Robustness of hydrological models for simulating impact of climate change on high and low streamflow in the Lesse

Master Thesis

Amy ten Berge

Enschede, April 2024

UNIVERSITY OF TWENTE.

Robustness of hydrological models for simulating impact of climate change on high and low streamflow in the Lesse

Master Thesis

Enschede, April 2024

UNIVERSITY OF TWENTE.

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

to be presented on May 3, 2024

Author:

A.A. (Amy) ten Berge a.a.tenberge@student.utwente.nl

Supervised by:

Dr. Ir. M.J. (Martijn) Booij

Department of Civil Engineering, Water Engineering & Management

Dr. Ing. T.H.M. (Tom) Rientjes

Department of Water Resources

University of Twente, the Netherlands

Cover picture: Warming stripes, annual mean temperature in the Lesse, 1968-2021

PREFACE

This thesis is the final piece to obtain my master's degree in Civil Engineering and Management at the University of Twente. In my research project, I have assessed the robustness of two hydrological models for simulating the impact of climate change on high and low streamflow. I hope that this thesis will show you that the ability of hydrological models to simulate climate change impacts should not be taken for granted, but that deliberate choices in the calibration procedure can contribute to a more certain projection of the impact of climate change in the future.

Setting up this research project and diving into the robustness of hydrological models has made me learn a lot about the complexities and uncertainties involved in hydrological modelling. This has deepened my enthusiasm for research in this field. This would not have been possible without the valuable guidance from my supervisors of the University of Twente. The precise feedback of Martijn Booij and critical view of Tom Rientjes have improved this thesis a lot. Also, the long discussions during the meetings were very interesting and helpful. Their confidence and enthusiasm motivated me during this research. I would like to thank Martijn and Tom for this and more.

This study would not have been possible without the data on the KNMI'23 climate scenarios. I would like to thank Frederiek Sperna Weiland from Deltares and Henk van den Brink from KNMI for providing the data and helping me to get started with it. Next to that, I would like to thank Jules Beersma and Leon van Voorst for providing the historic climatic timeseries of the Lesse.

Lastly, I would like to thank my family and friends, for their support and for making my time at Enschede wonderful, before and during my thesis. A special thanks to all people that joined the walks in the lunch break, which I enjoyed a lot!

Hopefully, you will enjoy reading this thesis as much as I enjoyed doing my research.

Amy ten Berge

Enschede, April 2024

SUMMARY

Extreme high and low streamflow commonly negatively affects societies. A concern is that climate change will increase the frequency and severity of problems due to extreme streamflow. Hydrological models are used to simulate the rainfall-runoff transformation to quantify the impact of climate change on extreme streamflow. However, parameters of hydrological models are optimized for historic conditions and may not be valid for future scenarios.

This study evaluated the robustness of the hydrological models HBV and GR6J for simulating impact of climate change on high and low streamflow in the Lesse catchment, Belgium. Models are defined to be robust when they do not show notable deterioration in performance under changing climatic conditions. This means that there is no need to recalibrate model parameters or update model structure. To evaluate the robustness, both models were evaluated on historic periods that resemble climatic conditions projected by the recently published KNMI'23 climate scenarios of the Royal Netherlands Meteorological Institute.

To determine the periods that resemble future conditions, meteorological indicators were defined that summarize the meteorological conditions leading to high and low streamflow. It was found that the 3-day precipitation sum that is exceeded 6 days per year is an indicator for annual maximum daily discharges. This meteorological indicator for high flows was expected to be higher in 2100 for KNMI'23 scenarios with high future greenhouse gas emissions. The 150-day potential precipitation deficit that is exceeded 14 days per year was determined as an indicator for annual minimum 7-day mean discharges. This meteorological indicator for low flows was expected to be higher in 2100 for all KNMI'23 scenarios. Based on the expected changes in the future, a number of historic years was selected that resemble future conditions for each KNMI'23 climate scenario, for normal years (median) and extreme years (5% exceedance probability).

Both models showed a loss in performance in validation periods that resemble future conditions compared to calibration periods with historic climatic conditions. However, both models still performed acceptable, possibly due to the use of a multi-objective function for calibration. The GR6J model showed better performance in the simulation of high flows but had difficulties with simulation of low flows. The HBV model showed better performance for simulating low flows but could be improved on its ability to capture summer peaks. The optimal parameter values of the HBV and GR6J model were proven to be different when calibrated on different periods. This indicates that parameter sets calibrated on historic conditions are suboptimal under changing climatic conditions and thus lose validity. This advocates the use of a calibration period that closely resembles future climatic conditions for climate impact studies.

The median change in annual maximum daily discharge was projected to be between -14% and +27% in 2100 compared to reference period 1991-2020. For low streamflow, the median projected change in annual minimum 7-day mean discharge was between -66% and +13%. These ranges covered uncertainties in climate scenario, model structure and calibration approach. Two different calibration approaches were used, namely calibration on historic periods and calibration on periods that resemble future climatic conditions. The uncertainty in the projected impact of climate change was mainly due to uncertainty in future greenhouse gas emissions and climate response. However, the uncertainty subject to selected model structures and calibration approaches should not be neglected.

Therefore, it is important that the ability of hydrological models to simulate climate change impact is not taken for granted. Improving model structures of conceptual hydrological models, focusing for example on the simulation of summer peaks, may improve robustness of these models and therefore contribute to a projection of the impact of climate change in the future with less uncertainties.

TABLE OF CONTENTS

LIST OF SYMBOLS

HBV Model parameters:

FC - Field capacity of the soil moisture storage of the HBV model [mm]
β - Parameter in the soil moisture routine of the HBV model [-]
LP - Limit for potential evapotranspiration as a fraction of FC in the HBV model [-]
α - Measure of non-linearity for quick runoff in the HBV model [-]
K_f - Recession coefficient for upper response box of the HBV model [day^{-1}]
K_s - Recession coefficient for lower response box of the HBV model [day^{-1}]
$PERC$ - Maximum percolation from upper to lower response box in the HBV model [mm/day]
$Cflux$ - Maximum value of capillary flux in the HBV model [mm/day]

GR6J Model parameters:

X_1 - Capacity of the production store of the GR6J model [mm]
X_2 - Catchment exchange coefficient for the GR6J model [mm/day]
X_3 - Capacity of the non-linear routing store $R1$ of the GR6J model [mm]
X_4 - Time base of the unit hydrographs in the GR6J model [day]
X_5 - Threshold for catchment exchange for the GR6J model [-]
X_6 - Exponential routing store depletion coefficient of the GR6J model [mm]

KNMI'23 climate scenarios:

2100Hd - Climate scenario based on high emissions and a trend towards a drier future climate.
2100Hn - Climate scenario based on high emissions and a trend towards a wetter future climate.
2100Ld - Climate scenario based on low emissions and a trend towards a drier future climate.
2100Ln - Climate scenario based on low emissions and a trend towards a wetter future climate.

Testing schemes:

DSST - Differential split-sample test
SST - Split-sample test

Codes of testing schemes - The codes of the different tests are made up of the type of flow (High (H), Low (L) or Both (B)), the type of test (SST or DSST), the type of year (Normal (N) or Extreme (E)) and the climate scenario (Figure 6). In the codes for the SSTs, '_I' and '_II' are added. '_I' means calibration on period 1996-2020 and validation on period 1970-1994. '_II' means the opposite.

Objective functions:

NS_w - Weighted form of the Nash-Sutcliffe criterion, used to give emphasis to high flows [-]
NS_{inv} - Nash-Sutcliffe criterion calculated on inverse transformed flows [-]
RVE - Relative volume error [%]
y_w - Multi-objective function for high flows that combines NS_w and RVE [-]
y_{inv} - Multi-objective function for low flows that combines NS_{inv} and RVE [-]
y_{comb} - Multi-objective function for both flows that combines NS_w, NS_{inv} and RVE [-]
Q^i_{sim} (q^i_{sim}) - (Inverse transformed) Simulated discharge at day i [m^3/s]
Q^i_{obs} (q^i_{obs}) - (Inverse transformed) Observed discharge at day i [m^3/s]
$\overline{Q_{obs}}$ ($\overline{q_{obs}}$) - Mean of (inverse transformed) observed discharge over n days [m^3/s]
n- Total number of timesteps [-]

Indices:

Q_{max} - Annual maximum daily discharge [m^3/s]
$Q7_{min}$ - Annual minimum 7-day mean discharge [m^3/s]
$P3\_rank6$ - Meteorological indicator for high flows; 3-day precipitation sum exceeded 6 days per year [mm]
$PPD150\_rank14$ - Meteorological indicator for low flows; 150-day potential precipitation deficit that is exceeded 14 days per year [mm]

1. INTRODUCTION

1.1. Problem Context

Extreme high and low streamflow commonly negatively affects societies. Most eye-catching are high flows, which may cause risk to human life and can negatively affect economic development (Malede et al., 2022). Low flows may negatively impact important river functions, such as water supply, power production and navigation, on a much larger spatial and time scale and may result in deterioration of water quality (Demirel et al., 2013a).

A concern is that climate change will increase the frequency and severity of these problems (Tongal & Booij, 2018). Gudmundsson et al. (2021) showed that climate change, indicated by higher temperatures and changes in rainfall intensity, affects the entire flow regime, including high and low flows. The impact of climate change on extreme high flows is highly linked to trends in extreme precipitation events, especially in smaller catchments (Do et al., 2017). In larger catchments, temperature increases leading to snow melt and decreases in soil moisture may be dominant (Hirabayashi et al., 2021). The increase in snow melt may increase the severity of high flow events, while the decrease in soil moisture may result in less severe high flow events. Temperature increases may also decrease low flows, due to drier summers and increases in potential evapotranspiration (Marx et al., 2018; Rameshwaran et al., 2021; Tohver et al., 2014; Trenberth, 1999). However, these effects are uncertain, due to the complexity of different hydrological processes (Prudhomme et al., 2014).

Quantifying the impact of climate change on extreme streamflow involves a number of steps. Commonly, a number of global circulation models (GCMs) is used and combined to represent a multi-model ensemble (Coron et al., 2012; Vormoor et al., 2018). GCMs simulate the climate based on one or more emission scenarios as input. As the resolution of a GCM is too coarse for use at regional catchment scale, the second step often involves downscaling the output of the GCM with the use of regional climate models (RCMs) or statistical downscaling. Lastly, a hydrological model is used to simulate the rainfall-runoff transformation and this simulated runoff is compared with the recorded runoff (Chiew et al., 2014; Vaze & Teng, 2011). Whereas this last step generally contributes less to the overall uncertainty compared to the other steps (Arnell, 2011; Kay et al., 2009; Prudhomme & Davies, 2009; Teng et al., 2012; Wilby & Harris, 2006), the uncertainty in the rainfall-runoff transformation due to changing catchment conditions can be substantial (Merz et al., 2011; Wilby, 2005).

In October 2023, the Royal Netherlands Meteorological Institute (KNMI) published its new climate scenarios (KNMI'23) for the Netherlands and the Rhine and Meuse basin. Subject to respective emission scenarios, the KNMI'23 scenarios describe how the future climate could look like and form the basis for research to the effects of climate change (Van Dorland et al., 2023). The impact of climate change on extreme streamflow in the Rhine and Meuse basin can be estimated using these KNMI'23 scenarios and hydrological models. However, parameters of hydrological models are optimized for historic conditions and may not be transferable in time (Thirel et al., 2015). The ability of hydrological models to simulate catchment runoff behaviour under changing climatic conditions should thus not be taken for granted. It is important to assess the quality of hydrological models, to improve impact projections for the future.

1.2. Theoretical Framework

1.2.1. Hydrological models

Hydrological models are simplified representations of the real-world system (Moradkhani & Sorooshian, 2009), which are mainly used to enhance knowledge and understanding about hydrological processes, or as tools to predict or simulate impacts of interventions or changes (Sitterson et al., 2018). The focus of this study is on rainfall-runoff models, a type of hydrological model that simulates the transformation

of precipitation into runoff in a catchment (Sitterson et al., 2018). Rainfall-runoff models are forced with hydro-climatic timeseries, while the catchment characteristics are described by model parameters.

Various types of rainfall-runoff models exist. Models can for example be classified based on how the models describe hydrological processes, being empirical, conceptual or physically-based (Dwarakish & Ganasri, 2015). Empirical models, or data-driven models, only make use of observed data and thus do not characterize physical processes in the hydrological system (Jaiswal et al., 2020). Due to their black-box nature, empirical models lack physical analogues and transparency (Razavi, 2021). Conceptual and physically-based models do represent hydrological processes and storages that are theoretically observable and are mass-conservative (Jaiswal et al., 2020). Conceptual models often simplify hydrologic processes with empirical relations, whereas physically-based models represent characteristics of the catchment more explicitly, by using momentum and energy conservation equations and more model parameters. Due to their simpler model structure, conceptual models do not need as many parameters and computational resources compared to physically-based models.

Due to a lack of knowledge on the functioning of the hydrological system and/or a lack of data, all models remain to some extent conceptual and thus will always contain some parameters that have to be estimated (Beven & O'Connell, 1982; Coron et al., 2012; Montanari & Koutsoyiannis, 2012). This is done via calibration, which means that parameters are tuned such that the model most closely and consistently simulates the hydrological response in a basin over some historical calibration period for which both forcing data and output are available (Moradkhani & Sorooshian, 2009).

For calibration of hydrological models, an objective function is selected based on the purpose of simulation. An objective function is a measure of the error between the model-simulated output and observations (Gupta et al., 1998). Parameters are tuned such that this objective function is optimized. Parameter sets optimized for a certain objective function may be suboptimal for other objective functions. This for example means that models calibrated with focus on the simulation of peak flows may poorly simulate low flows. Multiple single objective functions can be aggregated into one multi-objective function. Gupta et al. (1998) advocated to use a multi-objective approach, as this limits the loss of information that is observed when using a single objective function. For this reason, the multi-objective paradigm is widely used in the calibration procedure of hydrological models (De Vos & Rientjes, 2007).

1.2.2. Hydrological models under change

To quantify the impact of climate change, hydrological models are applied under transient hydro-climatic forcing, while keeping the model parameter values constant (Chiew et al., 2014). Stationarity is assumed, which means that the system is characterized based on past observations only (Chiew et al., 2014). However, it may not be valid to assume stationarity in a changing climate, as the optimal parameter set may be different for future periods (Blöschl & Montanari, 2010; Ji et al., 2023; Merz et al., 2011; Nicolle et al., 2021; Thirel et al., 2015). Parameter values calibrated on historic periods may thus be suboptimal for modelling impact of climate change.

Two reasons can be distinguished for the potential change of model parameter values in response to changes in climatic conditions. First of all, dominant processes and catchment conditions may fundamentally change. For example, climate change may lead to changes in land cover, as vegetation may adapt their root systems to respond to water stress in dry periods (Bouaziz, 2021). Secondly, the optimal parameter set may change for different calibration periods because parameters tend to compensate for model structure errors and data errors (Merz et al., 2011). Model parameters are tuned such that a predefined objective function is optimized, aiming to minimize the difference between simulated and observed runoff. The choice for an objective function determines which aspects of the

hydrograph receive most attention. To optimize the objective function, parameters may thus represent more than just the static catchment characteristics and may thus also compensate for errors in the model structure, hydro-climatic forcing data, initial conditions and observations (De Vos et al., 2010).

Osuch et al. (2015) found that the parameters of the HBV model are mostly influenced by mean precipitation, mean wet day precipitation and mean evapotranspiration, even though the general idea of parameters in hydrological models is that that they are independent of time-varying boundary conditions. Wagener et al. (2003) and Niel et al. (2003) showed that parameters sets are significantly different when calibrated on periods with different hydroclimatic conditions but could not find obvious relationships between the parameter sets and climatic conditions in the calibration periods. This shows that calibration parameters may not have a physical interpretation as they compensate for uncertainties in model structure. This parameter non-uniqueness is one of the well-known limitations of conceptual models (Beven, 2006; Wilby, 2005).

Parameter non-uniqueness, also parameter non-identifiability, means that multiple sets of parameter values can fit equally well to observed data (Beven, 2006). It may thus be difficult to determine which parameter set represents the underlying system most correctly. A parameter set that does not lead to the optimal value for the objective function may not be chosen, whereas this parameter set would qualitatively give more correct simulation of the response in a catchment. Beven & Binley (1992) recognized that multiple parameter sets may be equally likely and introduced the GLUE method to estimate the uncertainty associated with model predictions due to uncertainty in parameter set. In this method, each possible parameter set is assigned a likelihood of being the best simulator for observed behaviour. Based on this, a probability distribution function for each parameter can be obtained. De Vos & Rientjes (2007) let go the idea of time-invariant parameters and used dynamic model parameters for a rainfall-runoff model of a meso-scale catchment. Next to that, minimizing calibration to the past by using physically-based models is often tried, however, the parameter values in these more complex models may not be representative for the underlying processes as well (Blöschl & Montanari, 2010).

By keeping model parameters constant in climate change applications, hydrologists thus assume that parameters are well-identified and that parameter values remain valid for a future period. As this may not be true, model performance generally declines when models are used to predict behaviour outside the calibration period using the same parameter set (Ji et al., 2023). This is even more the case when models are run under drastically different hydroclimatic conditions than those used for calibration.

Thirel et al. (2015) emphasized the importance of assessing the robustness of models before applying them under changing conditions. In this study, a model is defined to be robust when a model does not show notable deterioration in performance under changing climatic conditions. This means that there is no need to recalibrate model parameters or update the model structure. Robust models are not only able to simulate closely to observations in a reference period (i.e. precision) but are also able to react to change in an accurate way (Thirel et al., 2015). A lack of model robustness could undermine the confidence in hydrological models (Blöschl & Montanari, 2010).

1.2.3. Tests for assessing the robustness

The most widely used tool to assess model robustness in time and space is the four-level testing scheme proposed by Klemeš (1986). The underlying principle of this testing framework supposes that when a model needs calibration, the model should be evaluated on both the calibration dataset and an independent dataset. This scheme consists of the split-sample test (SST) on two independent periods, the proxy-basin test on two neighbouring catchments, the differential split-sample test (DSST) on two contrasting independent periods and the proxy-basin differential split-sample test on neighbouring catchments and contrasting periods.

The DSST is most relevant for applying models under changing climatic conditions, as this test is recommended for simulating flows other than those observed in the existing flow record (Klemeš, 1986). In the DSST, the available data is divided in two parts based on their climatic differences. The first part of the data is used for calibration, the second part for validation. The idea is that when the calibration and validation is done over climatically contrasting past periods, the model faces the difficulties it will have to deal with in the future (Nicolle et al., 2021). In case the model shows performs substantially worse in the validation period than in the calibration period, this test can falsify a hydrological model for applying it under change. However, a satisfactory behaviour in the DSST does not mean that the model can always be confidently used for future periods (Nicolle et al., 2021). This is because the observed historic variability in climate may not be as extensive as the changes that are expected to occur in the future (Stephens et al., 2020).

Statistical analysis is used to partition the reference historical period into periods with contrasting climates. In most applications of the DSST, this discretization is based on mean annual precipitation (Chiew et al., 2009; Luo et al., 2012; Refsgaard & Knudsen, 1996; Ruelland et al., 2015; Tramblay et al., 2013; Vaze et al., 2010; Wilby, 2005; Wu & Johnston, 2007), sometimes in combination with temperature (Dakhlaoui et al., 2017; Seiller et al., 2012; Xu, 1999) or potential evapotranspiration (Coron et al., 2012). In some other applications of the DSST, the discretization is based on observed runoff characteristics only (Donnelly-Makowecki & Moore, 1999; Seibert, 2003; Vormoor et al., 2018).

A limitation of the DSST is that only a small number of contrasting periods is identified (Coron et al., 2012; Dakhlaoui et al., 2019). Several studies tried to overcome this by proposing variations of the original testing scheme of Klemeš (1986). For example, Coron et al. (2012) proposed a generalization of the standard SST and DSST by testing as many climatic configurations as possible, including similar and contrasting conditions between calibration and validation. This generalized split-sample test considers all possible configurations of continuous subperiods of equal length. Besides, a random bootstrap SST was proposed (Coron, 2013), in which sub-periods are generated from randomly selected combinations of years. Dakhlaoui et al. (2017) increased the number of validation exercises by using four combinations (hot/dry, hot/wet, cold/dry and cold/wet years) (4-sub-period DSST) and lastly, the general differential split-sample test was proposed in which a combination of the random bootstrap SST and the 4 sub-period DSST was used to efficiently identify a number of contrasting periods (Dakhlaoui et al., 2019).

Despite the limitations that the testing framework of Klemeš (1986) may have, Ji et al. (2023) concluded that the application of (variations of) this test may still be the best course of action today. Only two tests for assessing model robustness were proposed without implementing the testing framework of Klemeš (1986), which are the robustness assessment test (RAT; Nicolle et al., 2021) and the proxy for model robustness (PMR; Royer-Gaspard et al., 2021). The main advantage of these tests is that robustness can be evaluated without employing a multiple calibration process. However, also these tests have limitations, often being similar to the limitations of the differential split-sample test (Nicolle et al., 2021).

1.2.4. Findings on model robustness

In recent decades, there has been increasing interest in exploring robustness for hydrological models, due to the growing demand for climate change impact assessments (Ji et al., 2023) and due to the growing interest on hydrological modelling under change linked to the Panta Rhei decade of IAHS (Montanari et al., 2013). Various studies applied a model robustness test, either to test whether a hydrological model could be used for future conditions (Chiew et al., 2009; Ruelland et al., 2015; Tramblay et al., 2013; Wilby, 2005; Xu, 1999), to compare the robustness of different models (Dakhlaoui et al., 2017; Donnelly-Makowecki & Moore, 1999; Refsgaard & Knudsen, 1996; Seiller et al., 2012) or to

compare or improve robustness by advanced calibration schemes (Luo et al., 2012; Seibert, 2003; Wilby, 2005; Wu & Johnston, 2007). Based on these studies, some overarching conclusions can be derived.

First, performance losses were found to be higher for wet and cold to dry and hot transfers than vice versa (Ji et al., 2023). In general, models tend to underestimate runoff with a wetter and colder future climate only slightly, whereas runoff is overestimated significantly with a dryer and hotter future climate (Dakhlaoui et al., 2019). Model parameters may thus be less transferable under drier future conditions. The reason for this may be related to the more pronounced nonlinearities in hydrological processes in drier environments compared to wetter regimes (Van Esse et al., 2013). Second, degradation of performance under changed conditions is especially apparent when the model is run under hydroclimatic conditions that differ substantially from the hydroclimatic conditions of the calibration period (Ji et al., 2023). Numerous studies tried to determine transferability limits, being the values for changes in climatic variables for a future climate between which the calibrated parameter values are valid (Bastola et al., 2011; Dakhlaoui et al., 2017, 2019; R. Singh et al., 2011; Sleziak et al., 2018; Vaze et al., 2010). For example, Vaze et al. (2010) derived that models, when calibrated over 20 years of data, can generally be used for catchments in Australia when the future mean annual rainfall is between 15% drier or 20% wetter than the mean annual rainfall in the calibration period. However, generalizing interpretations on model robustness is challenging, as interpretations depend on catchment characteristics, model structure, calibration method, the range of climatic configurations tested and the specific requirements of users (Coron et al., 2012; Gupta et al., 2014; Ji et al., 2023). Acknowledging these limitations, Ji et al. (2023) proposed that a model can only be considered transferable if the annual mean precipitation and temperature vary within 10% and 1.75^{o}C, respectively.

To improve model robustness, Ji et al. (2023) advised to use a calibration period that closely resembles future climatic conditions, as long as a sufficiently long calibration period (5 to 10 years) is employed. This approach would yield parameter values based on the most relevant data, capturing future behavioural patterns effectively. This recommendation contrasts with the regular approach of calibration using the full-length dataset. This approach is normally regarded as the most robust option, as it provides a diverse and unbiased representation of climate variability without needing a priori knowledge on what the future climatic conditions will be (Bastola et al., 2011).

Despite the fact that many studies advice to use a calibration period that closely resembles future climatic conditions, only a few studies have put model robustness in the context of climate projections. Dakhlaoui et al. (2019) compared the limits of transferability of three conceptual hydrological models to climate projections under two representative concentration pathway scenarios (RCPs 4.5 and 8.5) for catchments in northern Tunisia. They found that for climate projections under RCP 8.5, the calibrated parameter set was invalid. The projected impact of climate change on mean annual runoff when the model was calibrated on the whole period was found to be 5 to 20% lower compared to when the model was calibrated on subperiods with mean annual precipitation and temperature closer to projections. Also Singh et al. (2011) found that, when using parameters calibrated with data resembling future conditions, streamflow projections for 394 watersheds in the United States could diverge significantly from those using parameters calibrated with the full dataset. Both studies focused on the robustness of models for simulating climate change impact on mean annual runoff but not on high or low flows. Wilby (2005) did look at high and low flows, by determining the change in the runoff exceeded 95%, 50% and 5% of the time, for numerous climate scenarios, hydrological models and calibration periods for the river Thames. He found that the uncertainty in projected change in runoff due to the choice of the calibration period was comparable to the uncertainty due to the future emission scenarios.

1.3. Research Gap

The previous discussions stressed the need for evaluating the robustness of hydrological models before applying them under changing conditions. Despite the increased attention to model robustness in literature, only a few studies (Dakhlaoui et al., 2019; R. Singh et al., 2011; Wilby, 2005) have aimed to develop advanced approaches for testing validity of models for climate change impact assessments. However, the conclusions of these studies may not be applicable to all catchments (Ji et al., 2023).

Ji et al. (2023) stated that additional methodological works are still required on the topic of robustness evaluation. One of their recommendations is to not only focus on mean conditions of climate characteristics but look at a wider range of possible changes. Some studies applied a model robustness test focusing on high flows (Donnelly-Makowecki & Moore, 1999; Seibert, 2003; Vormoor et al., 2018), however, no study was found that focused on the simulation of the combination of high and low flows.

1.4. Research Aim

This study aimed **to evaluate the robustness[1] of the HBV and GR6J model[2] for simulating the impact of climate change on high and low streamflow[3] in the Lesse catchment[4], by evaluating the models on historic periods resembling the climatic conditions projected by climate change scenarios[5].**

[1] In this study, a model is defined to be robust when a model does not show notable deterioration in performance under changing climatic conditions. This means that there is no need to recalibrate the model or update the model structure. A lack of model robustness can either be attributed to model structure or to model parameters.

Robustness was evaluated by testing if and how the performance of a model changed under changing climatic conditions. If the model showed a notable deterioration in performance under changing climatic conditions, this indicated a lack of robustness. Secondly, robustness was evaluated by testing the extent to which model parameter values changed when models were calibrated on different climatic conditions. In case parameter values remained approximately constant and the model performed well under changing climatic conditions, a model was robust.

[2] This study evaluated the robustness of the HBV and GR6J model. These are both conceptual rainfall-runoff models that are frequently used in model robustness assessment and climate change impact studies. The focus was on two models, as a multi-model approach can help in identifying model flaws, in understanding models and in proposing improvements (Dakhlaoui et al., 2019; De Boer-Euser et al., 2017; Thirel et al., 2015). Section 2.2 describes the choice for and the structure of these models.

[3] The focus of this robustness assessment was on how well the models were able to capture high and low streamflow. This comprises the magnitude and timing of high and low flows with a return period of approximately one year. This aligned with the recommendation of Ji et al. (2023) to explore a wider range of potential hydroclimatic changes, rather than focusing solely on average conditions. A return period of 1 year was chosen, because longer return periods asked for extreme value distributions, which would bring additional extrapolation errors (Booij, 2005) and would be likely to change due to human influences and climate change (Katz, 2013).

[4] The focus of this study was on the Lesse, Belgium, which is a tributary to the Meuse. In section 2.1, the choice for the Lesse is explained and a description of the catchment is given.

[5] This study used the KNMI'23 climate scenarios as projections of future conditions. Different climate scenarios for time horizon 2100 were used. This is described in section 2.3.3 in more detail.

1.5. Research Questions

To achieve the research aim, 5 research questions were formulated. The first research question (RQ1) provided insight in the projected changes of meteorological conditions that lead to high and low flows.

RQ1 – **Which meteorological conditions lead to high and low flows in the Lesse and how are these conditions expected to change in the future?**

The next step was to select historic periods that resemble future conditions projected by climate change scenarios. For this, the meteorological conditions determined in RQ1 were used.

RQ2 – **Which historic periods most closely resemble future conditions projected by climate change scenarios?**

Then, the robustness of the HBV and GR6J model was evaluated by testing whether the HBV and GR6J models showed deterioration in simulating high and low streamflow when validated under climatic conditions that resemble future conditions projected by the KNMI'23 scenarios (RQ3). A comparison was made between the performance in periods which resemble future conditions and the performance under historic climatic conditions. In RQ4, it was tested to what extent parameters of the HBV and GR6J model changed when these were calibrated on climatic conditions that resemble future conditions projected by the KNMI'23 scenarios, compared to calibration on historic climatic conditions. In this study, historic climatic conditions encompass all climatic conditions that were observed during a certain period in the past.

RQ3 – **What is the performance of the HBV and GR6J model in simulating high and low streamflow under climatic conditions resembling future conditions projected by climate change scenarios, compared to the performance of the models under historic climatic conditions?**

RQ4 – **What are the model parameter values of the HBV and GR6J model when calibrated on climatic conditions resembling future conditions projected by climate change scenarios, compared to when the models are calibrated on historic climatic conditions?**

Finally, the projected impact of climate change on high and low streamflow in the Lesse was determined. A comparison was made between the impact of climate change for different climate scenarios, hydrological model structures and calibration approaches. Two different calibration approaches were used, namely calibration on historic periods and calibration on periods that resemble future climatic conditions. This comparison gave insight in the uncertainty in the impact of climate change due to these choices.

RQ5 – **What is the projected impact of climate change on high and low streamflow in the Lesse, and how does the uncertainty of this impact due to different climate scenarios relate to the uncertainty due to different hydrological model structures and calibration approaches?**

1.6. Research Scope

This study focused on the robustness of models for simulating the impact of climate change. With climate change, global warming resulting in higher temperatures, extremer precipitation, prolonged dry periods and higher variability was understood. The focus of this study was thus on the temporal transferability of models. Transferability of models in space was outside the scope of this study.

Only the impact of climate change on high and low streamflow was examined. The impact of climate change on mean streamflow or other climate characteristics, such as evapotranspiration, was excluded from this study. It was assumed that catchment characteristics that may impact high and low streamflow do not change in the future.

1.7. Thesis Outline

Chapter 2 explains the study area, hydrological models and data that were used in this study. In chapter 3, the methods of this study are presented. This starts with a general overview of the methodology and continues with a detailed description of the methods used to answer each research question. The results of each research question are presented and discussed in chapter 4. An overarching discussion of the results of all research questions can be found in chapter 5. The effects of limitations in data and methods are described, the results are compared with literature and some generalizations are formulated. The thesis ends with the conclusions per research question, a general conclusion and the recommendations in chapter 6.

2. STUDY AREA, MODELS AND DATA DESCRIPTION

2.1. Study area

This study focused on the Lesse, which is a sub-basin of the Meuse. The choice for this study area, its main characteristics and the hydroclimatic conditions are described below.

2.1.1. Choice for the study area

The focus of this study was on the Lesse, a sub-catchment of the Meuse, which has the advantage of not being influenced by reservoirs, locks and weirs (RIWA, 2022). The recently published KNMI'23 scenarios were published for the combined area of the Netherlands and the upstream parts of the catchments of the Rhine and Meuse, as the amount of precipitation in the regions upstream is important for river discharges in the Netherlands (Van Dorland et al., 2023). Compared to the Rhine, the Meuse is less affected by snowmelt, making the discharge more dependent on rainfall. Because of this, the Meuse has more pronounced fluctuations in river flow. This study focused specifically on high and low flows, making the Meuse basin an interesting area to look at.

The sub-catchments of the Meuse located in the Belgian Ardennes have a relatively fast response, due to their impermeable soil and high slope (RIWA, 2021). This causes their discharge regimes to be relatively extreme when compared to those of other tributaries, which made them interesting to study.

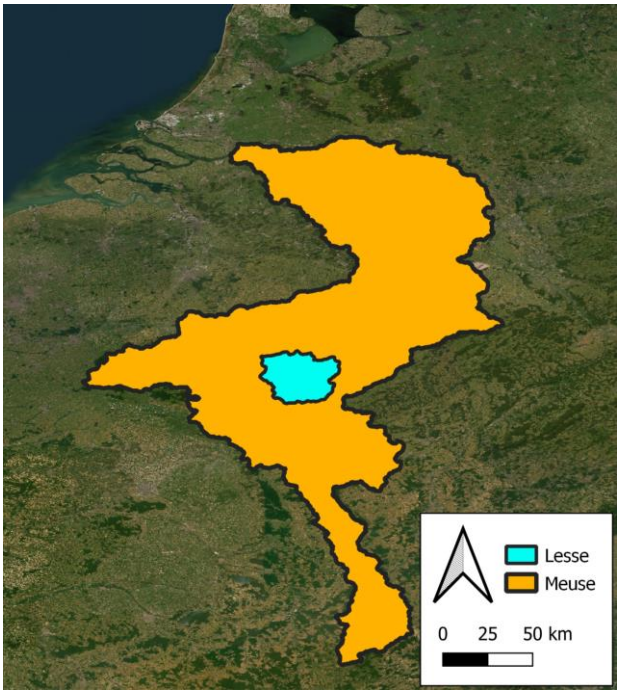

Figure 1 - Location of the Meuse catchment (orange) and the Lesse sub-catchment (blue) (Lehner & Grill, 2013).

2.1.2. Description of the study area

The Lesse river springs in Ochamps at an elevation of 403 m+NAP, and flows to Anseremme, at an elevation of 89 m+NAP, where it joins the Meuse (Berger, 1992). The elevation in the catchment ranges between 89 m+NAP and 589 m+NAP and is shown in Figure 2 (European Space Agency, 2021). The catchment is located in the Belgian Ardennes and covers an area of 1286 km^2 upstream the measurement station Gendron (Service Public de Wallonie, 2023). The length of the river is 83 km and its main tributaries are the Gembes, Lomme, Wimbe and Ywenne. The average gradient of the river is approximately $5 \cdot 10^{-3}$ m/m. This gradient is the highest near the source and decreases towards the mouth (Berger, 1992).

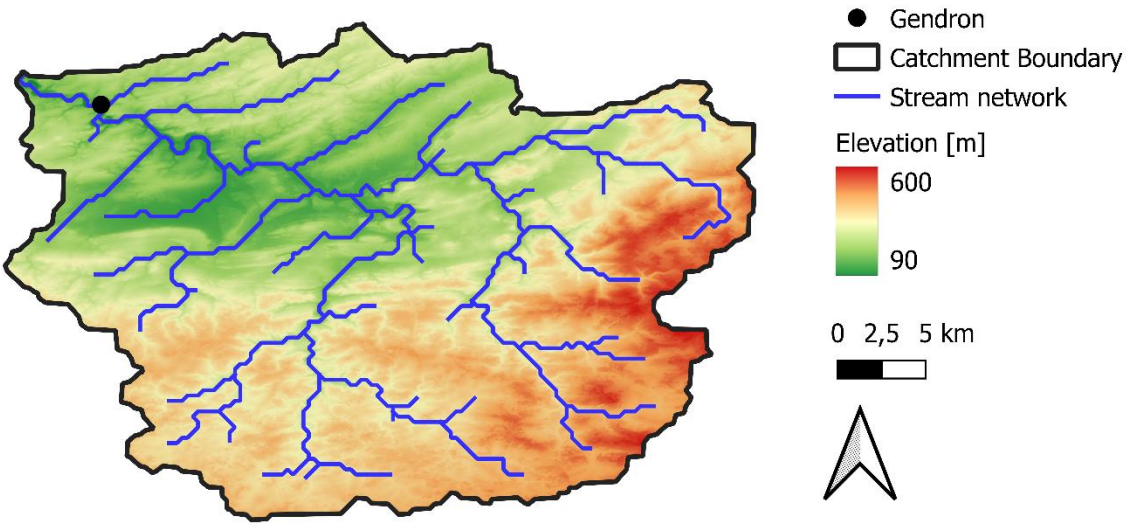

Figure 2 - Elevation [m] in the Lesse catchment (European Space Agency, 2021).

The main land-use class of the Lesse is forest (55%), followed by crops (29%), pastures (11%) and urban area (5%) (European Environment Agency, 2000). Due to the presence of limestone in the basin, karst features are found, resulting in some underground parts of the river, such as near Belvaux and Han (Berger, 1992). Due to the high gradient and impermeable soil, the Lesse is a fast-reacting river. As a result, the contribution of the Lesse to the Meuse is relatively larger during high flows than during low flows (RIWA, 2021).

2.1.3. Hydroclimatic conditions

Figure 3 shows the average precipitation, potential evapotranspiration and streamflow for each month in the Lesse catchment, based on the data described in section 2.3 for period 1968-2021. It is visible that the precipitation is relatively constant throughout the year, whereas the potential evapotranspiration has a clear peak in summer. As a result, the lowest discharges occur during summer and the highest discharges in winter. The mean annual precipitation of the Lesse catchment is 933 mm/year. The precipitation is the highest in the south and the smallest in the west (Berger, 1992). The mean annual runoff at the measuring station near Gendron is 427 mm/year.

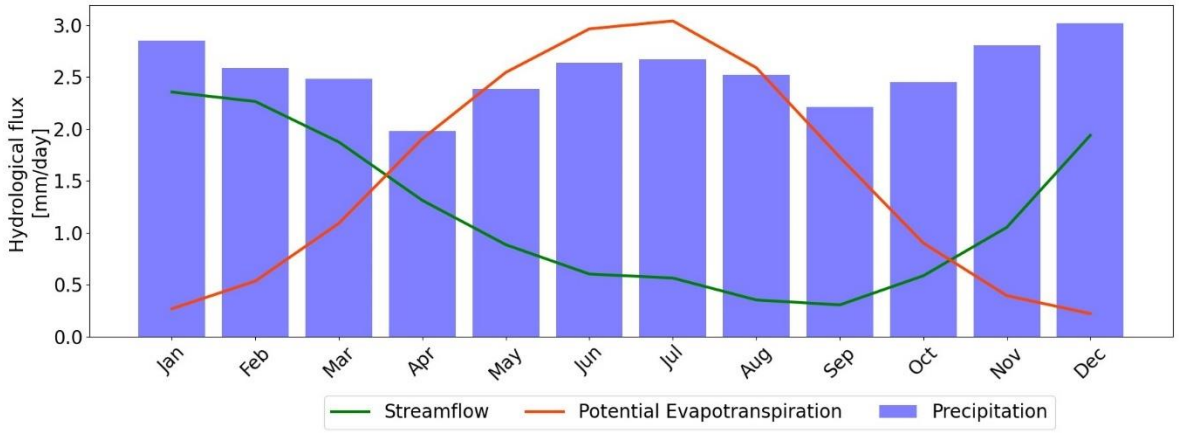

Figure 3 - Climatology of streamflow (green), potential evapotranspiration (red) and precipitation (blue) [mm/day] for the Lesse. Based on data described in section 2.3 (period 1968-2021).

2.2. Hydrological models

This study evaluated the robustness of two rainfall-runoff models, namely the HBV and GR6J model. Both models are parsimonious conceptual models (Dakhlaoui et al., 2019), meaning that the models are intentionally kept simple with at maximum 8 model parameters to be optimized for modelling rainfed catchments. These models were preferred over more physically-based models because they are more interpretable, easier to understand and less prone to overfitting. For simplicity, the models were lumped, which means that spatial variability within the catchment was not included. The models were run at a daily timestep. Both models were calibrated based on historical precipitation, potential evapotranspiration and runoff.

The models differ in the way they conceptualise the hydrological processes, their complexity and number of parameters. This made it interesting to compare both models. An important difference is that HBV had no-flux boundary conditions, meaning that precipitation is the only input and evapotranspiration and discharge are the only output. In contrast, the GR6J model includes groundwater exchange with other catchments, which can either enter or exit the system. This groundwater exchange is physically realistic in catchments with karst features such as the Lesse (Le Moine et al., 2007). However, calibration of this model may tune groundwater exchange parameters in such manner that simulated streamflow match to observed streamflow, while possibly leading to unrealistic values for groundwater exchange.

2.2.1. HBV

The HBV model (Bergström, 1976) has been widely used under various climatic conditions in climate change impact studies and robustness assessment studies (Dakhlaoui et al., 2019; De Wit et al., 2007; Stephens et al., 2019). Next to that, the HBV model forms the basis for flood forecasting systems in the Netherlands of the rivers Rhine and Meuse (Sperna Weiland et al., 2015).

2.2.1.1. Model description

The HBV model is a conceptual model which simulates river discharge (Q) using precipitation (P) and potential evapotranspiration (PET) as input. Several versions of the model exist (Bergström & Lindström, 2015). In this study, the HBV-96 version is used (Lindström et al., 1997; SMHI, 2006). Figure 4 shows the structure of the HBV model. The model consists of three storage boxes. The soil moisture (SM) box controls the runoff formation and the response function is described by the upper response (UZ) and lower response (LZ) box. The model is described in detail below.

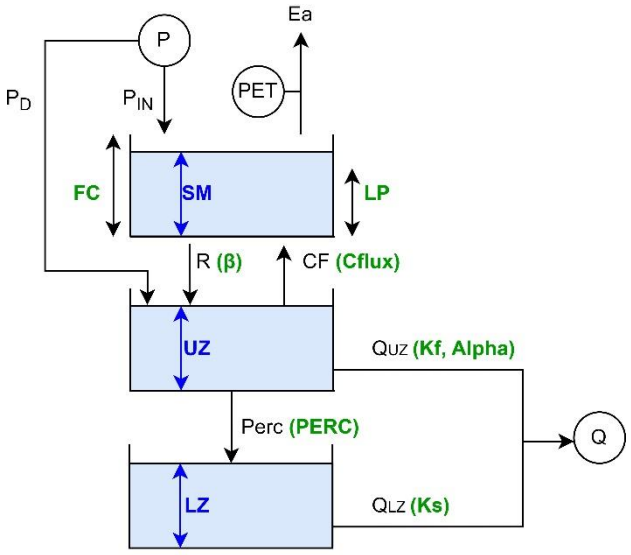

Figure 4 - Structure of the HBV-96 model, adopted from Lindström et al. (1997). Model parameters are given in green.

1. Soil routine

The soil moisture box is filled due to infiltration (P_{IN}) and emptied due to actual evapotranspiration (Ea). The inflow in the soil moisture box (P_{IN}) depends on the precipitation (P), the level in the soil moisture box (SM) and field capacity (FC), the maximum capacity of the soil moisture box (Equation 1).

$$P_{IN} = \begin{cases} P, & SM + P \leq FC \\ FC - SM, & SM + P > FC \end{cases} \quad \text{[Eq. 1]}$$

All remaining precipitation (P_D) is directly transferred to the upper response box (Equation 2).

$$P_D = P - P_{IN} \quad \text{[Eq. 2]}$$

The recharge from the soil moisture box to the upper response box is determined using Equation 3. This recharge depends on the infiltration and on the soil moisture content of the previous day. In Equation 3, β determines the relative contribution to runoff from precipitation at a given soil moisture deficit.

$$R = P \cdot \left(\frac{SM}{FC}\right)^{\beta} \quad \text{[Eq. 3]}$$

The actual evapotranspiration E_a depends on the soil moisture content (SM) of the previous day, the potential evapotranspiration (PET) and the parameter LP, which is given as a fraction of FC. LP is the limit at which the evapotranspiration reaches its potential value (Equation 4).

$$Ea = \begin{cases} PET \cdot \left(\frac{SM}{LP \cdot FC}\right), & SM < LP \cdot FC \\ PET, & SM \geq LP \cdot FC \end{cases} \quad \text{[Eq. 4]}$$

The soil moisture storage is updated by adding the infiltration (P_{IN}) and subtracting the recharge (R) and actual evapotranspiration (Ea) (Equation 5). The upper response box (UZ) is updated by adding the direct precipitation (P_D) and the recharge (R) (Equation 6).

$$SM = SM + P_{IN} - R - E_a \quad \text{[Eq. 5]}$$

$$UZ = UZ + P_D + R \quad \text{[Eq. 6]}$$

2. Response function

The upper response box (UZ) represents the quick runoff routine. This box is filled with recharge (R) and direct precipitation (P_D) and has as output the percolation ($Perc$, Equation 7), the capillary flux (CF, Equation 10) and the non-linear quick runoff (Q_{UZ}, Equation 13). The upper response box has no maximum capacity but cannot be negative. Therefore, no more water can leave the box than is present in the box and the boxes are updated after every flux.

$$Perc = \max(UZ; PERC) \quad \text{[Eq. 7]}$$

$$UZ = UZ - Perc\ ;\ LZ = LZ + Perc \quad \text{[Eq. 8 \& Eq. 9]}$$

$$CF = \max\left(C_{flux} \cdot \left(1 - \frac{SM}{FC}\right); FC - SM; UZ\right) \quad \text{[Eq. 10]}$$

$$SM = SM + CF;\ UZ = UZ - CF \quad \text{[Eq. 11 \& Eq. 12]}$$

$$Q_{UZ} = \max\left(k_f \cdot UZ^{1+\alpha}; UZ\right) \quad \text{[Eq. 13]}$$

$$UZ = UZ - Q_{UZ} \quad \text{[Eq. 14]}$$

The lower response box represents the slow runoff routine. This box is filled with percolation ($Perc$, Equation 8) and emptied with base flow (Q_{LZ}, Equation 15). This base flow depends linearly on the storage in the lower response box and cannot be higher than what is present in the lower response box. The lower response box is updated with Equation 16.

$$Q_{LZ} = \max(k_s \cdot LZ; LZ) \quad \text{[Eq. 15]}$$

$$LZ = LZ - Q_{LZ} \quad \text{[Eq. 16]}$$

3. Total discharge

The total discharge is the sum of the quick runoff (Q_{UZ}) and base flow (Q_{LZ}). No transformation routine is included, as water is discharged within one timestep (1 day) through the catchment.

2.2.1.2. Model parameters

The 8 parameters of the HBV model are summarized in Table 1.

Table 1 - Description of the parameters in the HBV model.

	Unit	Description		Unit	Description
FC	[mm]	Field capacity of the soil moisture storage	K_f	[day^{-1}]	Recession coefficient for upper response box
β	[-]	Parameter in soil moisture routine	K_s	[day^{-1}]	Recession coefficient for lower response box
LP	[-]	Limit for potential evapotranspiration (fraction of FC)	PERC	[mm/day]	Maximum percolation from upper to lower response box
α	[-]	Measure of non-linearity for quick runoff	Cflux	[mm/day]	Maximum value of capillary flux

2.2.2. GR6J

The GR6J model is a modified version of the widely-used GR4J model. The GR4J model, modèle du Génie Rural à 4 paramètres Journalier (Perrin et al., 2003), is a daily lumped four-parameter rainfall-runoff model which is widely used for climate change impact studies and robustness assessment studies (Dakhlaoui et al., 2019; Tian et al., 2013; Zeng et al., 2019). This model performs generally well on estimating high flows but lacks performance on low flows. For example for the Ourthe, Semois and Lesse (sub-catchments of the Meuse), the model generally performs well in simulating flood peaks but overestimates low flows (De Boer-Euser et al., 2017). The modifications of the model proposed by Le Moine (2008) and Pushpalatha et al. (2011) have led to the GR6J model, which improved low-flow simulation, while performance losses in high-flow conditions were avoided (Pushpalatha et al., 2011). Therefore, the GR6J model was used in this study.

2.2.2.1. Model description

Figure 5 shows the model structure of the GR6J model, which was described by Perrin et al. (2003) and adopted by Le Moine (2008) and Pushpalatha et al. (2011). The inputs to the model are areal catchment precipitation (P) and potential evapotranspiration (PET) and the output is discharge (Q). Groundwater exchange with other catchments is included as well, which can be either an input or output of the model. All water quantities are expressed in millimetres and all operations are on a daily timestep. The model is described in detail below.

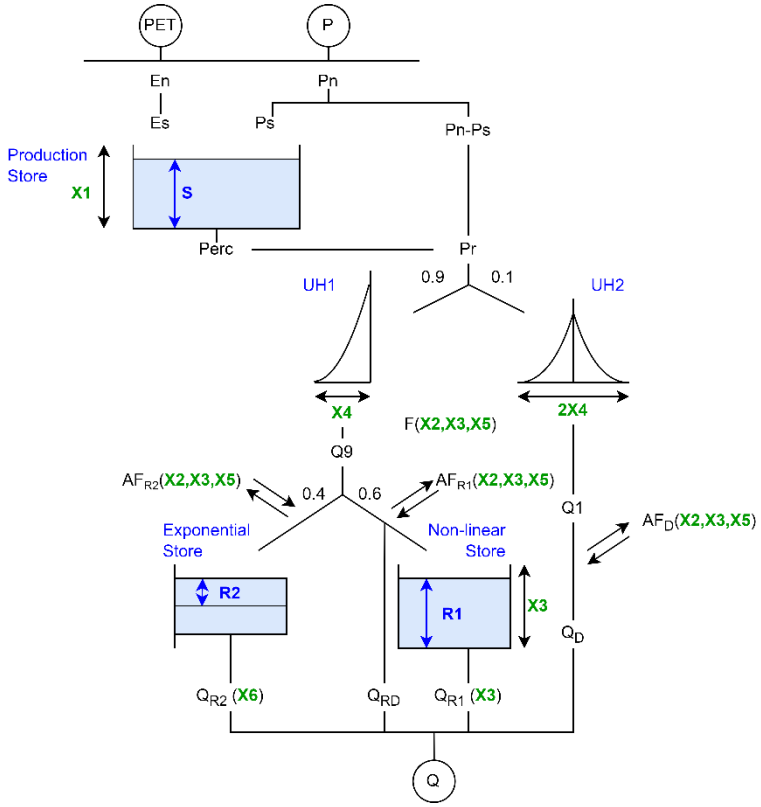

Figure 5 - Structure of the GR6J model, adopted from Perrin et al. (2003), Le Moine (2008) and Pushpalatha et al. (2011). Model parameters are given in green.

1. Determination of net rainfall or evapotranspiration capacity

First, the net precipitation (P_n) or net evapotranspiration capacity (E_n) is determined. In case the precipitation at a timestep is larger than the potential evapotranspiration, $P \geq PET$, there is net precipitation and the net evapotranspiration capacity is zero. In case the potential evapotranspiration is larger than the precipitation, $P < PET$, there is zero net precipitation and a net evapotranspiration capacity is determined. This is displayed in Equations 17 and 18.

$$\text{If } P \geq PET, \quad P_n = P - PET, \quad E_n = 0 \qquad \text{[Eq. 17]}$$

$$\text{If } P < PET, \quad P_n = 0, \quad E_n = E - PET \qquad \text{[Eq. 18]}$$

2. Production store or soil moisture accounting store

In case there is net precipitation ($P_n \neq 0$), a part of the net precipitation fills the production store, also called soil moisture accounting store. This part P_s is determined using Equation 19. In this equation, X_1 is the capacity of the production store and S is the level of the store.

$$\text{If } P_n \neq 0, \quad P_s = \frac{X_1\left(1-\left(\frac{S}{X_1}\right)^2\right)\tanh\left(\frac{P_n}{X_1}\right)}{1+\frac{S}{X_1}\tanh\left(\frac{P_n}{X_1}\right)} \qquad \text{[Eq. 19]}$$

In case of a net evapotranspiration capacity ($E_n \neq 0$), the actual evapotranspiration rate E_s is determined, which is a function of the level in the production store (S) and the maximum capacity of this store (X_1) (Equation 20). The evapotranspiration cannot be higher than what is present in the production store.

$$\text{If } E_n \neq 0: \quad E_s = \max\left(\frac{S\left(2-\frac{S}{X_1}\right)tanh\left(\frac{E_n}{X_1}\right)}{1+\left(1-\frac{S}{X_1}\right)tanh\left(\frac{E_n}{X_1}\right)}; S\right) \qquad \text{[Eq. 20]}$$

After this, the water content in the production store is updated using Equation 21. Due to the hyperbolic tangent functions in Equations 14 and 15, the level in the production store (S) can never exceed the maximum capacity X_1.

$$S = S - E_s + P_s \quad \text{[Eq. 21]}$$

After updating, a percolation leakage $Perc$ is determined, which is a power function of the reservoir content (Equation 22). Due to the power law in this percolation or infiltration function, the percolation does not contribute much to the streamflow and is thus mainly interesting for simulating low flows. After that, the content of the production store (S) is again updated (Equation 23).

$$Perc = S\left\{1 - \left[1 + \left(\frac{4S}{9X_1}\right)^4\right]^{-1/4}\right\} \quad \text{[Eq. 22]}$$

$$S = S - Perc \quad \text{[Eq. 23]}$$

3. Linear routing with unit hydrographs

The total quantity of water that reaches the routing functions (P_r) is determined using Equation 24.

$$P_r = Perc + (P_n - P_s) \quad \text{[Eq. 24]}$$

This quantity is divided into two flow branches. A fixed 90% of P_r is routed by unit hydrograph 1 ($UH1$), the remaining 10% is routed by unit hydrograph 2 ($UH2$). The water that is routed by unit hydrograph 1 is also routed by non-linear routing stores. By having two unit hydrographs, the time lag between the rainfall event and the resulting streamflow peak is simulated.

Parameter X_4 is the time-base of $UH1$. This means that, in a discrete form, unit hydrograph 1 has m ordinates, with m being the smallest integer exceeding X_4. The time-base of unit hydrograph 2 is $2X_4$ and has n ordinates, n being the smallest integer exceeding $2X_4$. The ordinates of the unit hydrograph are derived using the S-curves, which are given in Equation 26.

Unit Hydrograph 1

$$SH1(t) = \begin{cases} 0, & t \le 0 \\ \left(\frac{t}{X_4}\right)^{5/2}, & 0 < t < X_4 \\ 1, & t \ge X_4 \end{cases}$$

Unit Hydrograph 2

$$SH2(t) \begin{cases} 0, & t \le 0 \\ \frac{1}{2}\left(\frac{t}{X_4}\right)^{5/2}, & 0 < t \le X_4 \\ 1 - \frac{1}{2}\left(\frac{t}{X_4}\right)^{\frac{5}{2}}, & X_4 < t < 2X_4 \\ 1, & t \ge 2X_4 \end{cases} \quad \text{[Eq. 25]}$$

$$UH1(j) = SH1(j) - SH1(j-1) \quad \text{[Eq. 26]}$$
j is an integer

$$UH2(j) = SH2(j) - SH2(j-1)$$
j is an integer

The output of unit hydrograph 1, $Q9$, is calculated using Equation 27. The output of unit hydrograph 2, $Q1$, is calculated in a similar way, using Equation 28.

$$Q9(t) = 0.9 \cdot \sum_{k=1}^{m} UH1(k) \cdot P_r(t-k+1) \quad \text{[Eq. 27]}$$

$$Q1(t) = 0.1 \cdot \sum_{k=1}^{n} UH2(k) \cdot P_r(t-k+1) \quad \text{[Eq. 28]}$$

Maximum catchment water exchange

The water that reached the routing function is thus split into two branches. The first branch, the routing stores branch, is composed of a non-linear routing store and an exponential routing store. The branch is fed by $Q9$ from unit hydrograph 1. 60% of $Q9$ flows to the non-linear routing store, the remaining 40% to the exponential routing store. The exponential store is an additional (parallel) routing store compared to the GR4J model. Exponential stores are efficient for simulating long recession spells (Michel et al., 2003). The second branch, the direct branch, is fed by $Q1$ from unit hydrograph 2.

The next step is to determine the groundwater exchange for the non-linear routing store, the exponential routing store and the direct branch. The maximum possible exchange F is determined based on the content of the non-linear routing store R_1, the maximum capacity of the routing store X_3 and exchange parameters X_2 and X_5 (Equation 29). This groundwater exchange function was proposed by Le Moine (2008) to better account for interactions between surface and groundwater. The value can be either positive, zero or negative. This function replaces the old groundwater exchange function of the GR4J model.

$$F = X_2 \left(\frac{R_1}{X_3} - X_5\right) \quad \text{[Eq. 29]}$$

The actual catchment exchange of the routing store AF_{R1} equals the maximum possible exchange F if enough water is present in the routing store (Equation 30).

$$AF_{R1} = \begin{cases} F, & R_1 + 0.6Q9 + F \geq 0 \\ -R_1 - 0.6Q9, & R_1 + 0.6Q9 + F < 0 \end{cases} \quad \text{[Eq. 30]}$$

The actual catchment exchange of the exponential store AF_{R2} equals F, as the content in the exponential store can be negative as well (Equation 31).

$$AF_{R2} = F \quad \text{[Eq. 31]}$$

The actual catchment exchange of the direct branch AF_D equals the maximum possible exchange F if enough water is in the direct branch (Equation 32).

$$AF_D = \begin{cases} F, & Q1 + F \geq 0 \\ -Q1, & Q1 + F < 0 \end{cases} \quad \text{[Eq. 32]}$$

4. Routing stores

Non-linear routing store

The non-linear routing store content R_1 is updated by adding 60% of the water routed by unit hydrograph 1 ($Q9$) and the groundwater exchange of the non-linear routing store AF_{R1} (Equation 33). The routing store cannot exceed its capacity X_3. All water that does not fit in the non-linear routing store is discharge directly by Q_{R1d} (Equation 34).

$$R_1 = \max\left(R_1 + 0.6Q9 + AF_{R1}; X_3\right) \quad \text{[Eq. 33]}$$

$$Q_{R1d} = \max\left(R_1 + 0.6Q9 + AF_{R1} - X_3; 0\right) \quad \text{[Eq. 34]}$$

The output of the non-linear routing store Q_{R1} is determined using Equation 35. This flow depends on the level in the routing store R_1 and the maximum capacity X_3. The non-linear routing store is afterwards updated for the next timestep (Equation 36).

$$Q_{R1} = R_1 \left\{1 - \left[1 + \left(\frac{R_1}{X_3}\right)^4\right]^{-1/4}\right\} \quad \text{[Eq. 35]}$$

$$R_1 = R_1 - Q_{R1} \quad \text{[Eq. 36]}$$

Exponential routing store

The exponential routing store content R_2 is updated by adding 40% of the water routed by unit hydrograph 1 ($Q9$) and the groundwater exchange of the exponential routing store AF_{R2} (Equation 37). There is no maximum capacity of the exponential routing store.

$$R_2 = R_2 + 0.4Q9 + AF_{R2} \quad \text{[Eq. 37]}$$

The output of the non-linear routing store Q_{R2} is computed using Equation 38. The output only depends on the level of the routing store R_2 and an exponential store depletion coefficient X_6 (Michel et al., 2003). The routing store is updated afterwards (Equation 39).

$$Q_{R2} = X_6 \cdot \ln\left(1 + exp\left(\frac{R_2}{X_6}\right)\right) \quad \text{[Eq. 38]}$$

$$R_2 = R_2 - Q_{R2} \quad \text{[Eq. 39]}$$

5. Total streamflow

The water in the direct branch is fed by the output of unit hydrograph 1 $Q1$. Besides the catchment exchange AF_3, the water is discharged directly by Q_D (Equation 40).

$$Q_D = Q1 - AF_3 \quad \text{[Eq. 40]}$$

The total streamflow (Q) is finally obtained by adding the 4 flow components (Equation 41).

$$Q = Q_{R1} + Q_{R1d} + Q_{R2} + Q_D \quad \text{[Eq. 41]}$$

2.2.2.2. Model parameters

The 6 parameters of the GR6J model are summarized in Table 2.

Table 2 - Description of the parameters in the GR6J model.

	Unit	Description	P.	Unit	Description
X_1	[mm]	Capacity of the production store	X_4	[day]	Time base of unit hydrograph
X_2	[mm/day]	Catchment exchange coefficient	X_5	[-]	Threshold for catchment exchange
X_3	[mm]	Capacity of the non-linear routing store R1	X_6	[mm]	Exponential routing store depletion coefficient

2.3. Data description

2.3.1. Historic discharge timeseries

Daily discharge data for the period 1968-2021 was used. This is an observed discharge timeseries at the outlet point of the Lesse, near Gendron (50°12'40.8"N, 4°57'46.1"E). For the period 1968-1998, the discharge series was obtained from SETHY/WACONDAH. For 1998-2021, the discharge series was obtained from the annual reports from Service Public de Wallonie (2023). Because of unrealistic values, the data was pre-processed as described in Appendix A.

2.3.2. Historic climatic timeseries

Basin-averaged daily precipitation and potential evapotranspiration timeseries for the period 1968-2021 were used. These timeseries were provided by KNMI and were based on a combination of the HYRAS and E-OBS dataset. HYRAS data was used in the part of the study area for which this data was accessible. For the remaining part of the study area, E-OBS data was used. E-OBS is a daily high-resolution gridded dataset for various climate variables such as precipitation and minimum, maximum and mean surface temperature, which is based on measurements that are part of the European Climate Assessment & Dataset project (Cornes et al., 2018). Similarly, HYRAS is a high-resolution gridded daily dataset for various climate variables as well. This dataset is based upon measurements gathered for Germany and 9 neighbouring countries (Razafimaharo et al., 2020). The potential evapotranspiration timeseries were estimated using the Makkink formula (Hooghart & Lablans, 1988) by KNMI.

2.3.3. Future climatic timeseries

The recently published KNMI'23 climate scenarios were used as future climatic timeseries. In October 2023, the KNMI published its KNMI'23 scenarios (Van Dorland et al., 2023), which are a translation of the insights from the most recent IPCC report (IPCC, 2021) to the Netherlands and the Rhine and Meuse basin. These scenarios are plausible future climate states, containing sets of climatological variables that are to a certain degree physically consistent (Van Dorland et al., 2023). The KNMI'23 climate scenarios are based on climate simulations run with global circulation model EC-EARTH and downscaled with regional climate model RACMO (Van Dorland et al., 2023).

Four different scenarios were used, which all describe a possible climate state for time horizon 2100. The scenarios are constructed around two axes, which was done to decouple policy uncertainty from scientific uncertainty (Van Dorland et al., 2023). The first axis represents the uncertainty in future

greenhouse gas emissions, being either high (H) or low (L). The second axis represents uncertainty in the regional climate response, which could either be dry (d) or wet ('nat', n). This leads to four climate scenarios, which are the cornerstones between which the climate is likely to change (Figure 6). No probabilities or estimates of likelihood can be assigned to the scenarios.

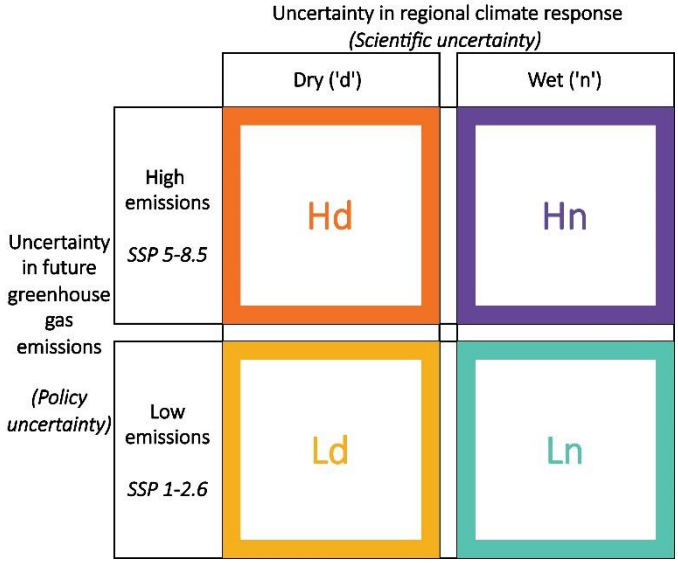

Figure 6 - Overview of the KNMI'23 climate scenarios considered in this study. Figure adopted from KNMI (2023).

The first axis thus represents the uncertainty in future greenhouse gas emissions, which is policy uncertainty. The amount of emission is uncertain because this depends on global climate policy, demographic, technological and socio-economic factors. Upper and lower boundaries for future emissions are used, to emphasize the consequences of international choices of mitigation policies (Van Dorland et al., 2023). The upper boundary is the SSP5-8.5 pathway, which is considered as the base line of not taking any mitigation measures at all, resulting in 50% likelihood of 4.9°C global warming in 2100 compared to the pre-industrial period. The lower boundary is the SSP1-2.6 pathway, which is close to the mitigation pathway that is aimed for in the Paris Agreement. This pathway is 50% likely to result in 1.7°C global warming in 2100 compared to pre-industrial conditions.

The second axis represents the uncertainty in regional climate response. Even though much progress has been made in understanding and quantifying climate feedbacks, the response of the climate to a doubling of the CO_2 concentration (climate sensitivity) is still very uncertain (IPCC, 2021). The output of climate models shows a large spread, because climate models differ in the way relevant processes are parametrized (Van Dorland et al., 2023). The KNMI'23 scenarios are based on two storylines of how precipitation can change. The storylines are designed such that they span the CMIP6 models spread in precipitation changes in the Netherlands and Rhine and Meuse basins. The CMIP6 models (IPCC, 2021) consistently project an increase in winter precipitation and decrease in summer precipitation but differ in their sign for annual precipitation. Therefore, the 33 CMIP6 models were partitioned based on their combined projected change of winter, summer and annual precipitation. The 'dry-trending group' contains the 11 models that projected the largest trends towards a drier future climate, which means a decrease in annual total precipitation, a strong decrease in summer precipitation and a small increase in winter precipitation. The 'wet-trending' group contains the 11 models with the largest trend towards a wetter future climate.

This study used the bias-corrected model output of the climate scenarios for time-horizon 2100. This is the output from climate model RACMO but corrected using Quantile Delta Mapping (Cannon et al., 2015) because the RACMO climatology differs from the observed climatology. The reasons for bias-

correction are further explained in section 3.1.2.1. For all scenarios, a dataset was used containing 8 timeseries of basin-averaged precipitation and potential evapotranspiration for the surrounding 30-year period (2086-2115). All 8 timeseries (ensemble members) were used to capture the climate variability. In the KNMI'23 scenarios, the potential evapotranspiration was calculated using the Makkink formula (Van Dorland et al., 2023).

For each climate scenario and time-horizon, bias-corrected RACMO model output was available for reference period 1991-2020 as well. The bias-corrected model output for reference period 1991-2020 of scenario 2050Md (time horizon 2036-2064, moderate emissions, dry-trending group) showed the best statistical match with observations and was therefore recommended by the KNMI (2023a) as the reference climate. This reference climate data consists of 8 ensemble members of 30 years as well. The use of the reference climate is further discussed in the chapter 3.

3. METHODS

An overview of the methodology used to evaluate the robustness of the HBV and GR6J model for simulating the impact of climate change on high and low streamflow in the Lesse is given in Figure 7.

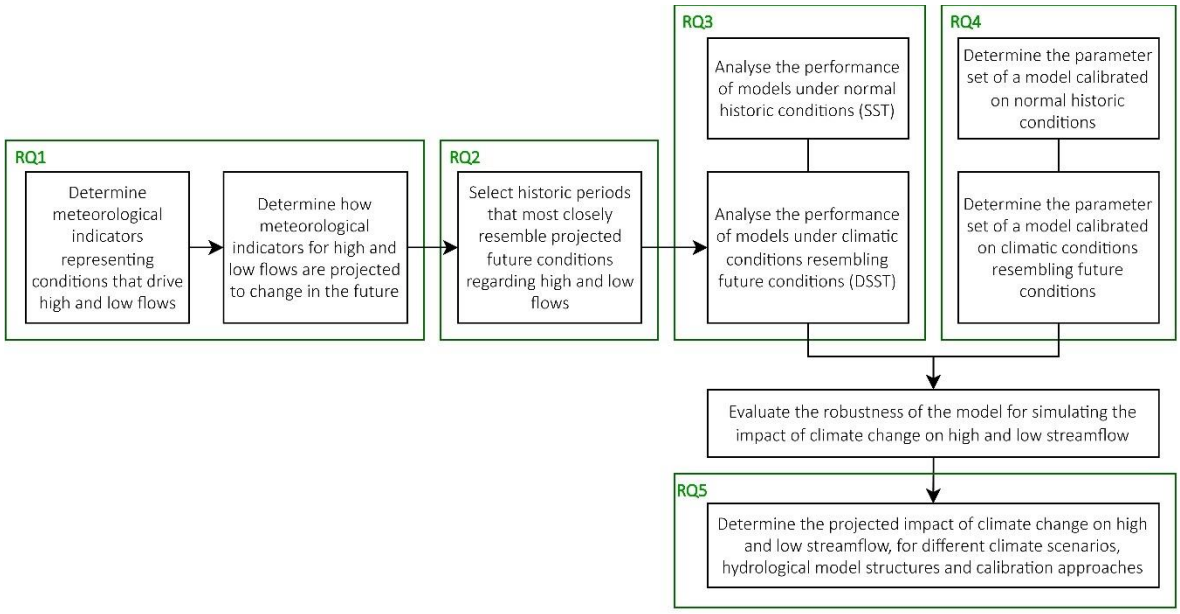

Figure 7 - Overview of methodology. Green frames indicate to which research question (RQ) each step belongs.

To evaluate the robustness, both models were evaluated under historic climatic conditions and under climatic conditions that resemble future conditions (RQ3). Next to that, model parameters were calibrated on historic conditions and conditions that resemble future conditions (RQ4). As a preparation for this, historic periods were selected that resemble future conditions (RQ2). For this, it was necessary to know which meteorological indicators correlate with high and low flows, and how these indicators were projected to change in the future (RQ1). Lastly, the impact of climate change on high and low streamflow in the Lesse was determined for various climate scenarios, hydrological model structures and calibration approaches (RQ5). The method per research question is explained in more detail below.

3.1. RQ1: Changes in important meteorological indicators

Evaluating the robustness of models in simulating the impact of climate change on high and low streamflow demanded for an understanding of the meteorological conditions that lead to these flow conditions and insight in how these meteorological conditions are expected to change in the future.

A single meteorological indicator for high flows was obtained that describes the meteorological conditions driving high flows, and a single meteorological indicator for low flows was obtained that describes the meteorological conditions that drive low flows. Section 3.1.1 describes the method used to obtain these meteorological indicators for high and low flows. Section 3.1.2 describes the method used to determine how these meteorological indicators are expected to change in the future.

3.1.1. Meteorological indicators for high and low flows

3.1.1.1. Step 1 – Description of high and low flows

High flows

High flows were described by the annual maximum daily discharge (Q_{max}), following studies on the implications of the KNMI climate scenarios for discharges of the Rhine and Meuse (Buitink et al., 2023; Sperna Weiland et al., 2015) and a model robustness study focused on high flows (Vormoor et al., 2018).

Annual maxima are of major importance for water managers, as series of annual maxima are often used for deriving design water levels for flood protection (Diermanse et al., 2010).

A series of annual maximum daily discharges was obtained by taking the maximum of all daily discharge values in one hydrological year. A hydrological year from the 1st of September to the 31st of August was used. The reason for this is explained in Appendix B. With 54 years of available data and a hydrological year from September to August, a series of 53 annual maximum daily discharges was obtained.

Low flows

For low flows, it is uncommon to use the annual minimum daily discharge, as this discharge is very sensitive to measurements errors (Smakhtin, 2001). An overview of low-flow indices was provided by Smakhtin (2001). These indices all describe different aspects of the low-flow regime, however, most of them are strongly intercorrelated. It was chosen to present the magnitude of a low flow by the annual minimum 7-day mean discharge ($Q7_{min}$), following the climate change impact studies by Buitink et al. (2023) and Sperna Weiland et al. (2015).

A series of annual minimum 7-day mean discharges was obtained by taking the moving average discharge over a period of 7 days and taking the minimum value in each hydrological year. The 7-day mean discharge at day t represents the average discharge over the surrounding 7-day period (day $t-3$ to day $t+3$). For low flows, the hydrological year was similar to the calendar year (January-December), as described in Appendix B. Thus, a series of 54 annual minimum 7-day mean discharges was obtained.

3.1.1.2. Step 2 – Correlation analysis

High flows

Correlation analysis was performed to find a meteorological indicator for high flows that best resembles the meteorological conditions that drive high flows. The correlation between several potential meteorological indicators for high flows and the annual maximum discharge was determined and the indicator with the highest correlation was selected. It was important that this meteorological indicator for high flows was not only justifiable from a statistical point of view (high correlation), but also hydrologically meaningful.

The most important meteorological indicator for annual maximum discharges was expected to be the precipitation preceding the discharge peak. High flows typically occur during periods of intense precipitation, leading to a sudden increase in water flow within rivers. This is especially the case when the soil is already saturated due to previous precipitation (saturation overland flow) or when precipitation exceeds the infiltration capacity of the soil (infiltration excess overland flow) (Kirkby, 1988).

The correlation between the annual maximum discharge and the preceding precipitation sum over various temporal windows was analysed. The temporal window was varied between 1 and 10 days, as this was physically the most meaningful range based on the scale of the catchment. These precipitation sums either included or excluded the day of the discharge peak. For example, a 3-day precipitation sum including the day of the annual maximum discharge (sum of day 0, 1 and 2, i.e. $P3i$) was computed, as well as a 3-day precipitation sum excluding the day of the peak discharge (sum of day 1, 2 and 3, i.e. $P3e$). This is illustrated in Figure 8.

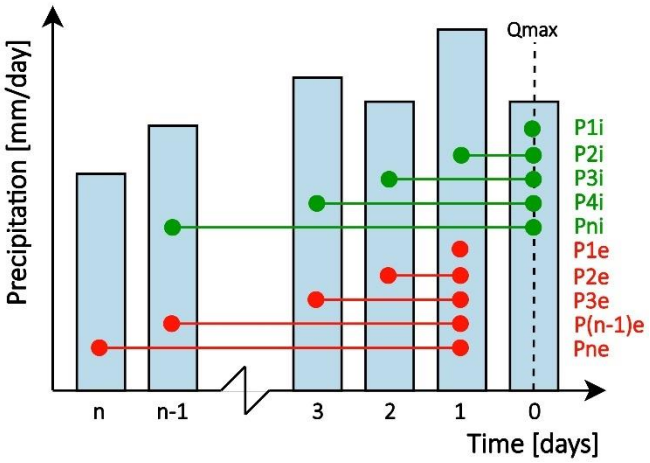

Figure 8 - Illustration of the different temporal windows of precipitation sums preceding the discharge peak. Green lines indicate the precipitation sums including (i) the day of the annual maximum discharge, red lines indicate the precipitation sums excluding (e) the day of the annual maximum discharge. The temporal windows varied between t=1 and 10 (n) days.

The Pearson correlation coefficient was computed to quantify the correlation between the series of annual maximum daily discharges and the series of each of the preceding precipitation sums. The Pearson correlation coefficient is the most common way for measuring a linear association between two variables (Helsel et al., 2020). A higher precipitation sum was expected to result in a higher annual maximum daily discharge, so a positive correlation was expected. The temporal window with the highest positive correlation was used for the next steps.

Low flows

Just as for high flows, correlation analysis was performed between the annual minimum 7-day mean discharge and several potential meteorological indicators for low flows. Again, the meteorological indicator for low flows had to be both hydrologically meaningful and statistically justifiable.

Demirel et al. (2013b) found that, in rainfed sub-basins of the Rhine, precipitation and potential evapotranspiration with large temporal scales (order of months) are important when predicting low flows. This is in line with the study of de Wit et al. (2007) on the Meuse basin, which hints at the importance of the preceding winter precipitation on low flows in summer. The combination of precipitation and potential evapotranspiration was thus important, as well as the consideration of larger temporal windows.

For these reasons, the correlation between the annual minimum 7-day mean discharge and the preceding potential precipitation deficit over various temporal windows was determined. The potential precipitation deficit (PPD) is the difference between the amount of potential evapotranspiration (PET) and the amount of precipitation (P) over a certain period of time (n days) (Equation 42). The potential precipitation deficit is a widely used indicator for characterizing meteorological droughts by the KNMI (Sluijter et al., 2018). A negative potential precipitation deficit indicates a potential precipitation surplus.

$$PPD = \sum_{i=0}^{n}(PET_i) - \sum_{i=0}^{n}(P_i) \qquad \text{[Eq. 42]}$$

The period started $n-1$ days before the day with the annual minimum 7-day mean discharge and ended at the day of the annual minimum 7-day mean discharge. The day of the annual minimum 7-day mean discharge is the middle day of the 7-day period featuring the minimum 7-day mean discharge, as stated in section 3.1.1.1. The temporal window of the potential precipitation deficit was varied between n=30 and 360 days, with steps of 30 days. This represented periods of 1 to 12 months (Figure 9).

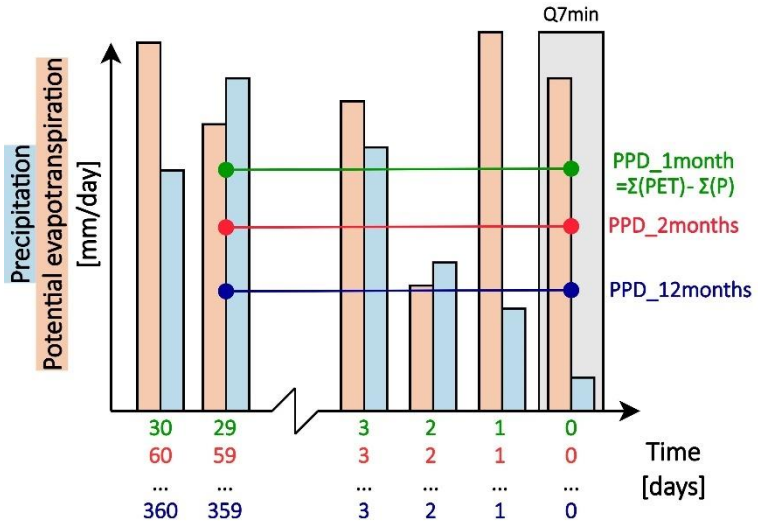

Figure 9 - Illustration of the different temporal windows considered for the potential precipitation deficit (PPD).

Just as for high flows, the Pearson correlation coefficient was used as a measure for the correlation. The correlation between the series of annual minimum 7-day mean discharges and the series of preceding potential precipitation deficits over various temporal windows was determined. A higher potential precipitation deficit was expected to result in a lower annual minimum 7-day mean discharge, therefore a negative correlation was expected. The temporal window with the most negative correlation was thus used for the next steps. Due to a lack of data, the preceding potential precipitation deficit could not be determined for all considered temporal windows in the first year. Therefore, only the observations of the period 1969-2021 (n=53) were used.

3.1.1.3. Step 3 – Exceedance frequency

High flows

In order to determine how the meteorological indicator for high flows obtained in the previous step was expected to change in the future, precipitation and discharge data would be needed for the future. However, as described in section 2.3.3, only precipitation and potential evapotranspiration timeseries were available for the future. An indicator thus had to be obtained that could be determined based on precipitation and potential evapotranspiration only. For this reason, a precipitation sum with a temporal window and an appropriate annual exceedance frequency was determined that corresponded the best with the precipitation sum preceding the annual maximum daily discharge.

The procedure for this was as follows: For each hydrological year, the precipitation sum with the temporal window of step 2 (t) preceding the annual maximum daily discharge was determined. This was called the precipitation sum of interest (Pt_oi). Then, the precipitation sum with that temporal window (Pt) was determined for each day of the year. The exceedance frequency (FE) is the number of days in a hydrological year for which the Pt was larger than the Pt_oi. The exceedance frequency thus was a number between 0 and 365. This is illustrated in Figure 10. Here, the orange boxes indicate the days with a smaller precipitation sum than the precipitation sum preceding the annual maximum discharge, and the blue boxes indicate days with a higher precipitation sum. The number of blue boxes was thus the exceedance frequency.

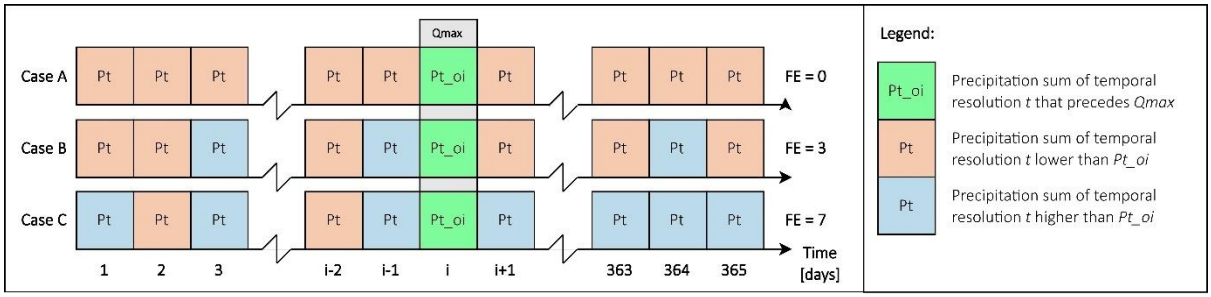

Figure 10 - Illustration of the procedure for determining the exceedance frequency FE of the precipitation sum of temporal window t that precedes Q_{max} (Pt_oi, green). The orange (blue) boxes indicate the days with a precipitation sum of temporal window t lower (higher) than the preceding precipitation sum Pt_oi. In this example, only 10 days are shown, and it is assumed that for the other days, the precipitation sum is not exceeded.

The exceedance frequency for each hydrological year (53 years in total) was determined and the median exceedance frequency was taken. The preceding precipitation sum with the temporal window of step 2 and the median annual exceedance frequency of this step was used as meteorological indicator for high flows. The median value was chosen over the mean because of the skewed distribution that was observed in exceedance frequencies. The median value is a more representative value, which is less affected by outliers or extreme occurrences.

Low flows

For low flows, the potential precipitation deficit with a certain temporal window that precedes the annual minimum 7-day mean discharge was determined as an indicator. However, just as for high flows, it was required that an indicator was obtained that could be determined based on precipitation and potential evapotranspiration only. Following the same approach as for high flows, a potential precipitation deficit over a certain temporal window was thus determined for every day in the year. The number of days for which the potential precipitation deficit was larger than the potential precipitation deficit preceding the annual minimum 7-day mean discharge was called the exceedance frequency. The preceding potential precipitation deficit with the temporal window of step 2 and the median exceedance frequency of this step was used as meteorological indicator for low flows.

3.1.2. Changes in the future

This section describes the method used to determine how the meteorological indicators for high and low flows are expected to change in the future. It was not possible to compare the future climatic timeseries and the observations directly, as this does not only show the impact of climate change, but also the possible climate model bias. The magnitude of the climate model bias was determined by comparing the observations with the reference climate. Based on this, the future climatic timeseries were bias-corrected and compared with the observations.

3.1.2.1. The reason for bias-correction

The reference climate should represent the climate as observed in the period 1991-2020. However, discrepancies between the reference climate and the observations exist for several reasons (Van Dorland et al., 2023). First of all, some meteorological processes are poorly represented in the RACMO climate model which was used for constructing the reference climate. Next to that, the land-use map and topography in RACMO and the climatological forcing used as input for the RACMO model may differ from reality, leading to biases in the climatology of RACMO. Thirdly, the resolution of the RACMO model is 12x12 km, while observations are local. Last but not least, the climatology in the reference climate will always differ from observations due to natural variability. To isolate the effects of climate change, Van Dorland et al. (2023) applied a bias-correction for each variable in the RACMO climate model using the Quantile Delta Mapping method. The bias-corrected model output was used in this study.

However, comparing the meteorological indicators computed from the observed dataset with the meteorological indicators computed from the reference climate data still showed a substantial difference. The reason for this could be that the bias-correction was not performed specifically on the meteorological indicators for high and low flows. As a result, the difference in meteorological indicators between observations and future scenarios may be due to the combination of climate change and a model bias. This is illustrated in Figure 11A. Therefore, a model bias-correction was applied to the meteorological indicators in the reference climate and future scenarios. These values were adapted such that the reference climate matched the observations. As a result, the difference between observed values and values in the future was due to climate change only. This is illustrated in Figure 11B.

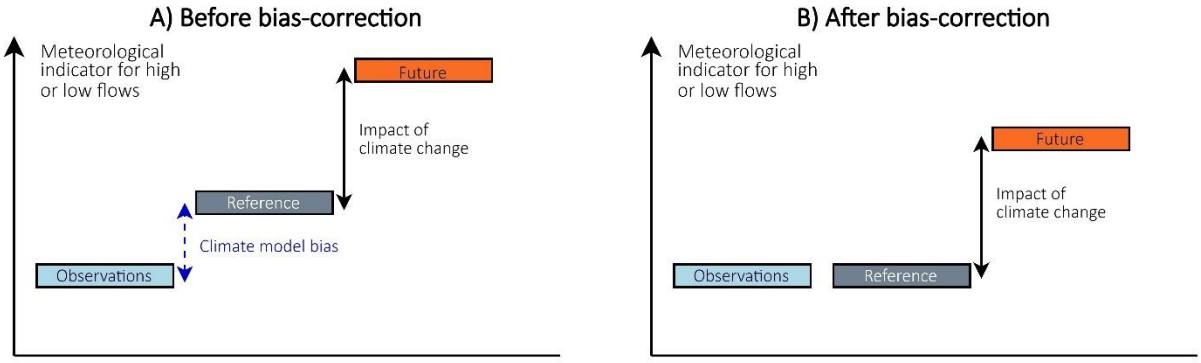

Figure 11 - Fictive illustration of the bias-correction. A) The fictive values of meteorological indicators in the observations (blue), reference climate (grey) and future scenario (orange) before bias-correction. The difference between observations and future is due to a climate model bias and the impact of climate change. B) The fictive values after bias-correction. Here, the differences between observations and future are due to the impact of climate change only.

3.1.2.2. Applying the bias-correction

To determine the magnitude of the model bias, the relative difference between the meteorological indicators for high flows from the observations and the reference climate was computed. This relative difference was used to correct the values of the meteorological indicator for high flows in the future scenarios. This means that, when the values in the observations were for example 5% lower than the values in the reference climate, the values in the future scenario were corrected by making them 5% lower. A relative bias was preferred over an absolute bias (i.e. applying the absolute difference between observations and reference climate on the future scenario), as a relative bias was used in most bias-correction methods for variables with an absolute zero (ratio variables), such as precipitation and potential evapotranspiration (Cannon et al., 2015).

High flows

The reference climate represents the climate of 1991-2020 and thus only the observations from the period 1991-2020 were used. This 30-year period contains 29 hydrological years, therefore 29 observed values of the meteorological indicator for high flows were computed. For the reference climate, 8 ensemble members with each 30-years of data were available. The meteorological indicator for high flows was determined for each available hydrological year in the reference climate, leading to (29·8=) 232 values of the meteorological indicator. To make it possible to compare 29 observations with 232 values in the reference climate, the 232 values of the reference climate were sorted and clustered into 29 groups of 8 indicators. The average value of each group was compared with the sorted 29 observations. This led to a percentual difference for 29 pairs of meteorological indicators. The average percentual difference was used as the relative bias. This procedure is illustrated in Figure 12.

Just as for the reference climate, 8-ensemble members with each 30 years of data were available for each of the future scenarios. Therefore, 232 values of the meteorological indicator for high flows were computed for each future scenario as well. These values were adjusted using the relative model bias.

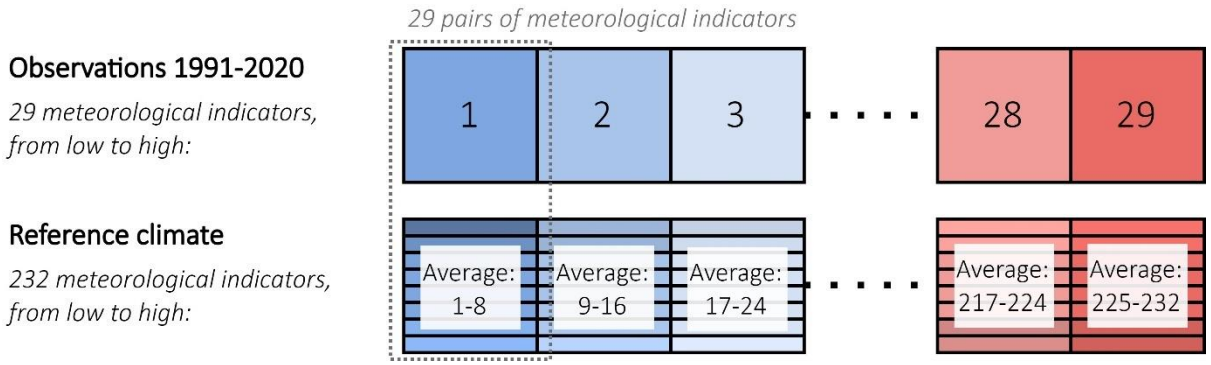

Figure 12 - Illustration of the procedure used to compute the average relative bias between observations and reference. The 29 values of observations were ranked from low to high. The 232 values of the reference climate were sorted from low to high as well, and clustered in 29 groups of 8 values. The average of each group was compared with the 29 values of the observations by taking a relative difference. The average of this relative difference was used for further steps.

Low flows

For the meteorological indicator for low flows, a similar procedure as for high flows was used. However, as the meteorological indicator for low flows was a potential precipitation deficit, consisting of a precipitation and potential evapotranspiration component, the relative bias was not determined directly. Instead, the relative bias was determined separately for the precipitation sum corresponding to the potential precipitation deficit, as well as for the potential evapotranspiration sum corresponding to the potential precipitation deficit. Then, the precipitation sum and potential evapotranspiration sum corresponding to the potential precipitation deficits in the future were corrected using these biases. With these new values, the corrected values of the potential precipitation deficits were determined.

It was consciously chosen not to use more complicated statistical bias-correction methods such as the Quantile Delta Mapping method used by Van Dorland et al. (2023), as only 29 values of the observations were available. This limited number of observations would lead to large uncertainties in the cumulative distribution function needed for the Quantile Delta Mapping method.

3.1.2.3. Changes in the future

After the bias-correction, boxplots were made from the series of meteorological indicators for high and low flows in the future, as well as for the indicators computed from observed data. Comparing these boxplots gave insight in the expected changes of these indicators in the future, focusing on normal and extreme years. Normal years were defined as the median of all years in the future and extreme years were years for which the meteorological indicator is exceeded 5% of the time.

3.2. RQ2: Selection of historic periods resembling future climatic conditions

This section describes the method used to select historic periods that resemble future conditions. A set of historic periods was selected for the 4 future scenarios and for normal and extreme years. This step thus led to (4·2=) 8 sets of selected historic periods for high flows and 8 sets for low flows. The historic periods were selected based on the selected meteorological indicators for high and low flows of RQ1.

High flows

A historic hydrological year in the period 1968-2021 was selected, when the meteorological indicator for high flows of that hydrological year was close to the meteorological indicator for high flows expected in the future. For selection of historic years resembling normal future years, the median of the

meteorological indicators of a future scenario was used. For extreme years, the meteorological indicator of a future scenario that is exceeded 5% of the time was used.

As the expected value for the meteorological indicator for high flows in a future scenario never perfectly matched the meteorological indicator observed in a historic hydrological year, a historic hydrological year was selected when it felt within the 10% bandwidth around the expected value for a future scenario. This means that the year was selected when the meteorological indicator for high flows in that year was between 5% lower or 5% higher than the expected future value.

Low flows

For selecting historic hydrological years that resemble future low flow conditions, a similar procedure was adopted. For low flows however, a 20% bandwidth was used. It is chosen to have a 20% bandwidth, because of the larger variability in meteorological indicators for low flows. This means that a historic year was selected when the meteorological indicator for low flows was between 10% lower and 10% higher than the expected value.

3.3. RQ3: Evaluation of performance for simulation under different climatic conditions

To assess the robustness of the HBV and GR6J model, it was first tested if and how the performance of a model changes under changing climatic conditions. A flowchart for this is given in Figure 13.

Various differential split-sample tests (DSSTs) were used to test if and how the performance of a model changes under changing climatic conditions. In these tests, the model was validated on historic periods that were selected in RQ2. The calibration was on non-selected historic periods. In case the performance in the validation period was as good as the performance in the calibration period, the model could be defined as robust. However, most hydrological models perform worse in a validation period than in a calibration period (Ji et al., 2023). Therefore, two split-sample tests (SSTs) were used as a reference. In the SSTs, the model was calibrated on the first half of the observed data and validated on the second half of the observed data, and vice versa. A model was robust in case the model did not show more notable deterioration in performance in the DSST than in the SST.

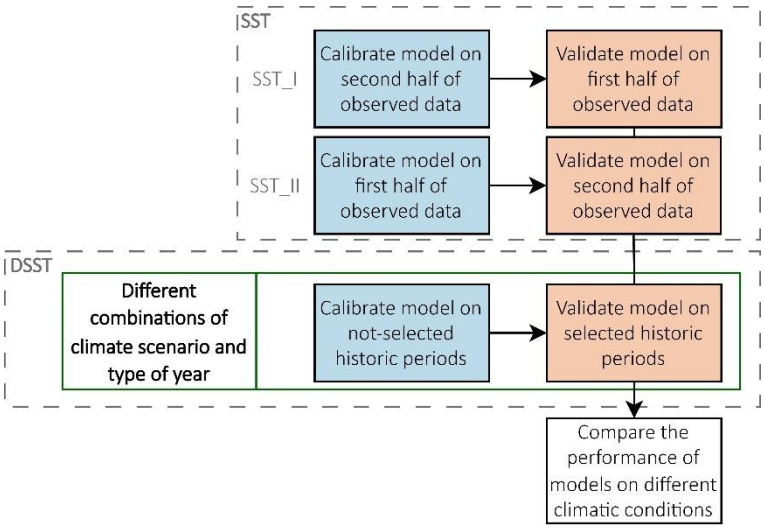

Figure 13 - Overview of methodology for RQ3. The procedure consisted of two SSTs ('SST_I' and 'SST_II') and a number of DSSTs. The green frame indicates that a number of DSSTs is formulated, related to the climate scenario and type of year.

The robustness of the HBV and GR6J model was evaluated for simulation of high flows, low flows and a combination of high and low flows. For each type of flow, the same testing methodology (Figure 13) was

used. However, the years selected for validation in the DSSTs differed per type of flow, as well as the selected calibration parameters and employed objective function. Section 3.3.1 provides an overview of the testing schemes for high flows, low flows and a combination of high and low flows. In section 3.3.2, the calibration procedure is described in more detail.

3.3.1. Testing schemes

High flows

For evaluating the robustness of models for simulating high flows, 2 SSTs and 4 DSSTs were formulated. An overview of the periods used for calibration and validation for each test is given in Figure 14.

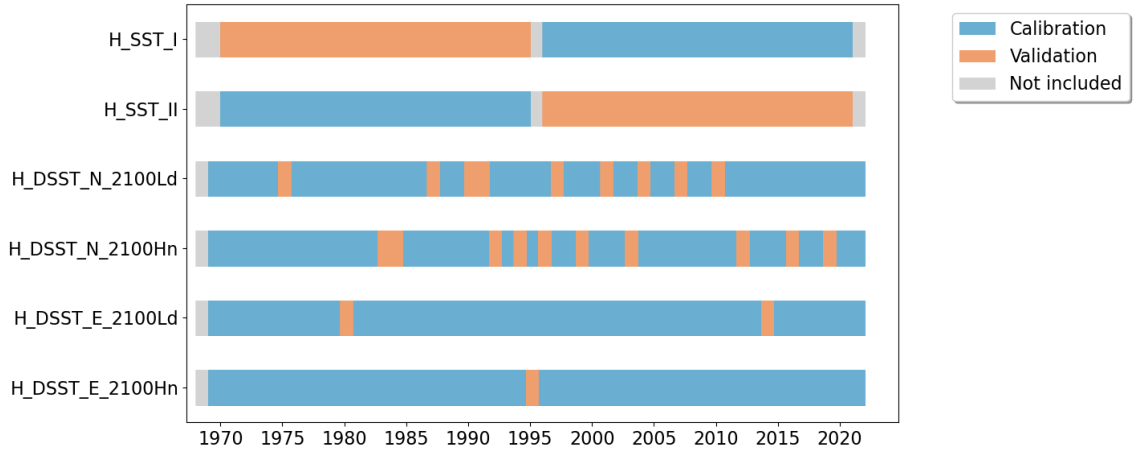

Figure 14 - Visualisation of the periods used for calibration (blue) and validation (orange) in the tests for simulation of high flows. Grey periods were not included in calibration or validation. The codes of the different tests are made up of the type of flow (High (H)), the type of test (SST or DSST), the type of year (Normal (N) or Extreme (E)) and the climate scenario (Figure 6). For the SSTs, the type of year and climate scenario is not relevant. In these tests, '_I' and '_II' are added. 'SST_I' means calibration on period 1996-2020 and validation on period 1970-1994. 'SST_II' means the opposite.

For the STTs, the period of 54 years for which historic data is available was divided in 2 periods of 25 years each. The first period was from 1/1/1970 to 31/12/1994 and the second period was from 1/1/1996 to 31/12/2020. The first two years (1968 and 1969) were used as warm-up period, to limit the effect of storage initialization. 1995 was not included in the calibration or validation, to make sure that both periods were independent of each other and 2021 was not included to make sure that both periods were of equal length.

Four DSSTs were formulated. In the first DSST, the validation periods matched the periods selected for the climate scenario with the smallest meteorological indicator for high flows in normal years, which is scenario *'2100Ld'*. This is the scenario for which the smallest high streamflow in normal years is expected. The validation periods of the second DSST were the periods selected for the climate scenario with the largest meteorological indicator for high flows in normal years, which is scenario *'2100Hn'*. This is the scenario for which the highest high streamflow in normal years is expected. Then, in the third and fourth DSST, the validation periods matched the periods for the climate scenarios with the smallest and largest meteorological indicator for high flows in extreme years, thus the climate scenarios for which the lowest and highest high streamflow are expected in extreme years. It was chosen to select two scenarios for normal years and two scenarios for extreme years, as these scenarios describe the limits between which high streamflow is likely to change for normal and extreme years in the future. A model was proven to be robust for all scenarios when the model was robust for these scenarios. In the DSSTs, all years that were not selected for validation were used in the calibration period, except 1968, which was used as warm-up period and thus excluded from calibration and validation.

An overview of the testing framework is given in Appendix C as well, which also shows the number of years selected for calibration and validation in each test.

Low flows

For evaluating the robustness of models for simulating low flows, 2 SSTs and 4 DSSTs were formulated. An overview of the periods used for calibration and validation for each test is given in Figure 15.

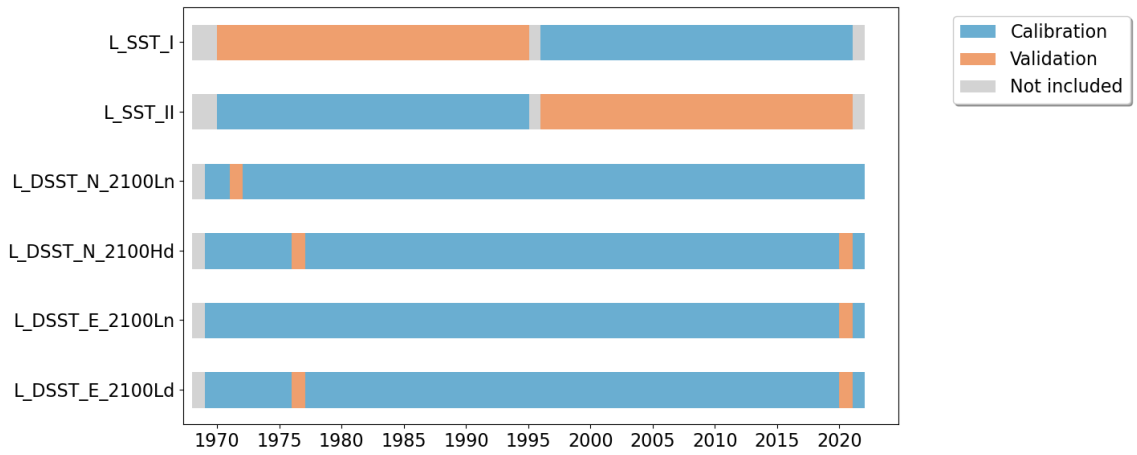

Figure 15 - Similar to Figure 14, but for the simulation of low flows. Type of flow is 'L' (Low).

The selected tests were based on the same reasoning as for high flows, which means the scenarios were selected which showed the lowest and highest meteorological indicator for low flows for both normal and extreme years. These scenarios thus represent the boundaries between which low streamflow is expected to change in normal and extreme years in the future. For the last DSST, scenario '*2100Ld*' was selected, as this was the scenario with the highest meteorological indicator for low flows in extreme years for which observed years were selected that resembled the conditions in this scenario. The validation period of test *'N_2100Hd'* was the same as the validation period of test *'E_2100Ld'*.

High and low flows

For evaluating the robustness of models for simulating both high and low flows, two SSTs and six DSSTs were formulated. An overview of the periods used for calibration and validation for each test is given in Figure 16.

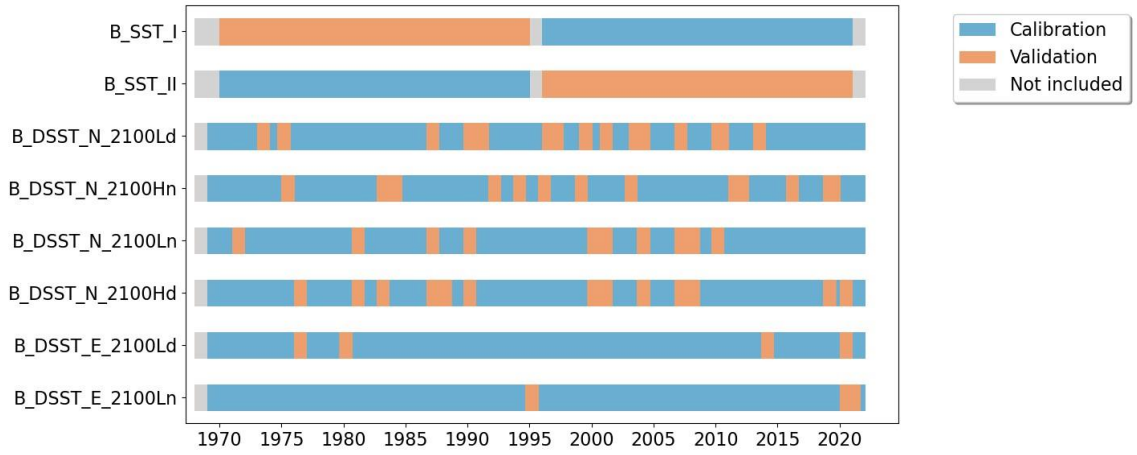

Figure 16 - Similar to Figure 14, but for the simulation of high and low flows. Type of flow is 'B' (Both).

The selected SSTs were based on the same reasoning as for high and low flows. Six DSSTs were formulated. Four tests were formulated for normal years (*'N_2100Ld'*, *'N_2100Hn'*, *'N_2100Ln'* and

'N_2100Hd'), one for each scenario that was included in the DSSTs for either high flows or low flows. For extreme years, DSSTs for scenarios *'E_2100Ld'* and *'E_2100Ln'* were formulated, as for these scenarios both years were selected for high flows and low flows. Appendix C shows the years selected for validation in all DSSTs, categorized into years selected based on high flows and based on low flows.

3.3.2. Calibration procedure

The HBV and GR6J models (section 2.2) were calibrated using the SCEM-UA algorithm (section 3.3.2.1). This algorithm uses a predefined objective function (section 3.3.2.2) and parameter space (section 3.3.2.3) to find the optimal parameter set. A sensitivity analysis was performed to select the most important parameters to be calibrated (section 3.3.2.4).

Parameters were calibrated on the different periods described in section 3.3.1. Although discontinuous sub-periods were used for model calibration and validation, the models were run continuously over the period 1968-2021. Only the periods corresponding to the calibration or validation periods were considered for calculating the objective function. The storages in the HBV model (SM, UZ and LZ) and GR6J model (S, R1 and R2) were initialized at 0 mm. The effect of this storage initialization was checked to be negligible because of the warm-up period that was included in all tests.

3.3.2.1. SCEM-UA algorithm

Automatic model calibration methods are objective and relatively easy to implement (Vrugt et al., 2003). Therefore, an automatic model calibration method called the Shuffled Complex Evolution Metropolis (SCEM-UA) algorithm (Vrugt et al., 2003) was used. This optimization algorithm was selected due to its proven fast converge to a global optimum (Osuch et al., 2015; Van Den Tillaart et al., 2013; Vrugt et al., 2003). The SCEM-UA algorithm is an automatic global searching method based on the SCE-UA algorithm of Duan et al. (1992). The SCEM-UA algorithm, a Markov Chain Monte Carlo (MCMC) sampler, uses the strengths of the Metropolis algorithm, controlled random search, competitive evolution and complex shuffling to find the optimal parameter set in a parameter space (Vrugt et al., 2003).

In short, a number (s) of parameter sets was randomly generated within the predefined parameter space. Based on the objective function for each parameter set, sets were ranked and partitioned in a number (q) of complexes and sequences. The Sequence Evolution Metropolis (SEM) algorithm was used to find new candidate parameter sets based on the objective function of previous sets. As a result, the algorithm continuously evolved the population towards parameter sets with the best objective function. This is called genetic drift. The algorithm stopped when the maximum number of iterations was reached. In this study, no convergence criterion was used to terminate the algorithm before reaching the maximum number of iterations. The SCEM-UA algorithm is extensively described in Vrugt et al. (2003).

Table 3 shows the settings that were used for the SCEM-UA algorithm and the explanation for choosing these settings. Most settings were based on the recommendations of Vrugt et al. (2003) for complex problems. The termination settings were configured such that convergence and a stable objective function value was found. This is shown in Appendix D.

Table 3 - Settings for the SCEM-UA algorithm.

	Setting	Description	Value	Explanation
Algorithmic settings	*q*	Number of sequences or complexes.	10	Recommended by Vrugt et al. (2003) for complex problems.
	s	Number of initial values.	250	Recommended by Vrugt et al. (2003) for complex problems.
Basic settings	*L*	Offspring value.	m/5	Value used in Van Den Tillaart et al. (2013)
	T	Predefined likelihood ratio.	10^6	Basic choice, used in the case studies of Vrugt et al. (2003).

	c_n	Jump rate.	$2.4/\sqrt{n}$	Basic choice, recommended by Vrugt et al. (2003).
Termination settings	*n*	Number of iterations.	5000	Turned out to be sufficient iterations to find convergence and a global optimum.
	GR	Gelman-Rubin convergence criterion. Score close to 1 indicates convergence.	Not included	Including this convergence criterion did not lead to a global optimum. Excluding the criterion still led to a fast computation.

3.3.2.2. Objective functions

Sorooshian & Gupta (1995) described the goal of calibration as "*selecting model parameters such that the model simulates the hydrological behaviour of the catchment as closely as possible".* Depending on the purpose of simulation, it was determined which hydrological behaviour should be simulated as closely as possible and thus which objective function was used. In general, four different objectives can be considered, namely (1) a good water balance, (2) a good overall agreement of the shape of the hydrograph, (3) a good agreement of high flows and (4) a good agreement of low flows (Madsen, 2000). This study used a multi-objective function, which combines multiple single objectives functions. The use of a multi-objective function resulted in a single-objective optimalisation problem which was solved with the SCEM-UA algorithm. The robustness of models for simulating high flows, low flows and a combination of high and low flows was evaluated. As the purposes for simulating high and low flows differ, a multi-objective function was selected for each type of flow. The multi-objective function was also used as performance criterion in the validation period.

High flows

For simulating high flows, model parameters were selected such that the model has a good agreement of peak flows. For this objective, a weighted form of the commonly used Nash-Sutcliffe criterion was used (NS_w, Equation 43), which was introduced by Hundecha & Bárdossy (2004). A weight w_i, equal to the observed flow at day i, was used to give emphasis to high flows. The mismatch between high observed and simulated discharge is penalized proportionally to the observed discharge value and therefore calibration on NS_w proves to provide good simulations of high flows in Vormoor et al. (2018) and Ott et al. (2013) as well. The value of NS_w ranges between -∞ and 1, with an optimum value of 1.

$$NS_w = 1 - \frac{\sum_{i=1}^{n}\left[w_i\left(Q_{sim}^i - Q_{obs}^i\right)^2\right]}{\sum_{i=1}^{n}\left[w_i\left(Q_{obs}^i - \overline{Q_{obs}}\right)^2\right]} \text{ with } w_i = Q_{obs}^i \qquad \text{[Eq. 43]}$$

Q_{sim}^i: Simulated discharge at day i
Q_{obs}^i: Observed discharge at day i
$\overline{Q_{obs}}$: Mean observed discharge over n days
n: The total number of timesteps

Next to that, a correct water balance was important. This objective was reflected in the objective function RVE, which is the relative volume error (Equation 44). The RVE varies between-∞ and ∞, with a perfect water balance leading to a RVE of 0. In case the RVE is shown as a percentage, Equation 44 is multiplied by 100.

$$RVE = \frac{\sum_{i=1}^{n}(Q_{sim}^i - Q_{obs}^i)}{\sum_{i=1}^{n}(Q_{obs}^i)} \qquad \text{[Eq. 44]}$$

Combining the objective functions, gave the multi-objective function y_w, defined in Equation 45. A comparable objective function was introduced by Akhtar et al. (2009). An optimal situation results in a NS_w of 1 and a RVE of 0, thus y_w of 1.

$$y_w = \frac{NS_w}{1 + |RVE|} \quad \text{with } RVE \text{ as a fraction} \qquad \text{[Eq. 45]}$$

Low flows

For the simulation of low flows, the model should have a good agreement of low flows. Therefore, the Nash-Sutcliffe criterion calculated on inverse transformed flows (NS_{inv}) was used (Equation 46). This objective function was introduced by Le Moine (2008) and recommended by Pushpalatha et al. (2012), as this criterion allows a focus on the lowest 20% of the flows over a study period. Just as the weighted Nash-Sutcliffe criterion, the NS_{inv} ranges between -∞ and 1, with an optimum value of 1. As the simulated and observed daily discharges are never equal to 0, the addition of a small constant ε described by Pushpalatha et al. (2012) was not necessary.

$$NS_{inv} = 1 - \frac{\sum_{i=1}^{n}\left(q_{sim}^{i} - q_{obs}^{i}\right)^2}{\sum_{i=1}^{n}\left(q_{obs}^{i} - \overline{q_{obs}}\right)^2} \text{ with } q^i = \frac{1}{Q^i} \qquad \text{[Eq. 46]}$$

A correct water balance was important for low flows as well. Therefore, a multi-objective function y_{inv}, defined in Equation 47, was used for the simulation of low flows. An optimal situation would result in a NS_{inv} of 1 and a RVE of 0, thus y_{inv} of 1.

$$y_{inv} = \frac{NS_{inv}}{1 + |RVE|} \qquad \text{[Eq. 47]}$$

Both flows

For the simulation of both flows, the model should have a good agreement on both high flows and low flows and should have a correct water balance as well. Therefore, the objective functions for high flows and low flows were combined into the multi-objective function y_{comb} (Equation 48). An optimal simulation would result in a NS_{inv} and NS_w of 1 and a RVE of 0, thus y_{comb} of 1.

$$y_{comb} = \frac{NS_w + NS_{inv}}{2(1 + |RVE|)} \qquad \text{[Eq. 48]}$$

3.3.2.3. Parameter space

In the SCEM-UA algorithm, model parameters were drawn from a predefined parameter space. The HBV model contains 8 parameters (section 2.2.1). The lower and upper limits of the model parameters of the HBV model are listed in Table 4. These parameter limits were adopted from Osuch et al. (2015), that based their parameter space on various studies that used the HBV model in different catchments (Abebe et al., 2010; Booij & Krol, 2010; Deckers et al., 2010; Seibert, 1999, 2003).

Table 4 - Parameter limits for the HBV model.

	FC [mm]	β [-]	LP [-]	α [-]	K_f [day^{-1}]	K_s [day^{-1}]	PERC [mm/day]	Cflux [mm/day]
Lower limit	10	0.01	0.1	0.1	0.0005	0.0005	0.001	0
Upper limit	1000	6	1	1	0.3	0.3	6	6

The GR6J model contains 6 parameters (section 2.2.2). The lower and upper limits of the model parameters of the GR6J model are listed in Table 5. These limits were based on the parameter values obtained for the GR6J model on a large set of catchments in France (Pushpalatha et al., 2011).

Table 5 - Parameter limits for the GR6J model.

	X_1 [mm]	X_2 [mm/day]	X_3 [mm]	X_4 [day]	X_5 [-]	X_6 [mm]
Lower limit	0	-5	0	0	-5	0
Upper limit	1500	5	500	10	5	20

3.3.2.4. Selection of calibration parameters

A univariate sensitivity analysis was conducted to select the most important parameters for calibration. By limiting the number of parameters to optimize, the calibration became more efficient and less computationally intensive. The objective function was insensitive for parameters when variations in parameters did not significantly affect the objective function. The parameters for which the objective function was insensitive were given a fixed default value (Pushpalatha et al., 2011; SMHI, 2006). As the

selection of parameters depended on the model structure and objective function, the sensitivity analysis was conducted for each model and type of flow separately. The results of the sensitivity analysis can be found in Appendix E. Table 6 shows for each model and type of flow which parameters were selected and which value was used in case a parameter was not selected.

Table 6 - Selected parameters per model and type of flow. Blue cells with 'Yes' indicate that the parameter was selected for calibration. For non-selected parameters, the default value is stated. Default values of HBV are from SMHI (2006) and for GR6J from Pushpalatha et al. (2011).

Selection of parameters HBV model								
Simulation of...	FC	β	LP	α	K_f	K_s	PERC	Cflux
High flows	Yes	Yes	Yes	Yes	Yes	0.005 day^{-1}	1 mm/day	1 mm/day
Low flows	Yes	Yes	Yes	1	0.005 day^{-1}	Yes	Yes	1 mm/day
High and low flows	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1 mm/day
Selection of parameters GR6J model								
Simulation of...	X_1		X_2	X_3	X_4		X_5	X_6
High flows	Yes		Yes	Yes	Yes		Yes	10 mm
Low flows	Yes		Yes	Yes	Yes		Yes	Yes
High and low flows	Yes		Yes	Yes	Yes		Yes	Yes

3.4. RQ4: Evaluation of parameter sets for simulation under different climatic conditions

To assess the robustness of the HBV and GR6J model, it was also tested to what extent model parameter values changed when the model was calibrated under changing climatic conditions. Large differences in model parameter sets indicate a low model robustness. In case model parameters remained approximately constant and the model performed well under changing climatic conditions, a model was robust. A flowchart is given in Figure 17.

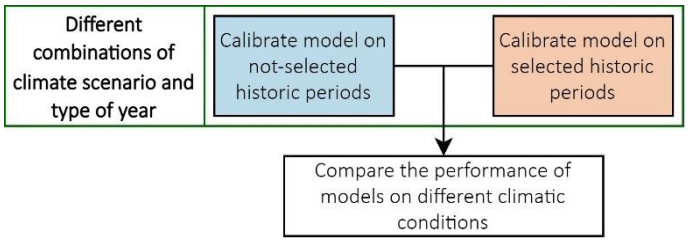

Figure 17 - Overview of methodology for RQ4. The models were calibrated on the periods not selected as periods resembling future conditions (blue) and on periods selected to resemble future conditions (orange). The green frame indicates that this was done for different combinations of climate scenario and type of year.

The parameter sets of the HBV and GR6J model were evaluated for simulation of high flows, low flows and a combination of high and low flows. The same testing schemes as described in section 3.3.1 were used. This means that the models were calibrated on the periods that do not resemble future conditions (blue) and periods that resemble future conditions (orange), shown in Figure 14, Figure 15 and Figure 16. The calibration procedure described in section 3.3.2 was used.

3.5. RQ5: Projected impact of climate change

This study focused on the robustness of hydrological models for projecting the impact of climate change on high and low streamflow in the Lesse. It was therefore interesting to see what the projected impact of climate change is and how this projected impact differed for different climate scenarios, hydrological model structures and calibration approaches. The projected impact of climate change on the annual maximum daily discharge (Q_{max}) and annual minimum 7-day mean discharge ($Q7_{min}$) in the Lesse was determined for 8 different scenarios. These scenarios were combinations of climate scenarios, hydrological model structures and calibration approaches (Table 7). The uncertainty of the impact due

to different climate scenarios was related to the uncertainty due to different hydrological model structures and calibration approaches by comparing the projected impact between the scenarios.

Table 7 - Overview of the 8 scenarios used to project the impact of climate change on high and low streamflow. The 8 scenarios consisted of combinations of a climate scenario (Figure 6), hydrological model structure and calibration approach. Calibration approach 'historic' means that models were calibrated on a 25-year historic period. Calibration approach 'future' means that models were calibrated on historic periods that resemble future climatic conditions.

# Scenario		1	2	3	4	5	6	7	8
Climate scenario	High flows	2100Ld	2100Ld	2100Ld	2100Ld	2100Hn	2100Hn	2100Hn	2100Hn
	Low flows	2100Ln	2100Ln	2100Ln	2100Ln	2100Hd	2100Hd	2100Hd	2100Hd
Model structure		HBV	HBV	GR6J	GR6J	HBV	HBV	GR6J	GR6J
Calibration approach		Historic	Future	Historic	Future	Historic	Future	Historic	Future

Climate scenarios

For determining the impact of climate change on high flows, climate scenarios *'2100Ld'* and *'2100Hn'* were used, as in these scenarios the smallest and largest changes in Q_{max} are expected, respectively. For low flows, climate scenarios *'2100Ln'* and *'2100Hd'* were selected, as here the smallest and largest changes in $Q7_{min}$ are expected.

Hydrological model structure

The uncertainty due to different hydrological model structures was determined by comparing the projected impact of the HBV and GR6J model.

Calibration approaches

Generally, hydrological models are calibrated on a sufficiently long historic period, with a wide range of climatic conditions. However, various studies advocated using a calibration period that closely resembles future climatic conditions (Ji et al., 2023; Stephens et al., 2020), as these are the conditions that are expected in the evaluation period. For both calibration approaches, the projected impact of climate change was determined. For calibration on historic periods, the models were calibrated on the period 1996-2020 (*'SST_I'* in section 3.3.1). Secondly, hydrological models were calibrated on the periods that were selected to resemble normal years in the future. These are the orange periods in section 3.3.1. For example, for high flows and climate scenario *'2100Ld'*, the model was calibrated on the periods that resemble future conditions (orange) for *'H_DSST_N_2100Ld'* (Figure 14).

All 8 ensemble members in the climatic timeseries for future scenarios (section 2.3.3) were used, each consisting of 30 years of data (period 2086-2115). The input of the hydrological models was thus a climatic timeseries with a length of 240 years. The first year was used as warm-up period. Therefore, Q_{max} was determined for 238 hydrological years for high flows and $Q7_{min}$ was determined for 239 years for low flows. Q_{max} and $Q7_{min}$ were determined for the 8 ensemble members of the reference climate (period 1991-2020) as well, using the HBV and GR6J model calibrated on historic period 1996-2020. A boxplot was made for each scenario. The relative change for each future scenario compared to the reference climate was determined by sorting the Q_{max} and $Q7_{min}$ values from high to low and calculating the relative difference between each pair or points. From the series of relative differences, boxplots were made.

4. RESULTS

4.1. RQ1: Changes in important meteorological indicators

4.1.1. Meteorological indicators for high and low flows

4.1.1.1. Correlation analysis

High flows

Figure 18 shows the Pearson correlation coefficients between the series of 53 annual maximum daily discharges and their preceding precipitation sums of different temporal windows. The correlation between the annual maximum daily discharge and the precipitation sum $P3i$ is the highest. This precipitation sum is the sum of the precipitation on the day with the annual maximum discharge and the two days before and has a Pearson correlation coefficient of 0.85.

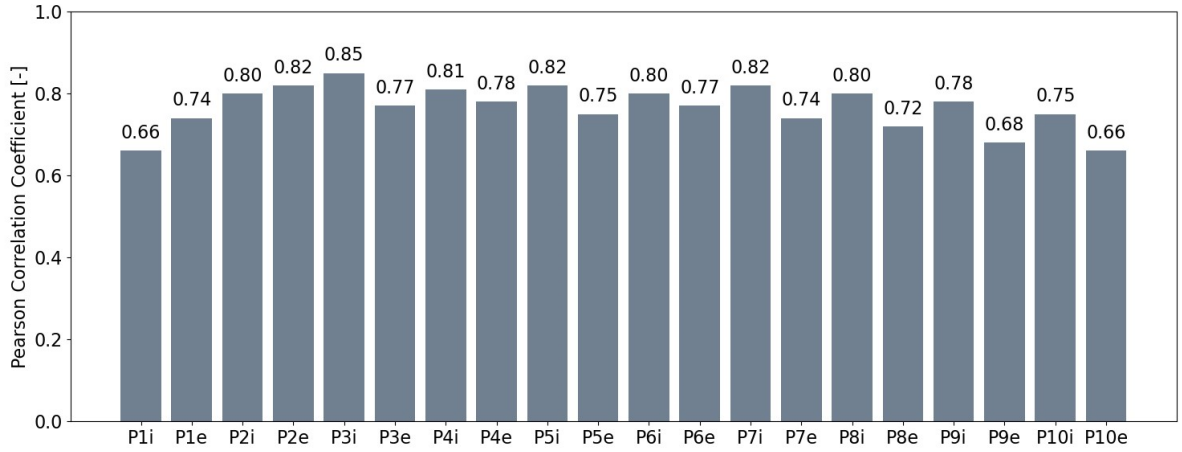

Figure 18 - Pearson correlation coefficients for the correlations between the timeseries of annual maximum daily discharges and preceding precipitation sums of various temporal windows.

For temporal windows of 3 days or larger, the precipitation sum including the day of discharge (PXi) shows a larger correlation than the precipitation sum excluding the day of discharge (PXe). For example, the correlation of $P4i$ is 0.81, while this correlation is 0.78 for $P4e$. This indicates that a part of the precipitation is discharged the same day. The positive correlation between $P3i$ and Q_{max} is visible in the scatterplot of Figure 20A. Here, a positive relationship between the annual maximum daily discharge and the preceding precipitation $P3i$ can be observed.

Low flows

Figure 19 shows the Pearson correlation coefficients between a series of 53 annual minimum 7-day mean discharges and the preceding potential precipitation deficits aggregated over periods from 1 to 12 months (a month is 30 days). Negative correlations are shown between the precipitation deficits and the annual minimum 7-day mean discharge, indicating that a larger precipitation deficit leads to a smaller minimum discharge. The correlation is the highest when the precipitation deficit was aggregated over 5 months (150 days), with a Pearson correlation coefficient of-0.76.

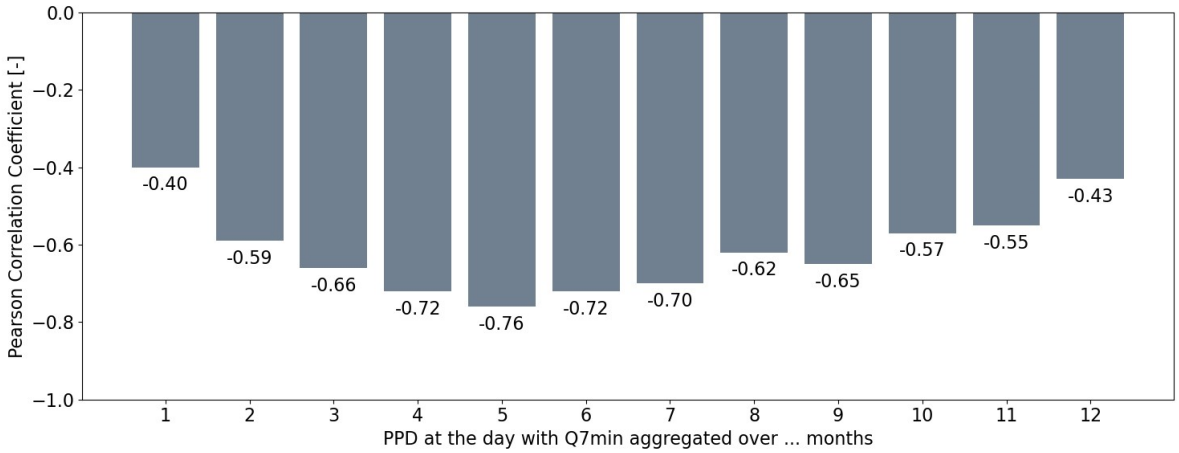

Figure 19 - Pearson correlation coefficients for the correlations between the timeseries of annual minimum 7-day mean discharges and preceding potential precipitation deficits aggregated over various temporal windows.

A scatterplot of the annual minimum 7-day mean discharge versus the preceding potential precipitation deficit aggregated over 150 days is shown in Figure 20B. For 13 out of 53 years, the potential precipitation deficit is negative, indicating a surplus. For the remaining years, a potential precipitation deficit is visible.

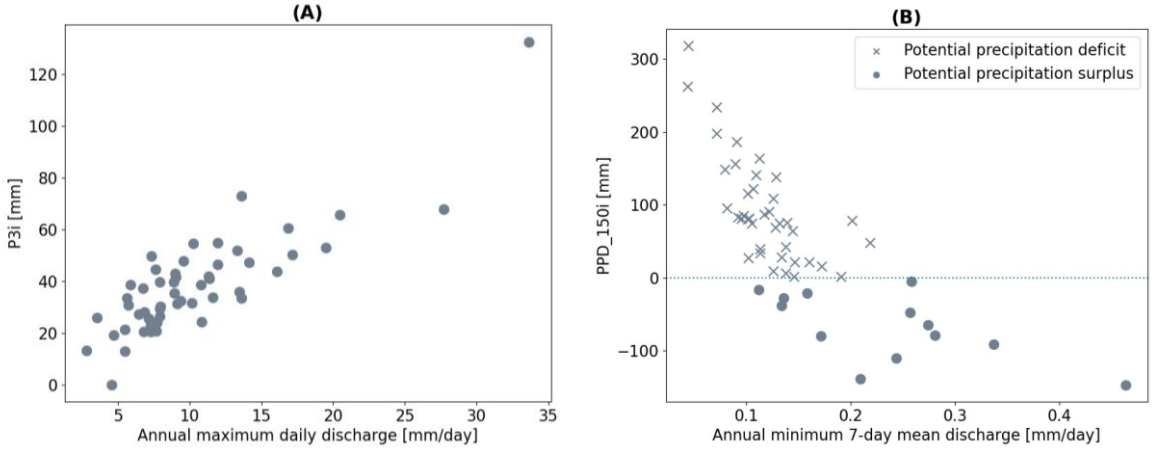

Figure 20 - A) Scatterplot of the annual maximum daily discharge versus the preceding precipitation sum with the highest correlation (P3i), B) Scatterplot of the annual minimum 7-day mean discharge versus the preceding potential precipitation deficit with the highest correlation (PPD_150i). The dots indicate a negative deficit, which is a surplus.

4.1.1.2. Exceedance frequency

High flows

Figure 21 shows the annual exceedance frequency of the $P3i$ value for the hydrological years 1969-2021. This exceedance frequency varies between 0 and 273. The median exceedance frequency is 6 days, which corresponds with an exceedance of 1.6% of the days in a year. Based on these results, the 3-day precipitation sum that is exceeded 6 times per year was chosen as a meteorological indicator for high flows.

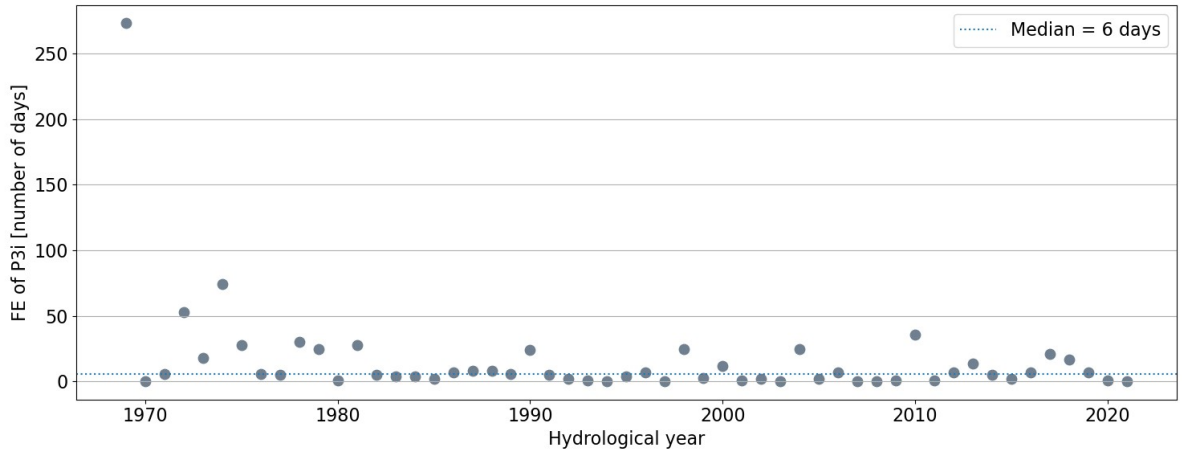

Figure 21 - Frequency of exceedance (FE) of the precipitation sum P3i preceding the annual maximum discharge. A frequency of exceedance of 0 means that the P3i was the highest P3 in that year. The dotted line represents the median exceedance frequency.

Low flows

Figure 22 shows the annual exceedance frequency of the preceding potential precipitation deficit with a temporal window of 150 days for the hydrological years 1969-2021. This frequency varies between 0 and 84. The median exceedance frequency is 14 days, which corresponds with an exceedance of 3.8% of the days in a year. Based on these results, the potential precipitation deficit over 150 days that is exceeded 14 times per year was chosen as a meteorological indicator for low flows.

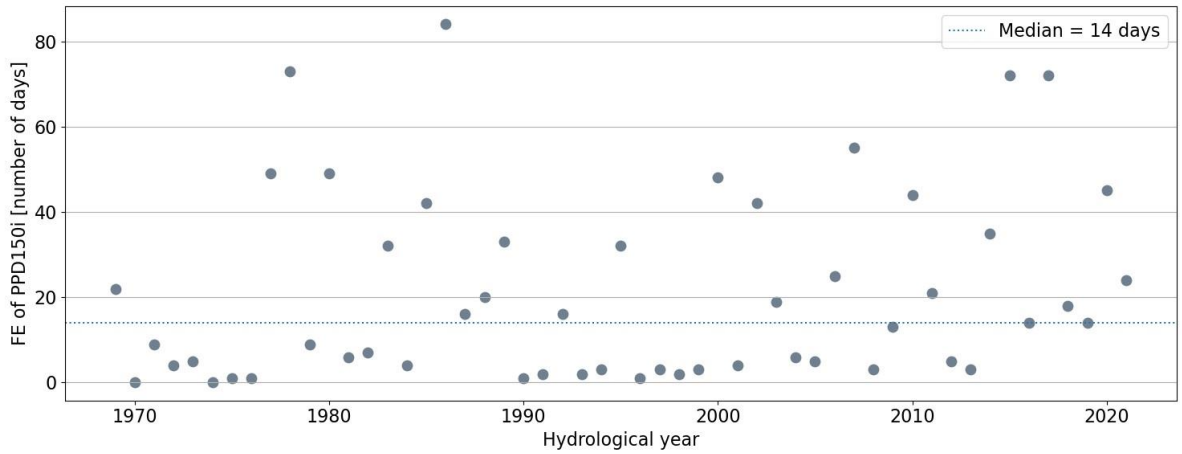

Figure 22 - Frequency of exceedance (FE) of the potential precipitation deficit over 150 days preceding the annual minimum 7-day mean discharge. A frequency of exceedance of 0 means that the PPD150i was the highest potential precipitation deficit over 150 days in that year. The dotted line represents the median exceedance frequency.

Compared to Figure 21, Figure 22 shows more variations in the exceedance frequency. This shows that it is more difficult to identify the rank for the preceding potential precipitation deficits for low flows than the preceding precipitation sums for high flows. The reason for this may be related to the fact that potential precipitation deficits were determined over a longer timespan (150 days), whereas the preceding precipitation sums were determined for 3 days. As a result, the exact rank of the potential precipitation deficit was more difficult to identify.

4.1.2. Changes in the future

4.1.2.1. Bias-correction

Figure 23A shows a boxplot of the meteorological indicator for high flows for the observations and the reference climate. The boxplot for the observations was made up of 29 values, the boxplot of the

reference climate was made up of 232 values. The median value of the observations is 36.0 mm, whereas the median value of the reference climate is 36.6 mm. This shows there is a small bias in the climate model; the precipitation was overestimated by the model.

Figure 23B shows a boxplot of the relative differences of the 29 pairs of meteorological indicators for high flows based on the observations and the reference climate. The meteorological indicator for high flows in the observed dataset is on average 4.74% lower than the meteorological indicator for high flows in the reference climate. Therefore, the meteorological indicators for high flows in the future scenarios were bias-corrected by making all values 4.74% lower.

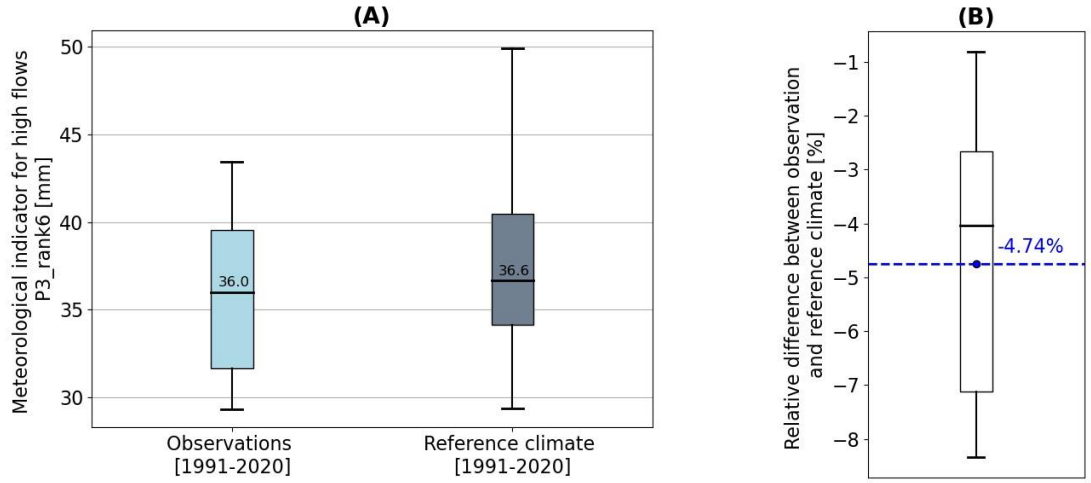

Figure 23 - A) Boxplot of the meteorological indicator for high flows for the observations (blue) and reference climate (grey). B) Boxplot of the relative differences of the 29 pairs of meteorological indicators based on the observations and reference climate. The horizontal lines in the boxes represent the median value, the boxes represent the 25-75% data range and the whiskers represent the 5-95% data range. The values in the boxes are the median values. The blue dotted line in Figure B shows the average value.

Figure 24 shows a boxplot of the meteorological indicator for low flows for the observations and the reference climate. The median value of the observations is 56.5 mm, whereas the median value of the reference climate is 85.5 mm. This shows that the climate model overestimated the potential precipitation deficit.

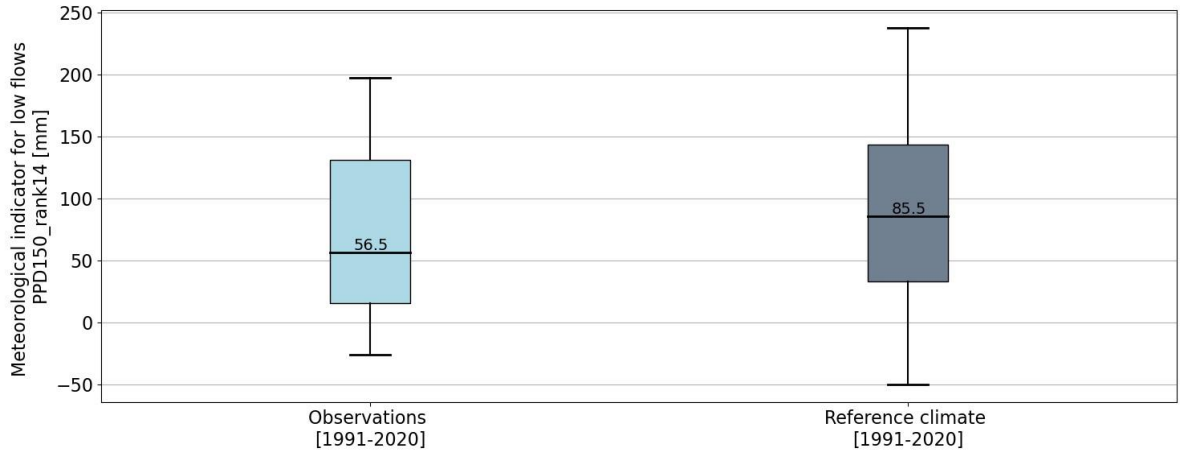

Figure 24 - Boxplot of the meteorological indicator for low flows for the observations (blue) and reference climate (grey). The horizontal lines in the boxes represent the median value, the boxes represent the 25-75% data range and the whiskers represent the 5-95% data range. The values in the boxes are the median values.

The bias can be either related to the corresponding precipitation sum, the corresponding potential evapotranspiration sum or a combination of both. Figure 25 shows that the climate model overestimates the corresponding precipitation sum as well as the corresponding potential evapotranspiration sum. Figure 26 shows that the precipitation sum corresponding to the meteorological indicator for low flows was overestimated on average with 4.09%, and the potential evapotranspiration sum was overestimated on average with 6.12%. The bias in potential evapotranspiration sum is thus larger than the bias in precipitation sum.

All future values of the precipitation sum and potential evapotranspiration sum were bias-corrected with 4.09% and 6.12%, respectively. The bias-corrected values of the potential precipitation deficit were obtained by subtracting the corrected values of the precipitation sum from the corrected values of the potential evapotranspiration sum.

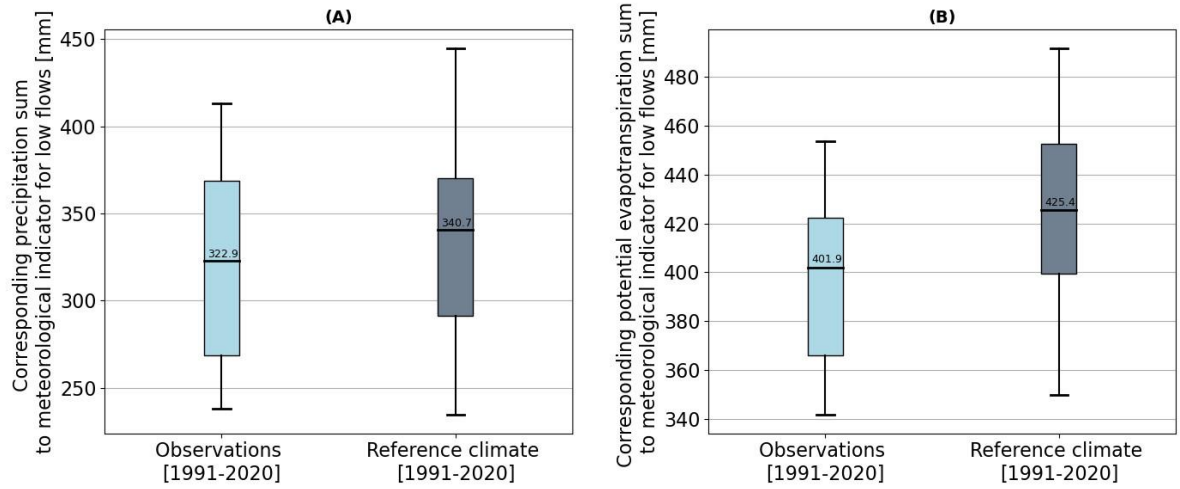

Figure 25 - Boxplots of (A) the corresponding precipitation sum and (B) the corresponding potential evapotranspiration sum to the meteorological indicator for low flows for the observations (blue) and reference climate (grey). The horizontal lines in the boxes represent the median value, the boxes represent the 25-75% data range and the whiskers represent the 5-95% data range. The values in the boxes are the median values.

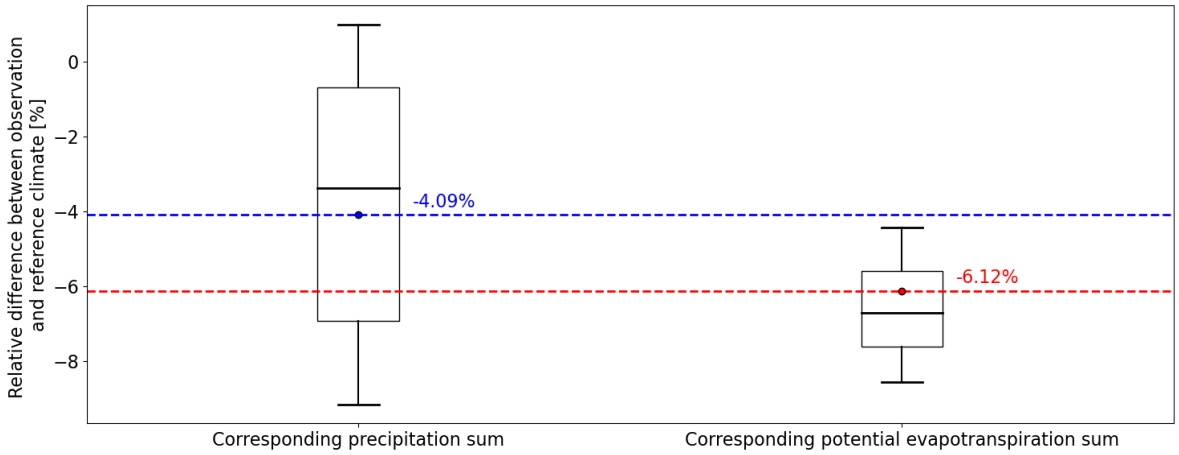

Figure 26 - Boxplots of the relative differences of the 29 pairs of corresponding precipitation and potential evapotranspiration sum to the meteorological indicator for low flows based on the observations and reference climate. The horizontal lines in the boxes represent the median value, the boxes represent the 25-75% data range and the whiskers represent the 5-95% data range. The values in the boxes are the median values. The blue and red dotted lines show the average values of the corresponding precipitation and potential evapotranspiration sum.

4.1.2.2. Changes in the future

Figure 27 shows the boxplots made up from a series of 232 bias-corrected meteorological indicators for high flows for each future scenario, as well as the boxplot made up from a series of 29 observed meteorological indicators for high flows. Normal years are defined as the median of all years in the future, extreme years are years for which the meteorological indicator is exceeded 5% of the time.

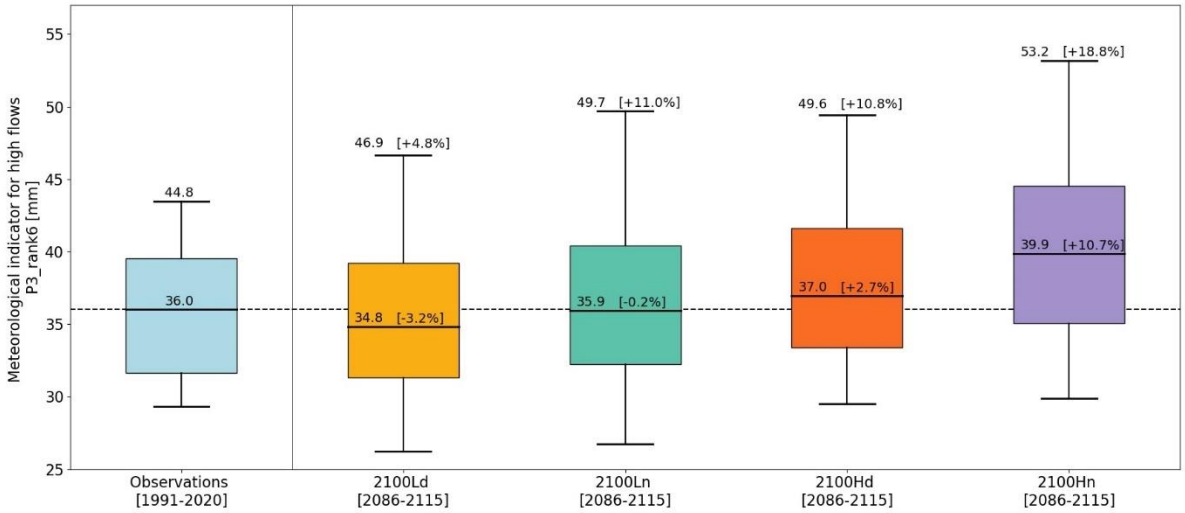

Figure 27 - Boxplot of the meteorological indicator for high flows, based on observations (blue) and the bias-corrected values of each future scenario (other colours). The horizontal lines in the boxes represent the median value, the boxes represent the 25-75% data range and the whiskers represent the 5-95% data range. The values in the boxes are the median values, with the relative change compared to observations in square brackets. The values at the whiskers are the absolute values in extreme years (5% exceedance probability), with the relative change compared to the observations in square brackets.

For normal years, the meteorological indicator for high flows increases compared to the observations in the high emission scenarios *'2100Hd'* and *'2100Hn'*. For the other scenarios, a slight decrease in the median value is visible. For extreme years, the meteorological indicator for high flows increases in all scenarios compared to the observations. The highest meteorological indicator for high flows in normal years is expected in scenario *'2100Hn'*, while the lowest is expected in scenario *'2100Ld'*. This means that it is expected that these scenarios will lead to the highest and lowest high streamflow, respectively. Climate scenario *'2100Hn'* is based on high future emissions and a trend towards a wetter future climate. It therefore makes sense that this climate scenario leads to the highest meteorological indicator for high flows. The relative change compared to observations is larger for extreme years than for normal years. This indicates that more variation in the meteorological indicator for high flows is expected in the future.

Figure 28 shows the boxplots made up from a series of 232 bias-corrected meteorological indicators for low flows for each future scenario, as well as the boxplot made up from a series of 29 observed meteorological indicators for low flows.

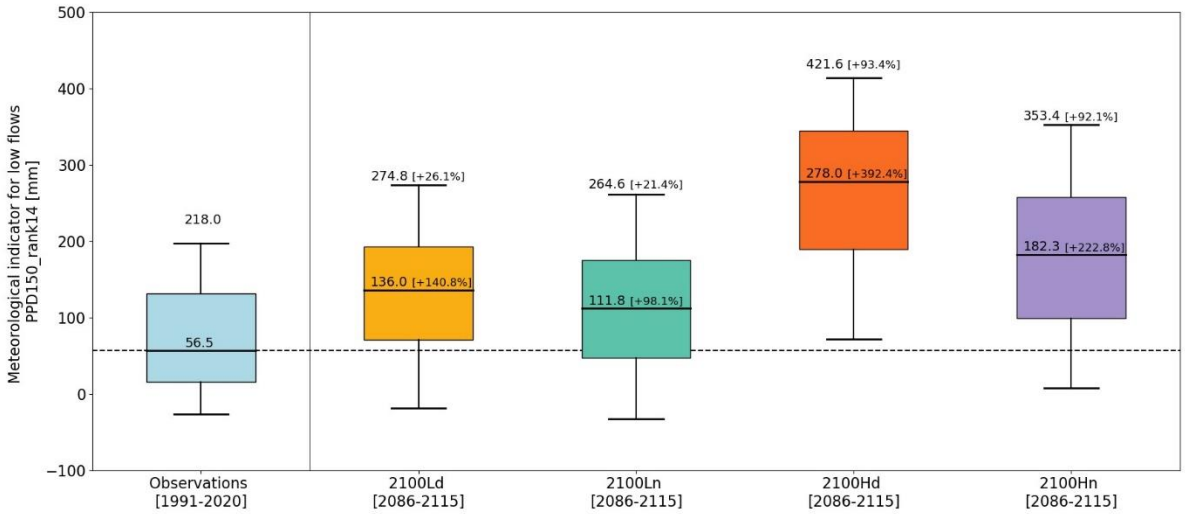

Figure 28 - Similar to Figure 27, but for the meteorological indicator for low flows.

An increase in the median value for the meteorological indicator for low flows compared to the observations is visible for all future scenarios, with the largest increase (392.4%) in scenario *'2100Hd'*. This scenario is based on high future emissions and a trend towards a drier future climate. It therefore makes sense that this climate scenario leads to the highest meteorological indicator for low flows. This means that it is expected that the lowest low streamflow is expected in scenario *'2100Hd'*. For extreme years, an increase is visible for all future scenarios as well, however this relative increase is smaller than for normal years. The increase in potential precipitation deficit is both due to an increase in potential evapotranspiration as well as a decrease in precipitation (Appendix F).

4.2. RQ2: Selection of historic periods resembling future climatic conditions

High flows

The sets of historic years selected to resemble future high flow conditions are shown in Figure 29 and Figure 30 for normal and extreme years, respectively. The upper row shows the observed values of the meteorological indicator for high flows in the period 1969-2021. The values range between 24 mm (2017) and 51 mm (1995). Then, the other 4 rows show the sets of hydrological years that are selected for each future scenario.

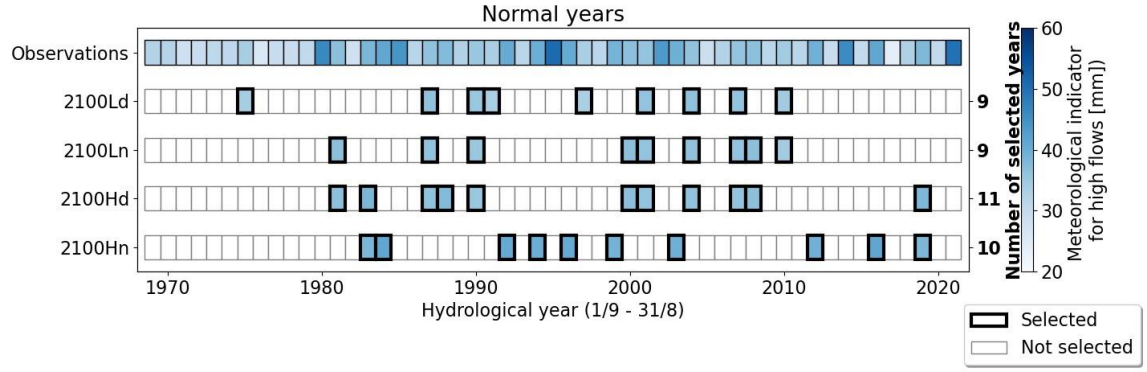

Figure 29 - Visualisation of the hydrological years (1/9/X-1 to 31/8/X) selected as historic period resembling future high flow conditions for each scenario, based on normal years. The colours in the boxes refer to the magnitude of the meteorological indicator for high flows. In the right of the figure, the number of selected years for each scenario is shown in bold.

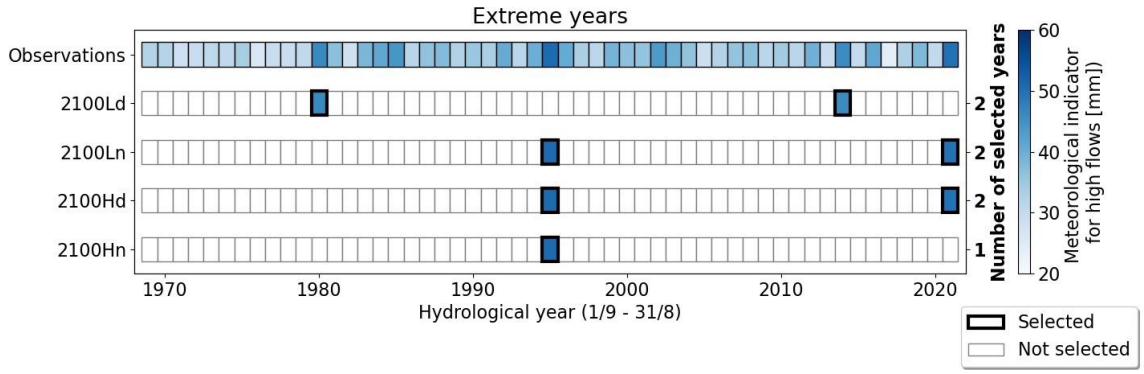

Figure 30 - Similar to Figure 29, but for extreme years.

For normal years, for each scenario at least 9 years are selected. There is some overlap in the years that are selected for each scenario. For example, year 2004 is selected for each scenario, except scenario *'2100Hn'*. The selection includes years with an above-average value for the meteorological indicator for high flows, however the years with the highest values for meteorological indicators are not included. For extreme years, only 1 or 2 years are selected. For scenarios *'2100Ln'* and *'2100Hd'*, these are 1995 and 2021, the years with the two highest values for the meteorological indicator. For scenario *'2100Hn'*, only 1995, the year with the highest meteorological indicator, is selected.

Low flows

The sets of historic years selected to resemble future low flow conditions are shown in Figure 31 and Figure 32 for normal and extreme years, respectively. The observed values of the meteorological indicator for low flows in the period 1968-2021 range between -141 mm (1987) and 299 mm (1976).

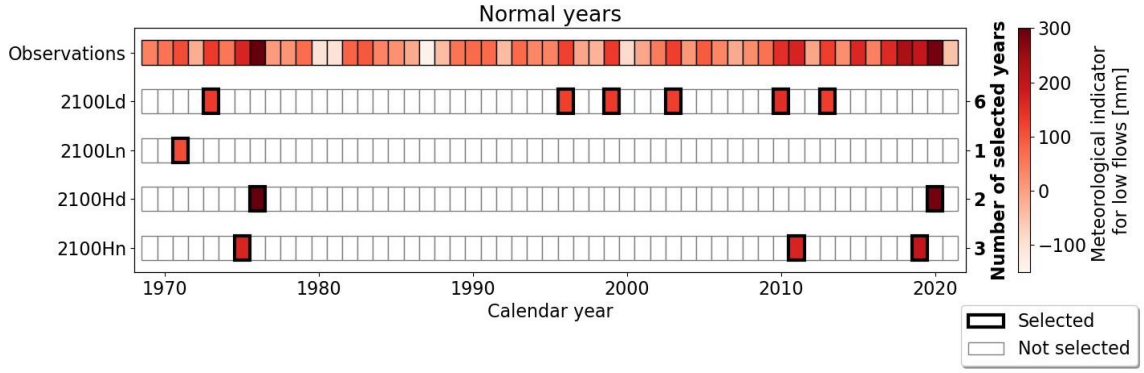

Figure 31 - Similar to Figure 29, but for low flows and normal years.

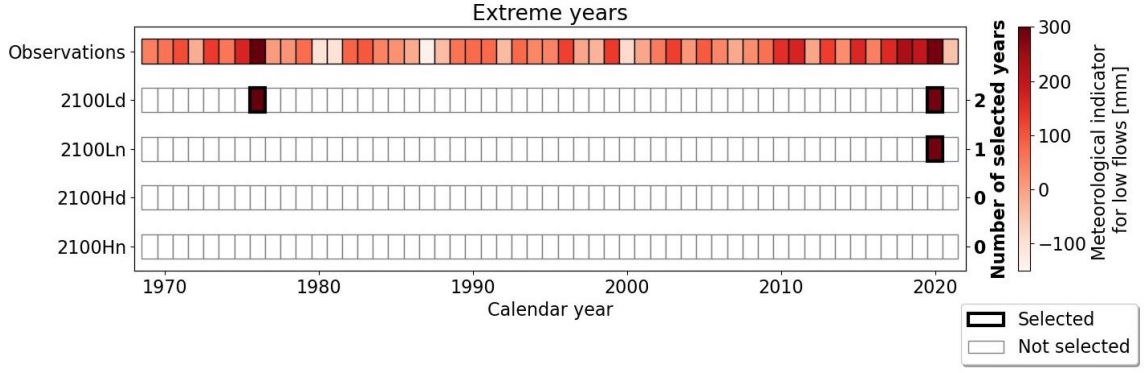

Figure 32 - Similar to Figure 29, but for low flows and extreme years.

For normal years, between 1 and 6 years are selected per scenario. Compared to the selected years for high flows, there is little overlap between the scenarios. For extreme years, only years are selected for *'2100Ld'* and *'2100Ln'*. For scenarios *'2100Hd'* and *'2100Hn'*, no observed years fall within the bandwidth and thus no years are selected.

4.3. RQ3: Evaluation of performance for simulation under different climatic conditions

This section shows and discusses the results for RQ3. First, sections 4.3.1 and 4.3.2 show the results and state the main observations for the HBV model and GR6J model, respectively. In section 4.3.3, the results of both models are compared, explained and discussed.

4.3.1. Results HBV model

The performance of the HBV model on the calibration and validation period of all testing schemes is shown for the simulation of high (Table 8A), low (Table 8B) and high and low flows (Table 8C).

*Table 8 - Performance of the **HBV** model in the testing schemes for simulation of A) high, B) low and C) high and low flows. The names of the testing schemes are explained in section 3.3.1. For high flows, performance is given by objective function y_w and corresponding NS_w and RVE. For low flows, performance is given by objective function y_{inv} and corresponding NS_{inv} and RVE. For simulation of high and low flows, performance is given by objective function y_{comb} and corresponding NS_w, NS_{inv} and RVE. Blue columns show the performance in the calibration period, orange columns the performance in the validation period. The last column (Δy) shows the difference in multi-objective function between validation and calibration. Red and green numbers indicate a degradation and increase of performance in the validation period compared to the calibration period, respectively.*

A) Simulation of high flows									
Testing scheme	Calibration period				Validation period				Δy_w [-]
	y_w [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	y_w [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	
HBV_H_SST_I	0.83	0.83	-	$-2.0 \cdot 10^{-4}$	0.80	0.82	-	-2.6	-0.03
HBV_H_SST_II	0.82	0.82	-	$+1.2 \cdot 10^{-4}$	0.80	0.83	-	+3.4	-0.02
HBV_H_DSST_N_2100Ld	0.82	0.82	-	$+3.2 \cdot 10^{-3}$	0.72	0.78	-	+8.0	-0.10
HBV_H_DSST_N_2100Hn	0.81	0.81	-	$+2.7 \cdot 10^{-3}$	0.81	0.82	-	+1.1	0.00
HBV_H_DSST_E_2100Ld	0.82	0.82	-	$-1.1 \cdot 10^{-3}$	0.75	0.75	-	$-8.9 \cdot 10^{-2}$	-0.07
HBV_H_DSST_E_2100Hn	0.81	0.81	-	$+2.6 \cdot 10^{-3}$	0.84	0.86	-	+2.6	+0.03
B) Simulation of low flows									
Testing scheme	Calibration period				Validation period				Δy_{inv} [-]
	y_{inv} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	y_{inv} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	
HBV_L_SST_I	0.86	-	0.86	$+2.5 \cdot 10^{-3}$	0.80	-	0.82	-2.2	-0.06
HBV_L_SST_II	0.86	-	0.86	$+2.3 \cdot 10^{-3}$	0.81	-	0.83	+2.7	-0.05
HBV_L_DSST_N_2100Ln	0.85	-	0.85	$-5.7 \cdot 10^{-3}$	0.72	-	0.83	-15	-0.13
HBV_L_DSST_N_2100Hd	0.84	-	0.84	$+7.3 \cdot 10^{-4}$	0.83	-	0.88	+5.6	-0.01
HBV_L_DSST_E_2100Ln	0.85	-	0.85	$+1.2 \cdot 10^{-3}$	0.91	-	0.92	-1.6	+0.06
HBV_L_DSST_E_2100Ld	0.84	-	0.84	$+5.4 \cdot 10^{-3}$	0.82	-	0.86	+5.3	-0.02

Table continues on next page

C) Simulation of high and low flows									
Testing scheme	Calibration period				Validation period				Δy_{comb} [-]
	y_{comb} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	y_{comb} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	
HBV_B_SST_I	0.84	0.82	0.87	$+5.9 \cdot 10^{-3}$	0.81	0.81	0.83	-2.2	-0.03
HBV_B_SST_II	0.85	0.83	0.87	$-4.1 \cdot 10^{-3}$	0.81	0.83	0.83	+2.8	-0.04
HBV_B_DSST_N_2100Ld	0.83	0.80	0.87	$+2.3 \cdot 10^{-3}$	0.78	0.83	0.80	+4.0	-0.05
HBV_B_DSST_N_2100Hn	0.82	0.79	0.86	$+1.7 \cdot 10^{-3}$	0.83	0.82	0.85	+0.33	0.01
HBV_B_DSST_N_2100Ln	0.83	0.80	0.87	$+3.1 \cdot 10^{-1}$	0.78	0.83	0.80	+4.4	-0.05
HBV_B_DSST_N_2100Hd	0.82	0.80	0.83	$+7.0 \cdot 10^{-3}$	0.81	0.83	0.88	+5.9	-0.01
HBV_B_DSST_E_2100Ld	0.81	0.79	0.84	$+1.6 \cdot 10^{-3}$	0.86	0.86	0.90	+2.7	0.05
HBV_B_DSST_E_2100Ln	0.84	0.82	0.85	$+9.6 \cdot 10^{-4}$	0.78	0.69	0.91	-2.9	-0.06

For all SSTs, the multi-objective function y in the validation period is slightly lower than in the calibration period, which is mainly due to an increase in RVE. In the calibration periods, RVE is close to 0%, while the absolute value of RVE increases on average to +/-2.7% in the validation period. NS_w is constant in the SSTs, while NS_{inv} is mainly lower in the validation period. As a result, the largest decrease in y is observed for the simulation of low flows.

In most DSSTs, the multi-objective function y is lower in the validation period than in the calibration period. However, in four cases y is higher in the validation period (*'HBV_H_DSST_E_2100Hn'*, *'HBV_L_DSST_E_2100Ln'*, *'HBV_B_DSST_N_2100Hn'* and *'HBV_B_DSST_E_2100Ld'*) or there is no difference between calibration and validation period (*'HBV_H_DSST_N_2100Hn'*). In all DSSTs, an increase in RVE is observed in the validation period. For the DSSTs with an increase in y, the increase in NS_w and/or NS_{inv} prevails over the increase in RVE.

In most DSSTs NS_w increases. Only in *'HBV_H_DSST_N_2100Ld'*, *'HBV_H_DSST_E_2100Ld'* and *'HBV_B_DSST_E_2100Ln'*, the NS_w decreases. NS_{inv} increases in the DSSTs on extreme years and scenario *'N_2100Hd'*. In all other cases, NS_{inv} decreases. *'HBV_L_DSST_N_2100Hd'* and *'HBV_L_DSST_E_2100Ld'*, which have the same validation and calibration period, show comparable results in both periods.

4.3.2. Results GR6J model

Table 9 shows the performance of the GR6J model on the calibration and validation period of all tests.

*Table 9 - Similar to Table 8, but for the **GR6J** model.*

A) Simulation of high flows									
Testing scheme	Calibration period				Validation period				Δy_w [-]
	y_w [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	y_w [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	
GR6J_H_SST_I	0.89	0.89	-	$-3.5 \cdot 10^{-3}$	0.83	0.87	-	-5.1	-0.06
GR6J_H_SST_II	0.89	0.89	-	$-7.3 \cdot 10^{-3}$	0.83	0.88	-	+5.6	-0.06
GR6J_H_DSST_N_2100Ld	0.90	0.90	-	$-2.1 \cdot 10^{-3}$	0.82	0.89	-	+8.5	-0.08
GR6J_H_DSST_N_2100Hn	0.91	0.91	-	$+3.1 \cdot 10^{-3}$	0.93	0.94	-	-0.43	+0.02
GR6J_H_DSST_E_2100Ld	0.92	0.92	-	$+5.0 \cdot 10^{-3}$	0.66	0.69	-	+4.7	-0.26
GR6J_H_DSST_E_2100Hn	0.91	0.91	-	$-3.3 \cdot 10^{-1}$	0.90	0.97	-	+8.7	-0.01

Table continues on the next page

B) Simulation of low flows									
Testing scheme	Calibration period				Validation period				Δy_{inv} [-]
	y_{inv} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	y_{inv} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	
GR6J_L_SST_I	0.83	-	0.83	$-2.5 \cdot 10^{-1}$	0.77	-	0.81	-5.3	-0.06
GR6J_L_SST_II	0.80	-	0.80	$+1.4 \cdot 10^{-3}$	0.75	-	0.79	+5.8	-0.05
GR6J_L_DSST_N_2100Ln	0.82	-	0.82	$+7.7 \cdot 10^{-3}$	0.72	-	0.80	-12	-0.10
GR6J_L_DSST_N_2100Hd	0.77	-	0.77	$-1.2 \cdot 10^{-2}$	0.83	-	0.84	+1.2	0.06
GR6J_L_DSST_E_2100Ln	0.79	-	0.79	$+4.5 \cdot 10^{-4}$	0.86	-	0.90	-4.5	0.07
GR6J_L_DSST_E_2100Ld	0.77	-	0.77	$-2.3 \cdot 10^{-2}$	0.83	-	0.83	+0.76	0.06
C) Simulation of high and low flows									
Testing scheme	Calibration period				Validation period				Δy_{comb} [-]
	y_{comb} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	y_{comb} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	
GR6J_B_SST_I	0.82	0.86	0.79	$-5.7 \cdot 10^{-3}$	0.76	0.82	0.79	-5.7	-0.06
GR6J_B_SST_II	0.81	0.83	0.80	$+2.3 \cdot 10^{-3}$	0.77	0.85	0.79	+5.9	-0.04
GR6J_B_DSST_N_2100Ld	0.86	0.90	0.83	$-1.2 \cdot 10^{-2}$	0.73	0.89	0.62	+3.7	-0.13
GR6J_B_DSST_N_2100Hn	0.85	0.91	0.79	$+1.3 \cdot 10^{-3}$	0.83	0.88	0.79	$-9.4 \cdot 10^{-2}$	-0.02
GR6J_B_DSST_N_2100Ln	0.85	0.90	0.79	$-3.0 \cdot 10^{-3}$	0.75	0.89	0.72	+7.4	-0.10
GR6J_B_DSST_N_2100Hd	0.80	0.85	0.75	$+4.2 \cdot 10^{-3}$	0.82	0.83	0.87	+3.3	0.02
GR6J_B_DSST_E_2100Ld	0.81	0.89	0.74	$-6.7 \cdot 10^{-1}$	0.87	0.88	0.92	+3.7	0.06
GR6J_B_DSST_E_2100Ln	0.80	0.83	0.78	$+6.1 \cdot 10^{-1}$	0.87	0.91	0.89	-3.3	0.07

For all SSTs, the multi-objective function y is lower in the validation than in the calibration period. The decrease in y is comparable for simulation of all types of flows. The decrease in y is mainly attributable due to an increase in RVE, but also due to a slight decrease in NS_w and/or NS_{inv}. The RVE increases from 0% in the calibration period to on average +/-5.6% in the validation period.

In 7 out of 14 DSSTs, the multi-objective function y is lower in the validation period than in the calibration period. In the 7 other DSSTs, y is higher in the validation period. The RVE increases in the validation period in all DSSTs. An increase in y means that the increase in NS_w and/or NS_{inv} efficiency prevails over the increase in RVE. NS_w increases in *'GR6J_H_DSST_N_2100Hn'*, *'GR6J_H_DSST_E_2100Hn'* and *'GR6J_B_DSST_E_2100Ln'*. NS_{inv} increases in the DSSTs on extreme years and scenario *'N_2100Hd'*. In all other cases, NS_{inv} decreases. *'GR6J_L_DSST_N_2100Hd'* and *'GR6J_L_DSST_E_2100Ld'*, which have the same validation and calibration period, show comparable results in both periods.

4.3.3. Discussion of the results RQ3

Performance of HBV and GR6J in the SSTs

For simulation of high flows, the HBV model showed a y_w of 0.83 in the calibration period of SST *'H_SST_I'* and a y_w of 0.80 in the validation period. The GR6J model showed a y_w of 0.89 and 0.83 in the calibration and validation period, respectively. This shows that the GR6J model performs better when simulating high flows. This is also visible in the hydrographs in Figure 33 below. These show the observed and simulated discharges for the HBV and GR6J model on a part of the calibration and validation period of *'H_SST_I'*. In both the calibration and validation period, it can be observed that models underestimate the annual maximum discharge. However, the GR6J model more closely simulates the peaks. As a result, the NS_w is higher, leading to a higher y_w. Next to that, the GR6J model better captures small peaks in summer, such as the peaks in July 2002. The HBV model is not able to capture these peaks well and shows a constantly decreasing discharge in summer.

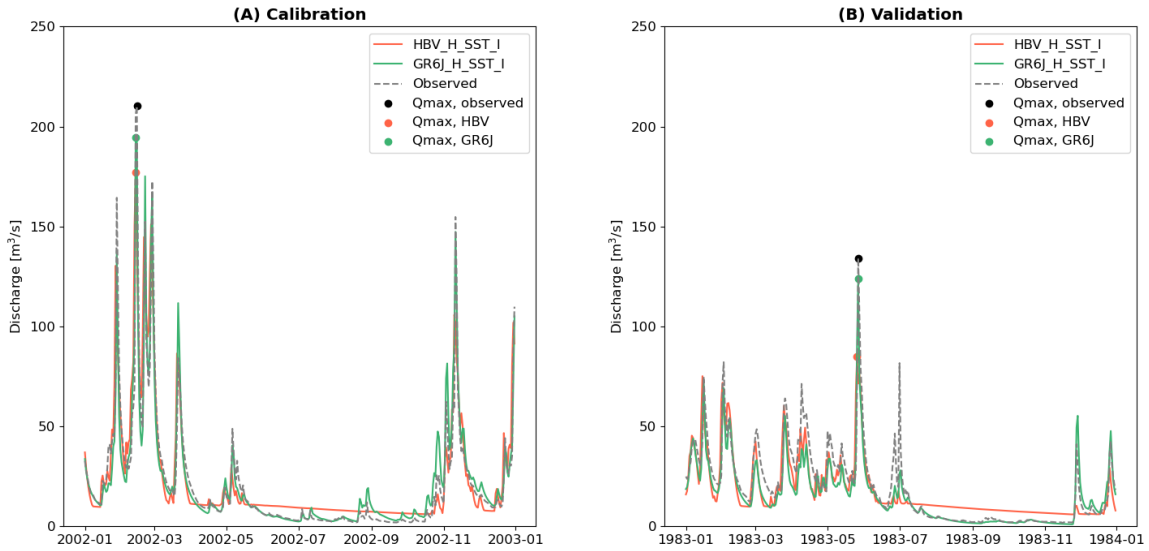

Figure 33 - Observed and simulated discharge in a part of A) the calibration period (2002) and B) validation period (1983). The green line shows the simulated discharge with the GR6J model, the red line shows the simulated discharge for the HBV model. Models are calibrated on the calibration period in testing scheme 'SST_I' with objective function y_w. The observed discharge is shown with the grey dotted line. The black, red and green dot show the annual maximum daily discharge which was observed, simulated by HBV and simulated by GR6J, respectively.

For simulation of low flows, the HBV model showed a y_{inv} of 0.86 in the calibration period of SST *'L_SST_I'* and a y_{inv} of 0.80 in the validation period. The GR6J model showed a y_{inv} of 0.83 and 0.77 in the calibration and validation period. This shows that the HBV model performs better when simulating low flows. This is also visible in the hydrographs in Figure 34 below. These show the observed and simulated discharges for the HBV and GR6J model on a part of the calibration and validation period of *'L_SST_I'*. In general, the HBV model simulates more closely to the observations and thus has a higher NS_{inv} and y_{inv}. Both models face difficulties with simulating the small peaks in summer: The GR6J model overestimates them, while the HBV model slightly underestimates them. Both models seem to underestimate the observed discharge just after a peak (April 2002), which shows that both models seem to have a too rapid decline in the recession part of the hydrograph.

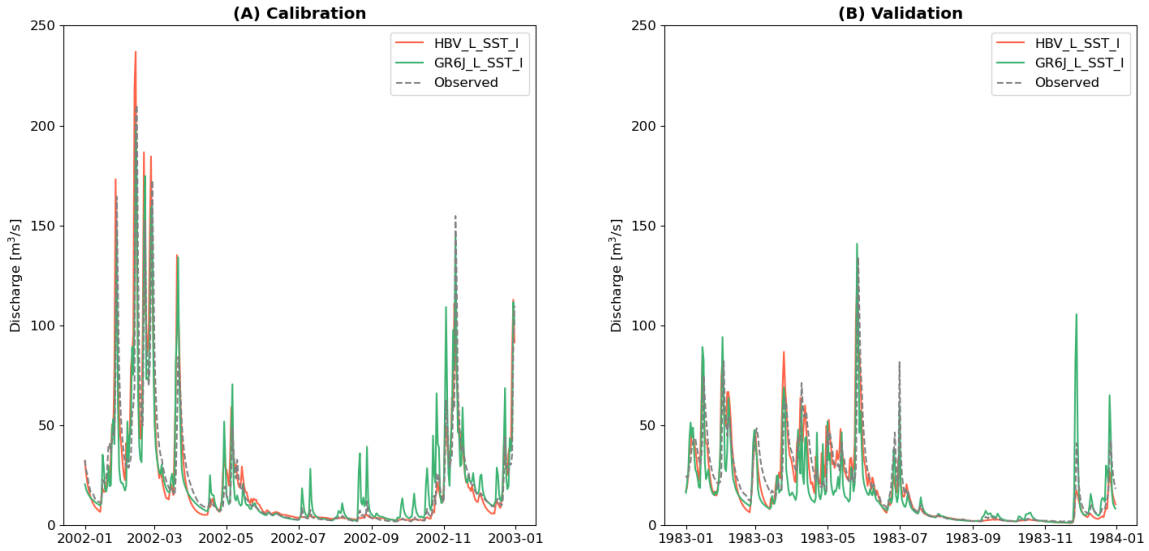

Figure 34 - Similar to Figure 33, but by calibration with objective function y_{inv} for low flows.

For simulating both high and low flows, in *'B_SST_I'*, the HBV model shows a y_{comb} of 0.84 and 0.82 in the calibration and validation period, while these values are 0.81 and 0.76 for the GR6J model. This

shows that the HBV model is better at simulating both high and low flows. This is mainly due to the better performance on low flows, which is represented by a higher NS_{inv}. The values of NS_w are comparable. Figure 35 confirms the better performance of the HBV model for simulating high and low flows. It is visible that the GR6J model overestimates peaks in summer and autumn, while the HBV model simulates these smaller peaks closer to observations.

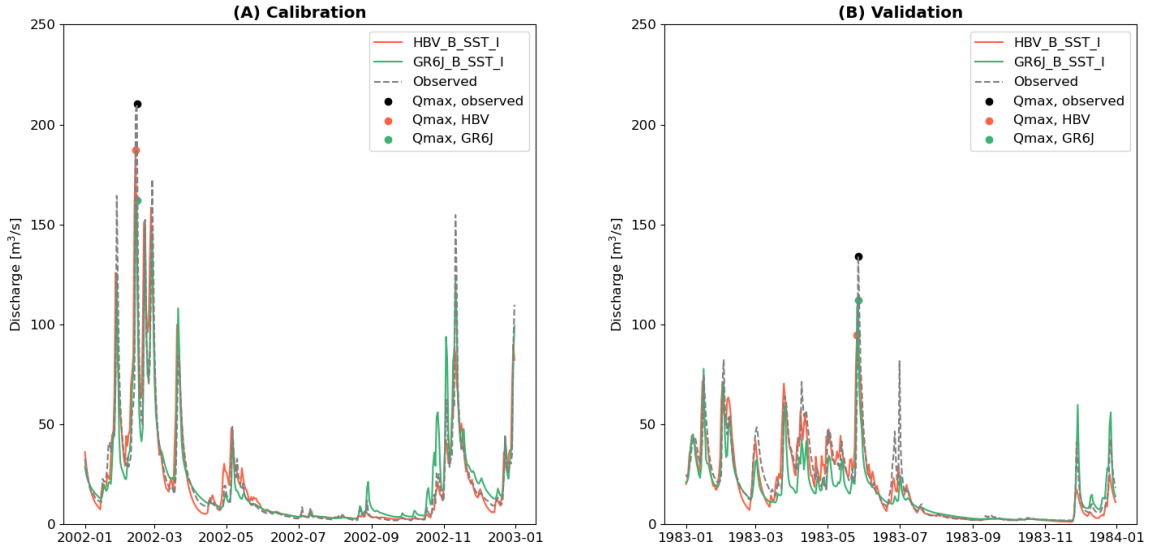

Figure 35 - Similar to Figure 33, but by calibration with objective function y_{comb} for both high and low flows.

For both models and all types of flow, the RVE is close to 0% in the calibration period. This implies that it is important that the water balance over the calibration period corresponds with observations for an optimal value of multi-objective function y. For HBV, the average RVE in the validation period of the SSTs is +/-2.7%, while this average RVE is +/-5.6% for the GR6J model. The increase in RVE is the main reason for a lower y in the validation period. This means that the model performance of both models decrease in the validation period, mainly due to a worse simulation of the water balance. Still, the model performs acceptable in the validation period, as the RVE in the validation period of the SSTs was 5.9% at maximum.

Performance of HBV and GR6J in the DSSTs for simulating high flows

In the DSSTs for simulation of high flows, both models show a large decrease in y_w in tests *'H_DSST_N_2100Ld'* and *'H_DSST_E_2100Ld'*. For both models, this is due to an increase in RVE and decrease in NS_w. Conversely, in DSSTs *'H_DSST_N_2100Hn'* and *'H_DSST_E_2100Hn'*, NS_w increases in the validation period. This increase in NS_w prevails over the increase in RVE, leading to a higher or similar y_w in the validation period.

As models were calibrated with multi-objective function y_w, including NS_w, the focus in the calibration was on good agreement on peak flows. In NS_w, a mismatch between high observed and simulated discharges is penalized proportionally to the observed discharge value. The mismatch between simulated and observed discharge in peak periods is thus weighted more than non-peak periods. In years with high peaks, the difference between weights in peak periods and non-peak periods is even more pronounced. This means that the mismatch between simulated and observed discharges in the rest of the year is weighted relatively less. As peaks are simulated rather well, years with a higher peaks show generally a higher NS_w.

This may be the reason for the fact that NS_w is very high in the validation period of DSST *'E_2100Hn'*. The NS_w is 0.86 for HBV and even 0.97 for GR6J. The validation period of this test consists of only one

hydrological year (1995), which includes a high peak which is simulated close to observations. The appointed reasoning may also explain the small changes in NS_w in *'N_2100Ld'* and *'N_2100Hn'*. The validation periods of these tests include years with peaks that are slightly lower and higher than in normal years. As a result, NS_w slightly decreases in *'N_2100Ld'* and slightly increases in *'N_2100Hn'*. However, the appointed reasoning does not explain the decrease in NS_w in *'E_2100Ld'*. Here, a large decrease in NS_w is observed, while the validation period includes two years with high observed peaks. The reason for this may be related to the overestimation of the maximum discharge in hydrological year 1980. The observed annual maximum discharge is 203 m^3/s, whereas the simulated discharge is 275 and 396 m^3/s for the HBV and GR6J model, respectively. This large overestimation of the peak discharge, which is not observed in any other year, may indicate a random error in the observed discharge timeseries or precipitation input.

The HBV and GR6J model show the same patterns in NS_w in the various DSSTs. However, the GR6J model is more pronounced in its changes. For example, whereas the HBV model has a NS_w of 0.82 and 0.75 in the calibration and validation period of DSST *'E_2100Ld'*, these values are 0.92 and 0.69 for the GR6J model. Regarding the RVE, both models show values in the same order of magnitude. The relative volume increases in the validation period, but within acceptable limits.

Performance of HBV and GR6J in the DSSTs for simulating low flows

For the simulation of low flows, again both models show comparable patterns regarding the change in multi-objective function y_{inv} in the DSSTs. In the first DSST, *'N_2100Ln'*, a large decrease in y_{inv} is observed in the validation period (-0.13 for HBV and-0.10 for GR6J). This is partly due to a small decrease in NS_{inv} and mainly due to a large increase in RVE, being-15% and-12% in the validation period for HBV and GR6J, respectively. In the other DSSTs, NS_{inv} increases in the validation period and the increase in RVE is relatively small (<6%).

The large decrease in RVE in DSST *'N_2100Ln'* suggests that both models have difficulties with simulating the water balance of the validation period of this DSST. The reason for this is that both models underestimate the observed small peaks in the summer period, leading to a large negative RVE. However, given that this validation period consists of only one year, such an observation may be coincidental. With a longer validation period, the relative volume error may have been closer to 0, as such a period contains periods in which discharge is overestimated and periods in which observed discharge is underestimated.

For all other DSSTs, the NS_{inv} increases substantially. The HBV model even shows an increase to 0.92 in the validation period of DSST *'E_2100Ln'*. The reason for this increase in NS_{inv} could be that the validation periods of these tests consists of 1976 and 2020 (*'N_2100Hd'* and *'E_2100Ld'*) or 2020 only (*'E_2100Ln'*). These years contain both long dry summers with nearly any summer peaks. These peaks are difficult to simulate, as these precipitation events have an intensity higher than the infiltration capacity of the soil and thus lead to infiltration excess overland flow. However, in the model, this precipitation is stored in the soil rather than discharged directly to the river. As these peaks have a low observed discharge value are weighted relatively much in NS_{inv}. As the validation periods of DSSTs *'N_2100Hd', 'E_2100Ln'* and *'E_2100Ld'* hardly contain any summer peaks, there is a smaller mismatch between observed and simulated inverse transformed discharges, leading to a higher value for NS_{inv}. For the GR6J model, this increase in NS_{inv} prevails over the increase in RVE, leading to an increase in y_{inv} in the validation period. For the HBV model, the increase in RVE is slightly higher, therefore the y_{inv} value does not increase in the validation period, except in *'E_2100Ln'*.

Performance of HBV and GR6J in the DSSTs for simulating high and low flows

For simulation of high and low flows, the results of the DSSTs show more differences between HBV and GR6J model. The HBV model shows smaller differences in y_{comb} between calibration and validation period compared to the GR6J model. The differences for HBV are between +0.01 and-0.06. For the GR6J model, these differences are larger, namely between +0.07 and-0.13. Both models show changes in RVE in the same order of magnitude for all DSSTs. The difference between both models can thus be attributed by the differences in NS_w and/or NS_{inv}.

The results for HBV show that, when calibrated on y_{comb}, the NS_{inv} is higher than NS_w in the calibration period of all DSSTs. The opposite is observed for GR6J. Here, NS_w is higher than NS_{inv} for all DSSTs. This relates to the finding that GR6J performs better at high flows and HBV at low flows.

Both models show decreases in NS_{inv} in the first three DSSTs, while increases are found for the last three DSSTs. This may be related to the relatively large proportion of years in the validation period that were selected to resemble future high flow conditions (Appendix C). This led to a validation period in which high flows dominate and as a result the increase in NS_{inv} in the validation period, which was observed for the simulation of low flows, does not apply here.

4.4. RQ4: Evaluation of parameter sets for simulation under different climatic conditions

This section shows and discusses the results for RQ4, for the HBV model (section 4.4.1) and the GR6J model (section 4.4.2). After that, in section 4.4.3, the results are discussed in general.

4.4.1. Results HBV model

The parameter sets of the HBV model calibrated on the periods in the testing schemes of high flows (Table 10A), low flows (Table 10B) and high and low flows (Table 10C) are shown below. The blue rows show the parameter sets when calibrated on the calibration period of RQ3 (blue in section 3.3.1), the orange rows show the parameter set when calibrated on the validation period of RQ3 (orange in section 3.3.1), which are years that were selected to resemble future conditions. The corresponding objective function values in the calibration periods can be found in Appendix G.

*Table 10 - Parameter sets of the **HBV** model when calibrated on the calibration and validation periods of the testing schemes for A) simulation of high flows, B) low flows and C) high and low flows. The names of the testing schemes are explained in section 3.3.1. Colours in the table refer to the colour of the periods in Figure 14, Figure 15 and Figure 16. Parameters for which no value is given, were fixed at α=1, K_f=0.005 day^{-1}, K_s=0.005 day^{-1}, $PERC$=1 mm/day and $Cflux$=1 mm/day.*

A) Simulation of high flows								
Testing scheme	Period	FC [mm]	β [-]	LP [-]	α [-]	K_f [day^{-1}]	K_s [day^{-1}]	PERC [mm/day]
HBV_H_DSST_N_2100Ld	Not-selected	113	5.97	0.351	0.403	0.040	-	-
	Selected	220	5.83	0.433	0.101	0.112	-	-
HBV_H_DSST_N_2100Hn	Not-selected	111	5.99	0.247	0.412	0.039	-	-
	Selected	266	5.20	0.700	0.320	0.056	-	-
HBV_H_DSST_E_2100Ld	Not-selected	110	5.99	0.214	0.454	0.032	-	-
	Selected	207	3.33	0.578	0.103	0.122	-	-
HBV_H_DSST_E_2100Hn	Not-selected	112	5.98	0.250	0.443	0.033	-	-
	Selected	306	4.68	0.570	0.146	0.116	-	-

Table continues on next page

B) Simulation of low flows								
Testing scheme	Period	FC [mm]	β [-]	LP [-]	α [-]	K_f [day^{-1}]	K_s [day^{-1}]	PERC [mm/day]
HBV_L_DSST_N_2100Ln	Not-selected	223	1.74	0.470	-	-	0.029	0.310
	Selected	157	1.11	0.470	-	-	0.042	0.380
HBV_L_DSST_N_2100Hd	Not-selected	191	1.46	0.353	-	-	0.028	0.348
	Selected	314	2.78	0.793	-	-	0.034	0.231
HBV_L_DSST_E_2100Ln	Not-selected	208	1.71	0.434	-	-	0.029	0.338
	Selected	191	0.89	0.103	-	-	0.031	0.432
HBV_L_DSST_E_2100Ld	Not-selected	193	1.35	0.344	-	-	0.027	0.323
	Selected	322	2.89	0.809	-	-	0.035	0.235
C) Simulation of high and low flows								
Testing scheme	Period	FC [mm]	β [-]	LP [-]	α [-]	K_f [day^{-1}]	K_s [day^{-1}]	PERC [mm/day]
HBV_B_DSST_N_2100Ld	Not-selected	188	1.81	0.431	0.630	0.015	0.024	0.336
	Selected	242	2.40	0.488	0.450	0.027	0.024	0.357
HBV_B_DSST_N_2100Hn	Not-selected	202	1.96	0.452	0.618	0.016	0.023	0.338
	Selected	189	1.41	0.318	0.638	0.013	0.026	0.323
HBV_B_DSST_N_2100Ln	Not-selected	185	1.81	0.434	0.622	0.015	0.024	0.333
	Selected	227	1.82	0.434	0.637	0.012	0.019	0.278
HBV_B_DSST_N_2100Hd	Not-selected	172	1.84	0.404	0.598	0.017	0.022	0.362
	Selected	290	1.58	0.408	0.654	0.013	0.037	0.295
HBV_B_DSST_E_2100Ld	Not-selected	184	1.49	0.346	0.730	0.009	0.023	0.315
	Selected	299	2.47	0.661	0.388	0.036	0.032	0.291
HBV_B_DSST_E_2100Ln	Not-selected	209	1.90	0.451	0.574	0.017	0.025	0.338
	Selected	105	2.40	0.107	0.264	0.092	0.013	0.550

Simulation of high flows

For simulation of high flows, the parameter sets obtained when calibrated on the periods that do not resemble future conditions (blue) are quite comparable between the DSSTs. Especially the differences between *'E_2100Ld'* and *'E_2100Hn'* are very small. This can be explained by the fact that the blue calibration periods of the DSSTs are for a large part the same (Figure 14).

Remarkable is the high value of β, which is close to 6 in each DSST. This means that β approaches the upper limit of its parameter space. Such a high value for β was not observed when calibrating the model on the calibration period of the SST (Table E1 in Appendix E, with $\beta \approx 3.7$). The reason for this may be the fact that hydrological year 2021 was included in the calibration period of all DSSTs, whereas this year was not included in the calibration period of the SSTs. Hydrological year 2021 showed an atypical extreme high flow event which took place in summer, which may have had large effects on the parameter set. In order to simulate this peak well, β is ideally very high and FC very small. A high value for β means that recharge R from soil moisture box to upper response box is smaller for a given soil moisture deficit. As a result, the soil moisture box is not emptied fast. In combination with a small FC, which means a low capacity of the soil moisture box, this leads to a large direct runoff component. It is expected that this direct runoff component was needed to simulate the 2021 peak most closely.

Calibrating the models on periods resembling future conditions (orange) leads to different parameter sets compared to calibration on periods that do not resemble future conditions (blue). This shows that optimized parameter values tend to be specific for the period and thus model forcing-time series they are calibrated on. Parameters calibrated on historic periods may thus be suboptimal for future conditions. For simulation of high flows, soil parameters FC and LP increase and β decreases. Comparing the values of FC and β between the different periods is difficult, because of the large influence of 2021 on these parameters. LP is higher when calibrated on periods that resemble future

conditions. This means that, given that FC also increases, the limit for potential evapotranspiration increases and thus the simulated actual evapotranspiration decreases. This leads to an increase in quick runoff response, which is reasonable, as the calibration data contains more peak flows.

Quick runoff parameter K_f increases when calibrated on future conditions. This means that quick runoff for a certain volume in the upper response box increases. The increase in K_f comes along with a decrease α, so the relation between quick runoff and upper response box volume becomes more linear.

Simulation of low flows

For simulation of low flows, the parameter sets obtained when calibrated on the periods that do not resemble future conditions (blue) are again quite comparable between the DSSTs. For example, parameter K_s deviates between 0.027 day^{-1} and 0.029 day^{-1}. Just as for high flows, this can be explained by the fact that the blue calibration periods of the DSSTs are for a large part the same (Figure 15).

The parameter sets obtained by calibration with objective function y_{inv} are substantially different from the parameter sets obtained by calibration with objective function y_w, even though the calibration periods for a large part overlap. This shows that parameter values optimized with objective function y_w may be suboptimal for simulating low flows. FC is higher for simulation of low flows than for simulation of high flows. This means the capacity of the soil moisture storage is higher, which means there is more possibility for buffering in case of large precipitation. This thus leads to less steep peaks, which is reasonable for a model that is not focused on simulating peaks due to objective function y_{inv}. A lower value for β means that for a given soil moisture deficit, the recharge is larger. This means that the upper response box is filled at a more constant pace, which again represents a slower responding system.

Calibration on periods that resemble future conditions (orange) leads to different parameter sets than calibration on periods that do not resemble future conditions (blue). This again shows that parameters are specific for the period on which they are calibrated. For DSST *'N_2100Ln'* and *'E_2100Ln'*, a decrease in FC, β and LP and an increase in $PERC$ is observed. For DSST *'N_2100Hd'* and *'E_2100Ld'*, the exact opposite is observed: An increase in FC, β and LP and a decrease in $PERC$. For all DSSTs, K_s increases.

In DSST *'N_2100Ln'* and *'E_2100Ln'*, $PERC$ thus increases, leading to a higher flux of percolation to the lower response box. Together with the increase in K_s this thus leads to more baseflow. The decrease in FC means that there is less possibility of buffering in case of large precipitation. The decrease in β means that there is a larger recharge for a given soil moisture deficit. The decrease in FC and β thus both lead to higher peaks after a precipitation event at the time there is a soil moisture deficit. The influence of the parameters is visible in Figure 36, which shows the simulated discharge in the summer period of 2020, which is the validation year of DSST *'E_2100Ln'*. It is visible that the model calibrated on future conditions (orange) shows a higher baseflow and captures peaks in summer slightly better.

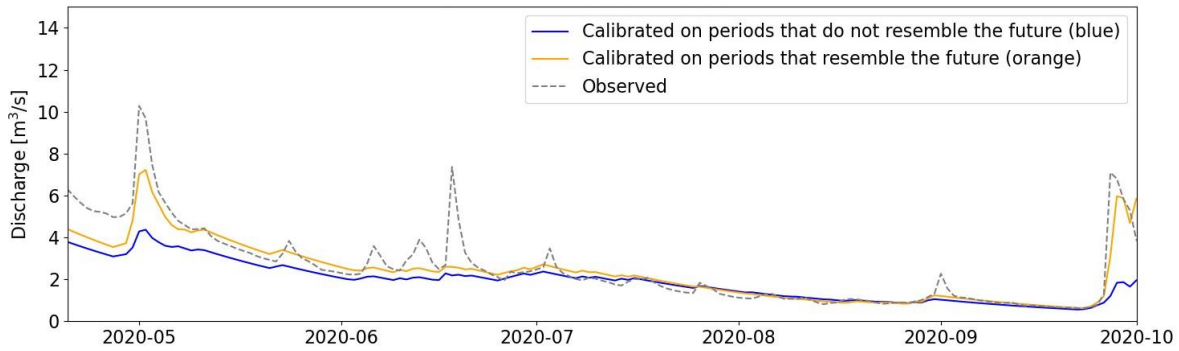

Figure 36 – Observed (grey) and simulated discharge in a part of the summer of 2020. Calibration is on the blue and orange periods of 'HBV_L_DSST_E_2100Ln'.

In DSST *'N_2100Hd'* and *'E_2100Ld*, $PERC$ decreases, leading to a lower flux of percolation to the lower response box. Despite the increase in K_s, this leads to less baseflow. The increase in FC and β thus both lead to lower peaks after a precipitation event at the time there is a soil moisture deficit. The influence of the parameters is visible in Figure 37, which shows the simulated discharge in the summer period of 1976, which is the validation year of DSST *'E_2100Ld'*. It is visible that the model calibrated on periods that resemble future conditions shows a lower baseflow.

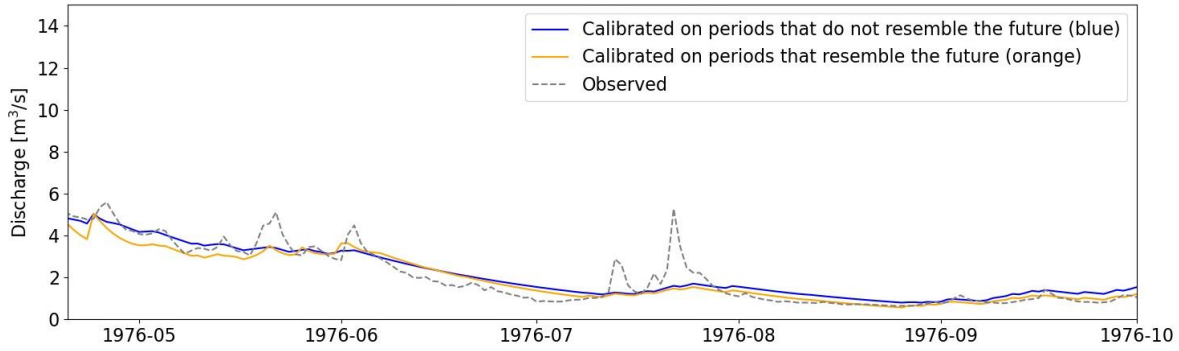

Figure 37 - Similar to Figure 36, but for the summer period of 1976 and scenario 'HBV_L_DSST_E_2100Ld'.

The parameter sets for DSST *'N_2100Hd'* and *'E_2100Ld'*, which have the same validation and calibration period, are comparable. This shows that the SCEM-UA algorithm was able to find a global optimum.

Simulation of high and low flows

For simulation of high and low flows, calibration on periods that do not resemble future conditions (blue) of the different DSSTs again lead to comparable parameter sets. Figure 16 shows that the blue calibration periods of the DSSTs largely overlap, which explains these similarities.

The parameter values deviate slightly from the parameter values obtained for models focused on simulation of high flows or low flows only. For simulation of high flows, parameters K_s and $PERC$ are fixed at default values and for simulation of low flows, parameters α and K_f are fixed at default values. These default values may not be the optimal values. For simulation of high and low flows, these parameters were included in the calibration and were thus not given the default value. As a result, due to dependencies between parameters, other parameters may be slightly different as well. The parameter values obtained are most comparable with the parameter values obtained for the simulation of low flows. This matches with the observation made when evaluating the results of RQ3, namely that the HBV shows higher values for NS_{inv} (low flows) than for NS_w (high flows) when calibrated on y_{comb}.

Similarly to simulation of high flows or low flows only, the parameter sets obtained when calibrated on periods that resemble future conditions (orange) are different than when calibrated on periods that do not resemble future conditions (blue). This indicates a lack of robustness. Compared to simulation of high flows or low flows only, it is more difficult to distinguish patterns in the parameter sets obtained by calibration on periods that resemble future conditions (orange). This is because the changes in parameter set may be related to the simulation of both high and low flows. For example, the parameter set obtained by calibration on *'E_2100Ln'* shows a large increase in K_f and decrease in K_s, whereas calibration on *'N_2100Hd'* shows a decrease in K_f and increase in K_s.

4.4.2. Results GR6J model

Table 11 shows the parameters sets of the GR6J model when calibrated on periods that do not resemble future conditions (blue rows) and periods that resemble future conditions (orange rows). The corresponding objective function values in the calibration periods can be found in Appendix G.

*Table 11 - Similar to Table 10, but for the **GR6J** model. For simulation of high flows, parameter X_6 was fixed at 10 mm/day.*

A) Simulation of high flows							
Testing scheme	Period	X_1 [mm]	X_2 [mm/day]	X_3 [mm]	X_4 [day]	X_5 [-]	X_6 [mm]
GR6J_H_DSST_N_2100Ld	Not-selected	102	-0.681	31	2.22	+0.34	-
	Selected	180	+0.915	35	2.12	+0.72	-
GR6J_H_DSST_N_2100Hn	Not-selected	114	+0.384	31	2.22	+0.95	-
	Selected	113	-0.196	47	2.09	-0.33	-
GR6J_H_DSST_E_2100Ld	Not-selected	98	+0.585	40	2.13	+0.79	-
	Selected	302	-0.225	19	2.38	-0.14	-
GR6J_H_DSST_E_2100Hn	Not-selected	144	+0.649	30	2.26	+0.78	-
	Selected	166	-0.834	26	2.23	+0.41	-
B) Simulation of low flows							
Testing scheme	Period	X_1 [mm]	X_2 [mm/day]	X_3 [mm]	X_4 [day]	X_5 [-]	X_6 [mm]
GR6J_L_DSST_N_2100Ln	Not-selected	216	-0.293	3.46	3.00	+0.22	10.84
	Selected	180	-0.042	2.06	4.15	-1.73	19.88
GR6J_L_DSST_N_2100Hd	Not-selected	73	-0.248	300	1.94	-0.30	0.19
	Selected	12	+0.152	336	2.02	+1.88	0.46
GR6J_L_DSST_E_2100Ln	Not-selected	83	-0.500	274	2.00	+0.05	0.13
	Selected	46	+0.116	297	1.40	+1.79	0.43
GR6J_L_DSST_E_2100Ld	Not-selected	80	-0.249	302	2.00	-0.29	0.16
	Selected	44	-0.729	263	2.08	+0.10	4.33
C) Simulation of high and low flows							
Testing scheme	Period	X_1 [mm]	X_2 [mm/day]	X_3 [mm]	X_4 [day]	X_5 [-]	X_6 [mm]
GR6J_B_DSST_N_2100Ld	Not-selected	158	-0.367	20	2.28	+0.20	9.22
	Selected	135	+0.101	284	2.42	+2.06	0.29
GR6J_B_DSST_N_2100Hn	Not-selected	165	-0.383	21	2.28	+0.20	9.22
	Selected	134	-0.322	20	2.29	+0.11	12.02
GR6J_B_DSST_N_2100Ln	Not-selected	148	-0.335	21	2.28	+0.16	10.16
	Selected	148	-0.534	21	2.37	+0.25	11.10
GR6J_B_DSST_N_2100Hd	Not-selected	93	-0.535	245	2.31	+0.10	0.11
	Selected	185	-0.440	23	2.26	+0.22	7.81
GR6J_B_DSST_E_2100Ld	Not-selected	168	-0.366	19	2.35	+0.19	9.66
	Selected	94	-0.349	54	2.16	-0.04	17.04
GR6J_B_DSST_E_2100Ln	Not-selected	69	-0.181	304	2.32	-0.55	0.34
	Selected	30	-0.214	250	1.92	-0.42	3.93

Simulation of high flows

The parameter sets obtained by calibration on periods that do not resemble future conditions (blue) show variations between the DSSTs, despite the fact that the blue calibration periods of the DSSTs are for a large part the same. This is especially apparent for parameters X_2 and X_5, which are related to the groundwater exchange. Both magnitude and signs of these parameters differ between the DSSTs. This finding matches with the observed variation in the parameter sets calibrated on *'SST_I'* in Appendix E.

For *'N_2100Ld'*, a negative X_2 is observed, which means that groundwater is extracted as long as $X_5 < R_1/X_3$ (Equation 29). X_5 has a relatively small value (0.34), which means that $X_5 < R_1/X_3$ for a large part of the time (99% of the calibration period). The other DSSTs show a positive X_2, which means groundwater is extracted as long as $X_5 > R_1/X_3$. As the values of X_5 in these DSSTs are large (>0.78), groundwater is extracted for most of the days as well (>0.96% of the calibration period). Despite the sign differences in X_2 and X_5, all DSSTs thus show a negative groundwater exchange flux. In the calibration period of each DSST, an average of 0.4 mm/day is extracted, which makes up 15% of the total

water balance. However, the groundwater exchange over time (Figure 38) shows that there are differences in the patterns of the groundwater exchange flux. For DSST *'N_2100Ld'*, groundwater exchange is larger (more negative) when R_1 is larger and closer to zero when R_1 is smaller. For the other DSSTs, the opposite occurs.

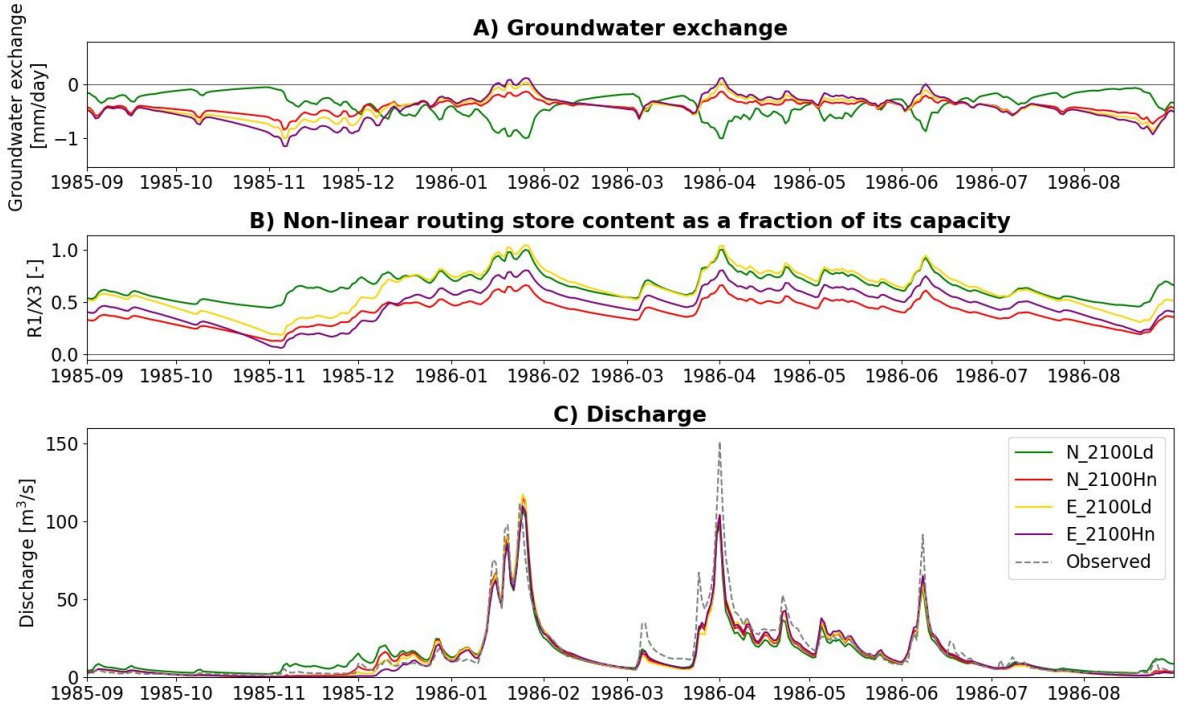

Figure 38 - A) Groundwater exchange over time for a part of the calibration period (hydrological year 1985), modelled with GR6J calibrated on the periods that do not resemble future conditions of the different DSSTs. B) Non-linear routing store content (R_1) as a fraction of its capacity (X_3) over time, for the same period and models. C) Discharge for the same period and models. Grey dashed line represents the observed discharge.

Thus, the parameter sets obtained when calibrated over periods that do not resemble future conditions (blue) are not comparable between the DSSTs, whereas the parameter sets are comparable for the HBV model. Despite the differences in parameter sets, all models simulate more or less the same discharge (Figure 38C). This indicates that there may be parameter equifinality, which means that multiple sets of parameter values can produce comparable simulations. Next to that, the parameters are interrelated. A higher value for X_1 comes with a lower value for X_3 and vice versa.

The parameter equifinality and interrelations between parameters make it challenging to attribute changes in parameter sets resulting from different calibration periods to underlying factors. As there are multiple parameter sets that can produce an equally good fit, it is difficult to determine whether changes in parameter values are due to differences in calibration period or due to parameter equifinality or a combination of both.

In general, parameter sets show a low value of X_1 and X_3. As these parameters are the capacity of the production store and non-linear routing store, low values indicate a fast responding system. These low values of X_1 and X_3 may be needed to simulate peaks well. Parameter X_4, the time-base of the unit hydrographs, has a relatively low value as well, which again matches with a fast-responding catchment.

Simulation of low flows

For simulation of low flows, again large variations in the parameter sets are observed and clear patterns are hard to distinguish. The variations in parameter sets obtained when calibrated on periods that do not resemble future conditions (blue) are even larger than for the simulation of high flows. This may be

explained by the larger number of parameters that is optimized. For the simulation of high flows, parameter X_6 was fixed at 10mm, while this was not the case for the simulation of low flows. The inclusion of X_6 may have impacted the other parameters in the routing module, which was observed by Pushpalatha et al. (2011) as well. This makes it hard to compare the parameter sets obtained when calibrated on y_w and y_{inv}.

X_6 is the depletion coefficient of the exponential routing store. A lower value of X_6 suggests a slower depletion, which means that the discharge from the non-linear routing store Q_{R2} is lower (Equation 38). A lower value for X_6 thus indicates a slower responding system. In all DSSTs except *'N_2100Ln'*, X_6 is close to 0 and X_3 is high compared to the values obtained for simulation of high flows. This means a slower reacting system. However, in these DSSTs also a very small value for X_1 is found, which indicates a fast reacting system. There is thus a relationship between parameters X_1, X_3 and X_6. For *'N_2100Ln'* for example, X_3 is smaller and X_6 is larger, indicating a faster responding system, but this may be compensated with a larger X_1.

Calibration on periods that resemble future conditions (orange) lead to different parameter sets than calibration on periods that do not resemble future conditions (blue). Also the parameter sets for DSSTs *'N_2100Hd'* and *'E_2100Ld'*, which have the same validation and calibration period, differ. This shows that multiple parameter sets can produce comparable simulations. Even though the SCEM-UA algorithm was able to find convergence (Appendix D), it may not have been able to find a global optimum. Just as for high flows, it is thus difficult to say whether these changes in parameters are due to the parameter equifinality or due to a different calibration period or both.

Simulation of high and low flows

For simulation of high and low flows, the parameter sets obtained by calibration on periods that do not resemble future conditions (blue) show more similarities compared to simulation of high flows or low flows only. For example, the parameter sets for *'N_2100Ld'* and *'N_2100Hn'* are comparable. The reason for this may be that models calibrated on y_{comb} should be able to capture both high and low flows well, and thus only a few parameter sets can produce these good simulations. Parameter equifinality may thus be less of a problem.

The parameter values for simulation of high and low flows have values mostly comparable with the parameter values obtained for the simulation of high flows. This matches with the observation made when evaluating the results of RQ3, namely that the GR6J model tends to focus on the calibration on high flows when calibrated on y_{comb}.

Parameter sets obtained when calibrated on periods that resemble future conditions (orange) are slightly different than when calibrated on periods that do not resemble future conditions (blue). This indicates a lack of robustness. Just as for the HBV model, it is difficult to distinguish patterns in the parameter sets obtained by calibration on periods that resemble future conditions (orange), as the changes in parameter set may be related to the simulation of both high and low flows. An analysis of the values of X_2 and X_5 showed that models calibrated on periods that resemble future conditions lead to a larger total loss of groundwater and comparable evapotranspiration, thus to a lower total discharge. This may be needed to simulate low flows well, while keeping a fast response to simulate high flows.

4.4.3. Discussion of the results RQ4

For both the HBV and GR6J model, parameter values proved to be different when calibrated on periods resembling future conditions. This indicates that parameters calibrated on historic periods may be suboptimal for future conditions, which is a lack of robustness. For the HBV model, changes in

parameter sets are for a large part explainable. For the GR6J model, due to aspects related to equifinality of calibrated parameter sets and interrelationships between parameters, this is more difficult.

4.5. RQ5: Projected impact of climate change

High flows

Figure 39 shows the relative change in annual maximum daily discharge in 2086-2115 compared to 1991-2020 for the 8 scenarios described in Table 7. The projected absolute values for annual maximum daily discharge in each scenario can be found in Appendix H.

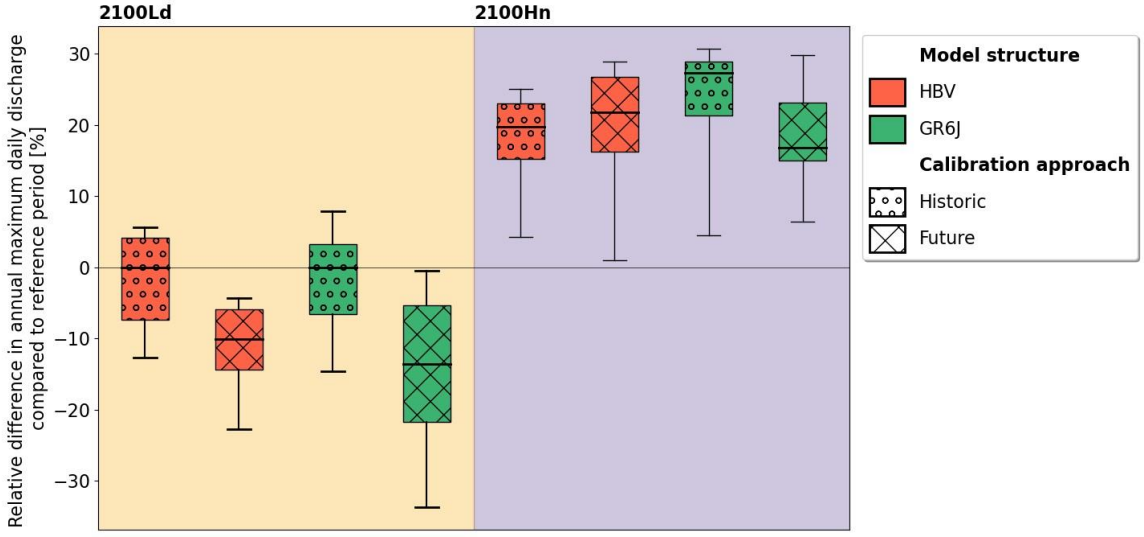

Figure 39 - Boxplot of the projected relative difference [%] in annual maximum daily discharge (Q_{max}) in the future (2086-2115) compared to the reference climate (1991-2020) for 8 scenarios (combinations of climate scenarios, hydrological model structures and calibration approaches (Table 7)). The boxplots in yellow and purple are based on climate scenarios '2100Ld' and '2100Hn', respectively. Red and green boxplots show the projected change in Q_{max} simulated with the HBV and GR6J, respectively. Dotted boxplots show the projected change in Q_{max} with models calibrated on historic periods, boxplots with diagonal stripes show the projected change in Q_{max} with models calibrated on periods resembling future conditions. The horizontal lines in the boxes represent the median value, the boxes represent the 25-75% data range and the whiskers represent the 5-95% data range.

The median change in annual maximum daily discharge compared to reference climate 1991-2020 fluctuates between-14% *('2100Ld, GR6J, Future'*) and +27% *('2100Hn, GR6J, Historic')*. The uncertainty in this projected impact is the largest due to different climate scenarios. Whereas an increase is projected with each model and calibration approach for climate scenario *'2100Hn'*, a decrease or comparable value as the reference climate for Q_{max} is projected for climate scenario *'2100Ld'*. Climate scenario *'2100Hn'* was based on high future emissions and a trend towards a wetter future climate. It therefore makes sense that this climate scenario leads to the highest annual maximum discharges.

Still, the uncertainty in the projected impact of climate change due to different hydrological model structures and different calibration approaches is substantial and can thus not be neglected. For example, within climate scenario *'2100Ld'*, the median change in Q_{max} deviates between-14% *('GR6J, Future')* and 0% *('HBV, Historic')* due to different hydrological model structures and calibration approaches. The uncertainty in projected impact due to different calibration approaches is slightly higher than the uncertainty in projected impact due to different hydrological model structures.

Figure 39 shows that calibration on periods resembling future conditions leads in most cases to a lower annual maximum daily discharge than when the model is calibrated on historic periods. This was unexpected, as models calibrated on historic periods showed to underestimate peaks. Models calibrated on periods resembling future conditions are expected to be trained more on peaks and to

simulate higher peaks. However, the atypical high flow event of 2021 was included in the period resembling historic conditions, which may have had large effects on the obtained parameter set, as was described in section 4.4.1.

The differences in projected impact of climate change due to different calibration approaches are more apparent for the GR6J model than for the HBV model. This shows that for simulation of Q_{max}, the GR6J model is more sensitive for the choice of the calibration approach than the HBV model. The quick runoff parameters of the GR6J model thus tend to be more specific for the calibration period.

Figure H1 in Appendix H shows that the variability of Q_{max} in the future is larger when simulated with the GR6J model than with the HBV model. The lower whiskers (5%) are comparable for both models, whereas the upper whiskers (95%) show more differences. This shows that the lower Q_{max} values in the future are quite comparable for both models, but that above-normal Q_{max} values are simulated differently by both models. The reason for this could be that the HBV model underestimates peak flows, whereas the GR6J model better captures high peaks. Especially in summer periods, the HBV model showed to have difficulties with simulating peaks, which is further elaborated in section 5.2.2. As a result, the projected above-normal Q_{max} values may be too low for the HBV model.

The variability of Q_{max} in the future is larger for climate scenario *'2100Hn'* than for climate scenario *'2100Ld'* (Figure H1 in Appendix H). This shows that for the climate scenario with high future emissions and a trend towards a wetter future climate, more variability in Q_{max} is expected.

Low flows

Figure 40 shows the relative change in annual minimum 7-day mean discharge in 2086-2115 compared to 1991-2020 for the 8 scenarios described in Table 7. The absolute values for the projected impact in each scenario can be found in Appendix H.

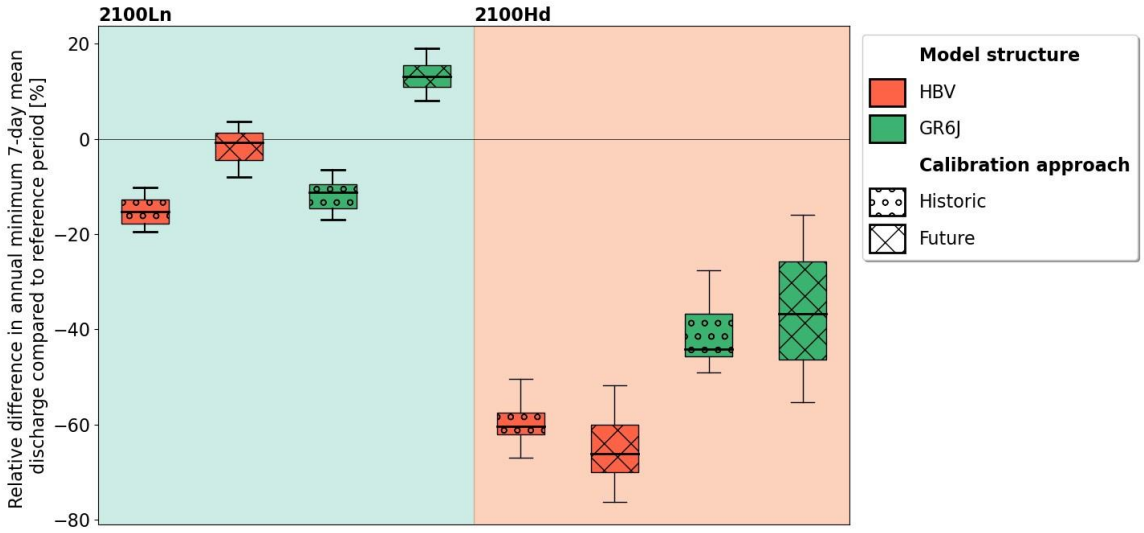

Figure 40 - Similar to Figure 39, but for the projected relative difference [%] in annual minimum 7-day mean discharge ($Q7_{min}$). Boxplots in green and orange are based on climate scenarios '2100Ln' and '2100Hd', respectively.

All scenarios project a decrease in $Q7_{min}$, except *'2100Ln, GR6J, Future'*, which projects a median increase of +13%. The largest decrease is projected for *'2100Hd, HBV, Future'*, with a projected median change of -66%. The changes in $Q7_{min}$ are thus expected to be larger than the changes in Q_{max}.

Just as for high flows, the uncertainty in projected impact is the largest due to different climate scenarios. Climate scenario *'2100Hd'* projects a larger decrease than climate scenario *'2100Ln'*. Climate scenario *'2100Hd'* was based on high future emissions and a trend towards a drier future climate. It

therefore makes sense that this climate scenario leads to the lowest annual minimum 7-day mean discharges. Especially for climate scenario *'2100Ln'*, the uncertainty in climate change impact due to calibration approaches cannot be underestimated.

For climate scenario *'2100Hd'*, large differences between projected impact with the HBV and GR6J model are observed. For this climate scenario, the uncertainty in projected impact is larger due to different hydrological model structures than due to different calibration approaches. The HBV model projects a decrease of approximately 70%, while this decrease is approximately 40% for the GR6J model. Figure H2 in Appendix H also shows that the GR6J model calibrated on periods that resemble future conditions predicts higher $Q7_{min}$ values. The GR6J model showed to be able to simulate $Q7_{min}$ values well in case of a prolonged drought. However, when these periods of drought are alternated with precipitation events in summer, the model tends to overestimate peaks and encounters difficulties with the subsequent recession after such a precipitation event. Low streamflow in summer periods in the future that include these precipitation events may thus be overestimated. As a result, the projected $Q7_{min}$ values may be too high for the GR6J model.

The variability of $Q7_{min}$ in the future is smaller for climate scenario *'2100Hd'* than for climate scenario *'2100Ln'* (Figure H2 in Appendix H). This shows that for the climate scenario with high future emissions and a trend towards a drier future climate, less variability in $Q7_{min}$ is expected. Years in the future will thus look more the same for climate scenario *'2100Hd'*.

5. DISCUSSION

This study evaluated the robustness of the HBV and GR6J model for simulating the impact of climate change on high and low streamflow in the Lesse. This chapter discusses the limitations in data and methods that may have had influence on the results of this study and puts the results in perspective in relation to existing literature. After that, some generalizations regarding the study area and hydrological models are made.

5.1. Limitations in data and methods

5.1.1. Limitations in data

In this study, historic discharge and climatic timeseries were used with a daily timestep. The Lesse has a relatively fast response and the results of RQ1 showed that a part of the precipitation is discharged at the same day. For simulation of high flows, it therefore would have been useful when timeseries with a sub-daily timestep were available. In this way, diurnal variability in precipitation could have been captured. This could have led to more insight in the meteorological conditions that lead to high streamflow and could have led to better simulations of high streamflow.

The historic discharge timeseries of this study has been used in multiple studies, indicating its quality and reliability. However, errors in the discharge timeseries may have had large effects on the model performance and parameters, especially for simulation of high flows (Van Den Tillaart et al., 2013). The same applies for errors in the climatic timeseries. Donat et al. (2014) showed that the gridded datasets that are used in this study lack reliability for precipitation extremes. For the simulation of high flows, errors in the discharge and climatic timeseries may thus have impacted the results of this study.

5.1.2. Limitations in methods

Meteorological indicators

This study focused on high and low streamflow with a return period of 1 year. Meteorological conditions that lead to this streamflow were based on historic data. This study assumed that the same meteorological conditions that led to high and low streamflow in the past will continue to do so in the future. However, this study did not consider other meteorological conditions that may lead to high and low streamflow in the future. This assumption has important implications for the results of this study. This is illustrated with the example of the high flow event that took place in 2021.

In 2021, an atypical high flow event took place that was caused by extreme precipitation. The event led to a lot of damage, amongst others in the Lesse (Task Force Fact-finding hoogwater 2021, 2021). The event, with a peak discharge of more than 500 m^3/s at Gendron, was atypical, because the event took place in summer. Whereas most historic peak discharges took place after a period of intense precipitation in winter with a saturated soil, this precipitation event was of such high intensity that infiltration was not possible (Task Force Fact-finding hoogwater 2021, 2021). This shows that other physical mechanisms played a role.

The high flow event of 2021 may have had large impacts on the results of this study. As the meteorological indicator for high flows was not high enough in 2021, this year was not selected for a period resembling future conditions. As a result, an extremely high flow event was included in the period that resembled historic conditions, whereas this high flow event is typically what is expected to happen more often in the future. Next to that, the model performance and parameters in RQ3 and RQ4 were substantially affected by high flow event 2021, as for example described in section 4.4.1.

This study mainly focused on the robustness of models for simulating the impact of climate change on high streamflow in winter that is caused by the same meteorological conditions as were observed in

history. However, year 2021 shows that high streamflow caused by different meteorological conditions was not covered by this study. This also implies that robust models should give physically realistic simulations. High and low streamflow in the future may be caused by different meteorological conditions than were used in this study. For projecting climate change impacts in the future, it is important that models can deal with other conditions as well.

Bias-correction

In this study, a relative bias-correction was applied to account for the discrepancies between the reference climate and observed data. This means that the relative difference between the median of a meteorological indicator in the reference climate and the observed median meteorological indicator was determined. This relative difference was then used to correct the meteorological indicators of the future climatic timeseries. It was consciously chosen to use a relative bias-correction on the median of the meteorological indicators, instead of using more complicated statistical bias-correction methods. However, the choice for this bias-correction method could have largely influenced the results. Using another bias-correction method would have resulted in different bias-corrected values for the future and thus a different selection of years.

Length of calibration and validation period

The historic discharge and climatic timeseries of this study covered a period of 54 years. From this period, historic years were selected that resembled the future, based on the meteorological indicator for high and low flows. These historic periods together made up one validation or calibration period for evaluating the robustness. Due to large expected changes in the meteorological indicator for low flows in the future, only a limited number of historic years was selected. As a result, the selected periods in RQ3 and RQ4 for the simulation of low flows have lengths of only one or two years. The same applies for the validation periods for simulation of high flows on extreme years. This means that the performance of the HBV and GR6J model was evaluated based on a small validation period (RQ3) and that the parameter sets were calibrated on a small calibration period (RQ4 and RQ5).

Perrin et al. (2007) showed that a calibration period of 1 year may be sufficient to obtain stable and robust parameter values, but this depends on the type of model and objective function. However, periods of 1 or 2 years, given the relatively small size of the study area with rapid rainfall-runoff responses, may not be necessarily sufficient for calibration and validation in this study, as the influence of one high or low flow event may have had large effects on the results of RQ3 and RQ4. This was for example the case when evaluating the performance of DSST *'H_DSST_E_2100Ld'*, which showed that the overestimation of the peak in 1980 largely affected the model performance. A too short calibration period could have led to nonunique solutions that are overfitted to the calibration period, meaning that true relationships in the data were not captured accurately. This could have made the projections for the future modelled with parameters calibrated on these periods more uncertain.

This shortcoming could be solved by expanding the bandwidth that was used to select a historic hydrological year. This bandwidth was chosen to be 10% and 20% for high and low flows, respectively. In case the bandwidth would have been larger, more years would have been selected. However, this would have meant that the selected years show less similarities with the future. A good balance is thus necessary.

Selection of years for simulation of high and low flows

For simulation of both high and low flows, the periods that resemble future conditions were a combination of periods that resemble future high flow conditions and future low flow conditions. Figure C1 in Appendix C shows that a relatively large proportion of years in the validation period was selected based on the meteorological indicator for high flows. Thus, years with high flows dominated these tests.

This may have impacted the performance of the models when validated on these years and the parameter sets obtained by calibration on these years. For example, calibration on these years was focused on tuning the quick runoff parameters which could have affected the other parameters. A more balanced period would have been better and could have been obtained by selecting as many years resembling high flows as years resembling low flows.

Selection of parameters

This study used a univariate sensitivity analysis to select parameters to which the objective functions were sensitive. This means that the parameters were varied one by one and parameter interactions were not considered. However, a univariate sensitivity analysis may have been too simple and thus unreliable for models with high interrelations between the parameters (Wagener & Kollat, 2007). It would have been useful to employ a multivariate sensitivity analysis as well, such that the entire parameter space would have been evaluated and parameter interactions would have been considered. Still, the selection of parameters seems reasonable. For example, parameters K_f and α, that are both related to the quick runoff in the HBV model, are always either jointly included (for high flows) or excluded (for low flows) from calibration. It is thus expected that a more complex sensitivity analysis method would have led to the same selection of parameters.

5.2. Results in perspective

Despite the limitations in the data and methods, the results of this study were largely relatable to what was found in literature. In this section, the results are put in perspective in relation to existing literature.

5.2.1. Selection of periods resembling future conditions

This study selected years that should resemble future high and low flow conditions based on meteorological indicators for high and low flows. These meteorological indicators should summarize the meteorological conditions that lead to annual maximum discharges and annual minimum 7-day mean discharges. As a meteorological indicator for high flows, the 3-day precipitation sum that is exceeded 6 days per year was determined. This meteorological indicator showed that the discharge depends on the precipitation of 3-days before, which is reasonable for a medium-sized basin with relatively fast response. As a meteorological indicator for low flows, the 150-day potential precipitation deficit that was exceeded 14 days per year was determined. The temporal window (150 days) is in line with findings of Demirel et al. (2013b) for rainfed sub-basins of the Rhine. In the end, the meteorological indicators showed to be good indicators for selecting periods that resembled the future. For example for low flows, the years with the highest meteorological indicator for low flows (1976 & 2020), which were selected as years that were selected to resemble future conditions, were also the years with the lowest annual minimum 7-day mean discharge.

5.2.2. Robustness of GR6J and HBV

Performance of GR6J and HBV model for simulating high and low flows

Both the GR6J and HBV model underestimated peak discharges (section 4.3.3). This was found by De Boer-Euser et al. (2017) in their model intercomparison study for amongst others the Lesse catchment as well. They also found that the GR4H model performed best for simulating high flows. As the GR4H and GR6J model are based on the same modelling concept, this corresponds with the results of this study where GR6J was better at simulating high flows than the HBV model.

For simulating low flows, the HBV model was preferred over the GR6J model. The HBV was well-capable at simulating the annual minimum 7-day mean discharge, especially in years with a long dry period. This was shown by the high value for NS_{inv} of 0.92 in the validation period of *'HBV_L_DSST_E_2100Ln'*. However, the HBV model had difficulties with simulating discharge peaks in the summer period. These summer peaks were underestimated by the model, as precipitation was stored in the soil rather than

discharged directly to the river. This is physically unrealistic, as these precipitation events in reality have an intensity higher than the infiltration capacity of the soil, which leads to infiltration excess overland flow. This quick runoff for summer events could have been modelled better with the addition of a very quick runoff component (De Boer-Euser et al., 2017). Future research may thus focus on improving model structures to improve hydrological modelling and understanding, especially for low flows.

The GR6J model, as a modified version of the widely-used GR4J model, was used in this study as the GR6J model was expected to provide better simulations on low flow compared to the original GR4J model without impacting the high-flow simulation ability (Golian et al., 2021; Pushpalatha et al., 2011). Still, the GR6J model did not perform as well as the HBV model did on low flows. Even though the annual minimum 7-day mean discharges were captured quite well, the model deviated from observations in the transition from high to low flows and in summer peaks. This is visualized in Figure 41, which shows the too steep recession curve in May 2001 and the overestimation of summer peaks. This may indicate that the GR6J model was calibrated such that a faster-responding system was represented than is actually the case.

However, the model's predictive power for simulating annual minimum 7-day mean discharges was still quite good. This may indicate that the groundwater exchange and/or evapotranspiration were adjusted such that this low streamflow was simulated well. Due to karstic features in the Lesse catchment, groundwater exchange is physically realistic (Le Moine et al., 2007). However, including groundwater exchange in hydrological models gives another uncertainty, as two sources of output fluxes (evapotranspiration and groundwater exchange) that are mostly unobservable have to be modelled. As this study calibrated the model parameters such that the parameters showed good fits on the discharge, the evapotranspiration and/or groundwater exchange may become physically unrealistic. The developers of the GR6J model acknowledged this limitation and focused on improving the model's predictive power rather than explicitly representing physical mechanisms (Pushpalatha et al., 2011).

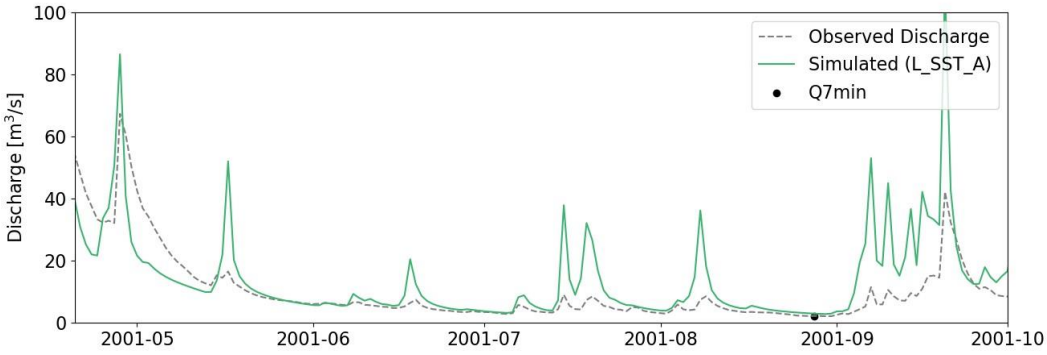

Figure 41 - Simulated discharge by the GR6J model, calibrated on test 'L_SST_I', and observed discharge in a part of the calibration period. Black dot shows the observed annual minimum 7-day mean discharge.

Calibration of the HBV model with y_{comb}, which focuses on simulation of both high and low streamflow, led to comparable values of NS_w and NS_{inv} as when calibrated on high or low streamflow only. This shows that the HBV model did not lose performance on the simulation of high or low streamflow when calibrated on this combined objective function. For the GR6J model, the performance on simulating low streamflow (NS_{inv}) slightly decreases when calibrated on both high and low streamflow. Both models are thus able to capture both high and low streamflow well within the same simulation.

Performance in periods resembling future conditions

The results of RQ3 showed that the performance of models in periods resembling future conditions was comparable with the performance in periods that do not resemble future conditions. Multi-objective

functions y_w, y_{inv} and y_{comb} were only slightly different between calibration and validation period in the DSSTs. This was both due to differences in RVE and Nash-Sutcliffe efficiencies.

The changes in RVE in the validation period compared to the calibration period in most of the DSSTs were in the same order of magnitude as the changes in RVE in the SSTs. This comparable increase in absolute RVE means that the water balance was captured worse in the validation period, but within acceptable limits. This contradicts with findings in other model robustness studies, such as the one of Dakhlaoui et al. (2019) that identified that the water balance was simulated worse in validation periods of their DSSTs. The reason for this could be that the validation periods used in this study, especially for simulating high flows, were not contrasting enough to identify these large model performance losses. The validation periods in this study were different from the calibration periods in their high and low flows. However, a year with a high flow did not have to include a totally different water balance in the rest of the year. As a result, the water balance could still be captured well. This point was discussed by Vormoor et al. (2018) as well.

In most DSSTs, the Nash-Sutcliffe efficiencies increased in the validation period. Reasons for these increases were given in section 4.3.3. Increases in Nash-Sutcliffe efficiencies were not observed in other model robustness studies. A difference with other model robustness studies, such as Dakhlaoui et al. (2017) is that this study used a multi-objective function instead of a single objective function. De Vos & Rientjes (2007) and Seibert (1999) showed that parameter sets may indicate a good fit for NS, but a poor fit for RVE and vice versa. By using a multi-objective function, parameter sets should give both a good fit for simulation of discharge extremes and a good water balance. The use of this objective function thus already made the models more robust, as physically more realistic parameter sets were obtained by calibration. When parameter sets were physically more realistic in calibration, these parameter sets may have been more realistic in validation periods as well. This thus shows that the choice of an objective function plays a large role in the evaluation of model robustness.

Changes in parameter sets

The results of RQ4 proved that both models showed different parameter sets when calibrated over different periods. This showed that parameter sets calibrated on historic conditions may be suboptimal for future conditions, which corresponds with the studies of Blöschl & Montanari (2010), Ji et al. (2023), Merz et al. (2011), Nicolle et al. (2021), Osuch et al. (2015) and Thirel et al. (2015).

This study used the SCEM-UA optimization algorithm for model calibration. With 5000 iterations, the algorithm showed to evaluate the parameter sets to a stable set (convergence, Appendix D). For the HBV model, a stable objective function and parameter set was found when calibrated multiple times on the same period. This means that the SCEM-UA algorithm could find a global optimum. For the GR6J model, a stable objective function was found, however, the parameter sets between multiple calibrations varied (Appendix E). This may indicate parameter equifinality, which means that multiple sets of parameter values produce comparable simulations. This problem was acknowledged by the developers of the model as well (Pushpalatha et al., 2011). The parameter equifinality made it challenging to attribute changes in optimal parameter values to differences in calibration period or to parameter equifinality. To get more insight in this, it would have been useful to do the same calibration exercises for each parameter separately, as parameter equifinality is expected to be less of a problem with a lower number of parameters. Next to that, it would have been interesting to calibrate on multiple smaller parts of the period that resembles future conditions. In case approximately the same parameter values are obtained for calibration on these subperiods with comparable conditions, the changes in parameter values in this study are mainly due to changes in calibration period. Pelletier & Andréassian (2022) showed that additional groundwater level data can improve the stability of parameters in the GR6J model. Further research into this field may thus improve the robustness of the model.

5.2.3. Effects of climate change

Due to an increase in potential evapotranspiration and a decrease in precipitation, the potential precipitation deficit is expected to increase, leading to a decrease in low streamflow. At the same time, more intense precipitation events are expected, leading to an increase in high streamflow. This study demonstrated that the projected effects of climate change on the annual minimum 7-day mean discharge are larger than the projected effects of climate change on the annual maximum discharge. This is in accordance with the projections of these discharges at Gendron with the KNMI'23 climate scenarios using the 'wflow_sbm' model Buitink et al. (2023), given in Appendix H. This stresses the need for a focus shift towards low flows.

The results of RQ5 showed that the uncertainty in projected impact due to different hydrological model structures and calibration approaches was not as large as the uncertainty due to different climate scenarios. This thus shows that the uncertainty in future emissions and regional climate response remains the largest source of uncertainty for climate change impact. However, the uncertainty in the climate change impact due to different hydrological model structures and calibration approaches cannot be neglected, which corresponds with findings of Merz et al. (2011) and Wilby (2005).

The projected change in annual maximum daily discharge was determined to be between-14% and +27% and to be between-66% and +13% in the annual minimum 7-day mean discharge. These ranges cover uncertainty in future emissions, climate response, model structure and calibration approach, which were expected to be the main sources of uncertainty. Still, climate change impacts may fall outside the given ranges, due to limitations in the hydrological modelling for example as a result of errors in calibration data or the uncertainty related to the choice for an objective function.

5.3. Generalizations

5.3.1. Study catchment

The focus of this study was on the Lesse, which is a medium-sized basin (1286 km^2) with a relatively fast response. The interpretations made for this catchment may be difficult to generalize, as basins with a different size or different catchment characteristics may lead to different outcomes. Future research into these catchments is thus needed.

Low streamflow in the Lesse showed to depend on the precipitation and potential evapotranspiration of the preceding 5-month period. This showed that the preceding winter does not have large effects on low streamflow, which is the case for the full Meuse basin (De Wit et al., 2007). Applying this methodology on larger sized basins may thus imply that periods longer than one year have to be selected to resemble conditions of the future.

5.3.2. Hydrological models

This study evaluated the robustness of two parsimonious conceptual hydrological models. This means the models are intentionally kept simple with a limited number of model parameters to be calibrated. Applying the methodology of this study on other parsimonious conceptual models is expected to result in similar outcomes. It is expected that using a different study basin would have more influence on the outcomes of this model robustness study than using a different conceptual model (Thirel et al., 2015).

Some models contain a relatively larger share of parameters derived from data on catchment characteristics rather than being obtained by calibration. For these models, the methodology applied in this study may be less relevant. For example, in the wflow_sbm model, most parameters are derived from data, but the parameter KsatHorFrac is calibrated and compensates for errors in model structure and data (Aerts et al., 2022). Calibration of this parameter on future conditions may have less effect on the simulated discharge, as the largest share of parameters are derived from data and are thus stable.

6. CONCLUSION AND RECOMMENDATIONS

This chapter gives the main conclusions per research question and an overarching conclusion. After that, practical recommendations and recommendations for future research are given.

6.1. Conclusion

RQ1 – Which meteorological conditions lead to high and low flows in the Lesse and how are these conditions expected to change in the future?

Annual maximum daily discharges in the Lesse were most often caused by precipitation events that took place in winter when the soil was already saturated. The 3-day precipitation sum that is exceeded 6 days per year showed to correlate with high flows. This meteorological indicator for high flows is expected to increase slightly in 2100 for climate scenarios with high future greenhouse gas emissions and decrease slightly for climate scenarios with low future greenhouse gas emissions.

Low streamflow in the Lesse correlated with a potential precipitation deficit over 150 days, which means that low flows were caused by a combination of high potential evapotranspiration and low precipitation. The 150-day precipitation deficit that is exceeded 14 days per year showed to be a good indicator for low flows. This meteorological indicator for low flows is expected to increase substantially in 2100 for all climate scenarios.

RQ2 – Which historic periods most closely resemble future conditions projected by climate change scenarios?

For most climate change scenarios, historic years were selected that resemble future conditions for high flows and for low flows. As the projected meteorological indicator for low flows in extreme years in the future for climate scenarios with high future greenhouse gas emissions was higher than has ever been observed, no historic periods were selected for these climate scenarios. The expected changes in the meteorological indicator for high flows were smaller, therefore a larger number of years resemble future conditions. For low flows, only one or two historic years resemble the future conditions projected by climate scenarios.

RQ3 – What is the performance of the HBV and GR6J model in simulating high and low streamflow under climatic conditions resembling future conditions projected by the KNMI'23 scenarios, compared to the performance of the models under historic climatic conditions?

The HBV and GR6J model performed well in simulating high and low streamflow in the Lesse. The GR6J model showed better performance on high streamflow, whereas the HBV model showed better performance on low streamflow. Both models faced difficulties with simulating summer peaks.

The performance of both models when simulating streamflow under historic conditions was different than the performance under conditions resembling future conditions, due to differences in the simulation of the water balance and the simulation of high and low streamflow. However, large degradation in performance that was observed in other studies was not observed in this study. This may be due to the use of a multi-objective function. For simulation of high flows, another reason may be the fact that climatic conditions were not as climatically contrasting between validation and calibration periods as in other studies.

RQ4 – What are the model parameters of the HBV and GR6J model when calibrated on climatic conditions resembling future conditions projected by the KNMI'23 scenarios, compared to when the models are calibrated on historic climatic conditions?

Model parameters of the HBV and GR6J model were different when calibrated on climatic conditions resembling future conditions compared to when calibrated on historic climatic conditions. This shows that parameter values calibrated on historic periods may have been suboptimal for simulation of the

future and thus indicates a lack of robustness. For the HBV model, the changes in parameter sets were for a large part explainable. For the GR6J model, parameter equifinality was observed, which made it more difficult to find underlying reasons for the observed changes in parameter values. It is thus advised to calibrate models on climatic conditions that resemble future conditions, as this leads to higher performances in these periods.

RQ5 – What is the projected impact of climate change on high and low streamflow in the Lesse, and how does the uncertainty of this impact due to different climate scenarios relate to the uncertainty due to different hydrological model structures and calibration approaches?

Climate change is projected to lead to a small increase in annual maximum daily discharges and a large decrease in annual minimum 7-day mean discharges. The uncertainty in future greenhouse gas emissions and regional climate response was the largest source of uncertainty for determining this climate change impact. However, the uncertainty of the projected impact due to different hydrological model structures and calibration approaches should not be neglected.

Aim – To evaluate the robustness of the HBV and GR6J model for simulating the impact of climate change on high and low streamflow in the Lesse catchment, by evaluating the models on historic periods resembling the climatic conditions projected by climate change scenarios

Overall, given the methodology used in this study, the HBV and GR6J model showed to be robust for simulating the impact of climate change on high and low streamflow in the Lesse catchment. The performance of both models when simulating streamflow in periods that resemble future conditions was only slightly different from performance under historic conditions. In case the same meteorological conditions that led to high and low streamflow in the past will lead to high and low streamflow in the future, the HBV and GR6J model can thus be used for simulating impact of climate change on high and low streamflow in the Lesse.

However, the optimal parameter values of the HBV and GR6J model proved to be different when calibrated on different periods. This indicates that parameter sets calibrated on historic conditions are suboptimal under changing climatic conditions and thus lose validity. This showed that the model robustness can still be improved, by improving model structures or calibration approaches. As long as a completely robust model does not exist, it is advised to calibrate conceptual models on periods that resemble the conditions projected in the future.

6.2. Recommendations

6.2.1. Practical recommendations for projecting climate change impact

The results of this study showed that projecting climate change impact on high and low streamflow in the Lesse is highly uncertain. Even though other uncertainties, for example due to the choice of an objective function or due to errors in calibration data, may have affected the project impact as well, it is expected that uncertainties in climate scenarios, hydrological model structures and calibration approaches are the largest sources of uncertainty. To project climate change impacts on high streamflow in the Lesse, it is recommended to use the GR6J model. This model showed good performance on high streamflow, also in periods with climatic conditions resembling future conditions. For simulating low streamflow in the Lesse, it is recommended to use the HBV model. Next to that, it is recommended to calibrate models on periods that resemble future conditions. Despite the rather short calibration period, these periods may be more suitable for calibration.

Both models showed changes in model performance and parameters when calibrated or validated on a period with climatic conditions resembling the future. However, compared to other studies, the degradation in model performance in validation periods was rather small. This may be due to the use

of a multi-objective function, which focuses on both a good simulation of the water balance and a good simulation of high and/or low streamflow. It is therefore recommended to use a multi-objective function for calibration of models in the future.

6.2.2. Recommendations for future research

Improving the model structures of the HBV and GR6J model could improve the performance of the HBV and GR6J model for simulating high and low flows. As a result, the robustness of these models for projecting impact of climate change may improve as well. The HBV model underestimated peaks in the summer periods. Precipitation was stored in the soil rather than discharged, whereas precipitation intensity in reality may be larger than the infiltration capacity of the soil and thus may lead to infiltration excess overland flow in reality. Future research could focus on the addition of a very quick runoff component to improve simulations of summer peaks with the HBV model. This is of importance, as high flows in summer, such as the high flow event in 2021, may occur more often in the future.

The GR6J model was able to simulate high and low streamflow rather well but showed physically unrealistic behaviour in the rest of the year. The inclusion of the groundwater exchange may be physically realistic due to the karstic feature in the Lesse, but this groundwater exchange may be the reason for the parameter equifinality and physically unrealistic simulations as well. Further research on the role of this groundwater exchange function for simulating low flows is recommended.

This study showed that the impact of climate change on high and low streamflow is substantial. Especially low flows are projected to decrease in the future. Given the societal relevance, it is thus of importance to conduct more model robustness assessment studies that specifically focus on high and low streamflow. This study focused on high and low streamflow with a return period of 1 year. It would be interesting to evaluate the robustness for simulating the impact of climate change on streamflow with longer return periods as well.

Lastly, this study evaluated the robustness of two hydrological models on one study catchment. The conclusions made for this study may be difficult to generalize for other types of catchments or other hydrological models. It would therefore be interesting to extend this study, by applying a similar methodology to other hydrological models or catchments. This can be every catchment, as long as a sufficiently long period of historic data and climate scenarios are available. As the Lesse is a fast-responding catchment, it would be interesting to look at a slower responding catchment.

BIBLIOGRAPHY

Abebe, N. A., Ogden, F. L., & Pradhan, N. R. (2010). Sensitivity and uncertainty analysis of the conceptual HBV rainfall-runoff model: Implications for parameter estimation. *Journal of Hydrology*, *389*(3–4), 301–310. https://doi.org/10.1016/j.jhydrol.2010.06.007

Aerts, J. P. M., Hut, R. W., Van De Giesen, N. C., Drost, N., Van Verseveld, W. J., Weerts, A. H., & Hazenberg, P. (2022). Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model. *Hydrology and Earth System Sciences*, *26*(16), 4407–4430. https://doi.org/10.5194/hess-26-4407-2022

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

Arnell, N. W. (2011). Uncertainty in the relationship between climate forcing and hydrological response in UK catchments. *Hydrology and Earth System Sciences*, *15*(3), 897–912. https://doi.org/10.5194/hess-15-897-2011

Bastola, S., Murphy, C., & Sweeney, J. (2011). Evaluation of the transferability of hydrological model parameters for simulations under changed climatic conditions Evaluation of the transferability of hydrological model. *Hydrology and Earth System Sciences Discussions*, *8*, 5891–5915. https://doi.org/10.5194/hessd-8-5891-2011

Berger, H. E. J. (1992). *Flow Forecasting for the River Meuse* [TU Delft]. https://repository.tudelft.nl/islandora/object/uuid%3Aacefdfb7-0ce8-464a-a372-2bbc4a50ffa9

Bergström, S. (1976). *Development and Application of a Conceptual Runoff Model for Scandinavian Catchments*. https://www.researchgate.net/publication/255274162

Bergström, S., & Lindström, G. (2015). Interpretation of runoff processes in hydrological modelling-experience from the HBV approach. *Hydrological Processes*, *29*(16), 3535–3545. https://doi.org/10.1002/hyp.10510

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

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

Beven, K., & O'Connell, P. E. (1982). *On the role of physically-based distributed modelling in hydrology*. https://nora.nerc.ac.uk/id/eprint/5847/1/IH_081.pdf

Blöschl, G., & Montanari, A. (2010). Climate change impacts- throwing the dice? *Hydrological Processes*, *24*(3), 374–381. https://doi.org/10.1002/hyp.7574

Booij, M. J. (2005). Impact of climate change on river flooding assessed with different spatial model resolutions. *Journal of Hydrology*, *303*(1–4), 176–198. https://doi.org/10.1016/j.jhydrol.2004.07.013

Booij, M. J., & Krol, M. S. (2010). Balance between calibration objectives in a conceptual hydrological model. *Hydrological Sciences Journal*, *55*(6), 1017–1032. https://doi.org/10.1080/02626667.2010.505892

Bouaziz, L. J. E. (2021). *Internal processes in hydrological models A glance at the Meuse basin from space* [TU Delft]. https://doi.org/10.4233/uuid:09d84cc1-27e2-4327-a8c7-207a75952061

Buitink, J., Tsiokanos, A., Geertsema, T., ten Velden, C., Bouaziz, L., & Sperna Weiland, F. (2023). *Implications of the KNMI'23 climate scenarios for the discharge of the Rhine and Meuse*.

Cannon, A. J., Sobie, S. R., & Murdock, T. Q. (2015). Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *Journal of Climate*, *28*(17), 6938–6959. https://doi.org/10.1175/JCLI-D-14-00754.1

Chiew, F. H. S., Potter, N. J., Vaze, J., Petheram, C., Zhang, L., Teng, J., & Post, D. A. (2014). Observed hydrologic non-stationarity in far south-eastern Australia: Implications for modelling and prediction. *Stochastic Environmental Research and Risk Assessment*, *28*(1), 3–15. https://doi.org/10.1007/s00477-013-0755-5

Chiew, F. H. S., Teng, J., Vaze, J., Post, D. A., Perraud, J. M., Kirono, D. G. C., & Viney, N. R. (2009). Estimating climate change impact on runoff across southeast Australia: Method, results, and implications of the modeling method. *Water Resources Research*, *45*(10). https://doi.org/10.1029/2008WR007338

Cornes, R. C., van der Schrier, Gl., van den Besselaar, E. J. M., & Jones, P. D. (2018). An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets. *Journal of Geophysical Research: Atmospheres*, *123*(17), 9391–9409. https://doi.org/10.1029/2017JD028200

Coron, L. (2013). *Are conceptual hydrological models robust in a changing climate? Diagnostic on a set of French and Australian catchments*. Institut des Sciences et Industries du Vivant et de l'Environnement .

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., & Hendrickx, F. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resources Research*, *48*(5). https://doi.org/10.1029/2011WR011721

Dakhlaoui, H., Ruelland, D., & Tramblay, Y. (2019). A bootstrap-based differential split-sample test to assess the transferability of conceptual rainfall-runoff models under past and future climate variability. *Journal of Hydrology*, *575*, 470–486. https://doi.org/10.1016/j.jhydrol.2019.05.056

Dakhlaoui, H., Ruelland, D., Tramblay, Y., & Bargaoui, Z. (2017). Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern Tunisia. *Journal of Hydrology*, *550*, 201–217. https://doi.org/10.1016/j.jhydrol.2017.04.032

De Boer-Euser, T., Bouaziz, L., De Niel, J., Brauer, C., Dewals, B., Drogue, G., Fenicia, F., Grelier, B., Nossent, J., Pereira, F., Savenije, H., Thirel, G., & Willems, P. (2017). Looking beyond general metrics for model comparison - lessons from an international model intercomparison study. *Hydrology and Earth System Sciences*, *21*(1), 423–440. https://doi.org/10.5194/hess-21-423-2017

De Vos, N. J., & Rientjes, T. H. M. (2007). Multi-objective performance comparison of an artificial neural network and a conceptual rainfall-runoff model. *Hydrological Sciences Journal*, *52*(3), 397–413. https://doi.org/10.1623/hysj.52.3.397

De Vos, N. J., Rientjes, T. H. M., & Gupta, H. V. (2010). Diagnostic evaluation of conceptual rainfall-runoff models using temporal clustering. *Hydrological Processes*, *24*(20), 2840–2850. https://doi.org/10.1002/hyp.7698

De Wit, M. J. M., Van Den Hurk, B., Warmerdam, P. M. M., Torfs, P. J. J. F., Roulin, E., & Van Deursen, W. P. A. (2007). Impact of climate change on low-flows in the river Meuse. *Climatic Change*, *82*(3–4), 351–372. https://doi.org/10.1007/s10584-006-9195-2

Deckers, D. L. E. H., Booij, M. J., Rientjes, T. H. M., & Krol, M. S. (2010). Catchment Variability and Parameter Estimation in Multi-Objective Regionalisation of a Rainfall-Runoff Model. *Water Resources Management*, *24*(14), 3961–3985. https://doi.org/10.1007/s11269-010-9642-8

Demirel, M. C., Booij, M. J., & Hoekstra, A. Y. (2013a). Impacts of climate change on the seasonality of low flows in 134 catchments in the River Rhine basin using an ensemble of bias-corrected regional climate simulations. *Hydrology and Earth System Sciences*, *17*(10), 4241–4257. https://doi.org/10.5194/hess-17-4241-2013

Demirel, M. C., Booij, M. J., & Hoekstra, A. Y. (2013b). Identification of appropriate lags and temporal resolutions for low flow indicators in the River Rhine to forecast low flows with different lead times. *Hydrological Processes*, *27*(19), 2742–2758. https://doi.org/10.1002/hyp.9402

Diermanse, F. L. M., Kwadijk, J. C. J., Beckers, J. V. L., & Crebas, J. I. (2010, September). Statistical trend analysis of River Rhine and Meuse discharges. *BHS 3rd International Conference*. https://doi.org/10.7558/bhs.2010.ic9

Do, H. X., Westra, S., & Leonard, M. (2017). A global-scale investigation of trends in annual maximum streamflow. *Journal of Hydrology*, *552*, 28–43. https://doi.org/10.1016/j.jhydrol.2017.06.015

Donat, M. G., Sillmann, J., Wild, S., Alexander, L. V, Lippmann, T., & Zwiers, F. W. (2014). Consistency of Temperature and Precipitation Extremes across Various Global Gridded In Situ and Reanalysis Datasets. *Journal of Climate*, *27*(13), 5019–5035. https://doi.org/10.1175/JCLI-D-13

Donnelly-Makowecki, L. M., & Moore, R. D. (1999). Hierarchical testing of three rainfall-runoff models in small forested catchments. *Journal of Hydrology*, *219*, 136–152.https://doi.org/10.1016/S0022-1694(99)00056-6

Duan, Q., Sorooshian, S., & Gupta, V. (1992). Effective and Efficient Global Optimization for Conceptual Rainfall-Runoff Models. *Water Resources Research*, *28*(4), 1015–1031. https://doi.org/10.1029/91WR02985

Dwarakish, G. S., & Ganasri, B. P. (2015). Impact of land use change on hydrological systems: A review of current modeling approaches. *Cogent Geoscience*, *1*(1), 1115691. https://doi.org/10.1080/23312041.2015.1115691

European Environment Agency. (2000). *Corine Land Cover*. https://www.eea.europa.eu/data-and-maps/data/clc-2000-raster-3

European Space Agency. (2021). *Copernicus Global Digital Elevation Model.* Distributed by OpenTopography. https://doi.org/10.5069/G9028PQB. Accessed 2024-01-06

Golian, S., Murphy, C., & Meresa, H. (2021). Regionalization of hydrological models for flow estimation in ungauged catchments in Ireland. *Journal of Hydrology: Regional Studies*, *36*, 100859. https://doi.org/10.1016/j.ejrh.2021.100859

Gudmundsson, L., Boulange, J., Do, H. X., Gosling, S. N., Grillakis, M. G., Koutroulis, A. G., Leonard, M., Liu, J., Schmied, H. M., Papadimitriou, L., Pokhrel, Y., Seneviratne, S. I., Satoh, Y., Thiery, W., Westra, S., & Zhang, X. (2021). Globally observed trends in mean and extreme river flow attributed to climate change. *Science*, *371*, 1156–1162. https://doi.org/10.1594/PANGAEA.887470

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*(2), 463–477. https://doi.org/10.5194/hess-18-463-2014

Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, *34*(4), 751–763. https://doi.org/10.1029/97WR03495

Helsel, D. R., Hirsch, R. M., Ryberg, K. R., Archfield, S. A., & Gilroy, E. J. (2020). Statistical Methods in Water Resources. In *U.S. Geological Survey Techniques and Methods* (Vol. 4). https://doi.org/10.3133/tm4a3

Hirabayashi, Y., Alifu, H., Yamazaki, D., Imada, Y., Shiogama, H., & Kimura, Y. (2021). Anthropogenic climate change has changed frequency of past flood during 2010-2013. *Progress in Earth and Planetary Science*, *8*(1). https://doi.org/10.1186/s40645-021-00431-w

Hooghart, J. C., & Lablans, W. N. (1988). *Van Penman naar Makkink: een nieuwe berekeningswijze voor de klimatologische verdampingsgetallen*.

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

IPCC. (2021). *Climate Change 2021: The Physical Science Basis*. https://www.ipcc.ch/report/ar6/wg1/

Jaiswal, R. K., Ali, S., & Bharti, B. (2020). Comparative evaluation of conceptual and physical rainfall–runoff models. *Applied Water Science*, *10*(1). https://doi.org/10.1007/s13201-019-1122-6

Ji, H. K., Mirzaei, M., Lai, S. H., Dehghani, A., & Dehghani, A. (2023). The robustness of conceptual rainfall-runoff modelling under climate variability – A review. *Journal of Hydrology*, *621*. https://doi.org/10.1016/j.jhydrol.2023.129666

Katz, R. (2013). Statistical Methods for Nonstationary Extremes. In *Extremes in a Changing Climate* (Vol. 65). Springer. http://www.springer.com/series/6689

Kay, A. L., Davies, H. N., Bell, V. A., & Jones, R. G. (2009). Comparison of uncertainty sources for climate change impacts: Flood frequency in England. *Climatic Change*, *92*(1–2), 41–63. https://doi.org/10.1007/s10584-008-9471-4

Kirkby, M. (1988). Hillslope runoff processes and models. *Journal of Hydrology*, *100*, 315–339. https://doi.org/10.1016/0022-1694(88)90190-4

Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24. https://doi.org/10.1080/02626668609491024

KNMI. (2023a). *Modeluitvoer KNMI'23-klimaatscenario's*. https://www.knmi.nl/kennis-en-datacentrum/achtergrond/modeluitvoer-knmi-23-klimaatscenario-s

Le Moine, N. (2008). *Le bassin versant de surface vu par le souterrain: une voie d'amélioration des performances et du réalisme des modèles pluie-débit ?* [Thesis]. Université Pierre et Marie Curie.

Le Moine, N., Andréassian, V., Perrin, C., & Michel, C. (2007). How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments. *Water Resources Research*, *43*(6). https://doi.org/10.1029/2006WR005608

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, *27*(15), 2171–2186. https://doi.org/10.1002/hyp.9740

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

Luo, J., Wang, E., Shen, S., Zheng, H., & Zhang, Y. (2012). Effects of conditional parameterization on performance of rainfall-runoff model regarding hydrologic non-stationarity. *Hydrological Processes*, *26*(26), 3953–3961. https://doi.org/10.1002/hyp.8420

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

Malede, D. A., Agumassie, T. A., Kosgei, J. R., Andualem, T. G., & Diallo, I. (2022). Recent Approaches to Climate Change Impacts on Hydrological Extremes in the Upper Blue Nile Basin, Ethiopia. *Earth Systems and Environment*, *6*(3), 669–679. https://doi.org/10.1007/s41748-021-00287-6

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

Merz, R., Parajka, J., & Blöschl, G. (2011). Time stability of catchment model parameters: Implications for climate impact analyses. *Water Resources Research*, *47*(2). https://doi.org/10.1029/2010WR009505

Michel, C., Perrin, C., & Andreassian, V. (2003). The exponential store: A correct formulation for rainfall-runoff modelling. *Hydrological Sciences Journal*, *48*(1), 109–124. https://doi.org/10.1623/hysj.48.1.109.43484

Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, *48*(9). https://doi.org/10.1029/2011WR011412

Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L., Koutsoyiannis, D., Cudennec, C., Toth, E., Grimaldi, S., Blöschl, G., Sivapalan, M., Beven, K., Gupta, H., Hipsey, M., Schaefli, B., Arheimer, B., Boegh, E., Schymanski, S. J., … Belyaev, V. (2013). "Panta Rhei-Everything Flows": Change in hydrology and society-The IAHS Scientific Decade 2013-2022. *Hydrological Sciences Journal*, *58*(6), 1256–1275. https://doi.org/10.1080/02626667.2013.809088

Moradkhani, H., & Sorooshian, S. (2009). General Review of Rainfall-Runoff Modeling: Model Calibration, Data Assimilation, and Uncertainty Analysis. In *Hydrological Modelling and the Water Cycle: Coupling the Atmospheric and Hydrological Models* (Vol. 63, pp. 1–25). Springer. www.springer.com/series/6689

Nicolle, P., Andreássian, V., Royer-Gaspard, P., Perrin, C., Thirel, G., Coron, L., & Santos, L. (2021). Technical note: RAT- A robustness assessment test for calibrated and uncalibrated hydrological models. *Hydrology and Earth System Sciences*, *25*(9), 5013–5027. https://doi.org/10.5194/hess-25-5013-2021

Niel, H., Paturel, J. E., & Servat, E. (2003). Study of parameter stability of a lumped hydrologic model in a context of climatic variability. *Journal of Hydrology*, *278*(1–4), 213–230. https://doi.org/10.1016/S0022-1694(03)00158-6

Osuch, M., Romanowicz, R. J., & Booij, M. J. (2015). The influence of parametric uncertainty on the relationships between HBV model parameters and climatic characteristics. *Hydrological Sciences Journal*, *60*(7–8), 1299–1316. https://doi.org/10.1080/02626667.2014.967694

Ott, I., Duethmann, D., Liebert, J., Berg, P., Feldmann, H., Ihringer, J., Kunstmann, H., Merz, B., Schaedler, G., & Wagner, S. (2013). High-resolution climate change impact analysis on medium-sized river catchments in Germany: An ensemble assessment. *Journal of Hydrometeorology*, *14*(4), 1175–1193. https://doi.org/10.1175/JHM-D-12-091.1

Pelletier, A., & Andréassian, V. (2022). On constraining a lumped hydrological model with both piezometry and streamflow: Results of a large sample evaluation. *Hydrology and Earth System Sciences*, *26*(10), 2733–2758. https://doi.org/10.5194/hess-26-2733-2022

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

Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C., & Mathevet, T. (2007). Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models. *Hydrological Sciences Journal*, *52*(1), 131–151. https://doi.org/10.1623/hysj.52.1.131

Prudhomme, C., & Davies, H. (2009). Assessing uncertainties in climate change impact analyses on the river flow regimes in the UK. Part 2: Future climate. *Climatic Change*, *93*(1–2), 197–222. https://doi.org/10.1007/s10584-008-9461-6

Prudhomme, C., Giuntoli, I., Robinson, E. L., Clark, D. B., Arnell, N. W., Dankers, R., Fekete, B. M., Franssen, W., Gerten, D., Gosling, S. N., Hagemann, S., Hannah, D. M., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Wada, Y., & Wisser, D. (2014). Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(9), 3262–3267. https://doi.org/10.1073/pnas.1222473110

Pushpalatha, R., Perrin, C., Le Moine, N., Mathevet, T., & Andréassian, V. (2011). A downward structural sensitivity analysis of hydrological models to improve low-flow simulation. *Journal of Hydrology*, *411*(1–2), 66–76. https://doi.org/10.1016/j.jhydrol.2011.09.034

Pushpalatha, R., Perrin, C., Moine, N. Le, & Andréassian, V. (2012). A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology*, *420–421*, 171–182. https://doi.org/10.1016/j.jhydrol.2011.11.055

Rameshwaran, P., Bell, V. A., Davies, H. N., & Kay, A. L. (2021). How might climate change affect river flows across West Africa? *Climatic Change*, *169*(3–4). https://doi.org/10.1007/s10584-021-03256-0

Razafimaharo, C., Krähenmann, S., Höpp, S., Rauthe, M., & Deutschländer, T. (2020). New high-resolution gridded dataset of daily mean, minimum, and maximum temperature and relative humidity for Central Europe (HYRAS). *Theoretical and Applied Climatology*, *142*, 1531–1553. https://doi.org/10.1007/s00704-020-03388-w/Published

Razavi, S. (2021). Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling. *Environmental Modelling and Software*, *144*. https://doi.org/10.1016/j.envsoft.2021.105159

Refsgaard, J. C., & Knudsen, J. (1996). Operational validation and intercomparison of different types of hydrological models. *Water Resources Research*, *32*(7), 2189–2202. https://doi.org/10.1029/96WR00896

RIWA. (2021). *Lage afvoeren in de Maas - Bijdrage zijrivieren*. https://www.riwa-maas.org/wp-content/uploads/2021/11/IDF2704-RIWA-MAAS-Rapport-Lage-afvoeren-in-de-Maas-digitaal-klein.pdf

RIWA. (2022). *Low river discharge of the Meuse*. https://www.riwa-maas.org/wp-content/uploads/2022/06/IDF2857-RIWA-MAAS-Low-River-Discharge-of-the-Meuse-Rapport-digitaal.pdf

Royer-Gaspard, P., Andreássian, V., & Thirel, G. (2021). Technical note: PMR- A proxy metric to assess hydrological model robustness in a changing climate. *Hydrology and Earth System Sciences*, *25*(11), 5703–5716. https://doi.org/10.5194/hess-25-5703-2021

Ruelland, D., Hublart, P., & Tramblay, Y. (2015). Assessing uncertainties in climate change impacts on runoff in Western Mediterranean basins. *IAHS-AISH Proceedings and Reports*, *371*, 75–81. https://doi.org/10.5194/piahs-371-75-2015

Seibert, J. (1999). Regionalisation of parameters for a conceptual rainfall-runoff model. *Agricultural and Forest Meteorology*, *98–99*, 273–293. https://doi.org/10.1016/S0168-1923(99)00105-7

Seibert, J. (2003). Reliability of Model Predictions Outside Calibration Conditions. *Nordic Hydrology*, *34*(3), 477–492. https://doi.org/10.2166/nh.2003.028

Seiller, G., Anctil, F., & Perrin, C. (2012). Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. *Hydrology and Earth System Sciences*, *16*(4), 1171–1189. https://doi.org/10.5194/hess-16-1171-2012

Service Public de Wallonie. (2023). *Observations Débit 8221 | GENDRON*. Cotes Limnimétriques et Débits de l'année 1998 - 2021. https://hydrometrie.wallonie.be/home/observations/debit.html?station=DGH%2F8221&topic=annual_report

Singh, R., Wagener, T., Van Werkhoven, K., Mann, M. E., & Crane, R. (2011). A trading-space-for-time approach to probabilistic continuous streamflow predictions in a changing climate-accounting for changing watershed behavior. *Hydrology and Earth System Sciences*, *15*(11), 3591–3603. https://doi.org/10.5194/hess-15-3591-2011

Sitterson, J., Knightes, C., Parmar, R., Wolfe, K., Avant, B., Overview, A., & Muche, M. (2018). An Overview of Rainfall-Runoff Model Types. *International Congress on Environmental Modelling and Software*, 41. https://scholarsarchive.byu.edu/iemssconference/2018/Stream-C/41

Sleziak, P., Szolgay, J., Hlavčová, K., Duethmann, D., Parajka, J., & Danko, M. (2018). Factors controlling alterations in the performance of a runoff model in changing climate conditions. *Journal of Hydrology and Hydromechanics*, *66*(4), 381–392. https://doi.org/10.2478/johh-2018-0031

Sluijter, R., Plieger, M., van Oldenborgh, G. J., Beersma, J., & de Vries, H. (2018). *De droogte van 2018 - Een analyse op basis van het potentiële neerslagtekort*. https://klimaatadaptatienederland.nl/@204145/de-droogte-van-2018/

Smakhtin, V. U. (2001). Low flow hydrology: a review. *Journal of Hydrology*, *240*, 147–186. https://doi.org/10.1016/S0022-1694(00)00340-1

SMHI. (2006). *Integrated Hydrological Modelling System - Manual version 5.10*.

Sorooshian, S., & Gupta, V. K. (1995). Model calibration. In V. P. Singh (Ed.), *Computer Models of Watershed Hydrology* (pp. 23–68).

Sperna Weiland, F., Hegnauer, W. M., Bouaziz, L., & Beersma, J. (2015). *Implications of the KNMI'14 climate scenarios for the discharge of the Rhine and Meuse*. Deltares. https://publications.deltares.nl/1220042_000.pdf

Stephens, C. M., Marshall, L. A., & Johnson, F. M. (2019). Investigating strategies to improve hydrologic model performance in a changing climate. *Journal of Hydrology*, *579*. https://doi.org/10.1016/j.jhydrol.2019.124219

Stephens, C. M., Marshall, L. A., Johnson, F. M., Lin, L., Band, L. E., & Ajami, H. (2020). Is Past Variability a Suitable Proxy for Future Change? A Virtual Catchment Experiment. *Water Resources Research*, *56*(2). https://doi.org/10.1029/2019WR026275

Task Force Fact-finding hoogwater 2021. (2021). *Hoogwater 2021 - Feiten en Duiding*. https://www.enwinfo.nl/publish/pages/183541/211102_enw_hoogwater_2021-dv-def.pdf

Teng, J., Vaze, J., Chiew, F. H. S., Wang, B., & Perraud, J. M. (2012). Estimating the relative uncertainties sourced from GCMs and hydrological models in modeling climate change impact on runoff. *Journal of Hydrometeorology*, *13*(1), 122–139. https://doi.org/10.1175/JHM-D-11-058.1

Thirel, G., Andréassian, V., & Perrin, C. (2015). On the need to test hydrological models under changing conditions. *Hydrological Sciences Journal*, *60*(7–8), 1165–1173. https://doi.org/10.1080/02626667.2015.1050027

Tian, Y., Xu, Y. P., & Zhang, X. J. (2013). Assessment of Climate Change Impacts on River High Flows through Comparative Use of GR4J, HBV and Xinanjiang Models. *Water Resources Management*, *27*(8), 2871–2888. https://doi.org/10.1007/s11269-013-0321-4

Tohver, I. M., Hamlet, A. F., & Lee, S. Y. (2014). Impacts of 21st-Century Climate Change on Hydrologic Extremes in the Pacific Northwest Region of North America. *Journal of the American Water Resources Association*, *50*(6), 1461–1476. https://doi.org/10.1111/jawr.12199

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

Tramblay, Y., Ruelland, D., Somot, S., Bouaicha, R., & Servat, E. (2013). High-resolution Med-CORDEX regional climate model simulations for hydrological impact studies: A first evaluation of the ALADIN-Climate model in Morocco. *Hydrology and Earth System Sciences*, *17*(10), 3721–3739. https://doi.org/10.5194/hess-17-3721-2013

Trenberth, K. E. (1999). Conceptual Framework for Changes of Extremes of the Hydrological Cycle with Climate Change. *Climatic Change*, *42*, 327–339. https://doi.org/10.1023/A:1005488920935

Van Den Tillaart, S. P. M., Booij, M. J., & Krol, M. S. (2013). Impact of uncertainties in discharge determination on the parameter estimation and performance of a hydrological model. *Hydrology Research*, *44*(3), 454–466. https://doi.org/10.2166/nh.2012.147

Van Dorland, R., Beersma, J., Bessembinder, J., Bloemendaal, N., Van Den Brink, H., Brotons Blanes, M., Drijfhout, S., Groenland, R., Haarsma, R., Homan, C., Keizer, I., Krikken, F., Le Bars, D., Lenderink, G., Van Meijgaard, E., Meirink, J. F., Overbeek, B., Reerink, T., Selten, F., … Van Der Wiel, K. (2023). *KNMI National Climate Scenarios 2023 for the Netherlands*.

Van Esse, W. R., Perrin, C., Booij, M. J., Augustijn, D. C. M., Fenicia, F., Kavetski, D., & Lobligeois, F. (2013). The influence of conceptual model structure on model performance: A comparative study for 237 French catchments. *Hydrology and Earth System Sciences*, *17*(10), 4227–4239. https://doi.org/10.5194/hess-17-4227-2013

Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R., & Teng, J. (2010). Climate non-stationarity- Validity of calibrated rainfall-runoff models for use in climate change studies. *Journal of Hydrology*, *394*(3–4), 447–457. https://doi.org/10.1016/j.jhydrol.2010.09.018

Vaze, J., & Teng, J. (2011). Future climate and runoff projections across New South Wales, Australia: Results and practical applications. *Hydrological Processes*, *25*(1), 18–35. https://doi.org/10.1002/hyp.7812

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

Vrugt, J. A., Gupta, H. V., Bouten, W., & Sorooshian, S. (2003). A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, *39*(8). https://doi.org/10.1029/2002WR001642

Wagener, T., & Kollat, J. (2007). Numerical and visual evaluation of hydrological and environmental models using the Monte Carlo analysis toolbox. *Environmental Modelling and Software*, *22*(7), 1021–1033. https://doi.org/10.1016/j.envsoft.2006.06.017

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., & Gupta, H. V. (2003). Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. *Hydrological Processes*, *17*(2), 455–476. https://doi.org/10.1002/hyp.1135

Wilby, R. L. (2005). Uncertainty in water resource model parameters used for climate change impact assessment. *Hydrological Processes*, *19*(16), 3201–3219. https://doi.org/10.1002/hyp.5819

Wilby, R. L., & Harris, I. (2006). A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames, UK. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2005WR004065

Wu, K., & Johnston, C. A. (2007). Hydrologic response to climatic variability in a Great Lakes Watershed: A case study with the SWAT model. *Journal of Hydrology*, *337*(1–2), 187–199. https://doi.org/10.1016/j.jhydrol.2007.01.030

Xu, C. (1999). Operational testing of a water balance model for predicting climate change impacts. *Agricultural and Forest Meteorology*, *98–99*, 295–304. https://doi.org/10.1016/S0168-1923(99)00106-9

Zeng, L., Xiong, L., Liu, D., Chen, J., & Kim, J. S. (2019). Improving parameter transferability of GR4J model under changing environments considering nonstationarity. *Water (Switzerland)*, *11*(10). https://doi.org/10.3390/w11102029

APPENDICES

Appendix A: Pre-processing of observed discharge timeseries

As described in section 2.3.1, observed discharge data for the period 1968-2021 was used. For the period 1968-1998, the discharge series was obtained from SETHY/WACONDAH. For 1998-2021, the discharge series was obtained from the annual reports from Service Public de Wallonie (2023).

Annual reports from Service Public de Wallonie for the period 1968 – 1998 were available as well. This made it possible to compare the discharge data of Service Public de Wallonie and SETHY/WACONDAH for the period 1968 – 1998. When doing this, large differences were observed on each first of July. The discharge on the first of July in the annual reports from Service Public de Wallonie was very high when compared with the data from SETHY/WACONDAH, as illustrated in Figure A1.

It was therefore decided to pre-process the data such, that each discharge value on the first of July is substituted by the average discharge on the 30th of June and 2nd of July.

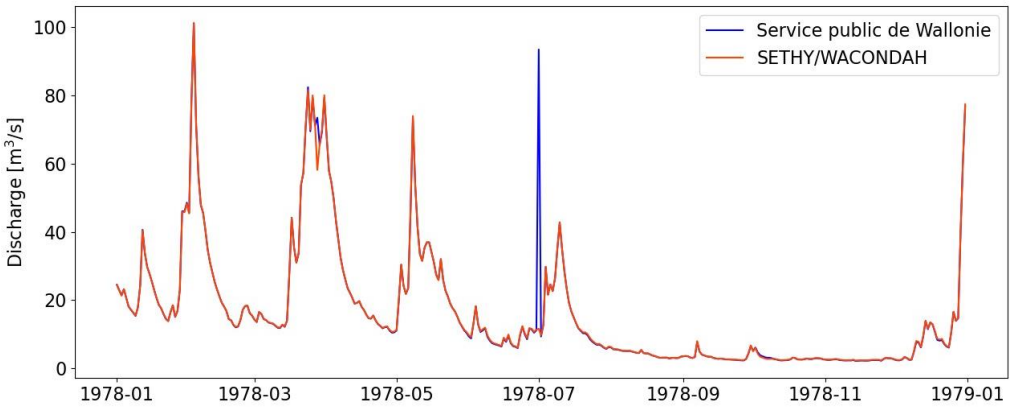

Figure A1 - A comparison of the observed discharge data for year 1978. In blue, the data from Service Public de Wallonie is shown. In red, the data from SETHY/WACONDAH is shown. A major difference was observed on the first of July.

Appendix B: Determination of hydrological year

For determining annual high flow statistics, a hydrological year from the 1st of September to the 31st of August was used. This hydrological year starts on the first day of the month with the lowest mean daily streamflow. As is visible in Figure B1, the lowest mean streamflow is on the 13th of September. Therefore, the hydrological year starts on the first day of September. In case a calendar year would have been used, there would be a larger probability that the annual maximum discharges in two different years are not independent but are related to the same discharge events. By taking a hydrological year from the 1st of September to the 31st of August, this probability is lower.

For annual low flow statistics, a calendar year was used. The year starts on the first day of the month with the maximum mean daily streamflow, which is on the 6th of January (Figure B1).

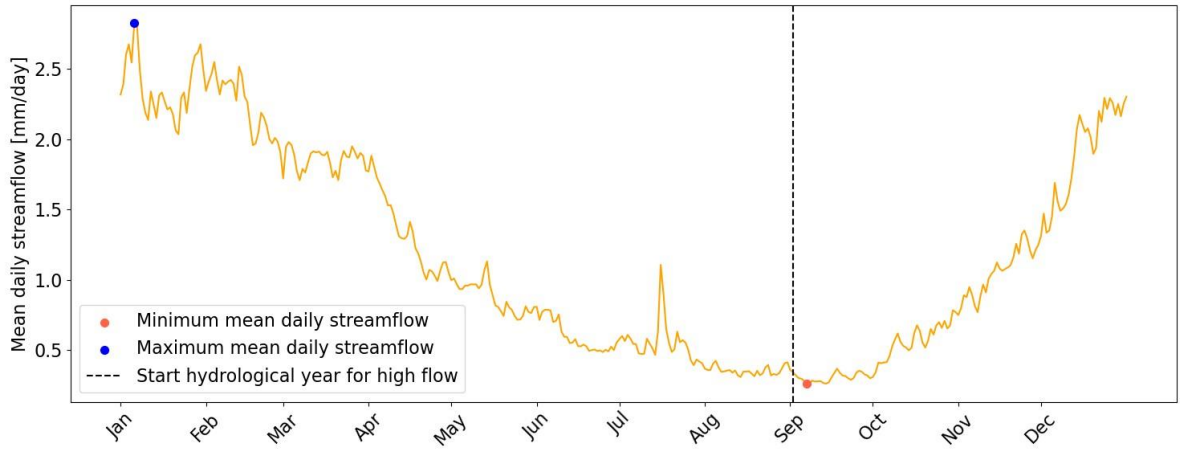

Figure B1 - Mean daily streamflow throughout the calendar year, based on observed streamflow data from 1968-2021. The minimum mean daily streamflow is indicated with a red dot (13th of September). Therefore, the hydrological year for high flows starts on the first of September. The maximum mean daily streamflow is indicated with a blue dot (6th of January).

Appendix C: Testing schemes

This appendix gives an overview of the testing schemes explained in section 3.3.1. Table C1 shows the (number of) years selected in the calibration and validation period for each test for simulation of high flows. For simulation of low flows, the overview is given in Table C2.

Table C1 - Overview of the testing framework used for high flows. The codes of the different tests are made up of the type of flow (High (H)), the type of test (SST or DSST), the type of year (Normal (N) or Extreme (E)) and the climate scenario (Figure 6).

Split-sample tests (SST)				
Code	Calibration period	Nr. of years	Validation period	Nr. of years
H_SST_I	1/1/1996 – 31/12/2020	25	1/1/1970 – 31/12/1994	25
H_SST_II	1/1/1970 – 31/12/1994	25	1/1/1996 – 31/12/2020	25
Differential split-sample tests (DSST)				
Code	Calibration period *All years except selected hydrological years*	Nr. of years	Validation period *Selected hydrological years from 1/9/X-1 to 31/8/X.*	Nr. of years
H_DSST_N_2100Ld	1/1/1969 – 31/12/2021 *Except selected years*	44	1975, 1987, 1990, 1991, 1997, 2001, 2004, 2007, 2010	9
H_DSST_N_2100Hn	1/1/1969 – 31/12/2021 *Except selected years*	43	1983, 1984, 1992, 1994, 1996, 1999, 2003, 2012, 2016, 2019	10
H_DSST_E_2100Ld	1/1/1969 – 31/12/2021 *Except selected years*	51	1980, 2014	2
H_DSST_E_2100Hn	1/1/1969 – 31/12/2021 *Except selected years*	52	1995	1

Table C2 - Similar to Table C1, but for low flows. Type of flow is 'L' (Low).

Split-sample tests (SST)				
Code	Calibration period	Nr. of years	Validation period	Nr. of years
L_SST_I	1/1/1996 – 31/12/2020	25	1/1/1970 – 31/12/1994	25
L_SST_II	1/1/1970 – 31/12/1994	25	1/1/1996 – 31/12/2020	25

Table continues on next page

Differential split-sample tests (DSST)				
Code	Calibration period *All years except selected calendar years*	Nr. of years	Validation period *Selected calendar years from 1/1/X to 31/12/X.*	Nr. of years
L_DSST_N_2100Ln	1/1/1969 – 31/12/2021 *Except selected years*	52	1971	1
L_DSST_N_2100Hd	1/1/1969 – 31/12/2021 *Except selected years*	51	1976, 2020	2
L_DSST_E_2100Ln	1/1/1969 – 31/12/2021 *Except selected years*	52	2020	1
L_DSST_E_2100Ld	1/1/1969 – 31/12/2021 *Except selected years*	51	1976, 2020	2

Table C3 shows an overview for the tests for simulation of high and low flows. Figure C1 shows the calibration and validation periods for simulation of high and low flows, categorized by years selected based on high flows and years selected based on low flows. Orange periods are periods that were selected based on both high flows and low flows. Because of this overlap and the different hydrological year used for high and low flows, the number of years in Table C3 is not the sum of the years selected for high flows and low flows (indicated by *).

Table C3 - Similar to Table C1, but for high and low flows. Type of flow is 'B' (Both). Blue years in the validation period are selected based on the meteorological indicator for high flows, red years based on the meteorological indicator for low flows.

Split-sample tests (SST)				
Code	Calibration period	Nr. of years	Validation period	Nr. of years
B_SST_I	1/1/1996 – 31/12/2020	25	1/1/1970 – 31/12/1994	25
B_SST_II	1/1/1970 – 31/12/1994	25	1/1/1996 – 31/12/2020	25
Differential split-sample tests (DSST)				
Code	Calibration period *All years except selected hydrological & calendar years*	Nr. of years	Validation period *Selected hydrological years from 1/9/X-1 to 31/8/X.* *Selected calendar years from 1/1/X to 31/12/X*	Nr. of years
B_DSST_N_2100Ld	1/1/1969 – 31/12/2021 *Except selected years*	39⅓	1973, 1975, 1987, 1990, 1991, 1996, 1997, 1999, 2001, 2002, 2004, 2007, 2009, 2010, 2012	9 + 6 = 13⅓*
B_DSST_N_2100Hn	1/1/1969 – 31/12/2021 *Except selected years*	41	1975, 1983, 1984, 1992, 1994, 1996, 1999, 2003, 2011, 2012, 2016, 2019, 2019	10 + 3 = 12*
B_DSST_N_2100Ln	1/1/1969 – 31/12/2021 *Except selected years*	43	1971, 1981, 1987, 1990, 2000, 2001, 2004, 2007, 2008, 2010	9 + 1 = 10
B_DSST_N_2100Hd	1/1/1969 – 31/12/2021 *Except selected years*	40	1976, 1981, 1983, 1987, 1988, 1990, 2000, 2001, 2004, 2007, 2008, 2019, 2020	11 + 2 = 13
B_DSST_E_2100Ld	1/1/1969 – 31/12/2021 *Except selected years*	49	1976, 1980, 2014, 2020	2 + 2 = 4
B_DSST_E_2100Ln	1/1/1969 – 31/12/2021 *Except selected years*	50⅓	1995, 2020, 2021	2 + 1 = 2⅔*

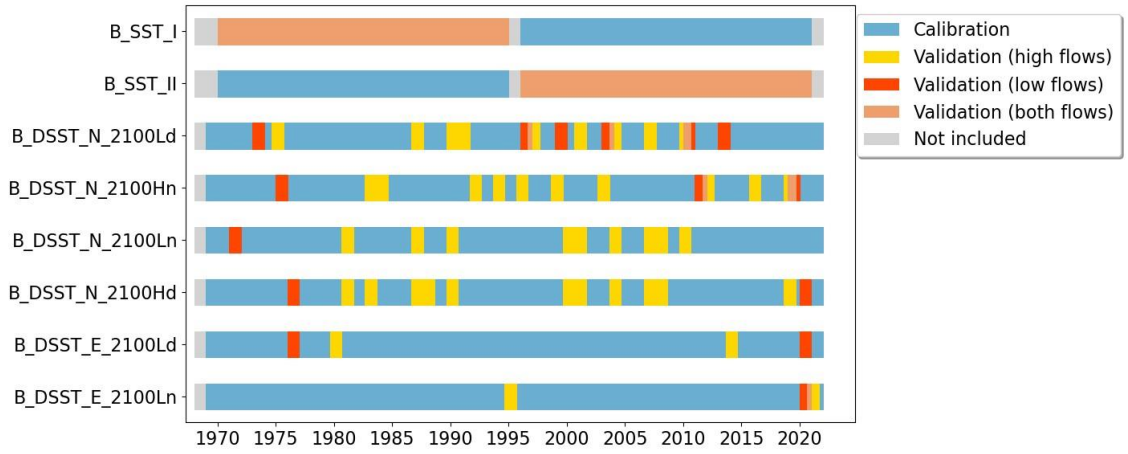

Figure C1 - Visualisation of the periods used for calibration (blue) and validation in the different tests for simulation of high and low flows. In yellow, validation periods selected based on meteorological indicators for high flows are visualized. Red periods are selected based on meteorological indicators for low flows. Periods that are selected based on the meteorological indicator for high flows and based on the meteorological indicator for low flows are given in orange. Grey periods are not included in calibration or validation. The codes of the different tests are made up of the type of flow (Both (B)), the type of test (SST or DSST), the type of year (Normal (N) or Extreme (E)) and the climate scenario (Figure 6).

Appendix D: SCEM-UA Calibration algorithm

The SCEM-UA optimization algorithm shows to the able to give a stable parameter set. After 5000 iterations, convergence and a global optimum is found.

Figure D1 and Figure D2 show the parameter values that were evaluated over each iteration for calibration of the HBV and GR6J model on testing scheme *'H_SST_I'*. It is visible that after 5000 iterations, the algorithm finds a stable model output, showing that convergence is found.

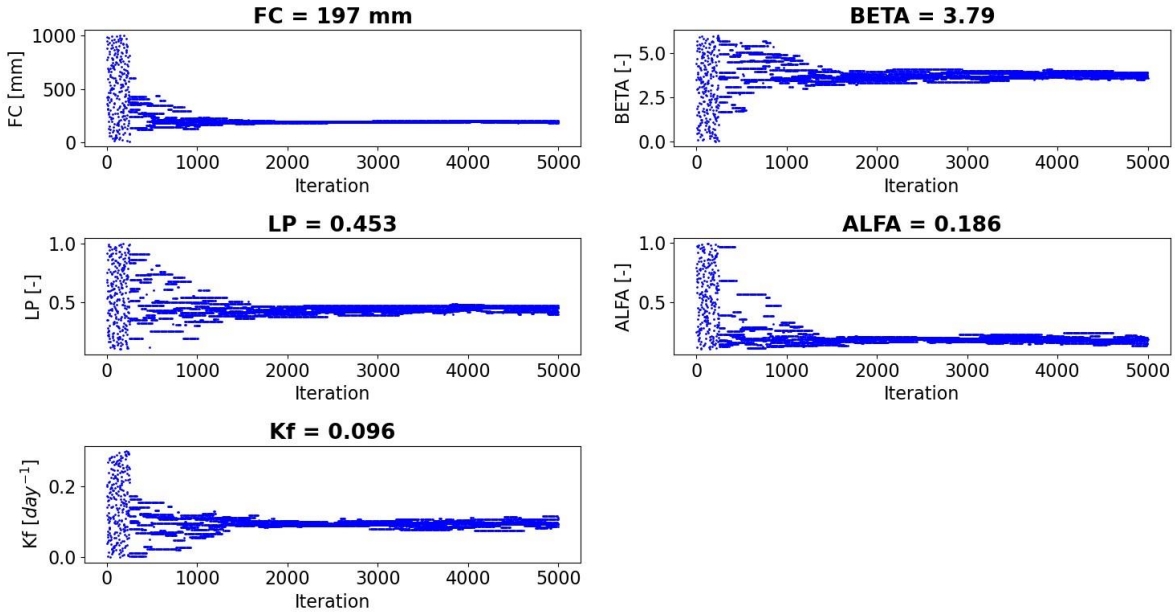

Figure D1 - Parameters sets for each iteration in the calibration of the HBV model on testing scheme 'H_SST_I' with calibration settings described in section 3.3.2.1.

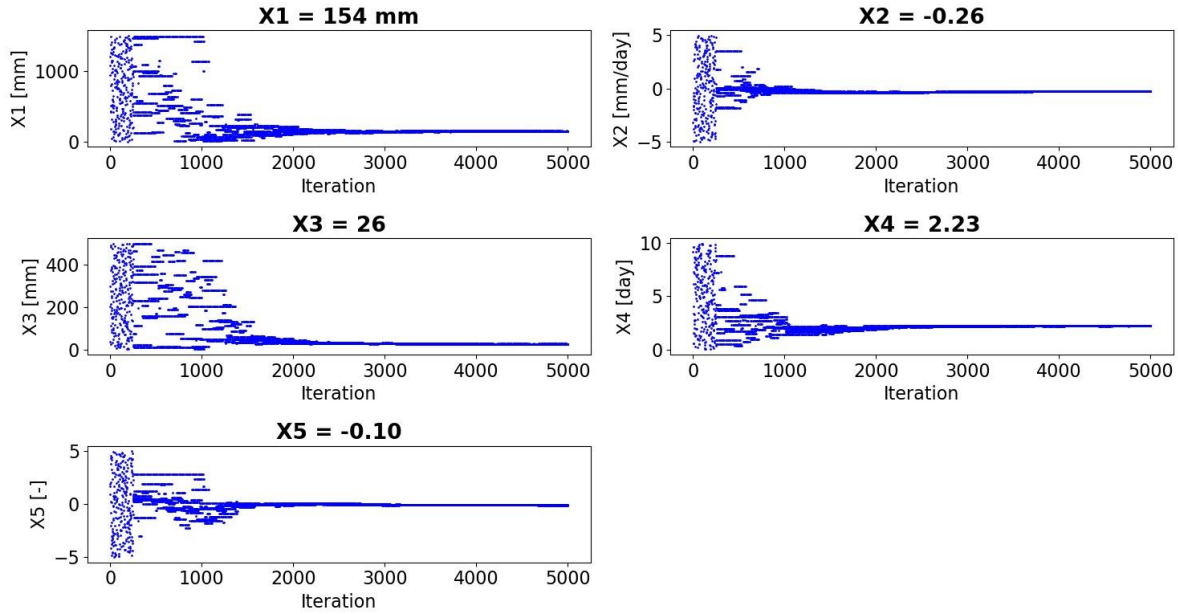

Figure D2 - Similar to Figure D1, but for the GR6J model.

It was tried to decrease calibration time by including the Gelman-Rubin convergence termination criterion in the SCEM-UA algorithm. A Gelman-Rubin convergence score (GR) close to 1 indicates that parameters have converged to a stable value. Following the recommendations of Vrugt et al. (2003), convergence is found when GR is lower than 1.2 for all parameters. Including the Gelman-Rubin convergence termination criterion thus means that calibration is stopped when all parameters have a GR<1.2.

Figure D3 shows the parameter sets evaluated over each iteration for the same calibration as Figure D1 (HBV model), but with the inclusion of the Gelman-Rubin convergence criterion. After 300 iterations, the algorithm was terminated, as GR<1.2 for each parameter. However, it is visible that convergence is not yet found; parameters still vary over a large part of the parameter space. It was therefore decided to exclude the Gelman-Rubin convergence termination criterion for model calibration.

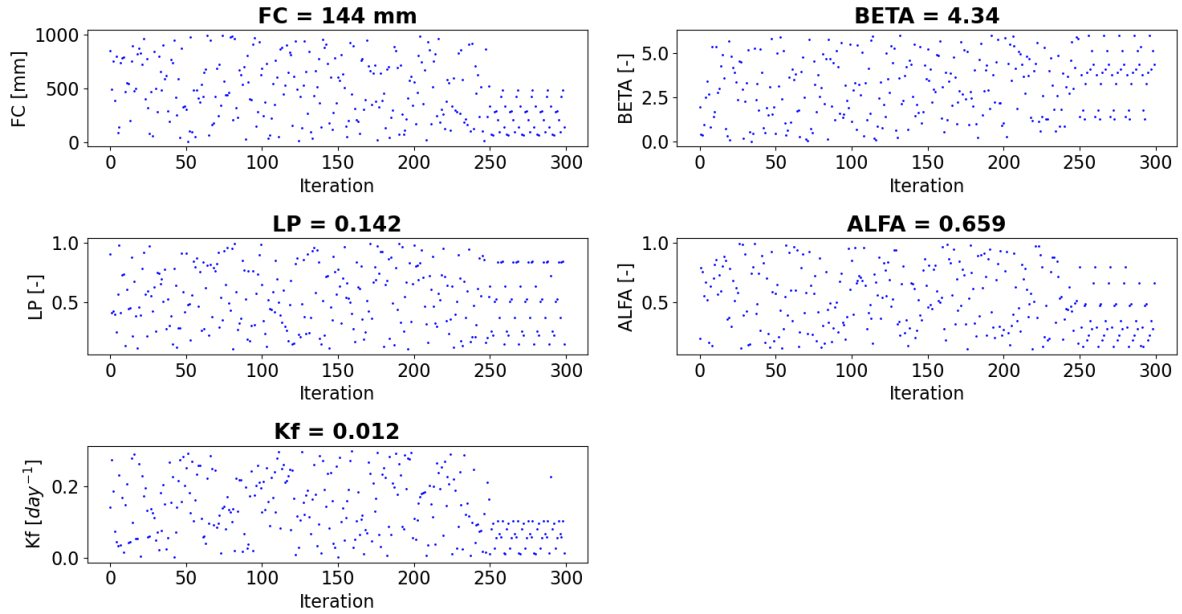

Figure D3 - Parameters sets for each iteration in the calibration of the HBV model on testing scheme 'H_SST_I' with calibration settings described in section 3.3.2.1, but with inclusion of the Gelman-Rubin convergence termination criterion.

Appendix E: Selection of calibration parameters

This appendix shows the results of the sensitivity analysis used to select the parameters for calibration. The sensitivity analysis provides insight in the parameters to which the objective function is insensitive, meaning that these parameters hardly impact model performance. Insensitive parameters were kept outside calibration, to limit the number of parameters in the optimization problem. As a result, the SCEM-UA algorithm was better able to find convergence efficiently.

A univariate sensitivity analysis was done, which means that parameters were varied one by one. First, calibration was done with all parameters. Then, each parameter was replaced by 1000 values ranging from 50% and 150% of the found optimal values. For each new parameter set, the multi-objective function was determined. Based on the sensitivity of the multi-objective function to a change in parameter, insensitive parameters were selected. The results of the sensitivity analysis were described per model and type of flow.

E.1. HBV, High flows

Figure E1 shows the results of the univariate sensitivity analysis of the HBV model, for simulation of high flows. Multi-objective function y_w hardly changes for a change in parameters K_s, $PERC$ and $Cflux$. These three parameters are thus determined to be the least sensitive. Therefore, these parameters were fixed at default values recommended by SMHI (2006), being 0.005 day^{-1}, 1 mm/day and 1 mm/day, respectively.

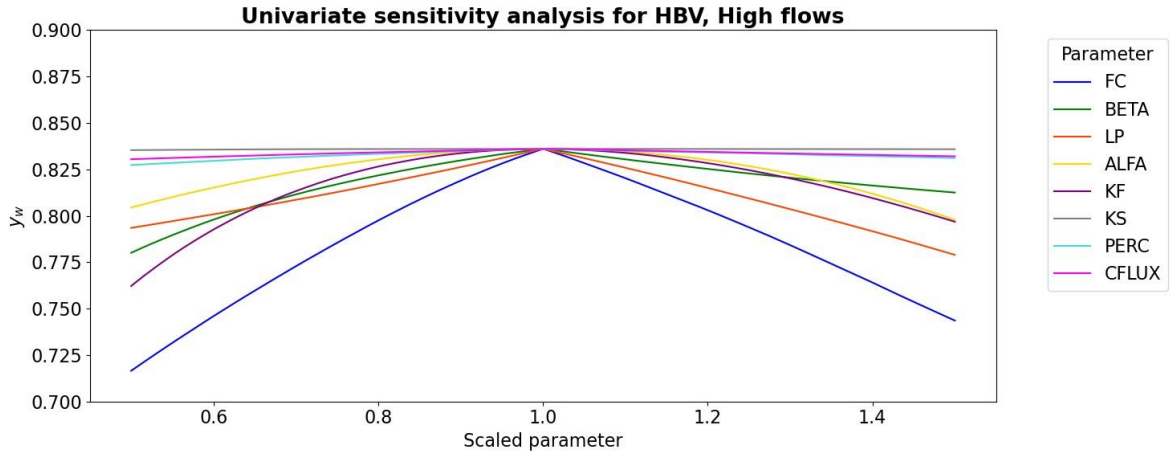

Figure E1 - Results of the univariate sensitivity analysis of the HBV model for simulation of high flows. The x-axis shows the parameter value scaled to its original value.

With fixed values for parameters K_s, $PERC$ and $Cflux$, calibration of the remaining 5 parameters shows a stable behaviour regarding objective function y_w and the resulting parameter set. Table E1 shows five outcomes of calibration for testing scheme *'SST_I'*. A stable objective function and parameter set is observed, indicating that the SCEM-UA found a global optimum.

Table E1 - Objective function y_w and parameter set of the HBV model, calibrated 5 times using testing scheme 'SST_I', objective function y_w. Parameters K_s, $PERC$ and $Cflux$ were fixed at 0.005 day^{-1}, 1 mm/day and 1 mm/day, respectively.

#	y_w	FC [mm]	β [-]	LP [-]	α [-]	K_f [day^{-1}]
1	0.83	206	3.95	0.488	0.178	0.098
2	0.83	193	3.64	0.430	0.203	0.088
3	0.83	199	3.77	0.458	0.172	0.101
4	0.83	195	3.71	0.441	0.180	0.097
5	0.83	196	3.82	0.449	0.198	0.089

E.2. HBV, Low flows

Figure E2 shows the results of the univariate sensitivity analysis of the HBV model, for simulation of low flows. Multi-objective function y_{inv} hardly changes for a change in parameters α, K_f and $Cflux$. These three parameters are thus determined to be the least sensitive. Therefore, these parameters were fixed at default values recommended by SMHI (2006), being 1, 0.005 day^{-1} and 1 mm/day, respectively.

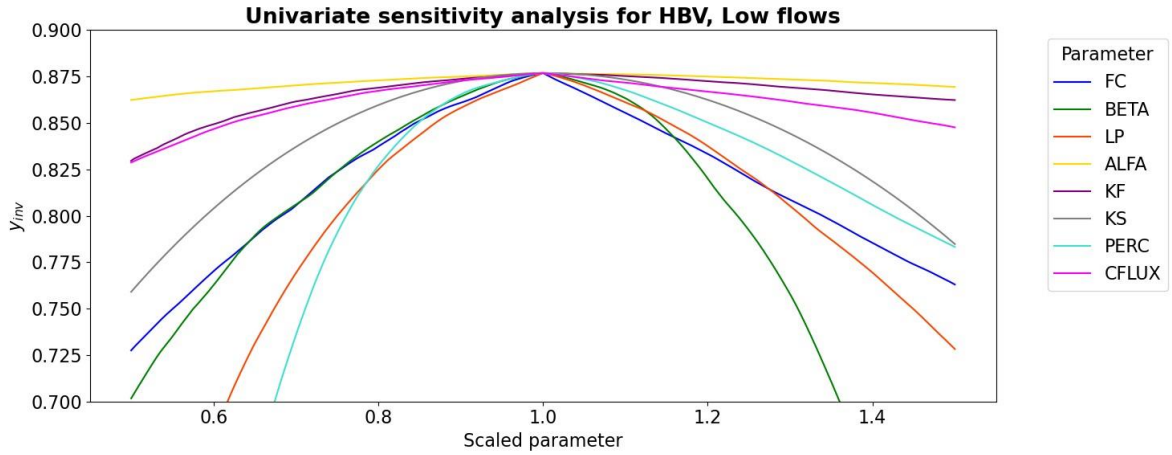

Figure E2 - Results of the univariate sensitivity analysis of the HBV model for simulation of low flows. The x-axis shows the parameter value scaled to its original value.

With fixed values for parameters α, K_f and $Cflux$, calibration of the remaining 5 parameters shows a stable behaviour regarding objective function y_{inv} and the resulting parameter set. Table E2 shows five outcomes of calibration for testing scheme *'SST_I'*. A stable objective function and parameter set is observed, indicating that the SCEM-UA found a global optimum.

Table E2 - Objective function y_{inv} and parameter set of the HBV model, calibrated 5 times using testing scheme 'SST_I', objective function y_{inv}. Parameters K_f, $Cflux$ and α were fixed at 0.005 day^{-1}, 1 mm/day and 1, respectively.

#	y_{inv}	FC [mm]	β [-]	LP [-]	K_s [day^{-1}]	PERC [mm/day]
1	0.86	210	1.20	0.270	0.0308	0.353
2	0.86	210	1.22	0.275	0.0308	0.352
3	0.86	210	1.20	0.269	0.0310	0.352
4	0.86	210	1.21	0.270	0.0309	0.351
5	0.86	210	1.21	0.274	0.0313	0.354

E.3. HBV, High and low flows

For simulation of high and low flows, no sensitivity analysis was done. Instead, the selection of parameters for high flows and for low flows was analysed. All parameters in the HBV model are important for simulation of high flows and/or low flows, except parameter $Cflux$ (Table E3). It was therefore expected that selecting every parameter except $Cflux$ would result in a global optimum. $Cflux$ was thus fixed at the default value of 1 mm/day (SMHI, 2006).

Table E3 - Selected parameters in the HBV model for high flows and for low flows.

Selected for...	FC [mm]	β [-]	LP [-]	α [-]	K_f [day^{-1}]	K_s [day^{-1}]	PERC [mm/day]	Cflux [mm/day]
High flows	Yes	Yes	Yes	Yes	Yes	No	No	No
Low flows	Yes	Yes	Yes	No	No	Yes	Yes	No

With a fixed value for parameter $Cflux$, calibration of the remaining 7 parameters shows a stable behaviour regarding objective function y_{comb} and the resulting parameter set. Table E4 shows five

outcomes of calibration for testing scheme *'SST_I'*. A stable objective function and parameter set is observed, indicating that the SCEM-UA found a global optimum.

Table E4 - Objective function y_{comb} and parameter set of the HBV model, calibrated 5 times using testing scheme 'SST_I', objective function y_{comb}. Parameter $Cflux$ was fixed at 1 mm/day.

#	y_{comb}	FC [mm]	β [-]	LP [-]	α [-]	K_f [day^{-1}]	K_s [day^{-1}]	PERC [mm/day]
1	0.84	213	1.40	0.314	0.539	0.0208	0.0282	0.356
2	0.84	229	1.67	0.410	0.627	0.0149	0.0285	0.348
3	0.84	214	1.40	0.316	0.544	0.0204	0.0283	0.353
4	0.84	213	1.39	0.312	0.551	0.0196	0.0282	0.358
5	0.84	214	1.39	0.314	0.548	0.0199	0.0282	0.354

E.4. GR6J, High flows

Figure E3 shows the results of the univariate sensitivity analysis of the GR6J model, for simulation of high flows. Multi-objective function y_w hardly changes for a change in parameter X_6. This parameter was thus determined to be the least sensitive and thus is fixed at 10mm, which is the median of the parameter space for this parameter (Pushpalatha et al., 2011).

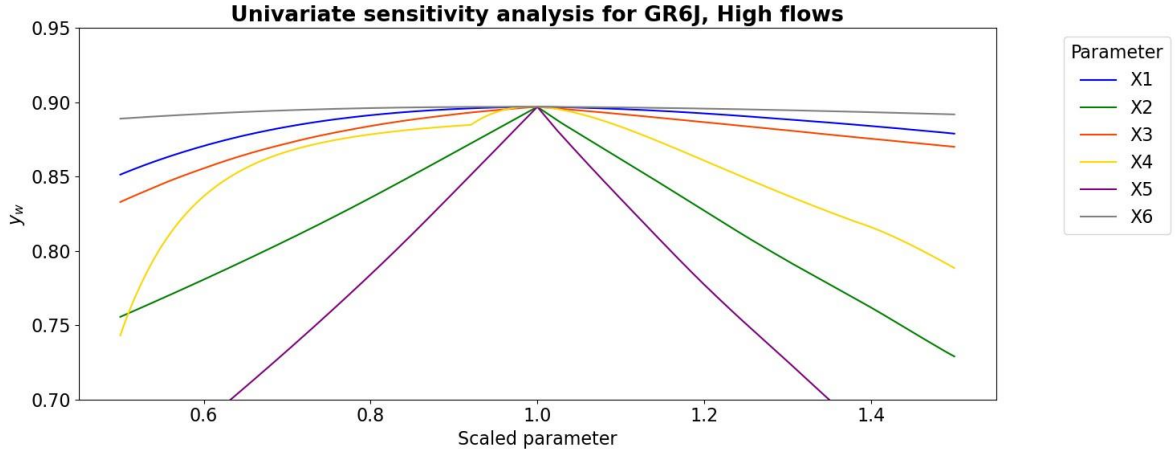

Figure E3 - Results of the univariate sensitivity analysis of the GR6J model for simulation of high flows. The x-axis shows the parameter value scaled to its original value.

With a fixed value for parameter X_6, calibration of the remaining 5 parameters shows a stable behaviour regarding objective function y_w. Table E5 shows five outcomes of calibration for testing scheme *'SST_I'*. However, the parameter sets show large variations. Especially parameters X_2 and X_5, related to groundwater exchange, show large variations. Despite this, still parameters X_1 to X_5 were selected for calibration.

Table E5 - Objective function y_w and parameter set of the GR6J model, calibrated 5 times using testing scheme 'SST_I', objective function y_w. Parameter X_6 was fixed at 10 mm.

#	y_w	X_1 [mm]	X_2 [mm/day]	X_3 [mm]	X_4 [day]	X_5 [-]
1	0.89	157	0.175	27.0	2.21	1.52
2	0.89	155	-0.0474	26.4	2.22	-3.03
3	0.88	178	-0.278	23.3	2.22	-0.0188
4	0.89	159	0.317	27.1	2.22	1.08
5	0.89	156	-0.151	26.6	2.21	-0.564

E.5. GR6J, Low flows

Figure E4 shows the results of the univariate sensitivity analysis of the GR6J model, for simulation of low flows. A substantial change in multi-objective function y_{inv} is visible for a change in each parameter. It was therefore chosen to select all parameters for calibration.

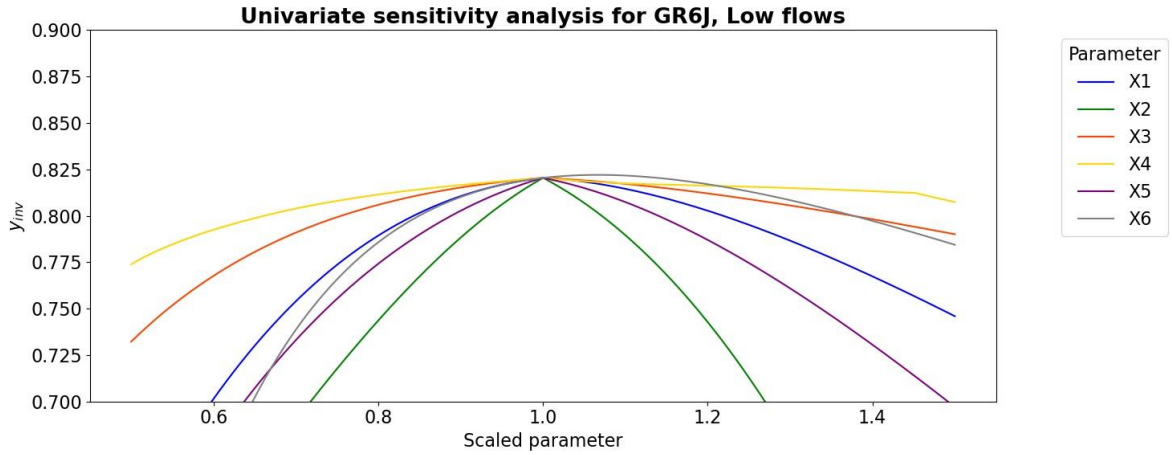

Figure E442 - Results of the univariate sensitivity analysis of the GR6J model for simulation of low flows. The x-axis shows the parameter value scaled to its original value.

Calibration results of 5 times calibration of the 6 parameters for testing scheme *'SST_I'* are shown in Table E6. A stable behaviour regarding objective function y_{inv} is observed. However, the parameter sets show large variations. Especially parameters X_2 and X_5, related to groundwater exchange, show large variations. Next to that, a low value for X_1 leads to a high value of X_3 and vice versa.

Table E6 - Objective function y_{inv} and parameter set of the GR6J model, calibrated 5 times using testing scheme 'SST_I', objective function y_{inv}.

#	y_{inv}	X_1 [mm]	X_2 [mm/day]	X_3 [mm]	X_4 [day]	X_5 [-]	X_6 [mm]
1	0.82	320	-0.327	5.93	2.06	0.226	7.98
2	0.83	231	-0.383	6.26	2.02	0.265	10.2
3	0.81	200	-0.171	7.15	1.90	-0.314	16.6
4	0.81	94.9	-0.577	280	1.87	0.0670	0.167
5	0.83	207	-0.373	6.03	2.08	0.245	11.7

E.6. GR6J, Both flows

Just as for the HBV model, the selection of parameters was based on the selected parameters for simulation of high flows and simulation of low flows (Table E7). All parameters in the GR6J model were important for simulation of high flows and/or low flow. Therefore all 6 parameters were selected for the simulation of both flows.

Table E7 - Selected parameters in the GR6J model for high flows and for low flows.

	X_1 [mm]	X_2 [mm/day]	X_3 [mm]	X_4 [day]	X_5 [-]	X_6 [mm]
Selected for high flows	Yes	Yes	Yes	Yes	Yes	No
Selected for low flows	Yes	Yes	Yes	Yes	Yes	Yes

Table E8 shows the calibration results when using the SCEM-UA algorithm with objective function y_{comb}. A stable objective function y_{comb} was found, varying between 0.81 and 0.84. Just as for simulation of high flows or simulation of low flows, variability is observed between the parameter sets.

Table E8 - Objective function y_{comb} and parameter set of the GR6J model, calibrated 5 times using testing scheme 'SST_I', objective function y_{comb}.

#	ycomb	X_1 [mm]	X_2 [mm/day]	X_3 [mm]	X_4 [day]	X_5 [-]	X_6 [mm]
1	0.82	51.5	0.155	386	2.21	1.610	1.45
2	0.82	63.2	-0.299	326	2.27	-0.264	1.29
3	0.84	195	-0.349	16.4	2.30	0.143	11.5
4	0.82	64.5	-0.318	325	2.24	-0.224	1.38
5	0.81	42.0	0.326	411	2.24	0.968	1.66

Appendix F: Changes in precipitation and potential evapotranspiration

The results described in section 4.1.2.2 showed that the meteorological indicator for low flows is expected to increase for all future scenarios. Figure F1 and Figure F2 show that this increase in potential precipitation deficit is both due to an increase in potential evapotranspiration as well as a decrease in precipitation.

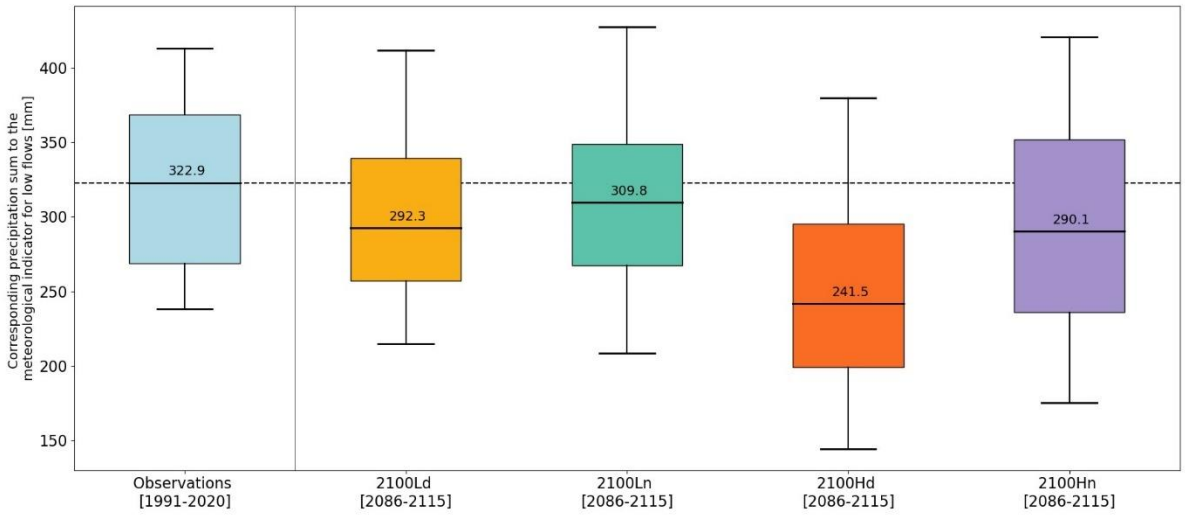

Figure F1 - Boxplot of the precipitation sum corresponding to the meteorological indicator for low flows, based on observations (blue) and the bias-corrected values for each future scenario (other colours). The horizontal lines in the boxes represent the median value, the boxes represent the 25-75% data range and the whiskers represent the 5-95% data range. The values in the boxes are the median values.

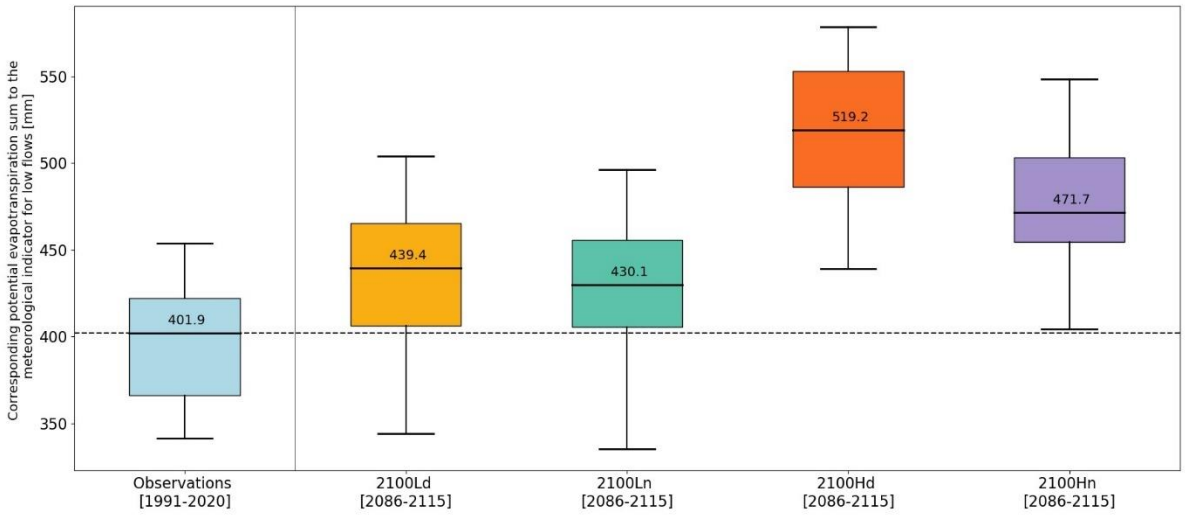

Figure F2 - Similar to Figure F1, but for the potential evapotranspiration sum.

Appendix G: Calibration performance

Table G1 shows the performance of the HBV model in the calibration period, when calibrated on the calibration period of RQ3 (blue in section 3.3.1) and when calibrated on the validation period of RQ3 (orange in 3.3.1). These are the periods that were selected to resemble future conditions. In general, model performance in the calibration period is higher when calibrated on the periods that resemble future conditions.

*Table G1 - Performance of the **HBV** model calibrated on the testing schemes for simulation of A) high, B) low and C) high and low flows. The names of the testing schemes are explained in section 3.3.1. For high flows, performance is given by objective function y_w and corresponding NS_w and RVE. For low flows, performance is given by objective function y_{inv} and corresponding NS_{inv} and RVE. For simulation of high and low flows, performance is given by objective function y_{comb} and corresponding NS_w, NS_{inv} and RVE. Blue and orange columns show the performance when calibrated on periods that do not resemble future conditions (blue) and periods that resemble future conditions (orange). The last column (Δy) shows the difference in multi-objective function. Red and green numbers indicate a degradation and increase of performance in the orange calibration period compared to the blue calibration period, respectively.*

A) Simulation of high flows									
Testing scheme	Calibration on not-selected years				Calibration on selected years				Δy_w [-]
	y_w [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	y_w [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	
HBV_H_DSST_N_2100Ld	0.82	0.82	-	$+3.2 \cdot 10^{-3}$	0.81	0.81	-	$-2.8 \cdot 10^{-3}$	-0.01
HBV_H_DSST_N_2100Hn	0.81	0.81	-	$+2.7 \cdot 10^{-3}$	0.83	0.83	-	$+1.6 \cdot 10^{-3}$	+0.02
HBV_H_DSST_E_2100Ld	0.82	0.82	-	$-1.1 \cdot 10^{-3}$	0.84	0.84	-	$-8.7 \cdot 10^{-3}$	+0.02
HBV_H_DSST_E_2100Hn	0.81	0.81	-	$+2.6 \cdot 10^{-3}$	0.87	0.87	-	$+1.3 \cdot 10^{-2}$	+0.06
B) Simulation of low flows									
Testing scheme	Calibration on not-selected years				Validation period				Δy_{inv} [-]
	y_{inv} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	y_{inv} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	
HBV_L_DSST_N_2100Ln	0.85	-	0.85	$-5.7 \cdot 10^{-3}$	0.89	-	0.89	$+2.2 \cdot 10^{-2}$	+0.04
HBV_L_DSST_N_2100Hd	0.84	-	0.84	$+7.3 \cdot 10^{-4}$	0.92	-	0.92	$-5.8 \cdot 10^{-3}$	+0.08
HBV_L_DSST_E_2100Ln	0.85	-	0.85	$+1.2 \cdot 10^{-3}$	0.97	-	0.97	$+3.4 \cdot 10^{-4}$	+0.12
HBV_L_DSST_E_2100Ld	0.84	-	0.84	$+5.4 \cdot 10^{-3}$	0.92	-	0.92	$-3.4 \cdot 10^{-3}$	+0.12
C) Simulation of high and low flows									
Testing scheme	Calibration on not-selected years				Calibration on selected years				Δy_{comb} [-]
	y_{comb} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	y_{comb} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	
HBV_B_DSST_N_2100Ld	0.83	0.80	0.87	$+2.3 \cdot 10^{-3}$	0.81	0.83	0.79	$+6.2 \cdot 10^{-3}$	-0.02
HBV_B_DSST_N_2100Hn	0.82	0.79	0.86	$+1.7 \cdot 10^{-3}$	0.84	0.81	0.87	$+1.0 \cdot 10^{-2}$	+0.02
HBV_B_DSST_N_2100Ln	0.83	0.80	0.87	$+3.1 \cdot 10^{-1}$	0.80	0.82	0.86	+4.1	-0.03
HBV_B_DSST_N_2100Hd	0.82	0.80	0.83	$+7.0 \cdot 10^{-3}$	0.85	0.81	0.89	$+3.7 \cdot 10^{-2}$	+0.03
HBV_B_DSST_E_2100Ld	0.81	0.79	0.84	$+1.6 \cdot 10^{-3}$	0.89	0.87	0.93	$+7.0 \cdot 10^{-1}$	+0.08
HBV_B_DSST_E_2100Ln	0.84	0.82	0.85	$+9.6 \cdot 10^{-4}$	0.87	0.84	0.90	$+6.2 \cdot 10^{-3}$	+0.03

Table G2 shows the performance of the GR6J model in the calibration period, when calibrated on the calibration period of RQ3 (blue in section 3.3.1) and when calibrated on the validation period of RQ3 (orange in 3.3.1). Again, in general, model performance in the calibration period is higher when calibrated on the periods that resemble future conditions.

*Table G2 - Similar to Table G1, but for the **GR6J** model.*

A) Simulation of high flows									
Testing scheme	Calibration on not-selected years				Calibration on selected years				Δy_w [-]
	y_w [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	y_w [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	
GR6J_H_DSST_N_2100Ld	0.90	0.90	-	$-2.1 \cdot 10^{-3}$	0.90	0.90	-	$-1.1 \cdot 10^{-2}$	0.00
GR6J_H_DSST_N_2100Hn	0.91	0.91	-	$+3.1 \cdot 10^{-3}$	0.94	0.94	-	$+5.3 \cdot 10^{-3}$	+0.03
GR6J_H_DSST_E_2100Ld	0.92	0.92	-	$+5.0 \cdot 10^{-3}$	0.89	0.89	-	$-1.3 \cdot 10^{-2}$	-0.03
GR6J_H_DSST_E_2100Hn	0.91	0.91	-	$-3.3 \cdot 10^{-1}$	0.96	0.96	-	$-1.3 \cdot 10^{-1}$	+0.05
B) Simulation of low flows									
Testing scheme	Calibration on not-selected years				Validation period				Δy_{inv} [-]
	y_{inv} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	y_{inv} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	
GR6J_L_DSST_N_2100Ln	0.82	-	0.82	$+7.7 \cdot 10^{-3}$	0.87	-	0.87	$+1.9 \cdot 10^{-3}$	+0.05
GR6J_L_DSST_N_2100Hd	0.77	-	0.77	$-1.2 \cdot 10^{-2}$	0.89	-	0.89	$+3.4 \cdot 10^{-2}$	+0.12
GR6J_L_DSST_E_2100Ln	0.79	-	0.79	$+4.5 \cdot 10^{-4}$	0.93	-	0.93	$+3.1 \cdot 10^{-2}$	+0.14
GR6J_L_DSST_E_2100Ld	0.77	-	0.77	$-2.3 \cdot 10^{-2}$	0.89	-	0.89	$+1.0 \cdot 10^{-2}$	+0.12
C) Simulation of high and low flows									
Testing scheme	Calibration on not-selected years				Calibration on selected years				Δy_{comb} [-]
	y_{comb} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	y_{comb} [-]	NS_w [-]	NS_{inv} [-]	RVE [%]	
GR6J_B_DSST_N_2100Ld	0.86	0.90	0.83	$-1.2 \cdot 10^{-2}$	0.78	0.85	0.71	$-1.7 \cdot 10^{-2}$	-0.07
GR6J_B_DSST_N_2100Hn	0.85	0.91	0.79	$+1.3 \cdot 10^{-3}$	0.84	0.87	0.80	$-2.4 \cdot 10^{-3}$	-0.01
GR6J_B_DSST_N_2100Ln	0.85	0.90	0.79	$-3.0 \cdot 10^{-3}$	0.83	0.88	0.77	$+1.2 \cdot 10^{-2}$	-0.02
GR6J_B_DSST_N_2100Hd	0.80	0.85	0.75	$+4.2 \cdot 10^{-3}$	0.88	0.88	0.88	$-7.8 \cdot 10^{-3}$	+0.08
GR6J_B_DSST_E_2100Ld	0.81	0.89	0.74	$-6.7 \cdot 10^{-1}$	0.88	0.88	0.88	$+4.6 \cdot 10^{-2}$	+0.07
GR6J_B_DSST_E_2100Ln	0.80	0.83	0.78	$+6.1 \cdot 10^{-1}$	0.92	0.95	0.89	$+7.2 \cdot 10^{-3}$	+0.12

Appendix H: Projected impact of climate change

Figure H1 and Figure H2 show the projected impact of climate change on the annual maximum daily discharge and annual minimum 7-day mean discharge in all scenarios described in Table 7, as well as in the reference climate. The projected impact simulated with the 'wflow_sbm' model by Deltares (Buitink et al., 2023) is given as well. These projected impacts were used as comparison.

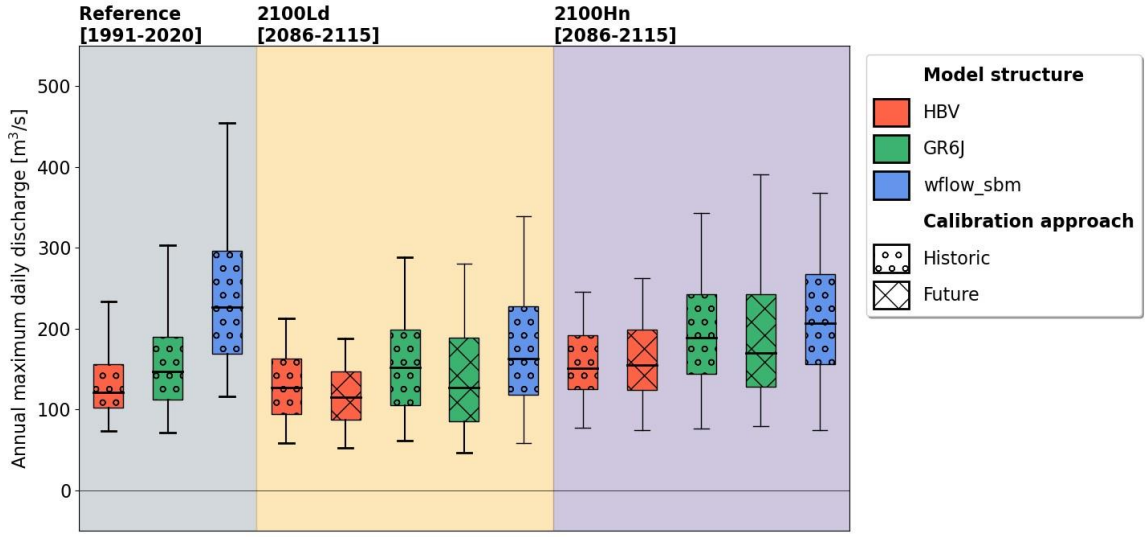

Figure H1 - Boxplot of the annual maximum daily discharge [Q_{max}, m³/s] projected for 8 scenarios (combinations of climate scenarios, hydrological model structures and calibration approaches (Table 7)), and the effects projected by Deltares using wflow_sbm (Buitink et al., 2023). The three most left boxplots (in grey) were based on the reference climate of KNMI (1991-2020). The boxplots in yellow and purple are based on climate scenario '2100Ld' and '2100Hn', respectively. These boxplots show the projected Q_{max} for the period 2086-2115. Red, green and blue boxplots show the projected Q_{max} simulated with the HBV, GR6J and wflow_sbm model, respectively. Dotted boxplots show the projected Q_{max} with models calibrated on historic period 1996-2020, boxplots with diagonal stripes show the projected Q_{max} with models calibrated on periods resembling future conditions. The horizontal lines in the boxes represent the median value, the boxes represent the 25-75% data range and the whiskers represent the 5-95% data range.

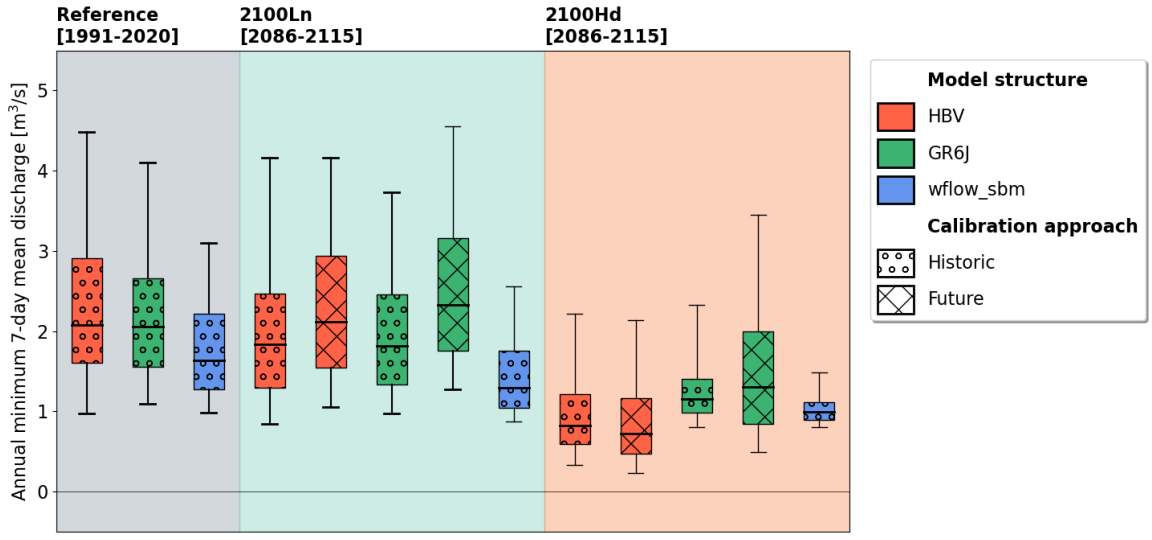

Figure H2 - Similar to Figure H1, but for the annual minimum 7-day mean discharge [$Q7_{min}$, m³/s]. Boxplots in green and orange are based on climate scenario '2100Ln' and '2100Hd', respectively.