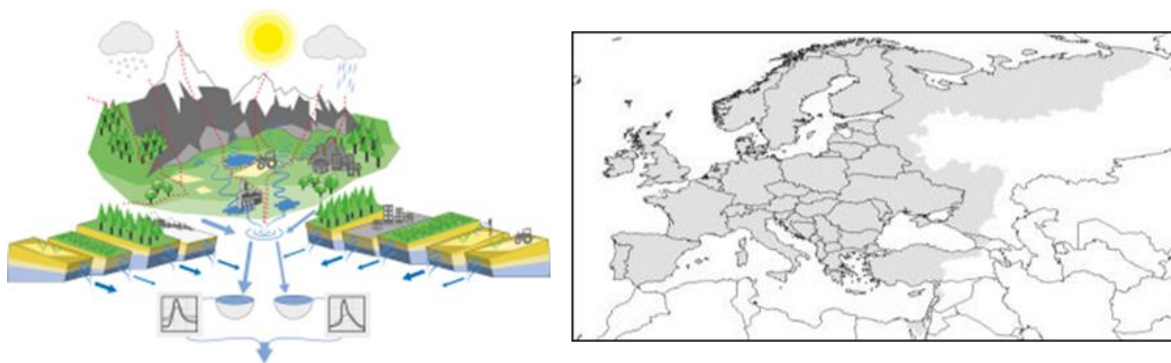
The report focuses on five European LLs, the Netherlands, Spain, Italy, Greece, Hungary (excluding Lesotho and Georgia), due to geographical domain covered by the hydrological impact model.



**Figure 2.1:** Locations of the I-CISK Living Labs.

## 2.2 Description of the hydrological model

The Hydrological Predictions for the Environment (HYPE) is a semi-distributed process-based model capable of simulating the hydrological processes from a single basin to global scale (Figure 2.2). The model has conceptual routines for most of the major land surface and subsurface processes. Snow accumulation and melt processes are modelled using the degree-day method with land use dependent parameters. HYPE simulates the water flow paths in soil, which is divided into three layers with a fluctuating groundwater table. A fraction of rainfall or snowmelt infiltrates into the topsoil, which is limited by a soil-type dependent maximum rate. If the soil moisture in the upper soil layer exceeds a threshold for macropore flow, part of the remaining water forms macropore flow. Potential evaporation (PET) is estimated using the modified Jensen-Haise model (Oudin et al., 2005), whilst PET is achieved only if either the actual soil moisture exceeds a large portion of the soil field capacity or the sub basin is defined as a waterbody. For soil moisture below this limit in non-waterbody areas, the actual evaporation, computed using the crop coefficient method in Allen et al. (1998), decreases linearly to zero at the wilting point. Runoff from the soil zone is computed when the soil moisture exceeds field capacity and it percolates from upper to lower soil layers when the soil moisture in the upper layers exceeds field capacity. The ground water level is estimated based on the level in the soil zone where the pore space is filled.



**Figure 2.2:** (Left) Schematic representation of the processes described in the HYPE model structure. (Right) Domain of the pan-European E-HYPE hydrological model.

In this deliverable, the pan-European E-HYPE hydrological setup is applied. The HYPE model is configured on a continental scale, covering the entire pan-European area of 8.8 million km<sup>2</sup> (Figure 2.2 right). This configuration is known as E-HYPE and divides the region into approximately 35,400 sub-basins, averaging 215 km<sup>2</sup> each, while it operates on a daily time-step (Hundecha et al., 2016). This setup (version 3.0) utilises open data, incorporating 8 soil types and 15 land use categories to define up to 75 Hydrological Response Units (HRUs). The model's construction leveraged a variety of open data sources for continental and global analysis, and details can be found in Hundecha et al., (2016). River networks and sub-catchments were outlined using WWF's Hydrosheds data (Lehner et al., 2008), while HRUs were established from land use and soil information from several databases. CORINE provided land use data, while lakes and reservoirs data were sourced from the GLWD (Global Lakes and Wetlands Database; Lehner and Döll, 2004) and Grand (Global Reservoir and Dam; Lehner et al., 2011) databases. Irrigated areas were identified using GMIA (Global Map of Irrigation Areas; Siebert et al., 2005; 2010) and MIRCA (Monthly Irrigated and Rainfed Crop Areas; Portmann et al., 2010) datasets, and soil types were derived from HWSD (Harmonised World Soil Database; Nachtergaele et al., 2012).

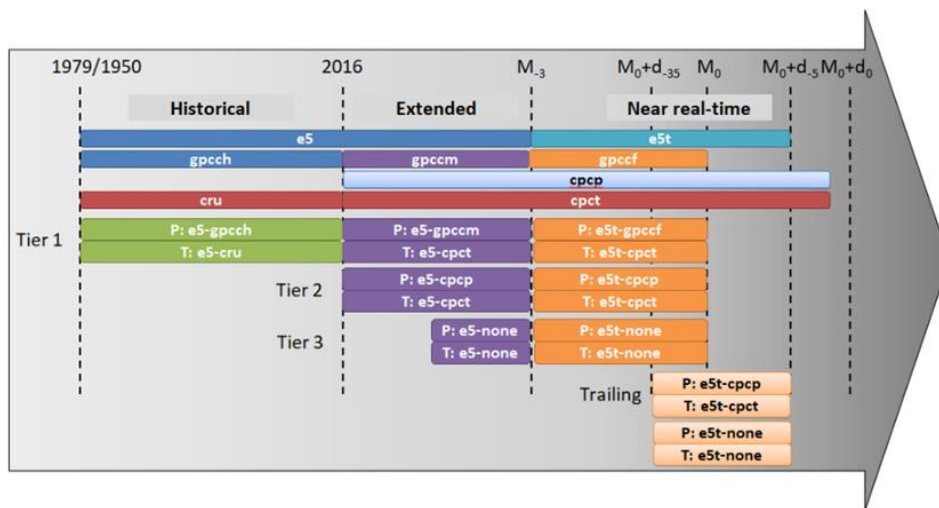
In I-CISK, the Living Labs that lie in the European domain have been investigated with the E-HYPE hydrological model, while those outside the European domain (Georgia and Lesotho) have been investigated with the World-Wide HYPE hydrological model (WWH; Arheimer et al., 2020). In this report, the investigation is done with the E-HYPE configurations.

### 2.3 Historical meteorological forcing

The E-HYPE hydrological model relies on the HydroGFD v3.0 meteorological dataset (Berg et al., 2021) as the forcing data to simulate soil, lake, and river conditions during both the spin-up phase and historical simulations. Due to the scarcity of daily meteorological data at large scales, this challenge has been addressed by adjusting reanalysis data with monthly global observations to ensure the monthly water balance aligns with observed data, while short-term processes are predicted by meteorological models (Weedon et al., 2014). However, such datasets are typically static, covering only historical periods until updated. Simulated streamflow from E-HYPE was obtained for the period 1961-2023.

In particular, the HydroGFD methodology, developed and maintained by the Swedish Meteorological and Hydrological Institute (SMHI), synergizes multiple global data streams to facilitate updates of meteorological information in near real-time (Figure 2.3). This operational framework takes advantage of global scale models from ECMWF with precipitation observations from the GPCP and temperature observations from both CRU and CPC. The integration of these data sources is dynamic, leading to the generation of HydroGFD over time. Specifically, the historical segment of the HydroGFD dataset, spanning the years 1979 to 2016, adjusts the

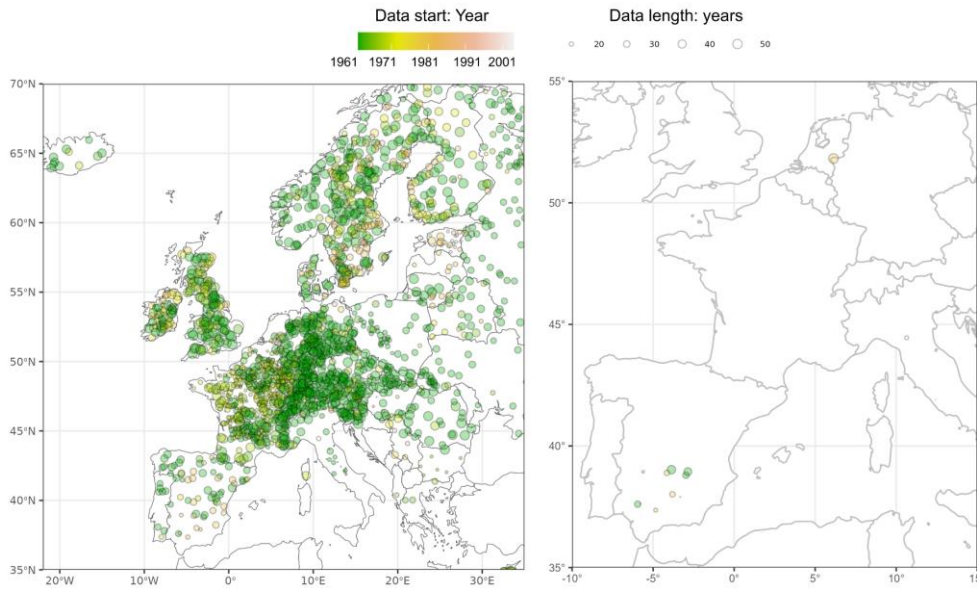
ERA5 reanalysis data by incorporating precipitation information from the GPCC and temperature data from the CRU dataset. For more recent segments, extending to three months prior to today, HydroGFD-Extended refreshes its database with the latest observations from the GPCC, CRU, and CPC. The most up-to-date dataset, covering the period from the past two to three months, is available through the HydroGFD-Near Real-time offering, which leverages forecast data from ERA5t, introduced with a lag of five days. This system delivers data at a spatial resolution of 0.25 degrees, equivalent to approximately 25 km. The HydroGFD-Historical and HydroGFD-Extended versions produce data outputs every three hours, while the HydroGFD-NearRealtime version does so every six hours. Derived from these outputs are daily cumulative values for precipitation and daily averages of other essential climate variables, including minimum and maximum temperatures. HydroGFD provides meteorological information every day across the world which makes it useful in operational settings.



**Figure 2.3:** The HydroGFD system that creates meteorological data for hydrological modelling, tracking both current and historical weather conditions.

## 2.4 Local streamflow observations

Streamflow observations have been collected across the pan-European domain from various data sources, including GRDC, EWA, SMHI, BHDC, Spanish and Italian authorities. Detailed information can be found in Hundedcha et al 2016. Around 2000 stations were selected based on data availability with at least 10 years of data, which is divided into training and testing periods with 80-20% split. Figure 2.4 illustrates the comprehensive spatial coverage of the stations across the entire European domain, with a higher concentration in central Europe and relatively fewer stations in the southern (e.g. Spain) and the eastern part of the continent. We note that this map does not represent the actual stations operated by the national or regional hydro-met services, but rather the stations that we had access to.



**Figure 2.4:** Location of the ~2000 streamflow gauges including the length of the observation data: (left) across the pan-European domain, and (right) the stations with available observation data in the Living Labs (the Netherlands, Spain, Italy, Greece and Hungary).

Extended observation data were provided from the I-CISK Living Labs, with daily streamflow observation data at Secchia (Italy, EHYPE SUBID 9780124) and Lobith (the Netherlands, EHYPE SUBID 9503374) and at different locations in Spain. The station locations were mapped to catchments in E-HYPE, and corresponding catchment simulation was extracted by the corresponding SUBID (SUBID is the unique ID of the sub-catchment in the E-HYPE model setup). A summary of the stations in the Living Labs is provided below, including the data length and starting year of the observations (Table 2.1). Some stations were excluded from this analysis due to the lack of streamflow observations which can further be used for local investigations as shown in Table 2.1 with grey background color.

**Table 2.1:** Information from the available streamflow gauges in each Living Lab. Stations with data length over 10 years are selected for further analysis, where stations with less than 10 years of data are denoted in grey colour.

#	Living Lab	Start year	Data length (years)	E-HYPE-based SUBID	Regulated
1	The Netherlands	1980	43	9503374	Yes
2	Italy	2003	20	9780124	No
3	Spain	1980	20	9559682	No
4	Spain	1980	17	9558645	No
5	Spain	1980	24	9559017	---
6	Spain	1961	17	9559950	Yes
7	Spain	1961	27	9560130	Yes
8	Spain	1961	27	9558813	No

9	Spain	1961	33	9558677	No
10	Spain	1980	18	9559312	---
11	Spain	1961	36	9558704	Yes
12	Spain	1980	24	9558946	Yes
13	Italy	1961	7	9780453	---
14	Spain	1975	8	9558040	---
15	Spain	1975	8	9558359	---
16	Spain	1975	8	9757777	---
17	Spain	1977	6	9558591	---
18	Spain	1977	6	9000898	---
	Hungary LL	---	---		---
	Greece LL	---	---		---

\* --- denotes no data available.

## 2.5 Simulated centennial projections

In addition to the historical investigation, WP3 aims to better understand the impacts of climate change at the local scale. To do so, daily precipitation and air temperature projections till the end of the century for the five Living Labs were additionally used. The projections were obtained from nine regional climate models that contribute to the Euro-CORDEX model ensemble (Berg et al., 2021). The ensemble consisted of three emission scenarios, three global circulation models (GCMs) and five regional climate models (RCMs); see Table 1. A total of 130 years of hydrological simulations have been conducted for each climate scenario (1971–2100). However, here, we split the dataset into distinct periods over which users and stakeholders commonly based their long-term decision and/or policy is made. This included a historical period (1971–2000) and three future periods: 2011–2040, 2041–2070, and 2071–2100 to represent the early, mid and late century periods, respectively. The simulations considered three representative concentration pathways (RCPs) - low, medium, and high emission scenarios (RCP2.6, 4.5, and 8.5 respectively). It is important to note that the ensemble does not cover all sources of uncertainty in the modelling chain. For instance, the EC-EARTH GCM and the CCLM4-8-17 (or HIRHAM5) RCM are over- and under-represented within the ensemble, respectively, yet we assume that all available climate models are equally probable in projecting future conditions and are free of systematic errors with acceptable historical performance to be accounted for in the ensemble.

**Table 2.2:** The ensemble of Euro-CORDEX projections used to produce hydrological impacts. All members are available for the RCP2.6, 4.5 and 8.5 emission scenario. Note that RCMs have a 0.11 degree (about 12.5 Km) horizontal spatial resolution. The GCM model MPI-ESM-LR has two realisations, r1i1p1 (r1) and r2i1p1 (r2), indicating different initial Earth system states of the GCM scenarios.

ID	GCM	RCM	Abbreviation
1	EC-EARTH	CCLM4-8-17	EC-EARTH+CCLM4-8-17

2		RACMO22E	EC-EARTH+RACMO22E
3		RCA4	EC-EARTH+RCA4
4		HIRHAM5	EC-EARTH+HIRHAM5
5	HadGEM2-ES	RACMO22E	HadGEM2-ES+RACMO22E
6		RCA4	HadGEM2-ES+RCA4
7	MPI-ESM-LR	RCA4	MPI-ESM-LR+RCA4
8		REMO2009	MPI-ESM-LR+REMO2009 r1
9		REMO2009	MPI-ESM-LR+REMO2009 r2

The precipitation and temperature projections were also bias-adjusted using a quantile mapping of two timescales. Bias-adjustment was conducted on the projections using a reference dataset (here the EFAS-Meteo gridded observational dataset). After bias-adjustment, the cumulative distribution of daily precipitation and temperature projections follows closely the one of the reference dataset. The nine bias-adjusted precipitation and temperature simulations were used to force the hydrological model E-HYPE and obtain daily streamflow simulations in the historical and the three future periods (Berg et al., 2021b). The streamflow simulations were used to identify and statistically characterise the extreme events (floods and droughts) in the Living Labs and their changes under future conditions (see section 3.2).

## 3 Methodology

### 3.1 Improving process understanding through hybrid hydrological modelling

#### 3.1.1 Statistical and ML-based post-processing of hydrological extremes

A major challenge in improving continental and global hydro-climate services is the insufficient integration of local knowledge and end-user datasets, which requires, as one way to move forward, application of advanced post-processing techniques. Although the process-based E-HYPE model captures the hydrological dynamics in the pan-European domain well, residuals remain between model simulations and local observations, with the residuals increasing particularly for the extremes. Recent advances have demonstrated the effectiveness of statistical and machine learning (ML) methods in increasing the accuracy and reliability of hydrological models (Slater et al., 2023), ensuring that “new” outputs can more accurately reflect specific local hydrological dynamics.

In this deliverable, we therefore investigate four post-processing strategies aimed at refining streamflow predictions. We applied different techniques (statistical and ML-based) to adjust streamflow simulations derived from the E-HYPE hydrological model to match local observations in the Living Labs. Our aim has been to improve the accuracy of volume estimates and the characterisation of hydrological extremes (accounting for floods and droughts), thereby bridging the gap between generic model outputs and local hydrological realities.

#### *Post-processing workflow*

In Figure 3.1, we illustrate the structure of our post-processing framework, which is applied to add value to the results from the process-based E-HYPE model. To address the discrepancies between the model outputs and the observations, the post-processing applies statistical and ML-based algorithms to reduce these residuals. Here, we used four methods, two statistical (Generalised Linear Model (GLM), Quantile Mapping (QM)) and two ML-based (Random Forest (RF) and Long Short-Term Memory networks (LSTM)). We assess the performance of these post-processors through various metrics focusing on both volume and extremes of the streamflow signal.

Moreover, we explore the importance of different features by examining how the post-processors' performance correlates with local and regional factors, such as climatology, topography, human activity, and hydrological regimes. These results will help us to build the fundamental knowledge for the next step, which is to incorporate the post-processing into the pre-operational climate services established in the I-CISK project.

The analysis in this deliverable focuses on implementing, evaluating and understanding the post-processing methods with the raw model. The evaluation is conducted at the Living Lab level, using the data from the LL stations. The understanding analysis of potential drivers is carried out at the pan European domain, in order to include various hydrological and climate conditions, ensuring a comprehensive conclusion.

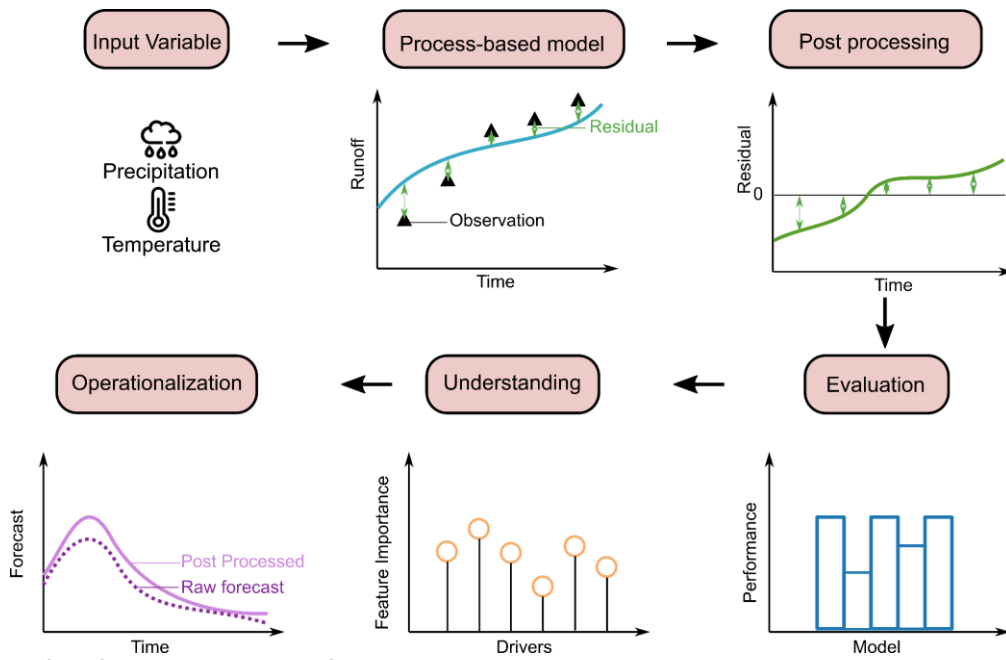


Figure 3.1: Workflow for post-processing of hydrological model outputs.

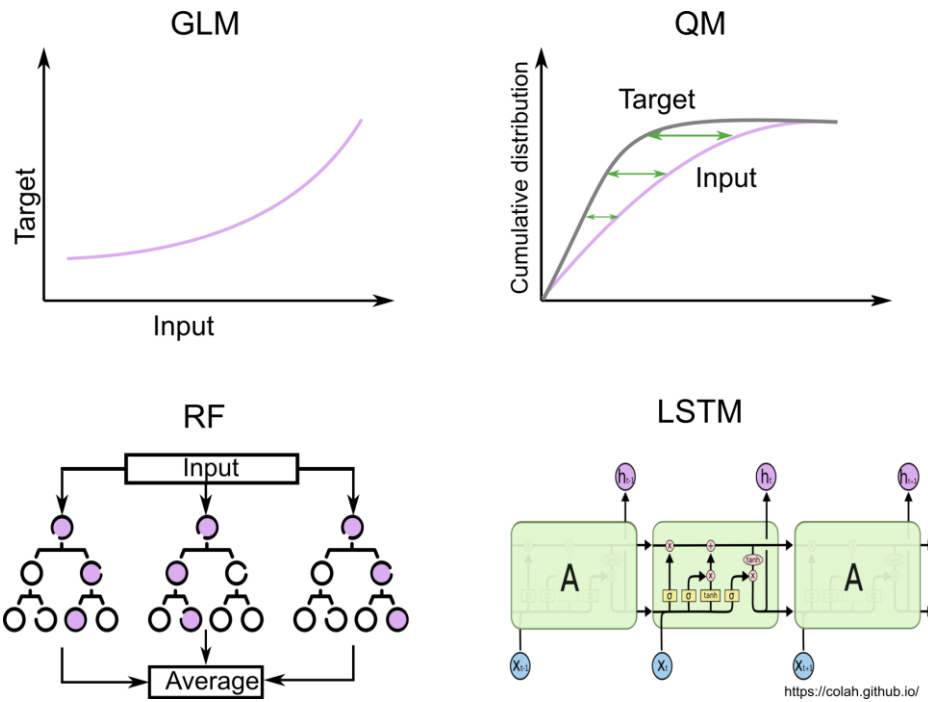
### Post-processing methods

The post-processing is based on four methods which are conceptually illustrated in Figure 3.2.

**Generalised Linear Model (GLM):** a statistical technique that extends linear regression to allow for non-normal distributions of error terms (Madsen and Thyregod, 2010). It allows the inclusion of different types of predictor variables and the modelling of response variables that follow non-normal distributions, such as Gaussian, to provide a flexible framework for understanding the relationships between variables.

**Quantile Mapping (QM):** a statistical technique used for bias correction by adjusting the distribution of the variable of interest (here simulated streamflow) to match the target variable (here observed streamflow) distribution, in order to correct systematic biases in model outputs (Gudmundsson et al., 2012). This method is particularly useful for improving the accuracy of hydrological predictions, as it ensures that the corrected model output maintains the statistical properties of the observed data across the entire distribution. The tricubic spline method is adopted here to allow for a smooth adjustment of the cumulative distribution functions, to improve the addressing of biases in the tails of the distribution.

**Random Forest (RF):** a supervised, non-parametric algorithm, where an ensemble of uncorrelated trees yields prediction for classification or regression. Multiple trees are built based on bootstrapping samples from the training data. In the case of regression, after all the trees are grown, the forests produce the final results by averaging predictions from the trees (Pham et al., 2021). The randomForest package in R was used for model training and testing. The same model configuration, regarding maximum node numbers (10) and minimum node size, is maintained across all stations. This ensures comparability throughout the study domain, allowing the analysis of potential influencing factors.



**Figure 3.2:** Conceptualised illustration of the statistical and ML-based post-processing methods: Generalised Linear Model (GLM), Quantile Mapping (QM), Random Forest (RF), and Long Short-Term Memory (LSTM).

**Long Short-Term Memory (LSTM):** a state-of-the-art model for time series, which is capable of learning long-term dependencies (Kratzert et al., 2018). For post-processing purposes, the LSTM in our framework is specifically designed with a 3-day lookback period, which has confirmed its capability of capturing temporal dependencies present in streamflow data by initially experimenting values between 1 to 215 lookback days. This model is structured with three layers containing different numbers of cells (i.e. 100-50-20), allowing an effective process and remembering information over extended periods.

For model training, the dataset was subsequently divided into training and testing periods, by applying an 80-20% data split. The model evaluation in the results section is conducted on the testing periods. To ensure the model's potential for generalisation and prevent overfitting, a portion of the training set (10%) is reserved as a validation set. The model training includes a monitoring mechanism where if the validation loss does not decrease over 10 consecutive steps, an early stopping criterion is triggered. This strategy ensures the model against overfitting by interrupting the training process when no further improvement is observed in the validation dataset, thereby allowing the model's performance to be optimised without compromising its ability to generalise to new, unseen data.

Normalisation is applied to the input data to scale the range of data points, facilitating smoother training process and more stable convergence. The target variable, representing the relative residual between observed and simulated streamflow, is calculated as below:

$$target = (y_{obs} - y_{sim}) / (y_{sim} + \epsilon) + 1$$

where  $\epsilon$  is a small constant introduced to prevent division by zero, particularly in scenarios of low streamflow, ensuring the target variable remains within a reasonable range.

To address data imbalances, particularly concerning extreme values critical for hydrological services, a sample weight technique is implemented. This method assigns weights to samples, emphasising the importance of accurately predicting extreme events, which are often underrepresented in the dataset but hold significant

importance for hydrological analyses and applications. Through this weighted approach, the model is better equipped to focus on and learn from these extremes. Weights are assigned by percentiles in the observations as in Table 3.1, where the 10<sup>th</sup>, 33<sup>rd</sup>, 66<sup>th</sup> and 90<sup>th</sup> percentiles were included for dividing the groups, representing low, lower than normal, higher than normal, and high extremes.

**Table 3.1:** Sample weights for the LSTM algorithm.

Range	Weight
> 90 <sup>th</sup>	0.2
66 <sup>th</sup> to 90 <sup>th</sup>	0.2
33 <sup>rd</sup> to 66 <sup>th</sup>	0.2
10 <sup>th</sup> to 33 <sup>rd</sup>	0.2
< 10 <sup>th</sup>	0.2

### 3.1.2 Process understanding through machine learning

Building on the evaluation of hydrological modelling and post-processing, it is also essential to understand potential drivers that influence the performance of this hybrid hydrological modelling, in order to provide enhanced CSs for local conditions. ML techniques have been proven to have good capability of identifying nonlinear and complex relationships between influencing factors and the target variables. Here, identifying key drivers that affect the performance of the E-HYPE model across Europe will assist the next step of incorporating the hybrid modelling into operations. Therefore, we adopted the Classification And Regression Trees (CART) method to assist in diagnosing the key drivers that can influence the model performance in terms of volume and high and low streamflow.

#### *Classification and Regression Trees (CART)*

CART is a non-parametric decision tree learning technique that models the prediction of a target variable by recursively partitioning the dataset and fitting a simple model within each partition (Breiman et al., 2017). Here, CART is used to identify the most important predictors of model performance and to model the complex, non-linear relationships between them. The algorithm splits the data into subsets based on the values of the input features that result in the largest reduction in heterogeneity of the target variable. This process continues until further splitting does not significantly improve the model's accuracy or until predefined stopping criteria are met, such as a minimum number of observations in each leaf of the tree. To avoid overfitting, the technique of pruning is used by removing branches that have little to no contribution to the model's predictive power, aiming to find the optimal balance between the tree's complexity and its accuracy on a validation set.

Next the descriptors' importance is calculated by summing changes in the probability of splitting on every descriptor and dividing the sum by the number of branch nodes (Pechlivanidis et al., 2020). This importance score is then standardised, spanning from 0 to 100 for comparability. The association between model performance and potential drivers were investigated using the method above, by calculating the feature importance of each potential driver. The considered potential drivers are listed in Table 3.2, including topography, climate, human impact and hydrological regimes. The hydrological regimes are described here by clusters of hydrologically similar responses previously identified across the pan-European. Details can be found in Pechlivanidis et al. (2020). We further note that some drivers are highly interdependent and could therefore introduce uncertainty to the CART analysis. Therefore, the highly interdependent drivers (with Pearson correlation coefficient greater than 0.6) were removed, and finally 8 potential drivers were kept for the CART analysis, as highlighted in italics in Table 3.2.

**Table 3.2:** Potential drivers considered in this study, including topography, climate, human impact and hydrological regimes. The column “Selected / Replaced by” denotes if the corresponding variable is selected and kept after removing the interdependency (✓), or replaced by other highly correlated variables.

Name	Abbreviation	Unit	Selected / Replaced by (->)
<i>Precipitation</i>	<i>Prec</i>	<i>mm</i>	✓
<i>Temperature</i>	<i>Temp</i>	°C	✓
Snow depth	Snow	cm	-> Temperature, Precipitation
Actual evapotranspiration	AET	mm	-> Temperature
Potential evapotranspiration	PET	mm	-> Temperature
Dryness index	PET/Prec	--	-> Temperature
<i>Evaporative index</i>	<i>AET/Prec</i>	--	✓
<i>Upstream Area</i>	<i>Area</i>	<i>km<sup>2</sup></i>	✓
<i>Elevation</i>	<i>Elev</i>	<i>m</i>	✓
<i>Relief ratio</i>	<i>Relief</i>	--	✓
Slope	Slope	%	-> Precipitation
<i>Degree of Regulation</i>	<i>DoR</i>	%	✓
<i>Hydrological Clusters</i>	<i>Cluster</i>	--	✓

### Comprehensive Rank Index (RI)

By applying the concept of feature importance, a comprehensive ranking index, as defined by Jiang et al. (2015), enables the evaluation and comparison of potential drivers' influence across various models. This ranking index (RI) is mathematically expressed as:

$$RI = 1 - \frac{1}{nm} \sum_{i=1}^n rank_i,$$

where  $m$  represents the total number of potential drivers, which in this study is 8, and  $n$  denotes the number of models, set at 5 (raw model and four post-processing methods) for this analysis.  $rank_i$  indicates the assigned rank of each potential driver, with 1 being the most critical and 8 the least. Thus, an RI value approaching 1 signals a more accurate and effective simulation outcome.

With RI, the analysis identifies the three most influential drivers across both the unprocessed model and the various post-processing methods. This approach can reveal the underlying drivers of the model performance and provide information on where post-processing methods can significantly refine the model's accuracy.

3.1.3 Evaluation framework

To evaluate the added value from post-processing, three evaluation metrics were used to assess the potential improvements with regard to errors in volume, high and low streamflow (Table 3.3).

Mean absolute error (MAE) calculates the average magnitude of errors in a set of predictions, without considering their direction (Willmott and Matsuura, 2005). MAE is a linear score which means that all individual differences are weighted equally in the average. The Scaled Mean Absolute Error (SMAE) serves to adjust MAE in relation to the average streamflow observed at each station, thus allowing the comparison of MAE values across stations that have varying streamflow magnitudes.

Nash-Sutcliffe Efficiency (NSE) is a normalised statistic that determines the relative magnitude of the residual variance ("noise") compared to the measured data variance ("signal") (Nash and Sutcliffe, 1970). NSE values range from  $-\infty$  to 1, where 1 indicates perfect model prediction. NSE has been widely used in hydrological modelling to assess how well a model can replicate high streamflow peaks, and hence flooding.

The logNSE modifies the traditional NSE to emphasise the performance of a model in predicting low streamflow conditions (Lamontagne et al., 2020). By using the logarithm of both observed and simulated values before calculating efficiency, logNSE sets particular sensitivity to differences in low streamflow, making it valuable for assessing model's adequacy under, for instance drought conditions.

**Table 3.3:** The evaluation metrics used to quantify the potential improvements for different characteristics of the streamflow time series.

Characteristic of the streamflow signal	Metric	Abbreviation	Equation
Volume	Scaled Mean Absolute Error	SMAE	$MAE = \frac{\sum_{t=1}^T  y_o^t - y_m^t }{T},$ $SMAE = \frac{MAE}{\bar{y}_o}$
High extremes	Nash-Sutcliffe Efficiency	NSE	$NSE = 1 - \frac{\sum_{t=1}^T (y_o^t - y_m^t)^2}{\sum_{t=1}^T (y_o^t - \bar{y}_o)^2}$
Low extremes	Logarithmic Nash-Sutcliffe Efficiency	logNSE	$\log NSE = 1 - \frac{\sum_{t=1}^T (\log(y_o^t) - \log(y_m^t))^2}{\sum_{t=1}^T (\log(y_o^t) - \log(\bar{y}_o))^2}$

Improvement at each station is further denoted by calculating the skill, which quantifies the efficacy of post-processing methods relative to raw E-HYPE simulations, with negative (positive) skill values indicating deterioration (improvements). A skill value approaching 1 signifies a greater enhancement in predictive performance, highlighting the effectiveness of the post-processing techniques in refining hydrological forecasts. The skill (over the historical simulation period) is expressed as:

$$Skill = \frac{Score_{pp} - Score_{raw}}{Score_{perfect} - Score_{raw}}$$

## 3.2 Understanding changes of hydro-climatic extremes under future conditions

### 3.2.1 Experimental setup

Here we present the approach followed to assess the impact of climate change on hydro-climatic extremes. To do so, we firstly proceeded on detecting the extremes and then focused on specific statistical indices and assessed how they are changing in time and under different emission scenarios.

#### *Detection of extreme events*

To detect the extreme events (which can lead to floods and droughts) in the Living Labs under future conditions we applied the threshold-level method (Pechlivanidis et al., 2017). This method is based on a threshold-level above (below) which a flood (drought) event is detected. The threshold method is used to compute various flood and drought properties, such as duration, deficit/surplus volume and number of events (Teutschbein et al., 2022, Quesada-Montano et al., 2018). The choice of the threshold level is subjective and can influence the estimates of these properties (Van Loon, 2015), which are typically derived from non-exceedance probabilities of the flow duration curve. Here, we used a fixed level threshold over the historical, early, mid and late century 30-year periods to identify the extreme events under present and future conditions. The thresholds for the two extremes were defined as 20% (Q80) and 90% (Q10) of non-exceedance probability in the historical period.

#### *Statistical properties of streamflow extreme events*

Extreme events were identified using the threshold-level method for each sub-basin of the Living Labs, and statistical properties of the extremes were then calculated. These properties or indicators included: (1) the duration of the extremes, (2) the surplus/deficit volume, and (3) the number of extreme events. To determine the duration of an event, the number of consecutive time steps (days) in which streamflow is below the pre-defined threshold, i.e., streamflow with a non-exceedance probability of 20% (Q80) and 90% (Q10), is calculated. The surplus/deficit volume (measured in mm) for a specific extreme event is found by adding up the deviations of streamflow from the threshold value during the analysed period. The number of drought events is given by counting the extreme events over the analysed period. These statistical properties were computed separately for each basin considered in each of the Living Labs on a 30-years basis in the historical and future periods using the thresholds computed in the historical period.

### 3.2.2 Evaluation framework

Given that hydro-climatic projections are available as an ensemble of equally probable members, the statistical properties of the extreme events were averaged over the ensemble in order to obtain a single value of the duration, deficit/surplus volume and number of events for each sub-basin of the Living Labs. These calculations were repeated for the historical, early, mid and late century periods and for the low, medium, and high emission scenarios (RCP2.6, 4.5 and 8.5, respectively). The changes in the statistical extreme event properties under future conditions were assessed as the difference between a given statistical property in the early, mid and late century period and the historical period and shown for all the emission scenarios.

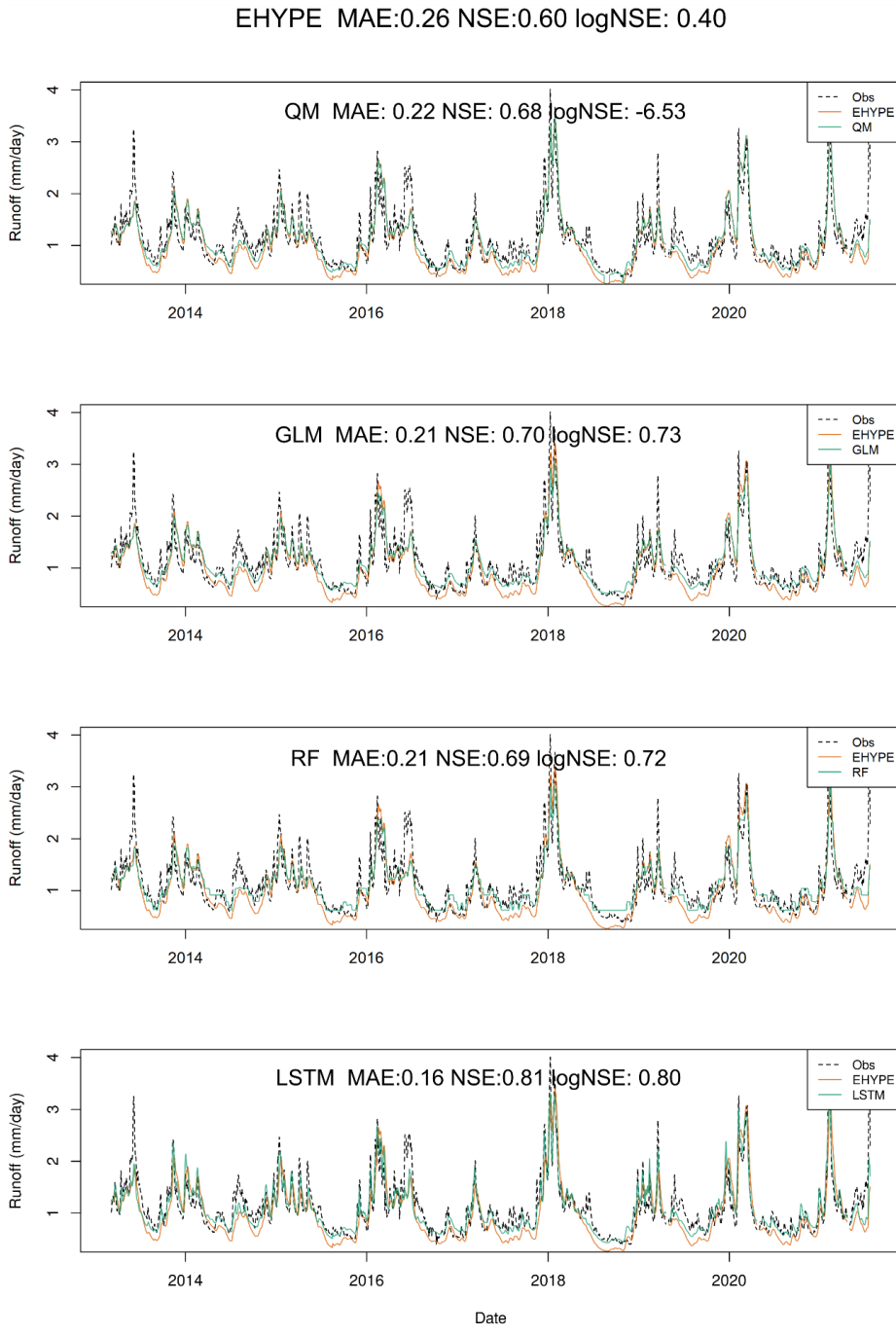
## 4 Results - Improving process understanding at the local scale through hybrid hydrological modelling

### 4.1 Improving historical model performance through post-processing

As explained in the Methodology section, four post-processing methods were applied to the streamflow data of the available stations in the LLs and their performance were evaluated using three performance metrics against the raw E-HYPE model performance. Figure 4.1 presents results from the station in the Netherlands in the Netherlands as an example. It shows the time series of the raw and post-processed simulations, and the corresponding performance metrics to quantify the potential improvement with regard to volume and extremes. Results show an overall enhancement of the raw E-HYPE model performance, specifically regarding volume and the accuracy of high flow extremes, shown by the reduction in SMAE and the increase in NSE. This indicates a more accurate representation of hydrological dynamics in total bias and for high streamflow.

Moreover, the investigation of low streamflow predictions reveals a special case, where QM significantly deteriorates  $\log\text{NSE}$  (-6.53), which is lower than that of raw simulation (0.4). This decrease in  $\log\text{NSE}$  is due to the near zero values produced by QM, which significantly influence the metric. This is due to the original underestimation from EHYPE hydrological model especially in the low flows, which has proposed challenges for the QM method when extrapolating the distribution. If considering the general pattern of low streamflow dynamics, QM still improves the raw simulation to a certain extent as seen from the time series plot, noting the potential marginal problems might occur for near-zero values. Another noticeable pattern occurs for RF, where occasionally a “staircase” pattern is observed for low streamflow, which indicates a certain model underfitting that requires a more complicated forest structure. We note here that the model structure and hyperparameters for the methods are consistent across different stations to allow comparison across the study domain, therefore, the model structure might not be the optimal setup for each catchment/station. Overall, the results illustrate the general success of the post-processing techniques in refining hydrological simulations and also emphasises the importance of method selection based on the specific hydrological conditions (volume and extremes) being modelled.

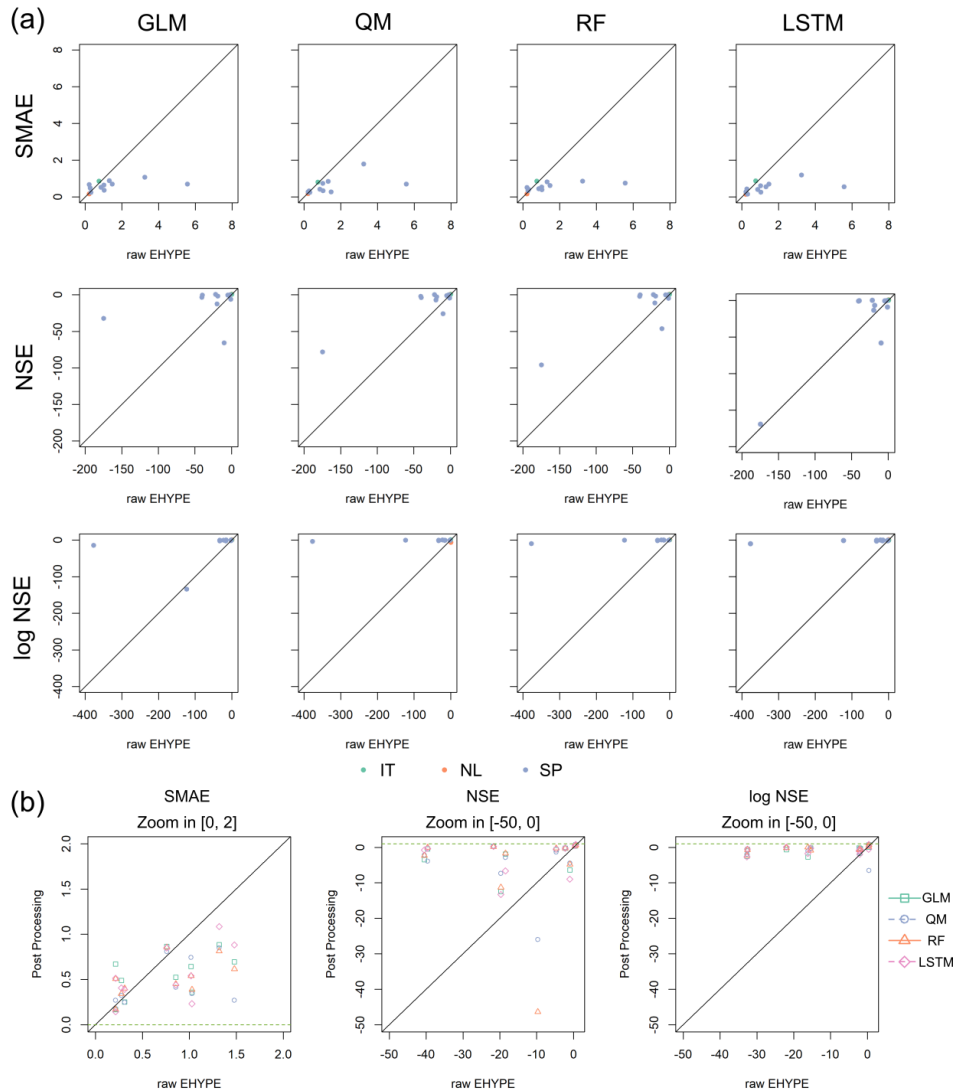
Similar results have been generated for each station in the living labs, and we next summarise them to better understand the overall patterns and tendencies for model improvements. Figure 4.2 shows the results of the three selected performance metrics for both raw and post-processed time series, allowing for a comparative analysis of their overall performance. For each station, the performance of a specific post-processor (y-axis) is plotted against the raw E-HYPE model performance (x-axis). The points below the 1:1 reference line denote improved performance for SMAE, while for NSE and  $\log\text{NSE}$  are the opposite, where points above the reference line show improved performance. For SMAE, most stations achieved a reduced error (lower SMAE) after post-processing, except two (or occasionally three) stations which include the station in the Italian LL and one (or two) station(s) in the Spanish LL, where SMAE is lower in the raw performance ( $< 1$ ). After post-processing, the range of SMAE reduced from up to 6, to lower than 2.



**Figure 4.1:** An example of post-processing results in the station from Netherlands Living Lab.

In general, post-processing improves NSE at most stations, with the exception at one or two stations in Spain. Particularly in Spain, several stations showed unsatisfied NSE and logNSE performance with large negative values using the E-HYPE simulations. This is due to the presence of reservoirs which heavily regulate streamflow, and makes the representation of streamflow dynamics very challenging for large-scale process-based hydrological models (Table 2.1). After post-processing, NSE and logNSE improve to a great extent, however some stations still achieve performance values below zero. This highlights the need to include more detailed local information, for example, (season-dependent) regulation schemes etc. For logNSE, improvements achieved from QM, RF and LSTM are comparable, with the points staying high above the

reference 1:1 line in Figure 4.2. Since logNSE represents the model performance with attention to low flows, which is an important aspect that the users in the Living Labs. From the zoomed-in plot, the improvements for low flow are considerable since the points stay here above the 1:1 reference line. Overall, different post-processing methods show a certain capability of improving streamflow simulation in different aspects. The varying performance across the stations requires an in-depth investigation to detect potential spatial patterns.



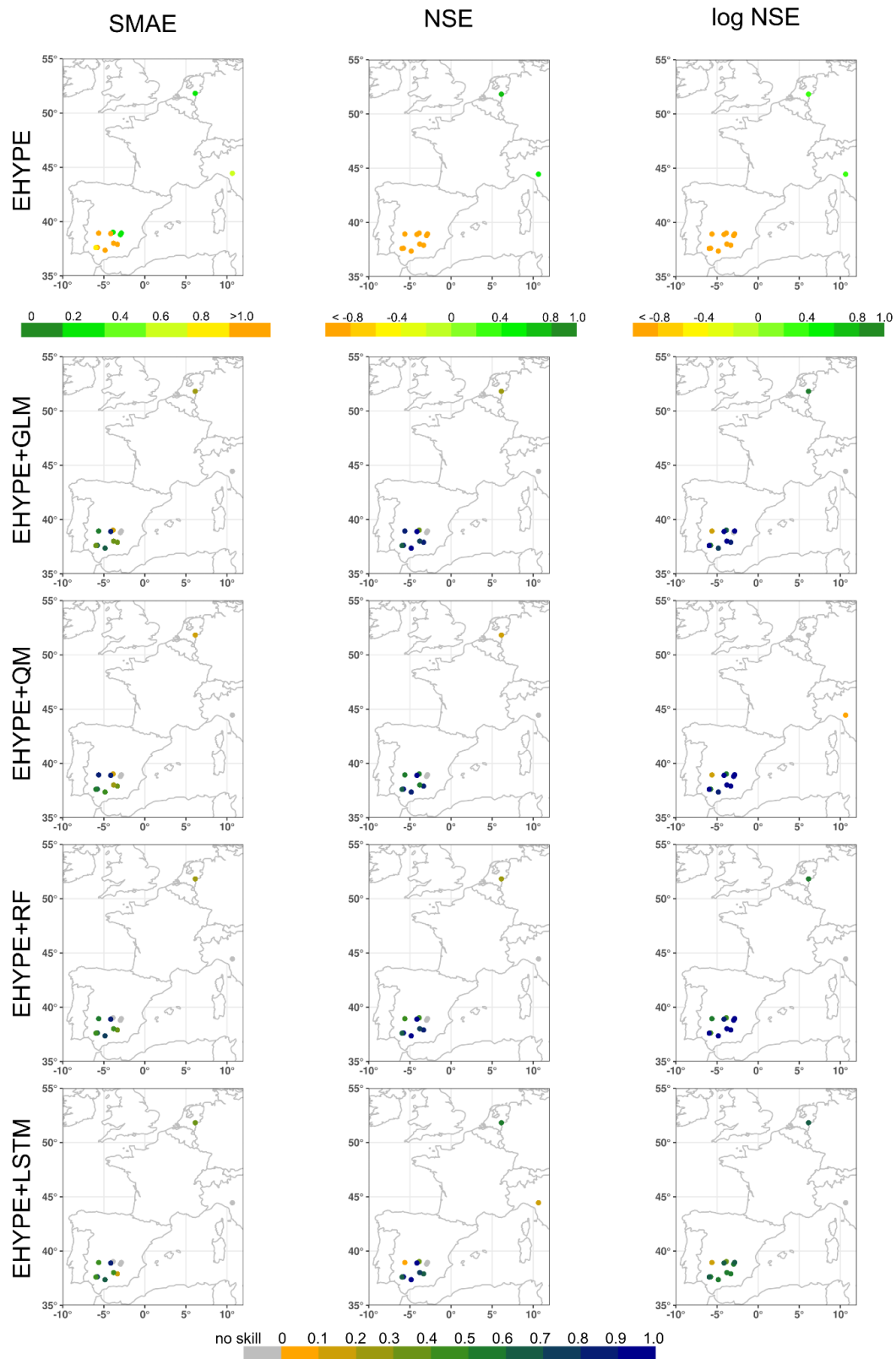
**Figure 4.2:** Scatterplot for the performance achieved after post-processing versus the raw E-HYPE performance: (a) performance for each method at the available streamflow stations in the Italian (IT), the Netherlands (NL) and Spanish (SP) Living Labs; (b) zoomed-in plots for the same metrics in (a) with four post-processing methods in the same plot. Green dotted lines denote the reference performance of a perfect prediction, 0 for SMAE and 1 for NSE and logNSE.

### 4.2 Spatial patterns of the post-processing results

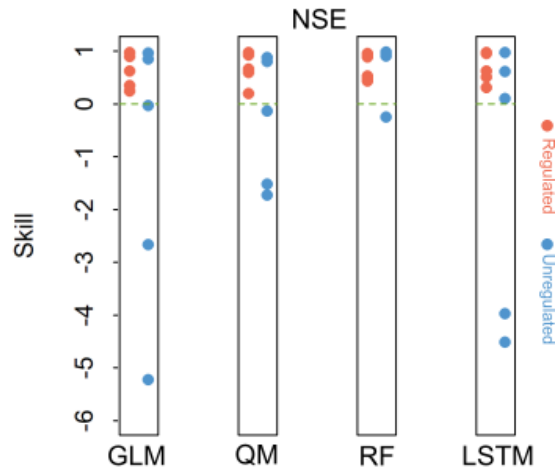
We next aim to investigate the added value provided by the different post-processing methods over the raw simulations. Figure 4.3 presents the skill for each station with green-blue shades indicating higher skill, yellow indicating lower skill, and grey indicating no improvement. A consistent pattern emerges across the different post-processing methods, with stations in Spain showing higher skill, especially for NSE and logNSE, which again highlights that post-processing significantly improves model performance for extremes in this part of

the domain. At the same time, spatial variations of the improvement across different post-processing methods also exist, as shown by the varying colour of skills at the same station.

When comparing these outcomes to the reference performance of the raw E-HYPE model simulations, it is observed that stations with very poor performance are precisely those where significant improvements are obtained through post-processing. This observation underscores the post-processing methods' effectiveness in enhancing model accuracy, particularly in areas initially having low predictive quality due to, for instance, anthropogenic influences. For a further investigation, we separated the stations into three categories: regulated, unregulated and no information (see Table 2.1), and compared the skills achieved at the regulated stations with the unregulated ones. Post processing skills achieved at the regulated stations are showing higher results than the stations without regulations, for the general bias (represented by SMAE) and high flows (represented by NSE). Taking the high flows as an example (NSE, Figure 4.4), skills achieved from different post processing methods for the regulated river systems (red dots) within the Living Labs are all positive values, with the median skills 0.63, 0.66, 0.62, 0.63 for GLM, QM, RF and LSTM method, respectively. For unregulated stations (represented by blue dots), post processing gave a wider range of skills, with the medians between -0.1 to 0.1 for different methods. Similar conclusions can also be drawn for the volume bias represented by SMAE. For low flows represented by logNSE, GLM and LSTM gave relatively similar skills for regulated and unregulated stations, while RF and QM provided higher skills in the unregulated stations. Nevertheless, the discussion regarding the performance difference at regulated/unregulated river systems needs more in-depth investigation at an extended spatial domain for a more concrete conclusion, while here we provided a potential direction that can be tested in the later phase of this project within the Living Labs.



**Figure 4.3:** Spatial distribution of the raw E-HYPE model performance (SMAE, NSE and log NSE) and the skills achieved after post-processing with the four different methods.



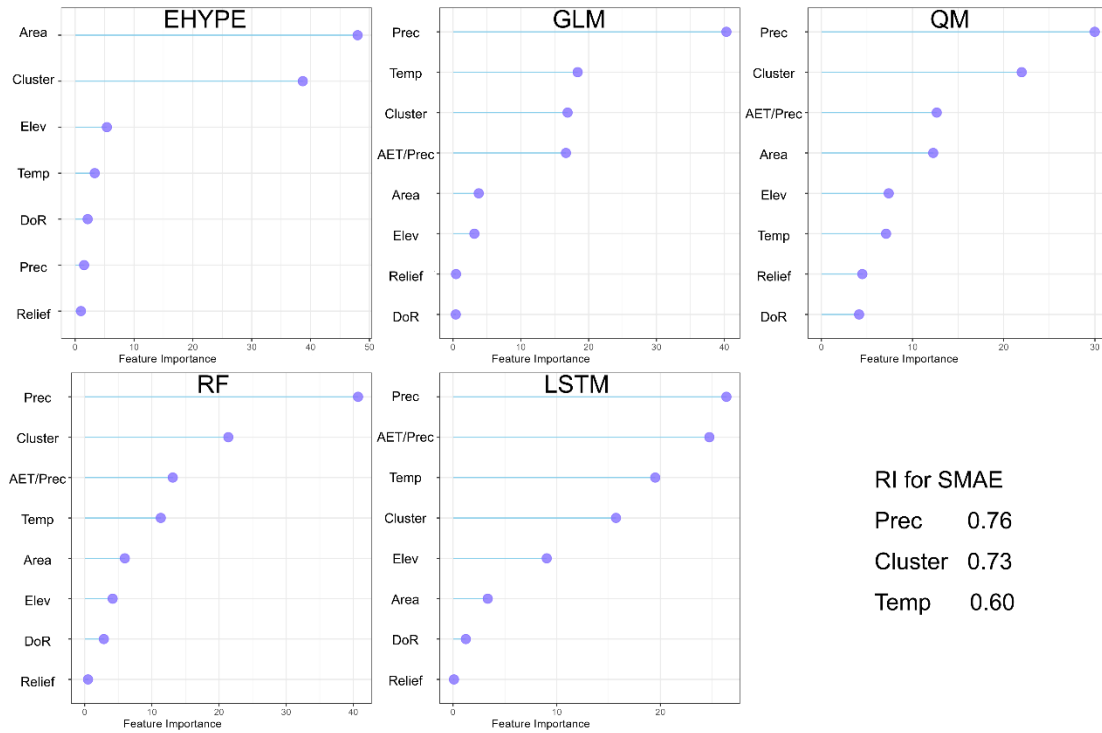
**Figure 4.4:** Improvements achieved from post processing methods (represented by skills) for regulated/unregulated river systems in the Living Labs.

Overall, the analysis suggests that there is no single superior model; each post-processing method presents varying degrees of skill across different locations and according to different performance metrics. This variability highlights the importance of selecting the appropriate post-processing technique based on specific regional characteristics and the particular aspects of hydrological behaviour being modelled.

#### 4.3 Performance attribution to basin descriptors

We next examined the potential drivers of the performance of hybrid model framework, including the raw EHYPE model, and the post-processing approach, with all of the about 2000 stations in the pan-European domain. This comparative analysis is needed in order to gain a deep understanding of the local dominant mechanisms linked to the model performance and post-processing skill. The outcomes will help us to build the solid ground for the next step, extending the post-processing to the ungauged basins through a regionalization approach. This step is particularly valuable for Living Labs, considering the insufficient observation data as shown in Table 2.1. By applying knowledge derived from an extended domain with similar hydrological regimes, we can effectively apply post-processing to the ungauged basins within the Living Labs, and therefore enhance our hydrological services of these areas.

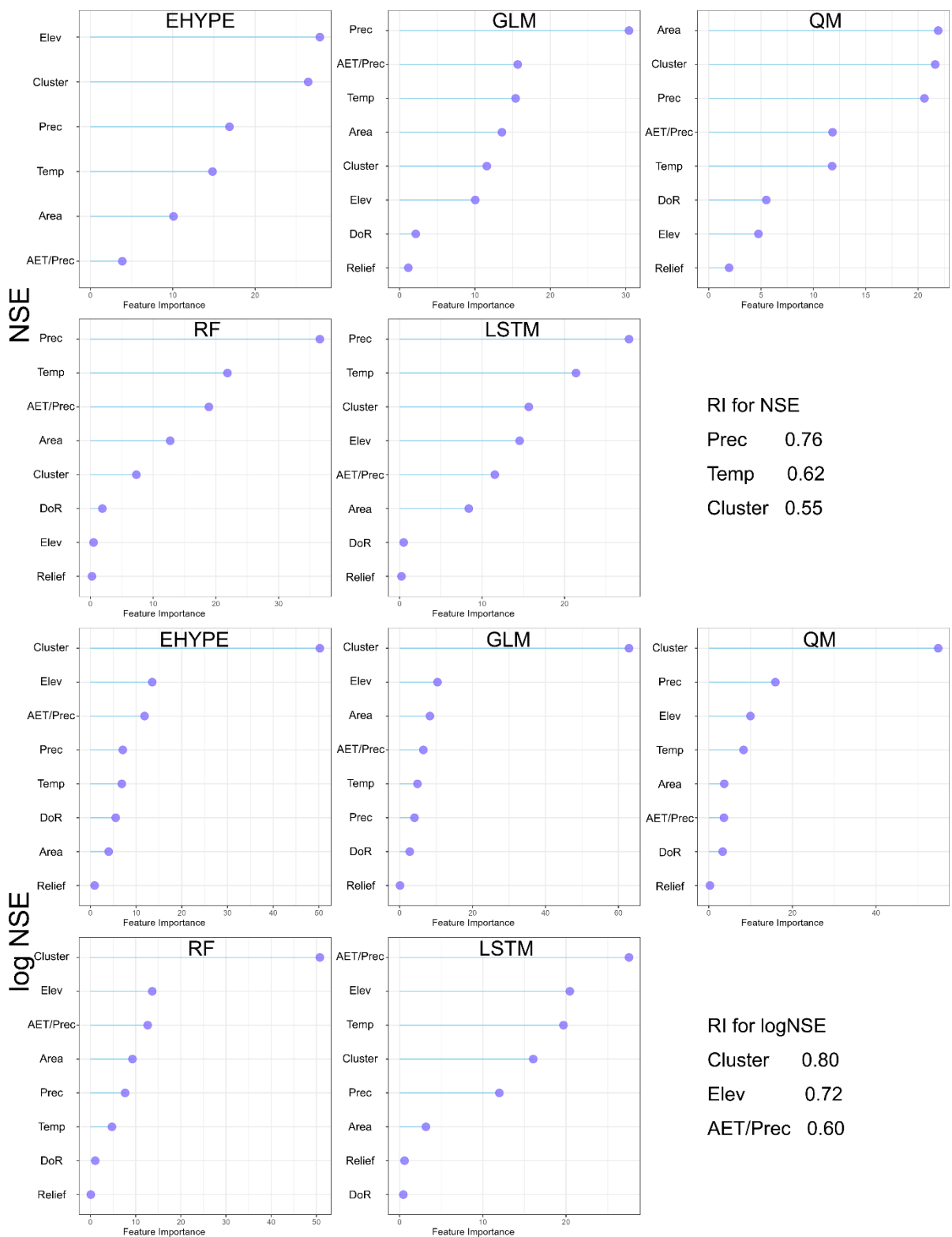
We used the CART method and explored the importance of various drivers by considering them as inputs, with the performance metrics and skill scores as the targets. Figure 4.5 shows the importance of each driver (feature) for the raw model and post-processing approach in terms of volume (represented by SMAE), with the comprehensive rank index of the three most influencing factors. From the analysis, several leading drivers were identified with the most important being the precipitation (Prec), hydrological regime (Cluster) and mean temperature (Temp) targeting model performance improvements in the total volume (described by the SMAE metric). For the raw E-HYPE model, the upstream area (Area) and the hydrological regime are the two most important factors, among others, while mean precipitation is the leading factor for all the four post-processing methods. This result suggests that the local climatic conditions and the local hydrological regime are strongly linked with the model/post-processing performance with regard to the total streamflow volume.



**Figure 4.5:** Key drivers identified for the raw E-HYPE model and post-processing performance based on the SMAE metric.

The same investigation was extended to include both NSE and logNSE, and further identify the most important drivers for each metric (see Figure 4.6). For NSE, mean precipitation, mean temperature and hydrological clusters emerge as the leading drivers. However, for logNSE, the leading factors are hydrological clusters, elevation and evaporative index (AET/Prec), each with its unique ranking across the different performance metrics. A pattern worthy of noting, identified for logNSE, is the leading feature importance of hydrological clusters, suggesting the strong link between the local hydrological regime and the model performance (both EHYPE and post-processing) with regard to the low streamflow extreme. One reason for the differences between NSE and logNSE, which reveal distinct levels of importance, is that for high flows, precipitation tends to dominate. For low flows (emphasized in logNSE), the hydrological characteristics of the basin become more prominent, since usually the low flows occur during the dry season. Therefore, in the context of logNSE, hydrological clusters hold greater importance, highlighting their significance in understanding and modelling these specific conditions.

The recurring presence of the hydrological cluster as one of the leading drivers in all performance metrics, accounting for both volume and extremes, further underscores its important role in understanding the model's performance and post-processing potential. Recognizing such key factors is essential for refining hydrological models, as it directs attention to the factors that most significantly impact the accuracy and reliability of hydrological simulations. Through this analysis, we gained deeper insights into the mechanisms driving model performance, allowing targeted improvements in climate services for the water sector.



**Figure 4.6:** Key drivers identified for the raw E-HYPE model and post-processing performance based on the NSE and logNSE metrics.

These results have confirmed the relationship between key drivers and the improvements in model performance derived from post-processing. This understanding will advance the regionalization efforts in the following work, allowing us to extend the post-processing techniques to ungauged basins, by using

hydrological clusters along with extra static inputs such as mean temperature and mean precipitation. This approach enables the transfer of insights gained from gauged basins to those that are ungauged, thereby broadening the applicability of our hydrological models in the Living Labs.

### 4.4 Summary of the results

Here, we applied four post-processing methods to streamflow simulations generated by the E-HYPE model and evaluated the added value from these methods using specific metrics (targeting different characteristics of the streamflow signal) and analysed the spatial distribution of their skill. Moreover, our investigation extends to identifying the primary factors driving performance enhancements gained in each post-processing method. The key findings include:

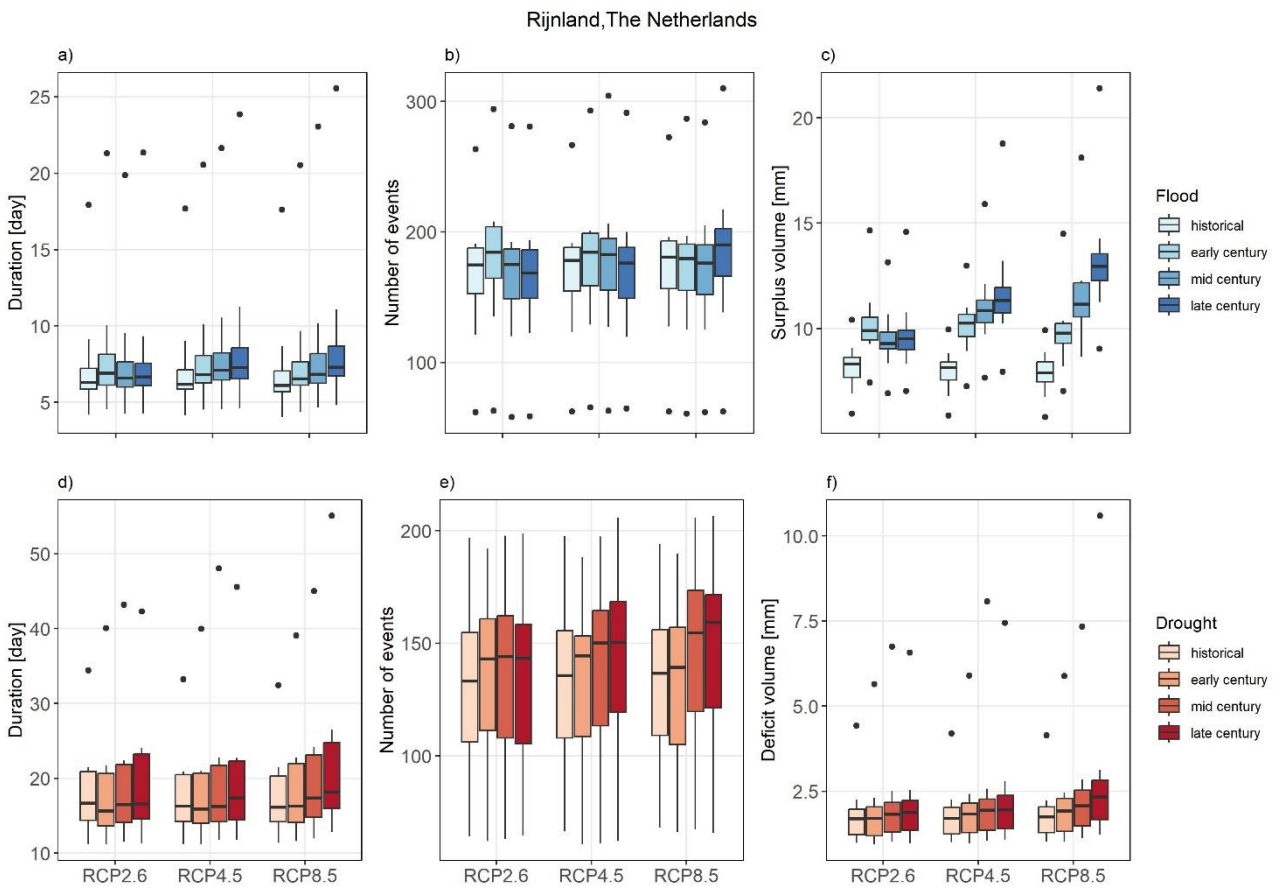
- The analysis reveals a notable improvement by post-processing the raw E-HYPE simulations, in terms of streamflow volume and extremes. This is evidenced by the decrease in SMAE and an increase in both NSE and logNSE metrics, which suggest that post-processors can provide a more accurate representation of the time series than the raw process-based model simulations.
- Across the different post-processing techniques, a similar spatial pattern of skill improvements is observed, showing higher skills in stations located in Spain. This pattern is especially pronounced in the context of extremes, indicating that post-processing enhances the model's ability to account for structural limitations in river systems affected by anthropogenic influence.
- Key drivers have been identified influencing the model performance and consequently the post-processing skill. These are the mean precipitation, mean temperature, hydrological regimes, elevation and evaporative index, with varying ranking across the metrics. This indicates their varying impact on model performance. Notably, the recurrent identification of hydrological regime (described here by the clusters) as a significant factor for both streamflow volume and extremes emphasises its importance in improving model performance.

## 5 Results - Understanding changes of hydro-climatic extremes under future conditions

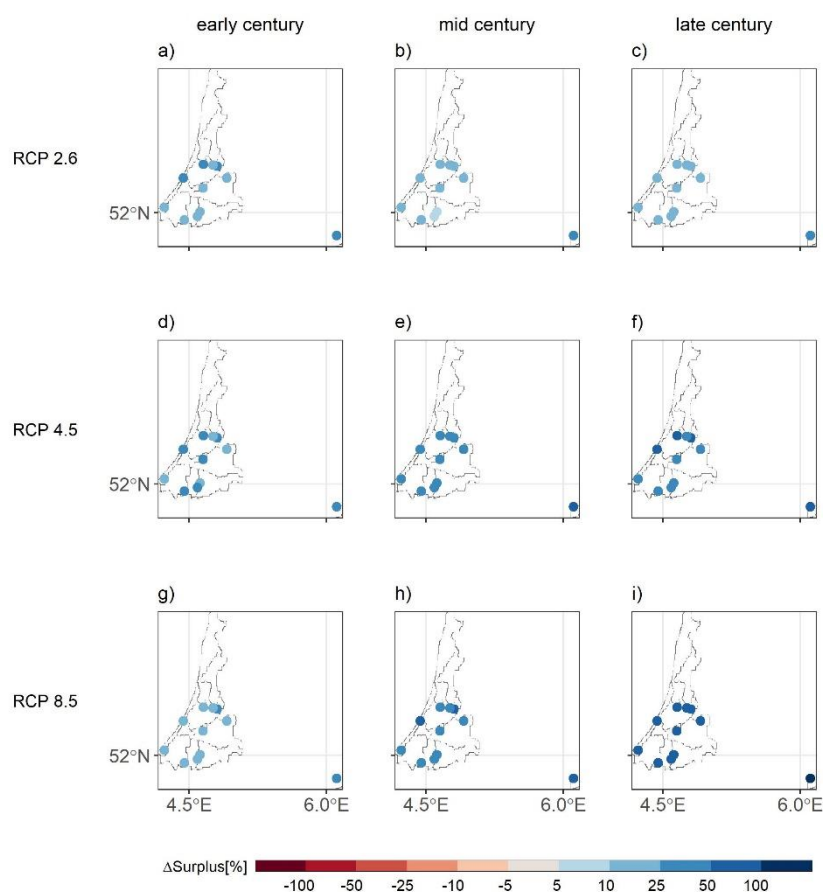
Here, the assessment of climate change on hydro-climatic extremes is conducted and presented for each living lab. Different indicators are used to describe the extremes and results are presented for different future periods and emission scenarios.

### 5.1 The Rijnland Living Lab - The Netherlands

The statistical properties (duration, number and surplus/deficit volume) of the streamflow extreme events calculated using the Euro-CORDEX based ensemble in the historical period, and the three future periods of early, mid and late century for the RCP2.6, 4.5 and 8.5 emission scenarios in Rijnland, The Netherlands, are shown in Figure 5.1. The results suggest that the duration of the flood events increases from the historical to the late century, especially under medium and high emission scenarios (Figure 5.1a). The number of events remains relatively stable with all the emission scenarios (Figure 5.1b), while the surplus volume, on average, shows changes. The increase in variability from the historical period to the late century indicates that the spatial variability of the surplus volume increases, especially under a high emission scenario (Figure 5.1c). This surplus especially increases along the western part of the Netherlands and in the Rhine river basin (Figure 5.2).

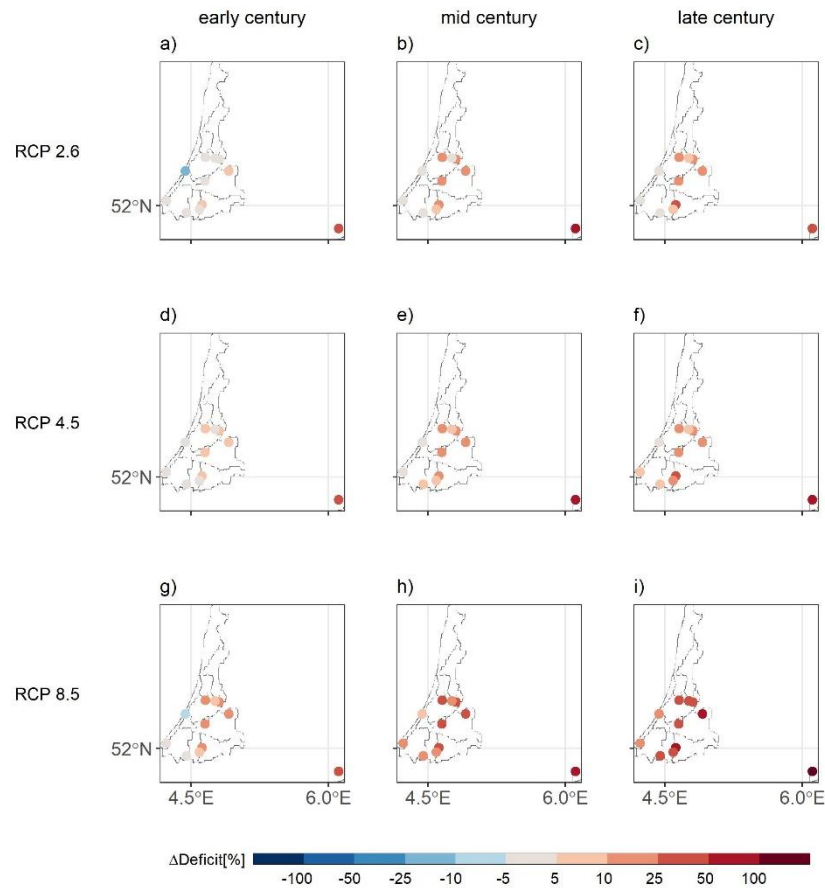


**Figure 5.1:** Boxplots representing the mean model ensemble duration, number and surplus/deficit volume of the flood (a,b,c) and drought (d,e,f) events in the historical, early, mid and late century for the RCP 2.6, 4.5, 8.5 emission scenarios in the Rijnland, The Netherlands. The boxplots are generated from the values of all the sub-basins in the Living Lab.



**Figure 5.2:** Change in surplus volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Rijnland, The Netherlands.

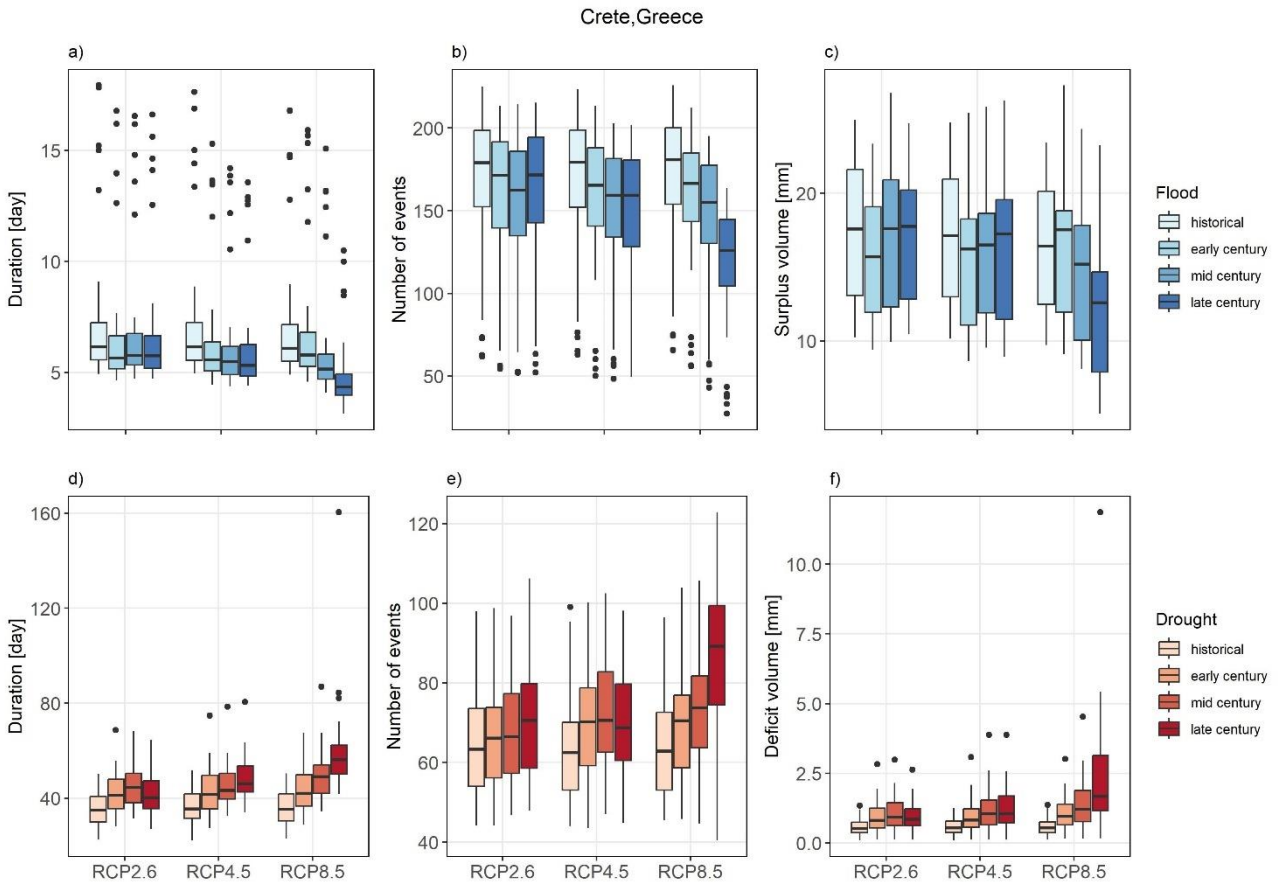
The drought events are more extended and of a smaller number (Figure 5.1d and 5.1e) compared to the flood events (Figure 5.1a and 5.1b). The drought duration and number of events increases from the historical period to the late century, especially under the high emission scenario (Figure 5.1d and 5.1e). On average, the deficit volume of the drought events shows higher value under the high emission scenario in comparison to the low one (Figure 5.1f). Moreover, in most of the basins, the deficit changes between the early, mid and late century periods and the historical period for all the emission scenarios (see Figure 5.3).



**Figure 5.3:** Change in deficit volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Rijnland, The Netherlands.

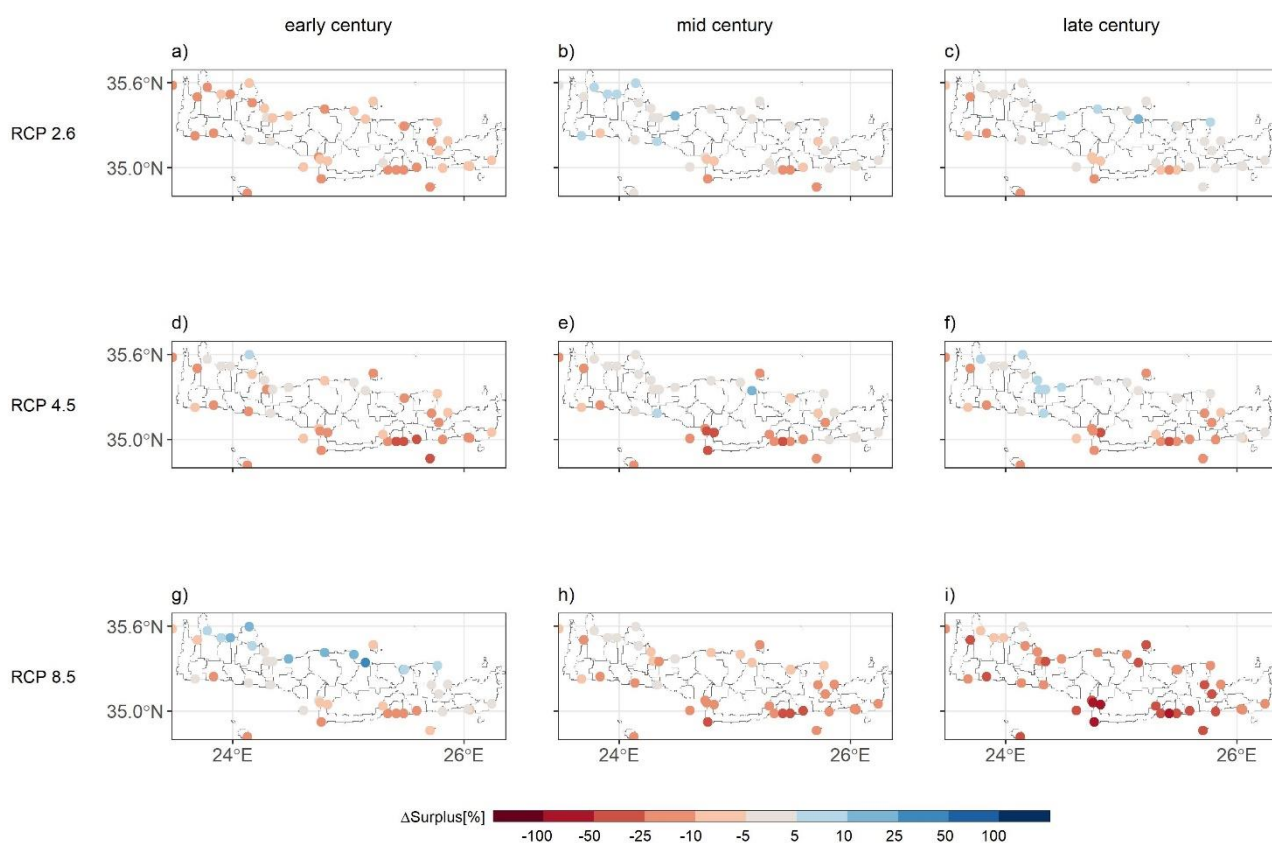
## 5.2 The Crete Living Lab - Greece

The duration, number of events and surplus/deficit volume of the streamflow extreme events in Crete, Greece, are shown in Figure 5.4. The duration of the flood events is relatively stable over time under the low and middle emission scenarios (RCP 2.6, 4.5), while it decreases in time under the high emission scenario, RCP 8.5 (Figure 5.4a). The number of flood events showed a similar behaviour to the flood duration (Figure 5.4b). The surplus volume decreases in time, especially with high emission scenarios, reflecting the changes in the duration and number of flood events (Figure 5.4c). In contrast, the drought duration and number of events increases with medium and high emission scenarios (Figure 5.4d and 5.4e). Consequently, an increase of the deficit volume is visible in time, especially with high emission scenarios (Figure 5.4f).

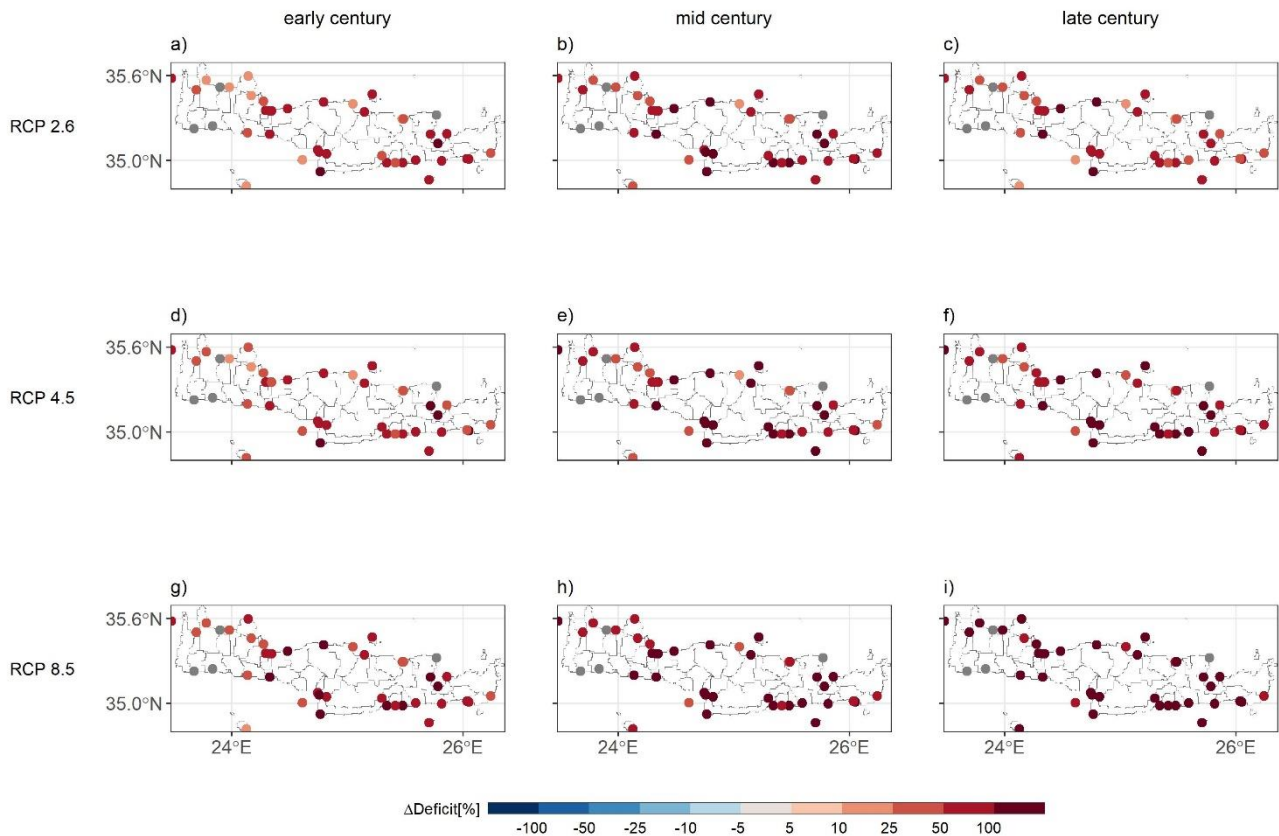


**Figure 5.4:** Boxplot representing the mean model ensemble duration, number and surplus/deficit volume of the flood (a,b,c) and drought (d,e,f) events in the historical, early, mid and late century for the RCP 2.6, 4.5 and 8.5 emission scenarios in the Crete basins. The boxplots are generated from the values of all the sub-basins in the Living Lab.

The spatial variability of the surplus volume changes in the early, medium and late century compared to the historical period for the RCP 2.6, 4.5 and 8.5 are shown in Figure 5.5. Most of the basins show little changes of the surplus volume in the early and mid-century for the low (Figure 5.5a-c) and medium emission scenarios (Figure 5.5d-f). However, the surplus volume shows decrease in most of the basins in the late century (Figure 5.5i). The deficit volume changes in the three future periods compared to the historical periods increase in most of the basins (Figure 5.6), especially in those basins located in the southern-eastern part of the island.



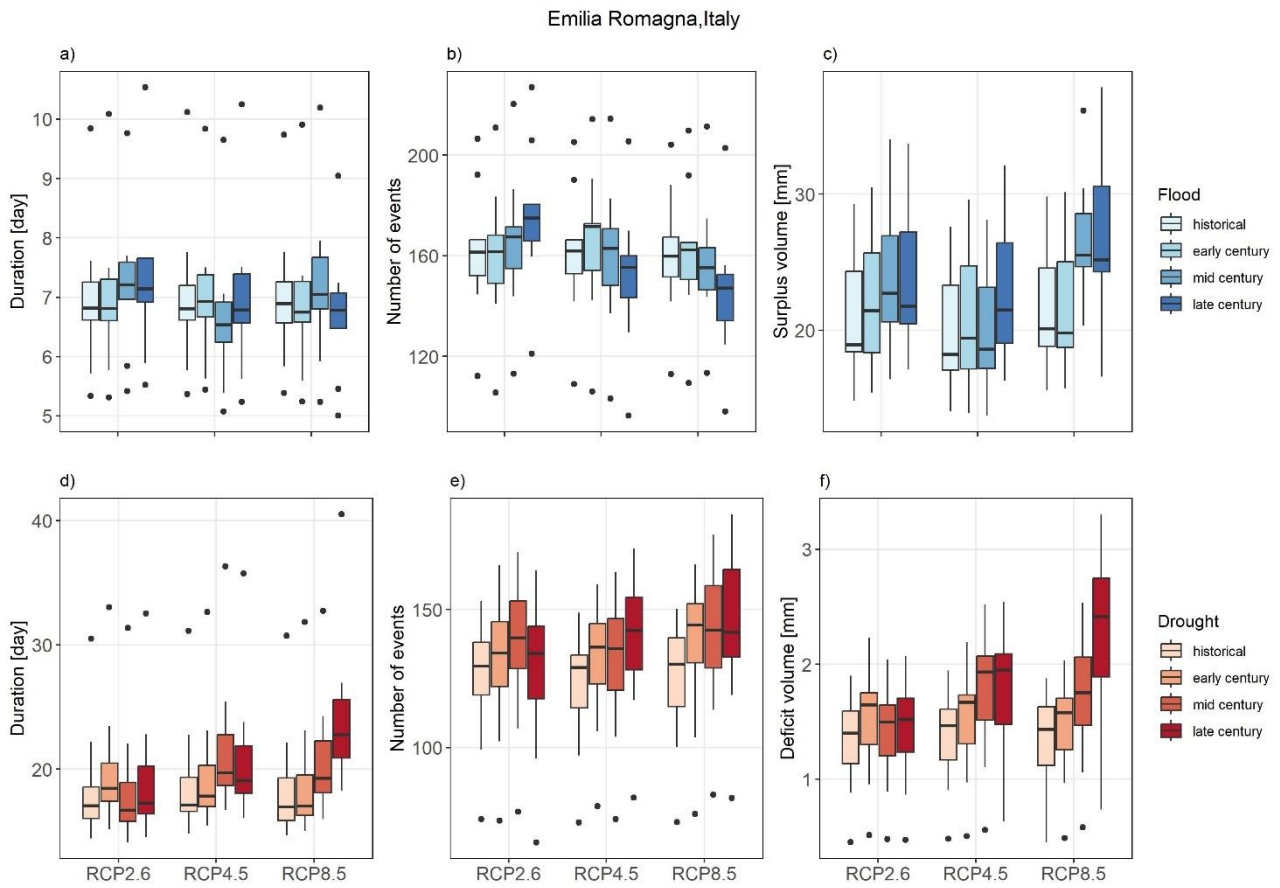
**Figure 5.5:** Change in surplus volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Crete, Greece.



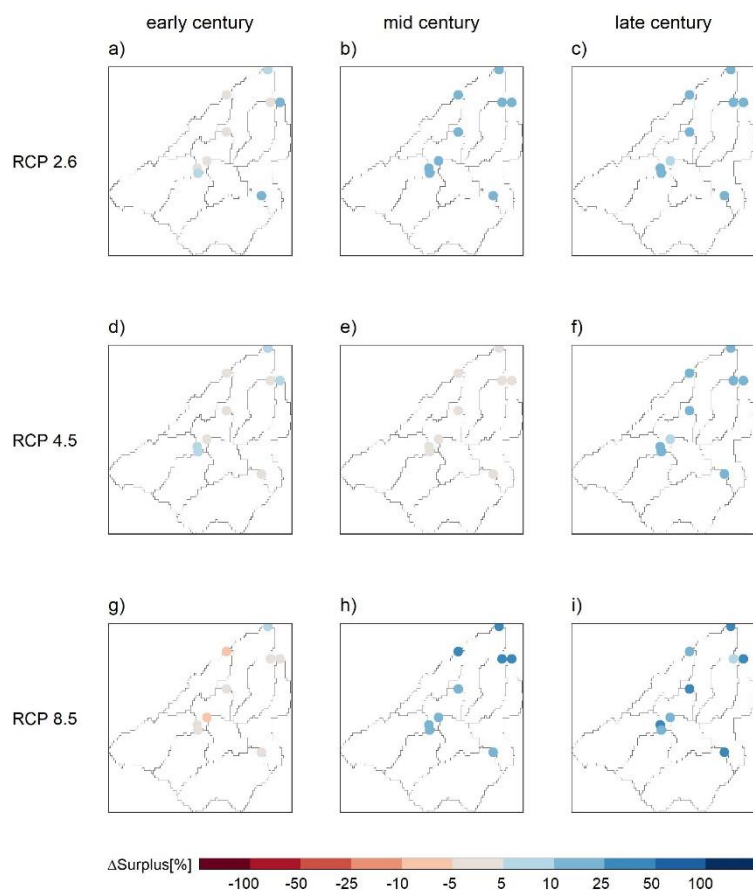
**Figure 5.6:** Change in deficit volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Crete, Greece. Basins with no drought events are in dark grey colour.

### 5.3 The Emilia Romagna Living Lab - Italy

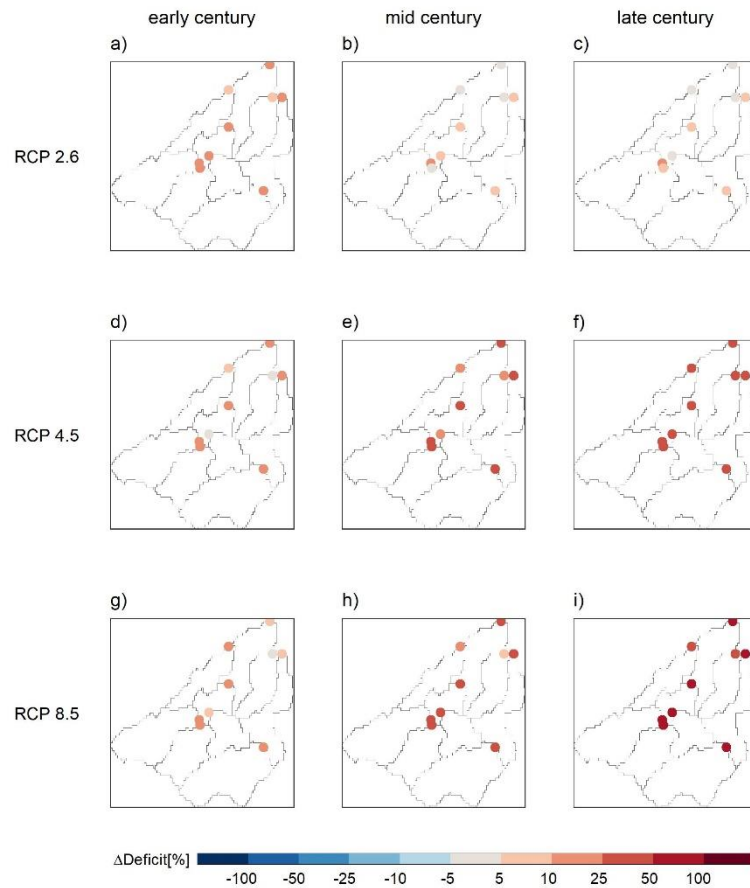
The results of the duration, number of events and surplus/deficit volume estimations of the streamflow extreme events in Emilia Romagna, Italy, are shown in Figures 5.7-5.9. The flood events, on average, are stable in time and under the emission scenarios in terms of duration (Figure 5.7a), while the number of flood events decreases in the late century under the medium and high emission scenarios (Figure 5.7b). The surplus value shows little change in time and with the low and medium emission scenarios (Figure 5.7c) indicating the decrease of the number of events does not affect this flood statistical property for these scenarios. However, the surplus volume increases in time with the high emission scenario, indicating larger events. In most of the basin, the surplus volume increases up to 5% in the early and mid-century compared to the historical period (Figure 5.8a and 5.8b) and up to 25-50 % in the late century in the high emission scenarios (Figure 5.8i). The drought duration and number of events increases in time with medium and high emission scenarios (Figure 5.7d and 5.7f). The deficit volume, on average, shows little change from the historical to the late century for low emission scenarios, while it increases with the medium and high emission scenarios (Figure 5.7f and Figure 5.9). The variability in deficit volume increases towards the late century under the medium and high emission scenarios, indicating that some basins are impacted by the changes in the other two statistical properties. These basins are located at the borders of this Living Lab (Figure 5.9f and 5.9i).



**Figure 5.7:** Boxplot representing the mean model ensemble duration, number and surplus/deficit volume of the flood (a,b,c) and drought (d,e,f) events in the historical, early, mid and late century for the RCP 2.6, 4.5, 8.5 emission scenarios in Emilia Romagna, Italy. The boxplots are generated from the values of all the sub-basins in the Living Lab.



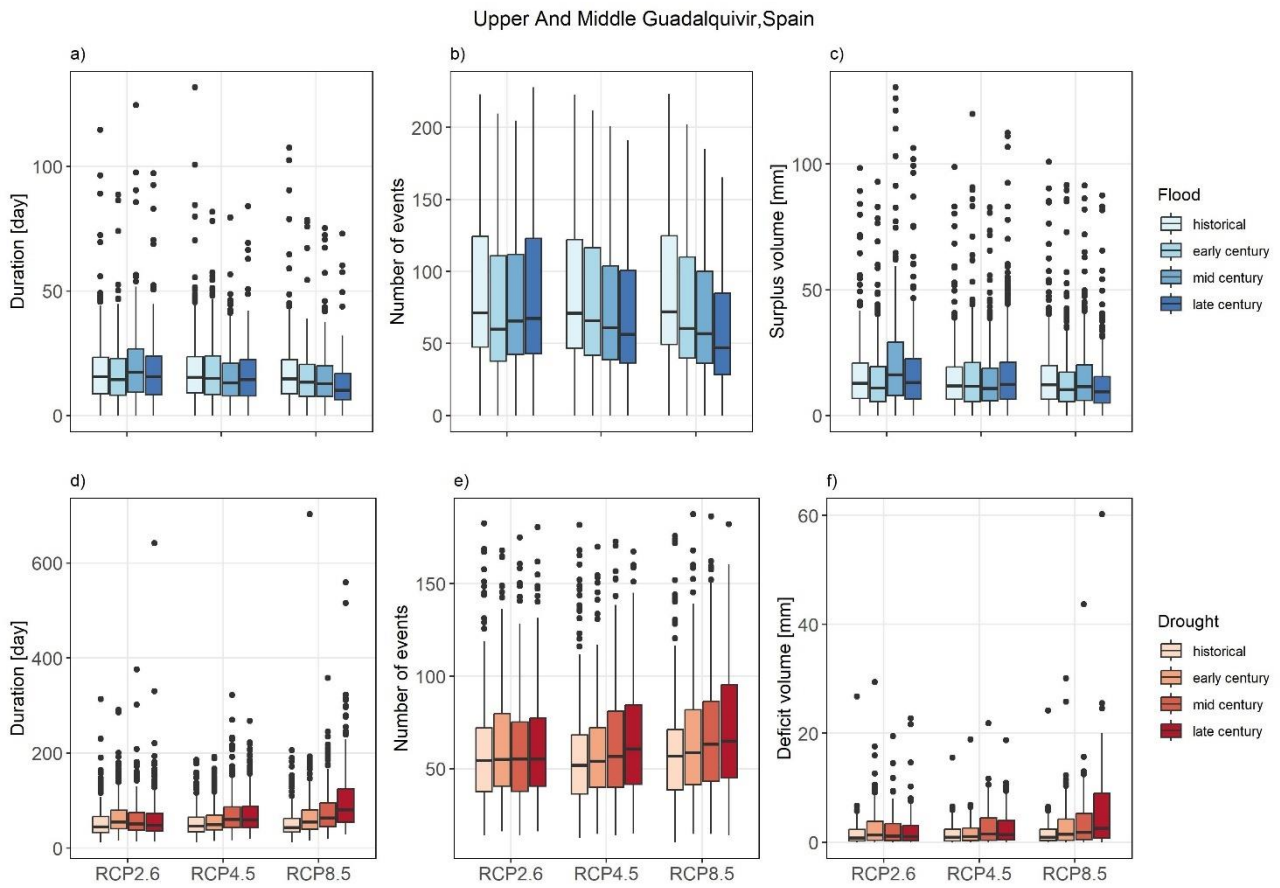
**Figure 5.8:** Change in surplus volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Emilia Romagna, Italy.



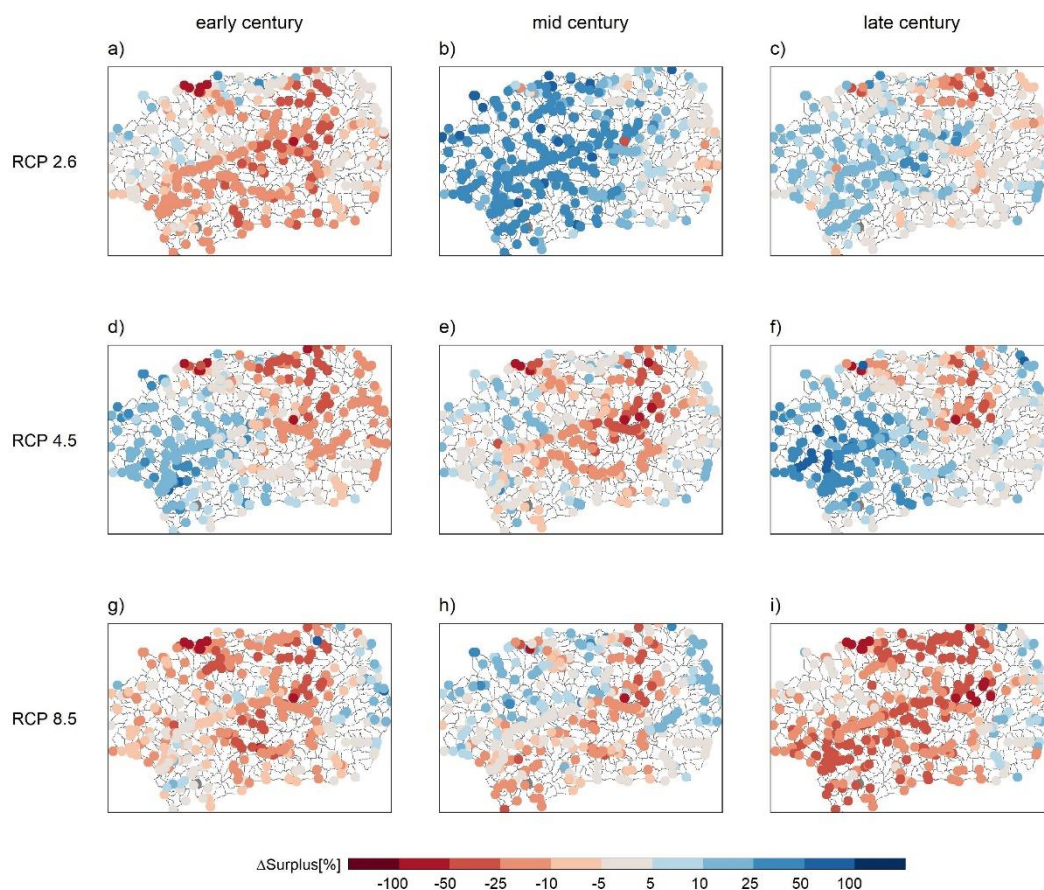
**Figure 5.9:** Change in deficit volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Emilia Romagna, Italy.

#### 5.4 The Guadalquivir Living Lab - Spain

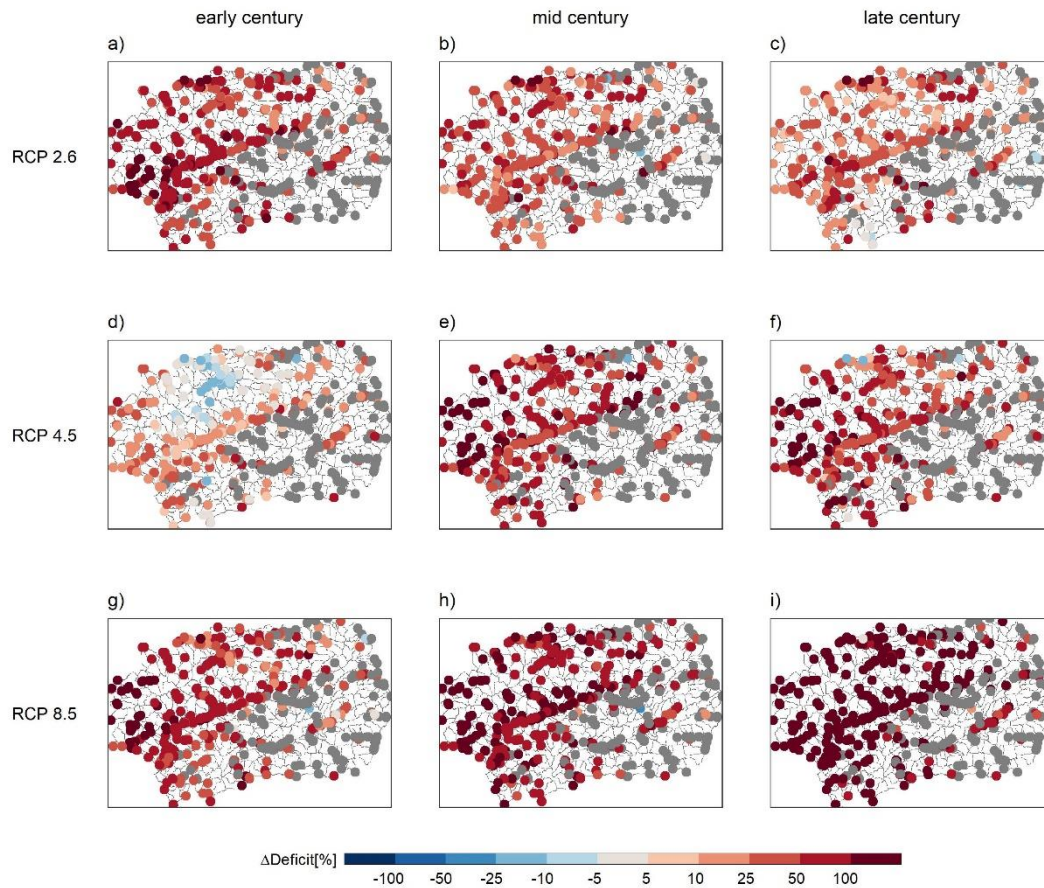
The statistical properties of the streamflow extreme events in the Upper and Middle Guadalquivir, Spain, are shown in Figure 5.10. The three statistical properties of the flood events show high variability among the basins in the historical and the future periods with all emission scenarios (Figure 5.10a-c and Figure 5.11). On average, the duration and number of events tend to decrease in the late century, while the surplus volume decreases in the late century compared to the historical period, especially under the high emission scenario (Figure 5.11i). The three statistical properties of the drought events are overall stable in time under the low and medium emission scenarios, while they increase towards the late century under the high emission scenario (Figure 5.10d-f and Figure 5.12).



**Figure 5.10:** Boxplot representing the mean model ensemble duration, number and surplus/deficit volume of the flood (a,b,c) and drought (d,e,f) events in the historical, early, mid and late century for the RCP 2.6, 4.5, 8.5 emission scenarios in the Upper and Middle Guadalquivir, Spain. The boxplots are generated from the values of all the sub-basins in the Living Lab.



**Figure 5.11:** Change in surplus volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in the Upper and Middle Guadalquivir, Spain.

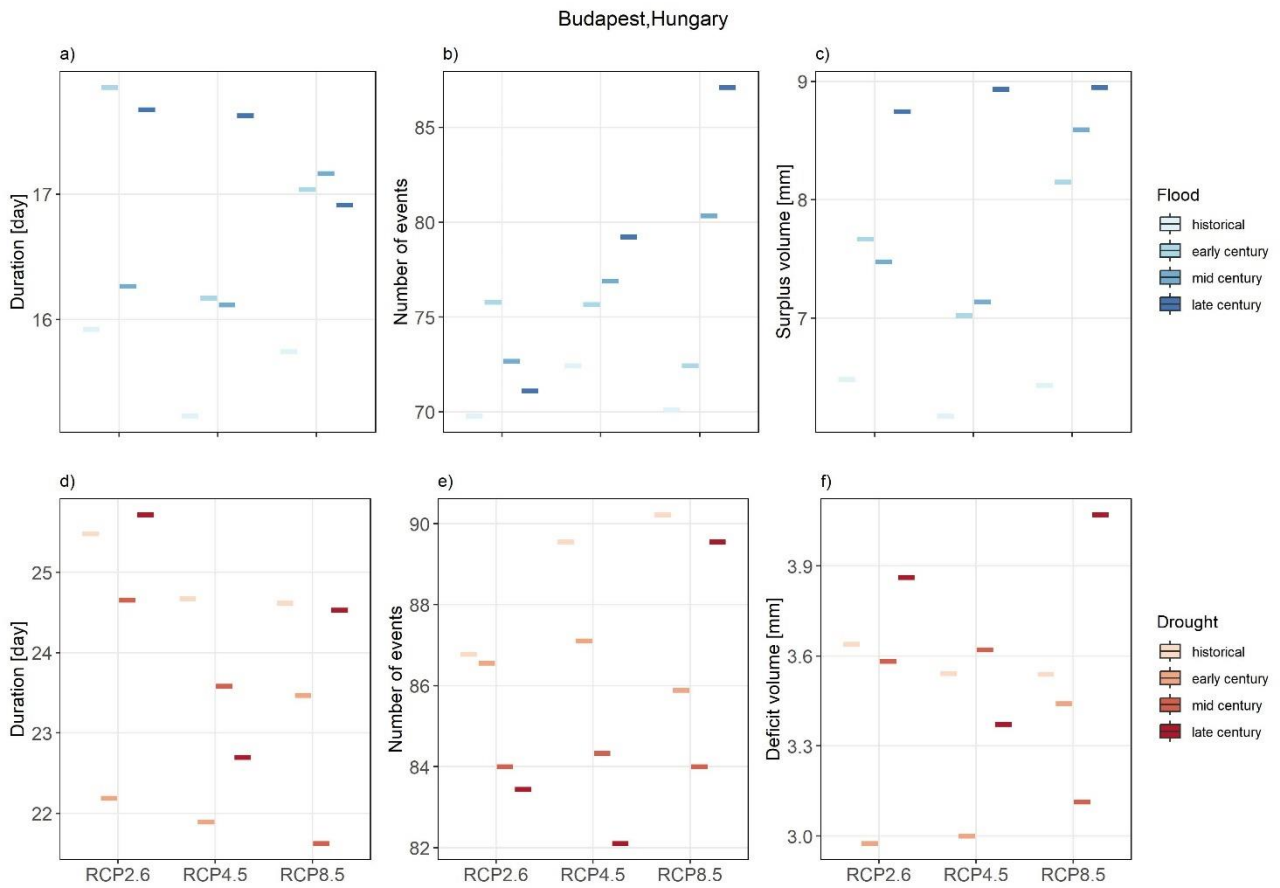


**Figure 5.12:** Change in deficit volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in the Upper and Middle Guadalquivir, Spain.

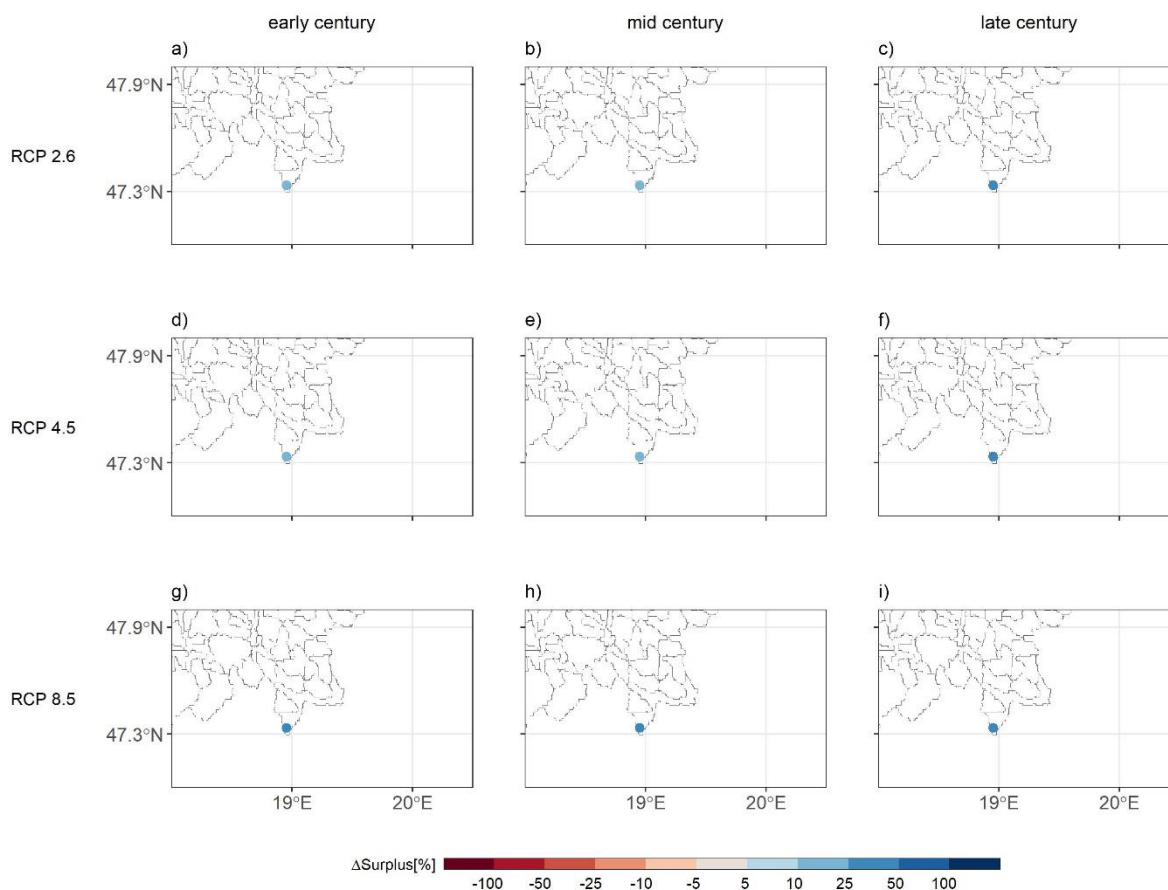
### 5.5 The Budapest Living Lab - Hungary

The statistical properties of the streamflow extreme events in Budapest, Hungary, are shown in Figure 5.13. The duration of the flood events is stable in time (Figure 5.13a), while the number of events increases in time, especially under the high emission scenario (Figure 5.13b). The surplus volume increases in time and with the emission scenario in a similar way in comparison to the number of events (Figure 5.13c and Figure 5.14). The duration of drought events shows a variability in time depending on the emission scenario (Figure 5.13d).

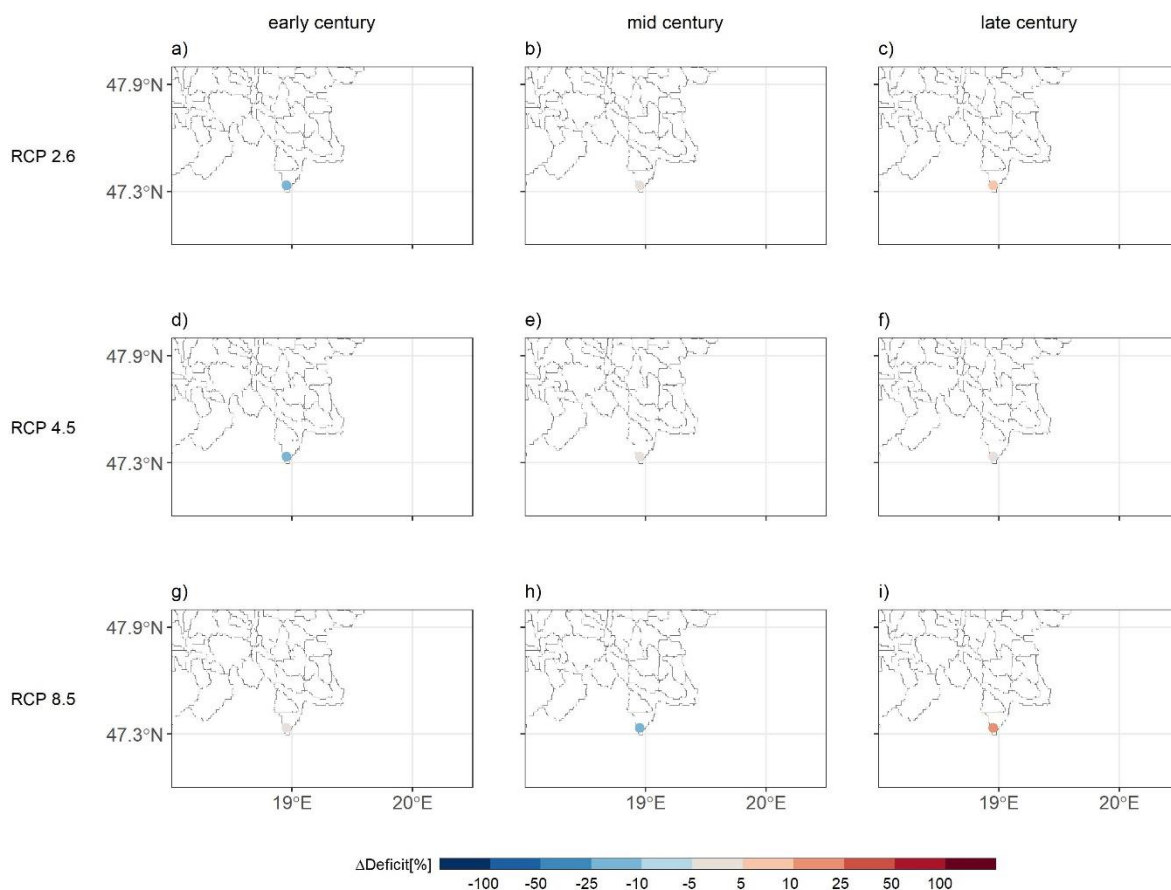
The duration decreases in the early century and tends to reach values close to historical in the late century under all emission scenarios. The number of events tend to decrease from the historical to the late century under the low and medium emission scenarios (Figure 5.13e). The deficit volume (Figure 5.13f) shows changes in time and with the emission scenarios similar to the drought event duration (Figure 5.13d). Moreover, the deficit volume decreases in the early century (Figure 5.15a, 5.15d and 5.15g) and increases in the late century under the low and high emission scenarios (Figure 5.15c and 5.15i).



**Figure 5.13:** Points representing the mean model ensemble duration, number and surplus/deficit volume of the flood (a,b,c) and drought (d,e,f) events in the historical, early, mid and late century for the RCP 2.6, 4.5, 8.5 emission scenarios in Budapest, Hungary. Note that this living lab is covered by the E-HYPE model setup with a single basin.



**Figure 5.14:** Change in surplus volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Budapest, Hungary.



**Figure 5.15:** Change in deficit volume in the early, mid and late century compared to the historical period for the RCP 2.6 (a,b,c), 4.5 (d,e,f) and 8.5 (g,h,i) emission scenarios in Budapest, Hungary.

## 5.6 Summary of the results

To provide an overview across all living labs, we collectively present here the changes for the different impact indicators focusing on the hydrological extremes (see Table 5.1).

**Table 5.1:** Summary of climate change impacts on the different indicators for the five living labs, 4 periods and 3 RCPs (RCP2.6/4.5/8.5).

Extremes	Indicators	Period	The Netherlands	Greece	Italy	Spain	Hungary
Floods	Duration [day]	Historical	6/6/6	6/6/6	6/6/6	15/15/14	16/15/16
		Early-century	7/7/6	5/5/6	6/7/6	14/15/13	18/16/17
		Mid-century	6/7/6	6/5/5	7/6/7	17/13/13	16/16/17
		End-century	6/7/7	6/5/4	7/7/6	15/14/10	18/17/17
	Number of events [-]	Historical	175/178/180	179/179/180	161/161/159	71/71/72	70/72/70
		Early-century	184/184/179	171/165/166	161/171/162	60/66/60	76/75/72
		Mid-century	174/182/176	162/159/155	167/162/155	65/61/57	73/77/80
		End-century	168/176/190	171/159/125	175/155/147	67/56/47	71/79/87
	Surplus volume [mm]	Historical	8.3/8.1/7.9	17.5/17.1/16.4	19/18.2/20.1	12.8/11.7/12.2	6.5/6.2/6.4
		Early-century	9.9/10.2/9.8	15.7/16.2/17.5	21.4/19.4/19.8	10.9/11.6/10.2	7.8/7.0/8.1
		Mid-century	9.3/10.8/11	17.6/16.5/15.1	22.7/18.6/25.5	16.1/10.7/11.4	7.5/7.1/8.6

		End-century	9.5/11/13	17.7/17.2/12.6	21.7/21.5/25.2	13.0/12.3/9.4	8.7/8.9/8.9
Droughts	Duration [day]	Historical	16/16/16	33/34/34	17/17/16	35/34/35	25/25/25
		Early-century	15/16/16	41/39/41	18/18/17	42/38/42	22/22/24
		Mid-century	16/16/17	42/41/46	16/19/19	41/42/48	25/24/22
		End-century	16/17/18	38/44/55	17/19/22	37/43/57	26/23/24
	Number of events [-]	Historical	133/135/136	59/58/60	129/129/130	42/37/42	87/89/90
		Early-century	143/144/139	63/68/68	134/136/144	42/41/44	86/87/86
		Mid-century	144/150/154	65/69/70	139/136/142	41/40/46	84/84/84
		End-century	143/150/159	68/67/88	134/142/141	42/42/49	83/82/89
	Deficit volume [mm]	Historical	1.7/1.7/1.7	0.4/0.5/0.5	1.4/1.5/1.4	0.3/0.3/0.3	3.6/3.5/3.5
		Early-century	1.7/1.7/1.9	0.7/0.8/0.9	1.6/1.7/1.6	0.4/0.3/0.5	3.0/3.0/3.4
		Mid-century	1.8/1.8/2.1	0.8/0.9/1	1.5/1.9/1.7	0.4/0.4/0.6	3.6/3.6/3.1
		End-century	1.9/1.9/2.3	0.8/1/1.4	1.5/1.9/2.4	0.4/0.4/0.9	3.9/3.4/4.1

## 6 Towards the future evolution of the I-CISK climate services

### 6.1 Conclusions from state-of-the-art investigations for understanding and predicting extremes

In this deliverable, we developed, adopted and investigated state-of-the-art scientific data and methods for better understanding and representing extremes in two components of the climate service at the living lab scale dealing with historical data and future projections: (1) improving process understanding through hybrid hydrological modelling, and (2) assessing the impact of climate change on hydro-climatic extreme impact indicators.

With regard to the hybrid hydrological modelling, we applied four post-processing methods to streamflow simulations from the continental scale process-based E-HYPE model at the scale of the Living Labs and evaluated the model performance using three different metrics, while also analysing the spatial pattern of the post-processing skill. The investigation moved a step forward, by identifying the primary factors that drive performance enhancements gained in each post-processing method. We concluded that:

- Post-processors have high potential to highly improve the quality of the raw simulations from large-scale process-based hydrological models used in continental climate services. This conclusion is valid for different characteristics of the streamflow signal, i.e. total volume and high/low extremes.
- The comparative analysis across the three living labs (the Netherlands, Italy and Spain) and across the different post-processing techniques, a similar spatial pattern of skill improvements is observed. In particular, results show higher skill in the Spanish living lab, which is especially pronounced in the context of extremes. This indicates that post-processing can add high value in river systems that are subject to reservoir management and in which large-scale models fail to represent the reservoir regulation schemes.
- The skill from post-processing is strongly linked to key physiographic and hydro-climatic drivers, such as mean precipitation, mean temperature, hydrological regimes, elevation and evaporative index. The ranking of these drivers varies across the different performance metrics, indicating sensitivity to the characteristics of the streamflow signal that is aimed to be improved.
- Overall, the recurrent identification of hydrological regimes, described here by the clusters, as a key driver for both total volume and high and low extremes, highlights the method's potential to apply post-processing at ungauged locations within and outside the living labs.

With regard to the assessment of the climate change on hydro-climatic extreme impact indicators, we applied a threshold-level method to detect the streamflow extreme events, accounting for floods and droughts, in each living lab under present and future conditions. We then characterise the extreme events by focusing on specific climate impact indicators that are statistical properties of duration, number of events and surplus/deficit volume. The extensive investigation using an ensemble of nine models from the large Euro-CORDEX ensemble of hydro-climatic simulations showed that (see also a summary in section 5.6):

- Among the extreme statistical properties, duration and surplus/deficit volume are the most informative indicators about the changes in time and under the different emission scenarios.
- The duration and surplus volume of the flood events, though with different magnitude, increases in the Rijnland and the Budapest Living Labs from the historical period to the late century, especially under the high emission scenario, while the Crete and the Guadalquivir Living Labs show a decrease of these two streamflow extreme properties.
- The duration and deficit volume increase in the Emilia Romagna and the Guadalquivir Living Labs from the historical period to the late century. However, these properties remain stable or decrease in the Crete and the Netherlands Living Labs.

## 6.2 Moving beyond the state-of-the-art operational climate services

It has been a user request to provide accurate hydrological predictions at the local scale and at the locations of interest. The hybrid hydrological modelling efforts aim to improve model outputs to be closer to real observations instead of the modelled reality, which further affects the users' trust to the climate service. The work presented here has been towards this direction and have showed significant improvements, consequently implying high added value to local decision-makers. The analysis at this stage focuses on the historical model performance of the simulated streamflow, while the ongoing efforts are targeting sub-seasonal to seasonal forecasts. We note that here we neither aim to address evaluation from the perspective of the user-oriented decision variables nor a co-evaluation – which is less a statistical analysis using performance metrics, but more to understand how decision-making has evolved after the delivery of the I-CISK climate services. These will be addressed in the upcoming deliverable D3.4.

The current results from hybrid hydrological modelling, where skills show spatial compensation among different post-processing methods, requires a further investigation on ensemble and averaging techniques, for instance, applying probabilistic multi-model-ensemble approaches to benefit from multiple model outputs. Therefore, the next step would be to explore possibilities of such techniques, such as copula-based Bayesian Model Averaging (Madadgar and Moradkhani, 2014), yet subject to research, to combine post-processing results from individual models and characterise the uncertainty induced. This would allow more reliable simulations and predictions by weighing and combining their individual results according to the bias/errors against the observations. In addition, we plan to apply this method in a sub-seasonal to seasonal forecasting context and consequently generate a new AI-enhanced hydrological forecasting service for the I-CISK climate services.

Moreover, on the way of operationalizing the post-processing techniques, a regionalization is necessary to transfer the knowledge gained from the gauged stations to the ungauged locations over which decisions are also made. This work can be built on the analysis of potential drivers, where hydrological regimes (clusters) were identified to be one of the most influencing factors for model performance and post-processing skills. Therefore, a regionalization can be conducted allowing generalised post-processing parameters within river systems that are characterised by similar hydrological response. Finally, the regionalized post-processing technique will be incorporated into the operational services, adjusting the seasonal streamflow forecasts currently generated without accounting for integration of local data.

The assessment of hydro-climatic extreme impacts shows that the duration and surplus/deficit volume were the most informative indicators with regard to the streamflow extreme changes under future periods in all the considered Living Labs. These indicators will be integrated in the I-CISK climate services as boxplots and maps for long-term decision making by the local users. A possible way forward would be to intercompare these results with other indicators currently used, if any, for policy making. We note that the current assessment, although local, has not considered local observations, due to the limited availability in space and time for trend analysis. Nevertheless, there is validity and usability of the results/insights given that the investigation was based on state-of-the-art high-resolution projections.

## References

- Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J. C. M., Hasan, A., and Pineda, L., (2020): Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation. *Hydrol. Earth Syst. Sci.*, 24, 535–559, <https://doi.org/10.5194/hess-24-535-2020>
- Berg, P., Photiadou, C., Simonsson, L., Sjökvist, E., Thuresson, J., and Mook, R., (2021): Temperature and precipitation climate impact indicators from 1970 to 2100 derived from European climate projections. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.9eed87d5
- Berg, P., Photiadou, C., Bartosova, A., Biermann, J., Capell, R., Chinyoka, S., Fahlesson, T., Franssen, W., Hundecha, Y., Isberg, K., Ludwig, F., Mook, R., Muzuusa, J., Nauta, L., Rosberg, J., Simonsson, L., Sjökvist, E., Thuresson, J., and van der Linden, E., (2021): Hydrology related climate impact indicators from 1970 to 2100 derived from bias adjusted European climate projections. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.73237ad6
- Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., 2017. *Classification and Regression Trees*. Chapman and Hall/CRC, New York. <https://doi.org/10.1201/9781315139470>
- Gudmundsson, L., Bremnes, J.B., Haugen, J.E., Engen-Skaugen, T., 2012. Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations &ndash; a comparison of methods. *Hydrol. Earth Syst. Sci.* 16, 3383–3390. <https://doi.org/10.5194/hess-16-3383-2012>
- Hundecha, Y., Arheimer, B., Donnelly, C., Pechlivanidis, I., 2016. A regional parameter estimation scheme for a pan-European multi-basin model. *J. Hydrol. Reg. Stud.* 6, 90–111. <https://doi.org/10.1016/j.ejrh.2016.04.002>
- Jiang, Z., Li, W., Xu, J., Li, L., 2015. Extreme Precipitation Indices over China in CMIP5 Models. Part I: Model Evaluation. *J. Clim.* 28, 8603–8619. <https://doi.org/10.1175/JCLI-D-15-0099.1>
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* 22, 6005–6022. <https://doi.org/10.5194/hess-22-6005-2018>
- Lamontagne, J.R., Barber, C.A., Vogel, R.M., 2020. Improved Estimators of Model Performance Efficiency for Skewed Hydrologic Data. *Water Resour. Res.* 56, e2020WR027101. <https://doi.org/10.1029/2020WR027101>
- Lehner, B. and Döll, P., 2004. Development and validation of a global database of lakes, reservoirs and wetlands, *J. Hydrol.*, 296, 1–22, doi:10.1016/j.jhydrol.2004.03.028
- Lehner, B., Verdin, K., and Jarvis, A., 2008. New global hydrography derived from spaceborne elevation data, *Eos, Trans. AGU*, 89, 93–94, doi:10.1029/2008EO100001
- Lehner, B., Liermann, C. R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J. C., Rödel, R., Sindorf, N., and Wisser, D., 2011. High-resolution mapping of the world’s reservoirs and dams for sustainable river-flow management, *Front. Ecol. Environ.*, 9, 494–502, doi:10.1890/100125
- Madadgar, S., Moradkhani, H., 2014. Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resour. Res.* 50, 9586–9603. <https://doi.org/10.1002/2014WR015965>
- Madsen, H., Thyregod, P., 2010. *Introduction to General and Generalized Linear Models*. CRC Press.

Moschini, F., Emerton, R., et al., 2022. Preliminary Report: Information on Climate Service Needs and Gaps, I-CISK Deliverable 2.1, Available online at [www.icisk.eu/resources](http://www.icisk.eu/resources)

Nachtergaele, F., van Velthuisen, H., Verelst, L., and Wiberg, D., 2012. Harmonized world soil database version 1.2, FAO, Rome and IIASA, Laxenburg, Austria.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. *J. Hydrol.* 10, 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., Loumagne, C., 2005. Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. *J. Hydrol.* 303, 290–306. <https://doi.org/10.1016/j.jhydrol.2004.08.026>

Pechlivanidis, I.G., Crochemore, L., Rosberg, J., Bosshard, T., 2020. What Are the Key Drivers Controlling the Quality of Seasonal Streamflow Forecasts? *Water Resour. Res.* 56, e2019WR026987. <https://doi.org/10.1029/2019WR026987>

Pechlivanidis, I.G., Arheimer, B., Donnelly, C. et al., 2017. Analysis of hydrological extremes at different hydro-climatic regimes under present and future conditions. *Climatic Change* 141, 467–481. <https://doi.org/10.1007/s10584-016-1723-0>

Pham, L.T., Luo, L., Finley, A., 2021. Evaluation of random forests for short-term daily streamflow forecasting in rainfall- and snowmelt-driven watersheds. *Hydrol. Earth Syst. Sci.* 25, 2997–3015. <https://doi.org/10.5194/hess-25-2997-2021>

Portmann, F. T., Siebert, S., and Döll, P., 2010. MIRCA2000 – Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling, *Global Biogeochem. Cy.*, 24, GB1011, doi:10.1029/2008GB003435

Quesada-Montano, B., Di Baldassarre, G., Rangecroft, S., Van Loon, A.F., 2018. Hydrological change: Towards a consistent approach to assess changes on both floods and droughts. *Advances in Water Resources* 111, 31–35. <https://doi.org/10.1016/j.advwatres.2017.10.038>

Siebert, S., Burke, J., Faures, J. M., Frenken, K., Hoogeveen, J., Döll, P., and Portmann, F. T., 2010. Groundwater use for irrigation – a global inventory, *Hydrol. Earth Syst. Sci.*, 14, 1863–1880, doi:10.5194/hess-14-1863-2010

Siebert, S., Döll, P., Hoogeveen, J., Faures, J.-M., Frenken, K., and Feick, S., 2005. Development and validation of the global map of irrigation areas, *Hydrol. Earth Syst. Sci.*, 9, 535–547, doi:10.5194/hess-9-535-2005

Slater, L.J., Arnal, L., Boucher, M.-A., Chang, A.Y.-Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., Villarini, G., Wilby, R.L., Wood, A., Zappa, M., 2023. Hybrid forecasting: blending climate predictions with AI models. *Hydrol Earth Syst Sci* 27, 1865–1889. <https://doi.org/10.5194/hess-27-1865-2023>

Teutschbein, C., Montano, B.Q., Todorović, A., Grabs, T., 2022. Streamflow droughts in Sweden: Spatiotemporal patterns emerging from six decades of observations. *Journal of Hydrology: Regional Studies* 42, 101171. <https://doi.org/10.1016/j.ejrh.2022.101171>

Van Loon, A.F., 2015. Hydrological drought explained. *WIREs Water* 2, 359–392. <https://doi.org/10.1002/wat2.1085>

Weedon, G.P., Balsamo, G., Bellouin, N., Gomes, S., Best, M.J., Viterbo, P., 2014. The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resour. Res.* 50, 7505–7514. <https://doi.org/10.1002/2014WR015638>

Willmott, C., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* 30, 79–82. <https://doi.org/10.3354/cr030079>

## Appendix 1 Glossary

Acronym	Definition
AET	Actual Evapotranspiration
AI	Artificial Intelligence
CII	Climate Impact Indicator
CS	Climate Service
GCM	Global Circulation Model
GLM	Generalised Linear Model
LL	Living Lab
LSTM	Long Short-Term Memory neural network
MAE	Mean Absolute Error
ML	Machine Learning
NSE	Nash-Sutcliffe Efficiency
PET	Potential Evapotranspiration
QM	Quantile Mapping
RCM	Regional Climate Model
RCP	Representative Concentration Pathway
RF	Random Forest
SMAE	Standardised Mean Absolute Error
WP	Work Package



# I-CISK

HUMAN CENTRED CLIMATE SERVICES

## Colophon:

This report has been prepared by the H2020 Research Project “Innovating Climate services through Integrating Scientific and local Knowledge (I-CISK)”. This research project is a part of the European Union’s Horizon 2020 Framework Programme call, “Building a low-carbon, climate resilient future: Research and innovation in support of the European Green Deal (H2020-LC-GD-2020)”, and has been developed in response to the call topic “Developing end-user products and services for all stakeholders and citizens supporting climate adaptation and mitigation (LC-GD-9-2-2020)”. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101037293.

This four-year project started November 1<sup>st</sup> 2021 and is coordinated by IHE Delft Institute for Water Education. For additional information, please contact: Micha Werner ([m.werner@un-ihe.org](mailto:m.werner@un-ihe.org)) or visit the project website at [www.icisk.eu](http://www.icisk.eu)

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101037293

