The impact of evaporation data calibration on regional hydrological model performance

E. Poppelier

Delft University of Technology

TU Delft

The impact of evaporation data calibration on regional hydrological model performance

A case study of the Geul, the Netherlands

by

E. Poppelier

Student number: 4552814
Date : May 23, 2023
Thesis committee: Dr. Markus Hrachowitz TU Delft
Dr. Laurène Bouaziz Deltares
Dr. Elisa Ragno TU Delft

Cover: 'De geul bij Cottessen', Pieter de Josselin de Jong, ca. 1890-1900.

TU Delft

Summary

Climate change-induced changes in weather patterns call for the development of hydrological models that perform well under increasingly extreme and varied conditions. Multiple research studies have demonstrated that hydrological models perform poorly when applied to climate conditions that differ from those during the calibration period of the model (Duethmann, Bloschl, & Parajka, 2020). The need for robust hydrological models was emphasised after the Geul catchment, and large parts of Belgium and Germany flooded in July 2021. During this flood, the hydrological model of the Geul was undergoing maintenance. However, the hydrologists from Waterboard Limburg (WL) indicated that the current model would not have been able to correctly forecast the flood under such an extreme rain event anyway (Expertise Netwerk Waterveiligheid, 2021).

The underlying assumption in this research is that hydrological models cannot perform to the same standard during changing weather conditions because they are overfitted during the calibration period. Meaning that the parameter values are the mathematical best fit for streamflow predictions, but do not represent the internal hydrological processes between precipitation and discharge. One way of trying to improve the internal processes of a hydrological model is to create more relations between these processes and external measurements (Kirchner, 2006). In this research, a secondary calibration data set is used, namely evaporation data.

There are two goals in this thesis, first building a hydrological model of the Geul which represents the streamflow response of the catchment satisfactory. Second, comparing if a difference in calibration methodology, specifically comparing calibration on discharge with calibration on both discharge and evaporation, would improve the predictive power of the model. These goals are set up to be able to answer the main research question of this thesis:

Is the predictive power of a discharge calibrated hydrological model of the Geul catchment in the Netherlands, greater than the predictive power of an evaporation calibrated model, in addition to discharge?

The methodology used in this thesis to calibrate and evaluate is Generalised Likelihood Uncertainty Estimation (GLUE). Within GLUE the concept of a single best calibration parameter set is abandoned in favour of selecting a number of the best behaviour models (Beven & Binley, 1992), as to represent the uncertainty in the input and calibration data and model design in the output of the model as well. The objective function used to determine these behaviour models was a combination of the Nash-Sutcliff efficiency (NSE) (Nash & Sutcliffe, 1970) and the NSE log for both streamflow and evaporation.

The first goal, building a hydrological model of the Geul which represents the stream flow response of the catchment satisfactory, is successfully reached. The model calibrated on streamflow (Model Q) and the model calibrated on both streamflow and evaporation (Model QE) were able to predict streamflow satisfactorily, with DeQ scores between 0.632 and 0.649 across all runs. Further analysis showed that both Model Q and Model QE scored good on the monthly runoff coefficient, achieving NSE scores between 0.733 and 0.801.

The second goal was to compare the performance of Model Q and Model QE, with which the main research question of this thesis could be answered. Model QE outperformed Model Q on monthly runoff coefficient, monthly average evaporation and cumulative evaporation, increasing the NSE score of Model Q of the monthly runoff coefficient from 0.774 to 0.801, the NSE log score of the monthly average evaporation from 0.884 to 0.891 and the cumulative evaporation from 0.760 to 0.774. However, Model QE did not outperform Model Q on streamflow.

Therefore, it cannot be concluded that additional calibration on evaporation data increases the predictive power for the streamflow of the model. However, the model now more accurately represents the observed evaporation data, without trading this in for less predictive power of streamflow. Therefore, Model QE can be assumed to be a more accurate description of the hydrological processes in the Geul catchment than Model Q.

Contents

1 Introduction

Models that can predict the streamflow[1] of a river have inherent importance to society. Not only for industries or nature but also to forecast regional flooding. The Geul catchment in the Netherlands and sections of Belgium and Germany were flooded in July 2021, with total damages estimated to range between 350 and 600 million euros (Expertise Netwerk Waterveiligheid, 2021). During this flood event, the hydrological model of the Geul was undergoing maintenance, therefore it was not in use. However, the hydrologists from Waterboard Limburg (WL) stated that the present model would not have been able to forecast the flood under such an extreme rain event correctly (Expertise Netwerk Waterveiligheid, 2021). More severe precipitation events are projected as a result of climate change, raising the need for a robust hydrological model that can manage these changing conditions.

1.1. Theoretical framework

A hydrological model is a simplified mathematical representation of the natural hydrological processes of a basin, which are visualised in figure 1.1. The purpose of creating such a model is to estimate streamflow based on rainfall data. Hydrology is well known to suffer from an overwhelming diversity of models (Horton et al., 2022). Within this multitude, Moges et al. (2021) describe the element that all these models have in common: uncertainty.

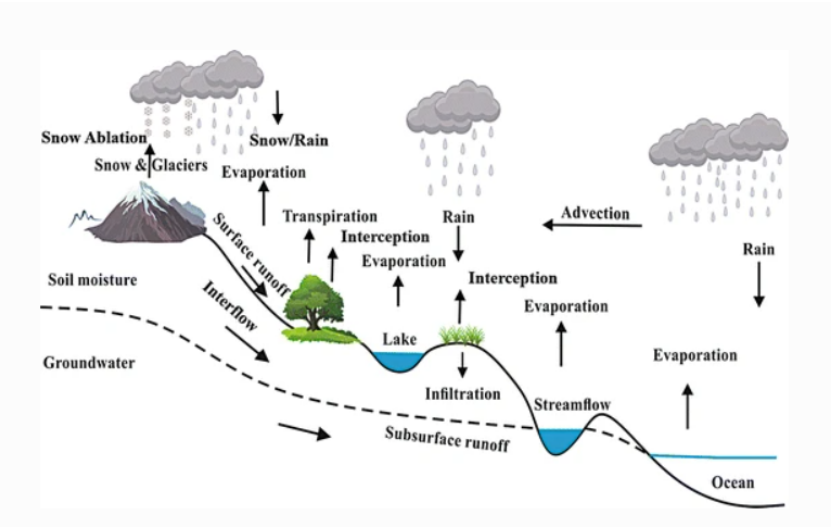


Figure 1.1: Overview of hydrological processes (Liu et al., 2017)

[1] In this thesis the terms streamflow, discharge and runoff are used interchangeably to describe the amount of water that flows through a river.

1.1.1. Uncertainty in hydrological modelling

Uncertainty in hydrological models can be caused by a number of factors, including the structure of the model, parameter values, and input or calibration data. As a result, the outputs of the model are also subjected to uncertainty. At the centre of this uncertainty rests the concept of equifinality, meaning that the same endpoint can be reached through various approaches (Khatami et al., 2019). In hydrological modelling, equifinality implies that many parameter sets can yield equally good results (Beven & Binley, 1992). This is visualised in figure 1.2, where the parameter values in red differ between model A and model B. However, the modelled streamflow is the same for both models and equal to the observed streamflow value. This discrepancy implies that at most one model is getting the right answer for the wrong reasons, as both models cannot simultaneously be close to reality and different from each other (Bouaziz, 2021).

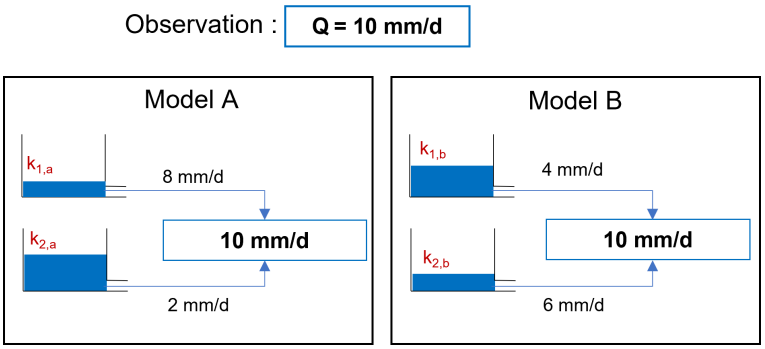


Figure 1.2: Which model is correct, if any? (Hrachowitz, 2021)

Equifinality starts to become a problem during the calibration of a hydrological model. Each model consists of parameters which are not yet known beforehand and which differ in values between catchments. The process of calibrating a model means that multiple parameter sets are tested and the parameter set that gives a model output which most closely represents the observed data, is thereafter used to make predictions. This is where the problem of equifinality becomes clear. An “optimal” set of model parameters may therefore appear to yield good predictions of the streamflow data but actually yield poor predictions for other fluxes and states, such as soil water storage, evaporation or groundwater levels (Whelan et al., 2019). This results in correct streamflow forecasts during the calibration period, but the internal processes of the model are not correct, and the model can fail to predict streamflow properly under varying conditions.

There have been multiple studies on minimisation strategies and quantification of different types of modelling uncertainty (e.g. Beven & Binley, 1992; Montanari & Brath, 2004; Kuczera et al., 2006). But it remains a major scientific and operational challenge (Renard et al., 2010). As equifinality is partly caused by a large number of parameters present in hydrological models and partly because of a lack of observations against which to compare the model outputs (Shokri et al., 2018), the reduction of the number of parameters or the inclusion of multiple calibration datasets may decrease the uncertainty around the modelled streamflow.

1.1.2. Evaporation as secondary calibration data set

There exist multiple fluxes and states within a hydrological model that have been utilized for calibration purposes, including soil moisture content, evaporation, or snow cover, in addition to streamflow (Sirisena et al., 2020). This present study implements evaporation data as a secondary calibration parameter due to its various benefits.

As highlighted by Zekters and Loaiciga (1993), evaporation is the second largest flux within the hydrological cycle, next to precipitation and streamflow. In the context of the Geul catchment specifically, more water leaves the catchment by evaporation than streamflow, thus making it an essential component of the hydrological cycle in Geul and important to be accurately represented in the model. Additionally, evaporation acts as the sole term that links land surface water balance and land surface energy balance (Zhao et al., 2013). This implies that by accurately representing evaporation in a model, both the water and energy balance can be closed, leading to more precise and reliable results.

Despite the limited number of measurement stations for evaporation in the Netherlands, with only two currently available (STOWA, 2010), global remote sensing evaporation datasets are readily accessible for calibration purposes. The downside of evaporation data is that large evaporation datasets are not made up of direct observations but are the results of a model with remote sensing data in combination with observations. In this thesis, GLEAM data is used as this data set, as it has been thoroughly validated throughout the years (Martens et al., 2017; Miralles et al., 2010, 2011).

Lastly, the inclusion of evaporation data in model calibration has been shown to better constrain hydrological models compared to streamflow alone (Becker et al., 2019; Jiang & Wang, 2019), leading to enhanced model performance.

1.1.3. Generalised Likelihood Uncertainty Estimation

To determine the parameter values in the model, which represent the streamflow observations as well as the evaporation data, the model needs to be calibrated. Beven and Binley (1992) introduce the idea of accepting all equally good performing models instead of one. They imply that because all model structures have some error and all observations and measurements on which model calibration is based must also have some error, there is no reason to believe that some calibration processes can find a single set of parameters without error. Rather, it is suggested that only an assessment of the likelihood or possibility of a specific parameter configuration being an acceptable simulator of the system is possible. As a result, rather than selecting the parameter set with the highest likelihood value, all parameter sets above a given likelihood value are selected. With GLUE, the model predictions are a combination of all the parameter sets that scored above the threshold likelihood value, thereby giving a prediction interval rather than a single line. With this interval, the uncertainty in input and calibration data as well as the model structure is also represented in the output of the model.

1.1.4. Multiple objective functions

Within this research there is no likelihood calculation, rather the models are scored on multiple object functions which represent the predictive power of the model. The predictive power of a hydrological model is the degree to which the hydrological model is capable of generating accurate streamflow prediction. This is determined by means of Nash Sutcliffe efficiency (Nash & Sutcliffe, 1970), see equation 3.29 which compares the modelled discharge values (Q_m) with the observed streamflow (Q_o) on each time step, and indexes this by the error of observed discharge (Q_o) on that time step with the averaged observed discharge($\overline{Q_m}$).

$$NSE_Q = 1 - \frac{\sum_{t=1}^{T}(Q_o^t - Q_m^t)^2}{\sum_{t=1}^{T}(Q_o^t - \overline{Q_m})^2} \tag{1.1}$$

$$NSE_E = 1 - \frac{\sum_{t=1}^{T}(E_o^t - E_m^t)^2}{\sum_{t=1}^{T}(E_o^t - \overline{E_m})^2} \tag{1.2}$$

This same calculation can be done for modelled evaporation and observed evaporation. Thereby calibration on only streamflow and on both streamflow and evaporation is possible.

1.2. Research framing

Remote sensing datasets as a whole have proven useful for calibration, evaluation, and assimilation in hydrological models in multiple studies over the last 15 years (e.g. Nijzink et al., 2018; Beck et al., 2009; Wanders et al., 2014). And also evaporation data sets specifically are being used to improve model performance (e.g. Sirisena et al., 2020; Immerzeel & Droogers, 2008; Zhang et al., 2009). Nevertheless, almost all studies that use evaporation as a secondary calibration set, are case studies. Lu et al. (2021) clarify that the quality of evaporation data varies significantly throughout the world, as evaporation datasets are created by a model itself, and calibrated by in-situ evaporation observations. This means that data quality is poorer in areas with fewer or non in-situ measurements and results vary around the world as these data points are not distributed equally over the earth, as can be seen in figure 1.3

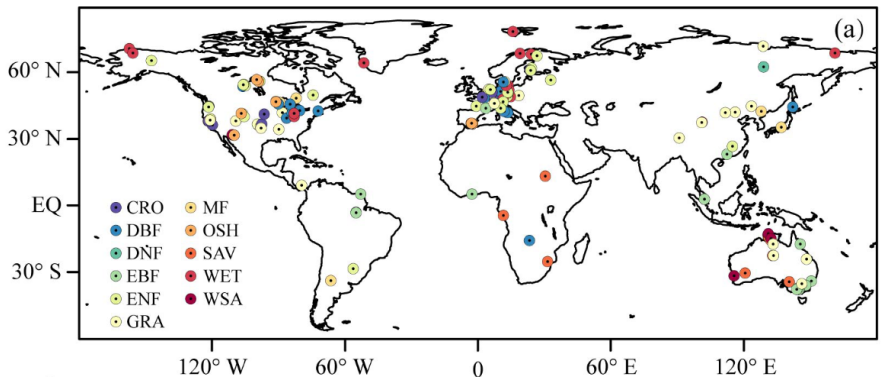


Figure 1.3: Spatial distribution of 181 in-situ flux EC sites across the world with 11 different land cover types (Lu et al., 2021)

1.2.1. Research gap

Because of the heterogeneity in evaporation data quality and case study-based research, the findings of existing studies on the value of using evaporation data as a secondary calibration data set cannot be assumed to be universally true. Furthermore, the addition of evaporation data does not always improve model performance. Evaporation datasets have a good potential for improving model calibration, but this is dependent on the calibration strategy, with distributed calibration strategies outperforming catchment average strategies (Dembélé et al., 2020). Throughout the world, evaporation calibration studies have taken place, but all in larger catchments, and mostly with distributed model setups (Jiang et al., 2020).

In the area of the Geul, there have been studies that use evaporation data as a calibration or evaluation dataset, on larger catchments, namely the Rhine (Huang et al., 2020) and the Meuse (Bouaziz, 2021). Both studies indicate that the evaporation flux in hydrological models can be used as a compensation flux. Indicating that when models are only calibrated on streamflow, the evaporation flux is used as the leftover precipitation output.

In this regional study, a lumped hydrological model is used, averaging out evaporation and precipitation data over the catchment to understand the impact calibration on evaporation can have on smaller-scale catchments. The lumped characteristic of the model decreases the chances of improving model performance, but the calibration limits the degree of freedom in the model and forces the model to select parameter values more carefully.

1.2.2. Research question

With the research gap in mind, the following research question has been drafted:

Is the predictive power of a discharge calibrated hydrological model of the Geul catchment in the Netherlands, greater than the predictive power of an evaporation-calibrated model, in addition to discharge?

The hypothesis of this research is that if the model is calibrated on evaporation data in addition to discharge data, it will yield better predictions during the evaluation period. As it will get the right answers for the right reasons.

1.3. Reading guide

Chapter 2 of this master thesis provides an overview of the Geul catchment and the data collection process that was used to gather relevant information. Chapter 3 expounds on the methodology used, covering the hydrological model as well as the calibration and evaluation process. The findings of this evaluation are presented in chapter 4, while assumptions, interpretations, and decisions that led to the results are discussed in chapter 5. Finally, chapter 6 includes the conclusions drawn from the study and recommendations for future research.

2

Case study description

In the following chapter, the Geul catchment is described in detail, including geography, climate, and special features of the area in section 2.1. Furthermore, the data collection of the Geul catchment and data quality assessment are specified in section 2.2.

2.1. The Geul catchment

The Geul is a river in the southernmost part of the Netherlands. Its origins lay in multiple springs in Belgium, the Netherlands, and Germany. As can be seen in figure 2.2, there are three main tributaries to the Geul, namely the Eyserbeek, the Selzerbeek and the Gulp. After passing through Valkenburg, the river connects with the Meuse, just north of Maastricht. The Geul, including tributaries, has a total length of 58 km and a catchment area of around 340 km^2 (de Moor et al., 2008).

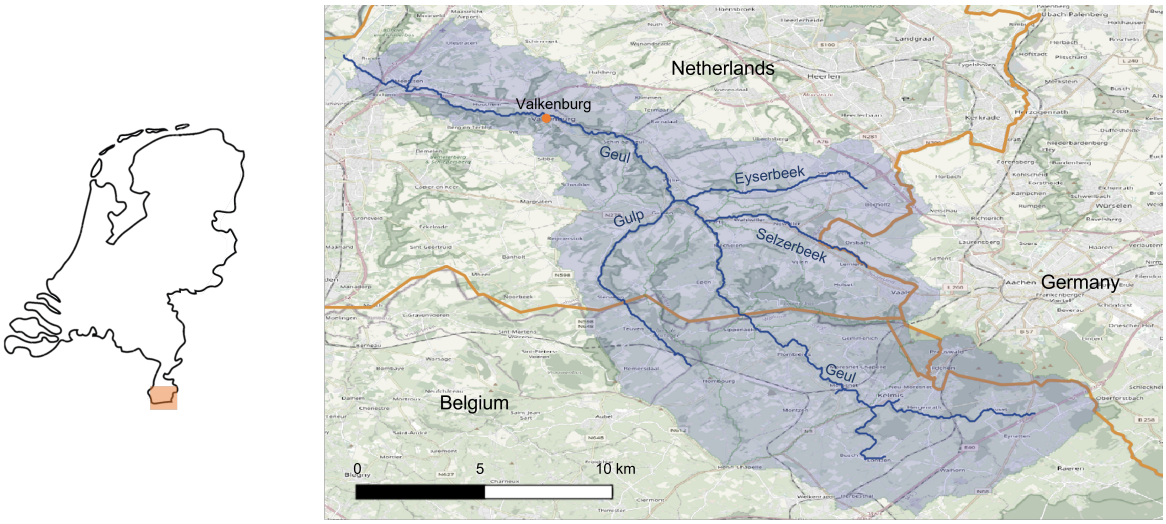


Figure 2.1: Location of the Geul catchment, including the Geul and its tributaries

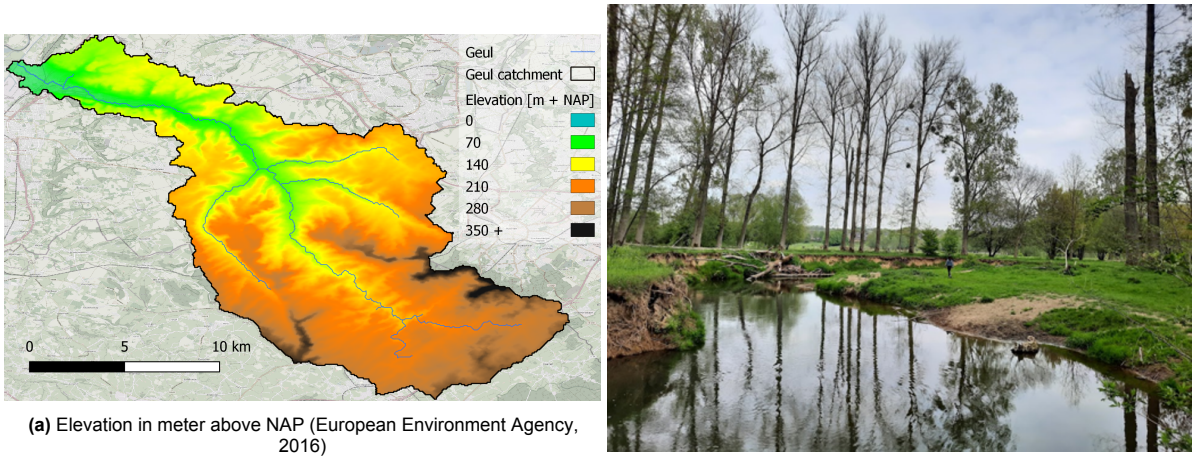
2.1.1. Climate

The Geul is a rain-fed river, with average yearly precipitation in the catchment varying from more than 900 mm per year near its headwaters to about 750 mm per year near its confluence with the Meuse (de Moor & Verstraeten, 2008; Provincie Limburg, 2021). The Netherlands overall has a sea climate, however, the south of Limburg is the region with the warmest summers, with an average of 17.9 °C and the coldest winters with an average of 3.5 °C, of the country (KNMI, 2020). The average yearly evaporation in the Geul catchment varies between 620 mm per year in the more downstream areas,

up to 720 mm per year in the upstream areas (Planet, 2022). It does snow, but only 18 days per year on average, calculated between 1991 and 2020 (Huirne, 2020), with a decreasing trend.

2.1.2. Unique features

The Geul has three unique features compared to other rivers in the Netherlands, it has a higher than average elevation difference and discharge, and it is free to meander. The river varies in slope from about 2% near its sources to 0.15% at its confluence with the Meuse (de Moor & Verstraeten, 2008) as it lays in an area with elevation differences of more than 250 m, which is uncommon for the Netherlands, as it is a very flat country. Furthermore, the average discharge is 2.82 m^3/s, measured just before the confluence with the Meuse, with peak discharges of more than 40 m^3/s during local floods (data from Waterboard Limburg). Compared with other rivers in the Netherlands, the relatively steep gradient and high peak discharges make it a very dynamic river, capable of severe erosion and fast lateral channel migration (Vandenberghe, de Moor, & Spanjaard, 2012). This erosion and migration are only possible because the Geul is one of the few free-meandering rivers in the Netherlands when it does not flow through an urban area.



(a) Elevation in meter above NAP (European Environment Agency, 2016)

(b) Free meandering section of the Geul, own picture

Figure 2.2: Two unique river features for the Netherlands

2.1.3. Water management

Even though the Geul is mostly free meandering, there are still water management efforts in the area. The management in the catchment is focused on retaining precipitation during wet periods to prevent flooding and therefore averaging out peak discharges. In the Dutch part of the catchment alone, there are over 500 rainwater buffers and numerous small weirs to contain the water in the Geul and its tributaries (Frenken, 2021). Furthermore, there are a number of weirs, both with stationary levels and with adjustable levels (Waterschap Limburg, 2019).

2.2. Data

A well-known expression in the modelling world is: 'Garbage in, means garbage out', this expression portrays the idea that the quality of your output data can only ever be as good as the quality of your data. To ensure that a model makes good predictions, the input and calibration data have to be correct. Therefore, in this section, the data collection process is described first, after which the quality of the data is discussed.

The data for this study were acquired with the intention of distributing the hydrological model into 1 km by 1 km grid cells on an hourly timescale. Therefore, a lot of effort has been put into finding data with the appropriate spatial and temporal resolution. Nevertheless, this model has in the end not been used, but has been worked out fully, see Appendix F including the reasoning behind why the model has not been used. The model used going forward has a single grid cell for the whole catchment and a daily time step. This means that all data with an hourly time step is re-sampled to daily data by taking

the average of all hourly data points. Furthermore, all raster data is averaged out over the catchment and reduced to one value for the whole catchment. An overview of all the collected data can be seen in table 2.1.

Table 2.1: Overview of collected data

	Data type	Timescale	Grid size	Source	Time interval
Forcing data	Precipitation	Hourly	1km x 1km	WL, KNMI, SPW	1963/01/01 - 2021/12/15
	Potential evaporation	Daily	100m x 100m	Planet	2012/07/24 - 2022/09/30
Calibration data	Actual evaporation	Daily	100m x 100m	Planet	2012/07/24 - 2022/09/30
	Discharge	Hourly	Point data	WL	2012/01/01 - 2021/01/01

2.2.1. Forcing data

Forcing data is the input data of a hydrological model, and represents the meteorological condition during each time step. Forcing data is usually collected at weather stations, and where possible interpolated between weather stations to acquire a raster data set. The timescale and grid size of the forcing data determines the maximum temporal and spatial resolution of the model. These, therefore, need to be on the right scale to capture the hydrological processes in the catchment. However, due to the change in model choice, the data has been aggregated over time and space.

Precipitation

The location of the catchment, partly in The Netherlands, partly in Belgium and partly in Germany, makes the gathering of precipitation data more difficult as each country has its own meteorological institution. There is no existing data set available for the region. The precipitation data set raster is created by L. Bouaziz, combining precipitation records from measurement stations in the Netherlands and Belgium, see figure 2.3. The specific location and measurement period of each measurement station can be found in Appendix A.

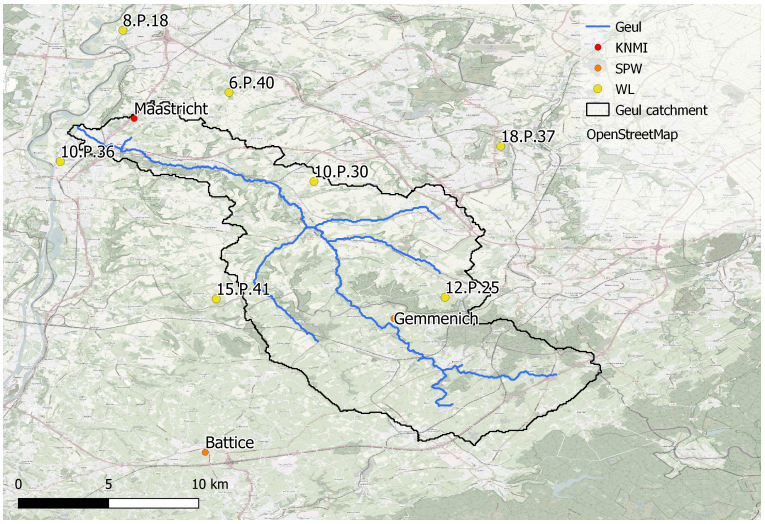


Figure 2.3: Overview measurement stations, own figure

Rain gauge measurements are relatively accurate at the point scale. However, the majority of uncertainty is introduced during the interpolation of these points because measurement stations are typically too sparsely distributed to accurately capture the high spatial and temporal variability of precipitation systems (Villarini et al., 2008). A recent study suggest an optimum rain gauge network density between 14 km2 per gauge and 38 km2 per gauge (Gyasi-Agyei, 2020), even though this study is a comprehensive review of pre vious research on the topic, there is also case study data used, which make these result biased for a more tropical area. In the Geul catchment the average density is 38 km2/gauge, distribution is lop-sided with more measurement points in the Dutch part of the catchment.

There are multiple interpolation methods, with some of the more well known being Thiessen polygon, inverse distance weighted and ordinary Kriging. In this research Thiessen polygon is used as interpolation method, as it is a relatively easy method of interpolation, essentially assigning precipitation values by nearest neighbouring measurement station, data points outside the catchment can be used, and it does not give 'false accuracy' by creating intermediate data points, as can be seen in figure 2.4. One of the downsides of this interpolation technique is the neglect of spatial variability and more specifically orographic effects (Jain & Singh, 2003). It is assumed that this effect can be neglected in the Geul catchment, as for Dutch standard there is a lot of elevation gain, but objectively the height differences are not large enough to lead to orographic rainfall.

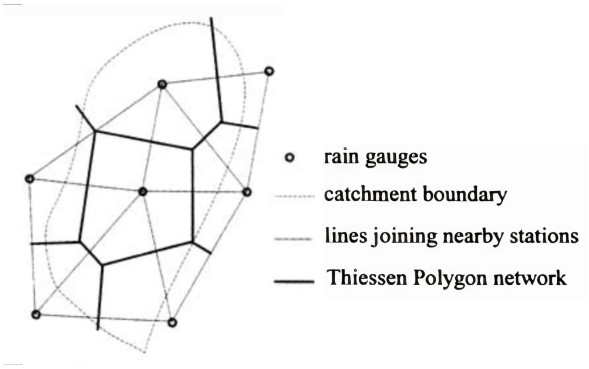


Figure 2.4: Thiessen Polygon (Schumann, 1998)

The rainfall measurements as the KNMI weather station in Maastricht start already in the 1960s, but the measurement points of Waterboard Limburg are only usable after 2012. Before 2012 there was no distinction between zero precipitation or a missing value in the data set, making the data of little use. Only after 2012, yearly precipitation totals correlated well with yearly precipitation amounts at the KNMI measurement point in Maastricht, indicating that missing values are zero precipitation after 2012. The measurement stations in Belgium started recording in 2012. Furthermore, not all stations were active each day from 2012 going onward, if a measurement stations did not have data for a specific day, the Thiessen polygons were redrawn without that measurement station.

Potential evaporation

Potential evaporation is the amount of water that could evaporate if there was an unlimited amount of water available in an area. On the other hand, actual evaporation is the amount of evaporation with the real amount of water available. Actual evaporation is a major component of the catchment water balance, and therefore it is important to have a good estimation of the potential evaporation, as this restricts the actual evaporation (Oudin, Michel, & Anctil, 2005). However, there are many models to calculate potential evaporation. Oudin et al. (2005), has compared 27 of these models, including the most well known methods Penman, Penman-Monteith and Priestly-Taylor. The Priestly-Taylor formula, see equation 2.1, which represent the energy-based term of the Penman formula outperforms the latter and appears slightly more efficient in rainfall-runoff models, however regional differences may change these results.

The potential evaporation is calculated by Planet (used to be VanderStat) (2022) with the Priestly Taylor equation, where the resistance factor (α) is calculated using their own hydrological model in combination with satellite images to determine root zone storage amounts.

$$E_{pot} = \alpha * \frac{\delta * (R_n - G)}{\lambda_v * (\delta * \gamma)} \tag{2.1}$$

α represents a resistance factor and varies from less than 1 in humid condition to almost 2 in arid condition, with a typical value of 1.26 for open water bodies (Priestley & Taylor, 1972). δ is the slope of the saturation vapour pressure curve. R_n is net radiation. G is the soil heat flux. λ_v is the volumetric latent heat of vaporisation of water, and γ is the psychometric constant.

The KNMI station in Maastricht also calculates daily potential evaporation with the Makkink equation, which is most the commonly used equation for potential evaporation in the Netherlands (van Kraalingen & Stol, 1997). The difference between the average daily potential evaporation from the Priestly-Taylor equation across the entire catchment and the point calculation at Maastricht is 13.3 %, calculated between 2014 and 2021, as can also be seen in figure 2.5 and the monthly difference is 15.3% and yearly 11.3%, see figure 2.6.

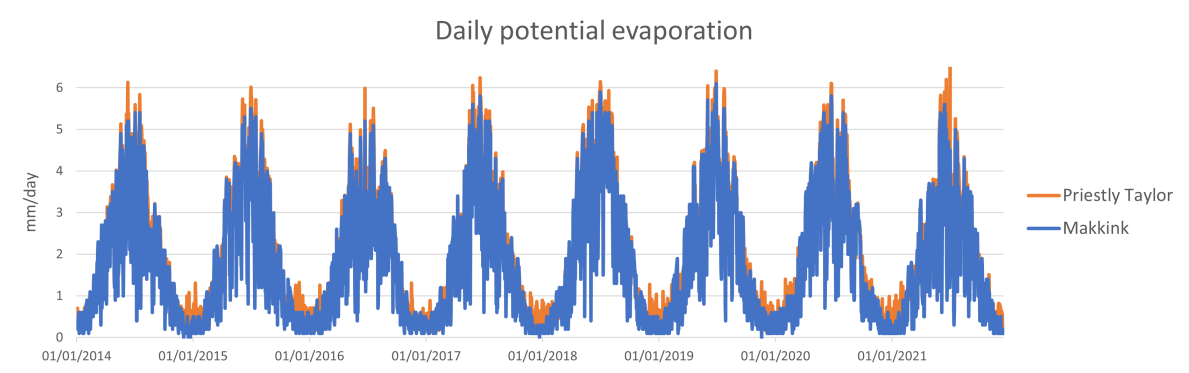
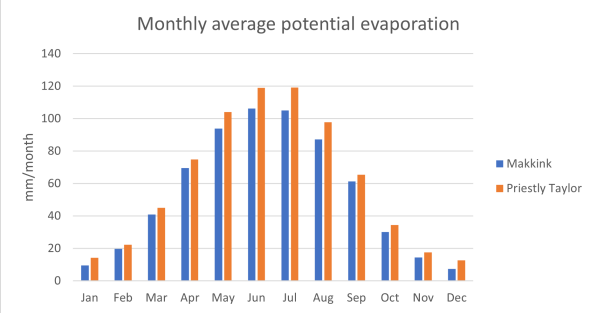
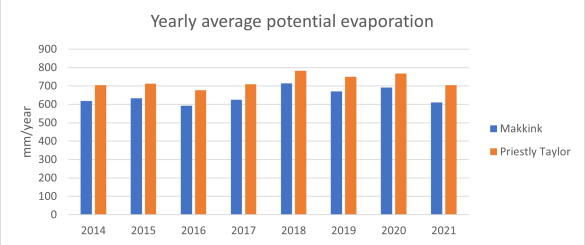


Figure 2.5: Daily potential evaporation



(a) Monthly potential evaporation



(b) Yearly potential evaporation

Figure 2.6: Monthly and yearly potential evaporation

It is assumed that this difference is caused by a number of reasons. First, the variability of the landscape of the Geul is taken into account with the raster Priestly Taylor calculation, where the Makkink evaporation is reference evaporation for short grass. Second, evaporation and precipitation are highest at the headwaters of the Geul in Belgium, see section 2.1.1, which is the area of the catchment that lays the farthest away from the KNMI measurement station in Maastricht. Last, evaporation is proportional to net radiation in the Priestley-Taylor equation, but Makkink evaporation is proportional to short-wave radiation (van Kraalingen & Stol, 1997).

Planet has informed its users that evaporation data from 2013 is erroneous, since the resultant potential evaporation is lower than expected and out of pattern with other years. Therefore, only potential evaporation from 2014 onward is used.

2.2.2. Calibration data

Calibration data is data with which the output of the model is compared, to determine how well the model has performed. Calibration data is usually only discharge data, but in this study, discharge as well as actual evaporation data is used to compare model output values with measured values.

Discharge data

The most common calibration parameter in hydrological modelling is discharge data, sometimes measured at different points along the river. A map of the location of these measurements points, can be seen in figure 2.7.

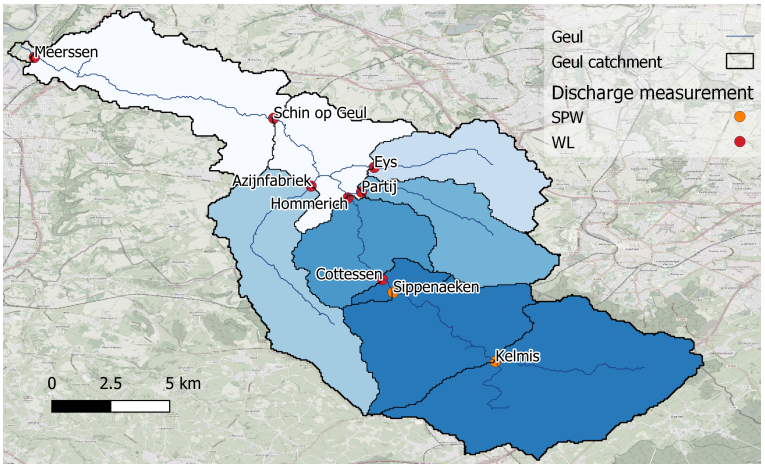


Figure 2.7: Discharge measurement locations

Discharge [m3/s] is the amount of water [m3] that flows through a river per second [/s]. This can be represented by the intersection of the river [m2] at some point and the speed with which the water passes through this intersection [m/s]. This can therefore normally not be measured directly, as the cross-section of the river is not constant along the length of a river, and the water speed is not constant throughout the cross-section.

At the measurement point in Meerssen, see figure 2.8, horizontally-oriented Acoustic Doppler Current Profiler (H-ADCP) is used since 2010. Acoustic Doppler current profilers (ADCPs) can be mounted horizontally at a riverbank, yielding single-depth horizontal array observations of velocity across the river (Hoitink, Buschman, & Vermeulen, 2009). With these velocities over the depth of the river, discharge can be calculated. The main advantage of this measurement methodology is the continuous measurement of stream flow, apart from maintenance operations. The disadvantage of H-ADCP is that is works best during normal and high flows, according to Waterboard Limburg. Therefore, a rating curve is still used to determine discharges during low flows, meaning below 1 m^3/s.

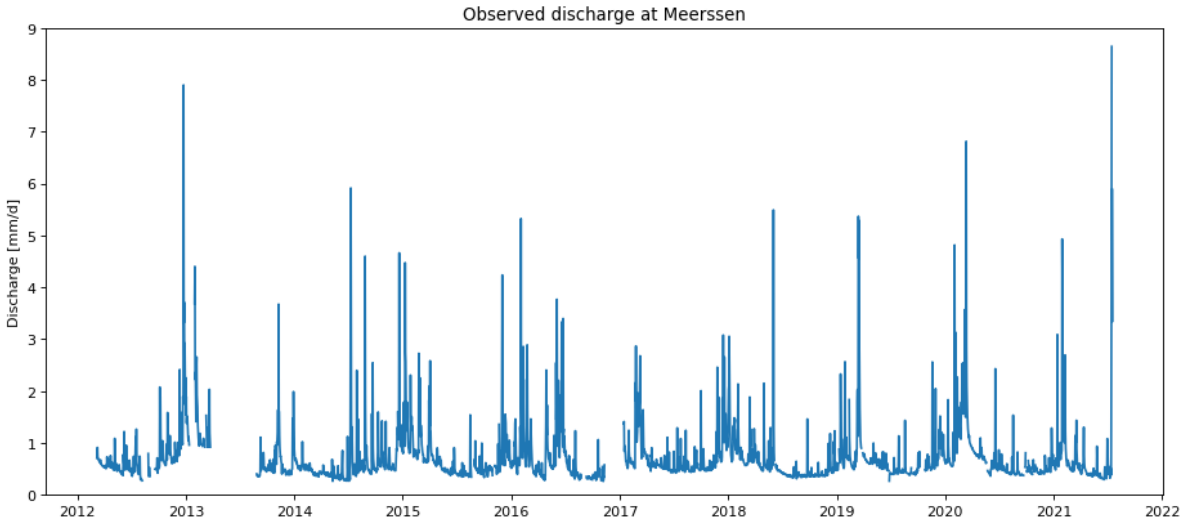


Figure 2.8: Observed discharge at Meerssen

A rating curve is an empirical relationship between stage in the river and discharge. Although it is the most common discharge 'measurement' methods, the method is prone to errors. These include, measurement errors, interpolation and extrapolation errors, changes in intersection during high flows and seasonal variations in roughness (vegetation) (Baldassarre & Montanari, 2009). Waterboard Limburg (P. Hulst, 2023) estimate that during flows under 1 m^3/s, the discharge in the Geul can be overestimated by up to 40% at times. However, every month the Waterboard does manual measurements

at measurement station Meerssen and during these measurements, they have never measured a discharge lower than 1 m^3/s. Which is also the lowest measurement of the time series available. So it seems that the discharge does not drop out of the measurement reach of the H-ADCP, but this is of course not certain.

The measurement data at Meerssen started only in the third month of 2012 and even though H-ADCP is a continuous measurement device, more than half of 2013 does not have a measurement point. This is because the river bed became unstable, and the measurement device needed to be recalibrated. Therefore, it has been decided to start the use of the discharge measurement from 2014 onward for calibration purposes.

Actual evaporation data

The actual evaporation is used as secondary calibration parameter, meaning that modelled evaporation will be compared with the actual evaporation data. The actual evaporation comes from Planet, see figure 2.9. Planet uses the Global Land Evaporation Amsterdam Model (GLEAM) (Miralles et al., 2011). GLEAM has a grid size of 27 km by 27 km, and this is re-projected to a 100 m by 100 m grid size by Planet with the use of land use and soil moisture satellite measurement.

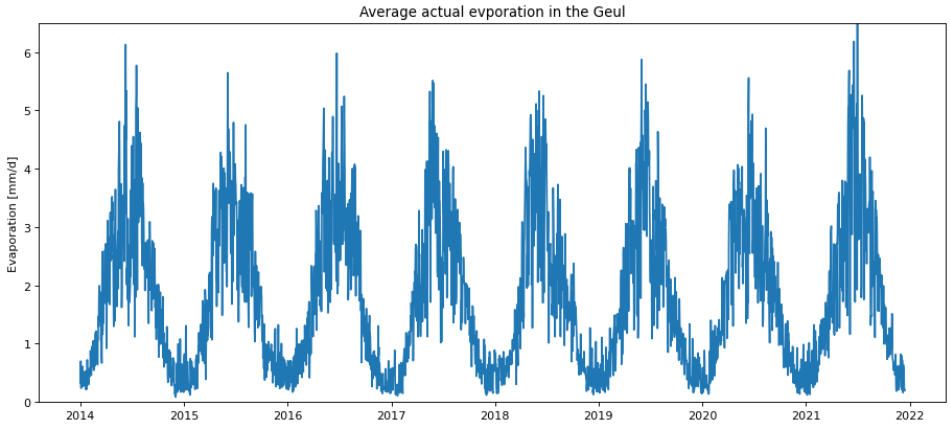


Figure 2.9: Actual evaporation

This evaporation data is not measured actual evaporation but GLEAM is a set of algorithms dedicated to the estimation of terrestrial evaporation and root-zone soil moisture from satellite data (Martens et al., 2017). Meaning that it is the output of a model which consists of four modules: potential evaporation, vegetation stress module, precipitation, an interception module and a soil moisture module, as is visualised in figure 2.10.

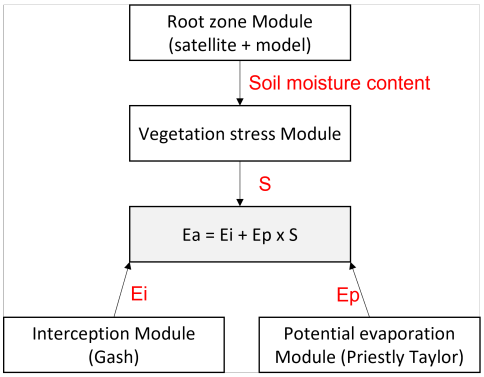


Figure 2.10: Calculation of actual evaporation by Planet (2022)

The interception evaporation is based on the analytical model of Gash (1979) derived by satellite observed rainfall. The potential evaporation is calculated by GLEAM with the Priestly Taylor equation, see equation 2.1, on the basis of temperature and radiation satellite observations. The soil moisture

content is determined by satellite images in combination with a hydrological model to determine available water in the root zone storage. The vegetation stress factor is determined with this available water in the root zone storage. The actual evaporation is calculated by potential evaporation and vegetation stress factor from the vegetation stress module and added to interception evaporation. Just as the potential evaporation, it is advised by Planet not to use the data from 2013 as it is erroneous. Therefore, only actual evaporation from 2014 onwards is used.

As the evaporation data is not directly measured, the GLEAM data is thoroughly validated against multiple measured datasets. The two main intermediate products of GLEAM are validated separately, namely rainfall interception (R = 0.86) and root-zone moisture (R = 0.60 and R = 0.69 for surface and deep layers respectively) (Miralles et al., 2010). Furthermore, the final evaporation estimations have been validated against eddy covariance measurements from 43 Fluxnet stations, both at a daily (R = 0.83) and a monthly (R = 0.90) time scale (Miralles et al., 2011). Fluxnet stations are mainly located in the EU and the US but cover the most common vegetation types and climates (Baldocchi et al., 2001). Moreover, no systematic bias for specific vegetation types or rainfall conditions has been detected. Therefore, even though GLEAM data is not observed evaporation itself, it has been determined to represent observed evaporation and intermediate production well enough, and it has been used in many hydro-meteorological applications ranging from climate trend to drought prediction and horological model calibration (e.g. López et al., 2017; Forzieri et al., 2017; López-Ballesteros et al., 2019).

2.2.3. Data quality check

The quality of the individual data sets has been reviewed in earlier sections. Yet, it is also critical to determine if the combination of data sets is physically possible and if they close the long-term water balance.

$$\frac{dS}{dt} = P - E_A - Q = 0 \tag{2.2}$$

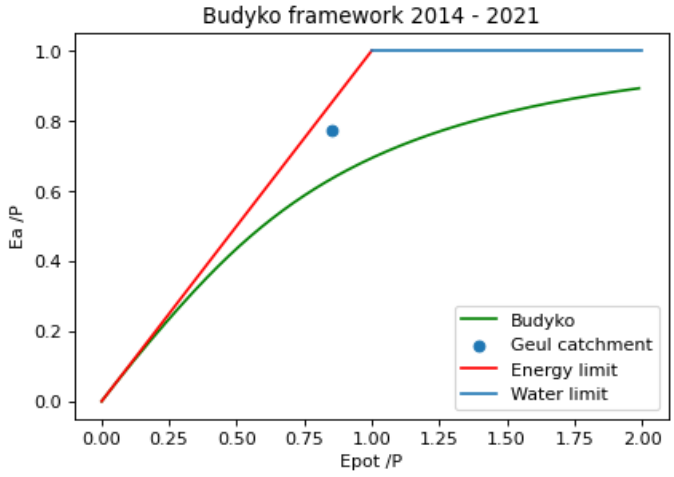


Figure 2.11: Budyko framework, long-term averages from 2014 until 2021

The Budyko framework does this by taking into account two criteria the combined data set need to fit. First, potential evaporation has to be larger than actual evaporation, it should never be possible that more water is evaporating from the catchment, than physically possible by the energy available. This criterion is represented by the energy limit, in figure 2.11. Meaning that a catchment should not be able to plot to the left of the energy limit. The second criterion is the water limit. The water limit represents that water leaving the catchment, in actual evaporation and discharge combined, can never be larger than the amount of water going into the catchment, see equation 2.4. The water limit is represented by the blue line in figure 2.11. A catchment should not be able to plot above the water limit. The axis of the Budyko framework, is a re-written form of a closed water balance, as can be seen in equation 2.3 and 2.4.

$$P = E_A + Q \tag{2.3}$$

$$1 = \frac{E_A}{P} + \frac{Q}{P} \tag{2.4}$$

The green line in figure 2.11 is the line on which a catchment is expected to plot, the Budyko (1948) line, see equation 2.5. It is an empirical relation and most major river basins in the world plot on or around this line.

$$\frac{E_A}{P} = \sqrt{\frac{E_{pot}}{P} * tanh(\frac{1}{\frac{E_{pot}}{P}}) * (1 - e^{-\frac{E_{pot}}{P}})} \tag{2.5}$$

The Geul long-term average precipitation, actual evaporation and potential evaporation data combined plot within the Budyko framework, albeit a bit higher than the expected location on the Budyko curve. This means that relatively more water evaporates, around 75% of all precipitation evaporates and less water than expected results in runoff, around 25%. The catchment also plots to the left of the imaginary line where E_{pot}/P = 1. This means that the catchment is energy-limited and not water limited. That is to say that if there would be more radiation, more water would evaporate. This aligns with the wet weather and limited sunshine the Netherlands is known for.

2.2.4. Conclusion

The previous sections described the many data sets that are used to force and calibrate the model. The various data sets have issues at different time periods, however after the years with missing values and erroneous data are deleted, the catchment's long-term water balance closes. This indicates that combining these datasets is physically possible.

Moreover, raster data has been aggregated across the catchment, and hourly data has been averaged throughout the day to provide the data sets shown in table 2.2. Which gives an overview of all the data used in this research. The time span in which all the data sets are of sufficient quality to be used determines the time interval in which the model will be calibrated and evaluated, which implies that for this study, the total time interval is between 2014/01/01 and 2021/12/15.

Table 2.2: Used data overview

	Data type	**Timescale**	**Grid size**	**Source**	**Time interval**
Forcing data	Precipitation	Daily	Point data	WL, KNMI, SPW	2012/01/01 - 2021/12/15
	Potential evaporation	Daily	Point data	Planet	2014/01/01 - 2022/09/30
Calibration data	Actual evaporation	Daily	Point data	Planet	2014/01/01 - 2022/09/30
	Discharge	Daily	Point data	WL	2014/01/01 - 2021/12/15

3

Methodology

This study's methodology is separated into four sections, as shown in figure 3.1.

In the conceptual model, the hydrological processes are selected and for each process, the corresponding equations are used. The type of catchment that the model must depict determines which processes and therefore equations are employed. The conceptual model is described in section 3.1.

The hydrological model of this research has a distinct characteristic, namely the use of landscape classification. The process of determining the landscape classes is described in section 3.2.

The procedural model converts the conceptual model into code, in this example Python, to allow for quick computations and a high number of runs in a short amount of time. Furthermore, the water balance evaluates the procedural model. The conversion of code into Python is not depicted in this chapter as it applies the same equations as the conceptual model, but the water balance is represented in section 3.3.

Model calibration is the process of selecting values for the parameters in the model with the purpose of creating a model whose output represents the observed streamflow and evaporation data. The calibration of the procedural model is elaborated upon in section 3.4.

The model evaluation procedure determines how well the models perform. How well do the model forecast discharge and other performance indicators? The results of these evaluations are presented in the next chapter, chapter 4. The methodology of these evaluations is described in section 3.5

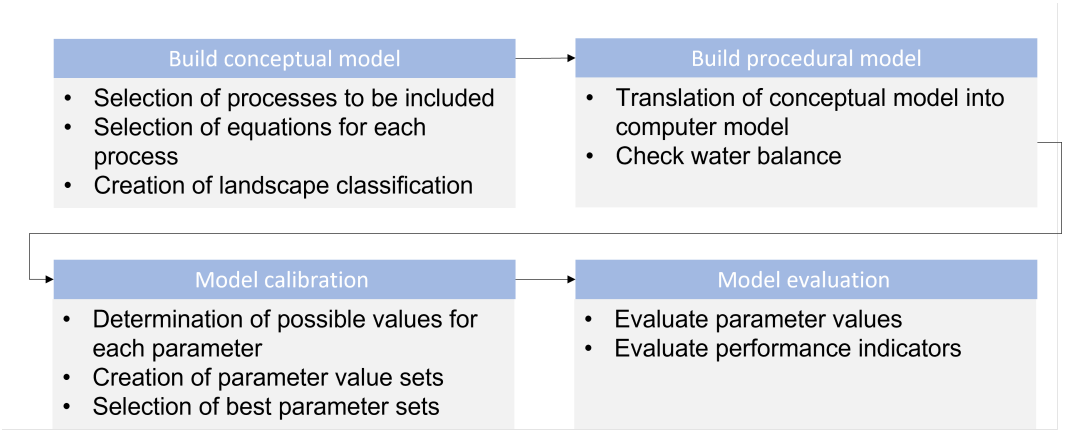


Figure 3.1: Methodology overview

3.1. Conceptual model

The model is named FLEX HBV, and consists of two parts. HBV in the name refers to the underlying modelling methodology, which is a lumped Hydrologiska Byrns Vattenbalansavdelning (HBV) model. Bergström (1973) developed the HBV model, which is still one of the most widely used hydrological models in the world today. The rainfall-runoff model uses a bucket approach to mimic the many hydrological processes that occur in the catchment. This bucket approach is clearly visible in the schematic overview of the model in figure 3.2. The FLEX element of the model refers to the use of flexible topography, meaning that there is a classification of landscapes within the model, which will be further explained in section 3.2.

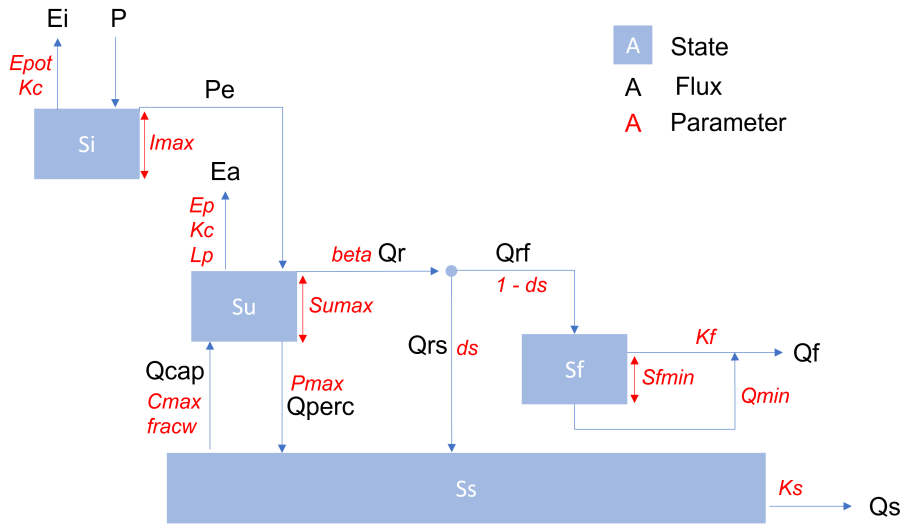


Figure 3.2: Conceptual model Flex HBV, own figure

3.1.1. Selected processes to represent the catchment

The simplest possible model would consist of only one bucket, with precipitation as input and evaporation and discharge as outputs. This bucket would be able to make some predictions of streamflow but would not be able to create multiple different timed reactions of the catchment on precipitation. The goal while creating the model is to keep it as simple as possible while including all dominant flow processes in the catchment. Dominant flow processes differ per landscape, which is described in section 3.2. figure 3.2 depicts the conceptual model with four main buckets. All symbols represented in the conceptual model in figure 3.2 are described in table 3.1.

The interception process is represented by the S_i bucket. Interception is all precipitation that does not directly reach the ground but is intercepted by leaves, plants, or buildings. Water can evaporate (E_i) off of these surfaces or run off to the ground once the maximum storage capacity (I_{max}) is reached. The S_u bucket depicts the ground's unsaturated root zone. Water can evaporate (E_a) from this storage in the top layer of the earth. There is also a connection to deeper groundwater, from which water can enter the root zone via capillary rise (Q_{cap}) or exit by deep percolation (Q_{perc}). Not all the water that exits the interception storage can be stored, and depending on how full the root zone storage is, water runs off (Q_r). Water that runs off from the unsaturated root zone storage (Q_r) is divided between the fast storage ($Q_{r,f}$) and the slow storage ($Q_{r,s}$). The S_f bucket represents fast groundwater runoff and storage. There is a continuous minimal discharge into the river Q_{min} and after a minimal amount of water is stored ($S_{f,min}$) the minimal discharge is supplemented with an outflow linear to the storage amount, giving the entire fast discharge response (Q_f). The S_s bucket represents the slow groundwater storage, all water entering the groundwater storage will eventually become the slow discharge response (Q_s) as a set amount (K_s) of the storage quantity is discharged at each time step.

Not all processes and fluxes are present in all landscapes. For example, capillary rise occurs only in wetlands because the groundwater table is close enough to the surface. Section 3.2 goes into greater detail on how the conceptual model is divided into the landscapes.

Two processes have been kept out of the model, namely the snow and Horton processes. Snow does occur in the Netherlands, as explained in section 2.1.1, but on average only for 18 days a year. The addition of a process brings along a number of new parameters that need to be calibrated. Therefore, consideration has been given to not including a snow process, thereby getting better calibrated parameters for the other processes. Horton processes are also left out for two reasons. First, the model is on a daily time step, and Horton processes happen very quickly on a scale of minutes to hours (Loon, 2015). Therefore, the model probably would not have been able to capture this process correctly. Furthermore, the addition of a non-linear parameter α to the fast storage bucket to represent non-linear discharge as with the Horton process did not increase the model's predictive value. Lastly, as indicated in section 2.1.3, there are over 500 rainwater buffers in the catchment, which collect the water that would otherwise cause this Horton flow.

Table 3.1: Overview of all symbols

Name	Description	Unit	Type
I_{max}	maximum interception storage	mm	parameter
$S_{r,max}$	maximum root zone storage capacity	mm	parameter
$S_{f,min}$	minimum fast storage for fast runoff	mm	parameter
P_{max}	maximum percolation flux	mm Δt^{-1}	parameter
C_{max}	maximum capillary rise	mm Δt^{-1}	parameter
L_P	evaporation reduction threshold	-	parameter
K_s	recession constant slow storage	Δt^{-1}	parameter
K_f	recession constant fast storage	Δt^{-1}	parameter
β	measure of non-linearity of soil runoff	-	parameter
d_s	splitter parameter for slow/fast storage	-	parameter
$frac$	fraction of class in the catchment	-	parameter
Kc_{ini}	crop coefficient initial growing stage	-	parameter
Kc_{mid}	crop coefficient mid season growing stage	-	parameter
Q_{min}	minimal constant streamflow	mm Δt^{-1}	parameter
P	precipitation	mm Δt^{-1}	forcing
P_e	effective precipitation	mm Δt^{-1}	flux
E_{pot}	potential evaporation	mm Δt^{-1}	forcing
E_i	evaporation from interception storage	mm Δt^{-1}	flux
E_P	potential soil evaporation ($E_{pot} - E_i$)	mm Δt^{-1}	flux
E_a	rootzone evaporation	mm Δt^{-1}	flux
Q_r	flux from the root zone storage	mm Δt^{-1}	flux
Q_{rf}	preferential recharge to fast storage	mm Δt^{-1}	flux
Q_{sf}	preferential recharge to slow storage	mm Δt^{-1}	flux
Q_{cap}	Capillary flux	mm Δt^{-1}	flux
Q_{perc}	Percolation flux	mm Δt^{-1}	flux
Q_f	runoff from fast storage	mm Δt^{-1}	flux
Q_s	runoff from slow storage	mm Δt^{-1}	flux

S_i	interception storage	mm	state
S_u	root zone storage	mm	state
S_f	fast storage	mm	state
S_s	slow storage	mm	state

3.1.2. Parametrisation of selected processes

The processes that are being described in the conceptual model, such as evaporation, capillary rise, deep percolation, and different runoff processes, are all described using different equations to make sure the mathematical response represents the natural one. In the next section, the parametrisation of the processes of each bucket of the conceptual model is depicted.

Interception

Precipitation (P) enters the interception storage (S_i), and water evaporates from the interception storage (E_i) by the amount available (1-K_s) from the potential evaporation (E_{pot}). Water runs off once the storage has reached its maximum storage capacity (I_{max}).

$$dS_i/dt = P - E_i - P_e \tag{3.1}$$

$$E_i = min(E_{pot} * (1 - K_c), S_i) \tag{3.2}$$

$$P_e = max(0, S_i - I_{max}) \tag{3.3}$$

The crop coefficient (Kc) has a different value throughout the year. This parameter represents the growing stages of the plants in the catchment. A full description of the crop coefficient values throughout the year and the specification of the growing stages can be found in section 3.4.1.

Unsaturated root zone storage

The effective precipitation (P_e) from the interception storage and capillary rise (Q_{cap}) from the slow storage enter the root zone storage. Water leaves the storage through percolation (Q_{perc}), runoff (Q_r) and evaporation (E_a). The evaporation amount is once again dependent on the crop coefficient (K_c) and no water runs off if there is no incoming effective precipitation.

$$dS_u/dt = P_e + Q_{cap} - Q_{perc} - Q_r - E_a \tag{3.4}$$

$$Q_{cap} = (1 - S_u/S_{u,max}) * C_{max} * frac_{wet} \tag{3.5}$$

$$Q_{perc} = S_u/S_{u,max} * P_{max} \tag{3.6}$$

$$Q_r = \begin{cases} (S_u/S_{u,max})^{\beta} * P_e & \text{if } P_e > 0 \\ 0 & \text{if } P_e = 0 \end{cases} \tag{3.7}$$

$$E_a = E_p * K_c * min(1, S_u/(S_{u,max} * L_p)) \tag{3.8}$$

$$E_p = E_{pot} - E_i \tag{3.9}$$

Fast storage
The inflow of the fast storage is a fraction (1-d_s) of the overland outflow of the root zone storage (Q_r). Water leaves the fast storage as a fraction (K_f) of the amount of water in the storage, but only if the storage holds more water than the minimum threshold value ($S_{f,min}$). Otherwise, there is only the minimum outflow (Q_min), this additional threshold ($S_{f,min}$) and minimum outflow (Q_min) are further explained in section 3.1.3.

$$dS_f/dt = Q_{r,f} - Q_f \tag{3.10}$$

$$Q_{r,f} = (1 - d_s) * Q_r \tag{3.11}$$

$$Q_f = \begin{cases} Q_{min} & \text{if } S_f < S_{f,min} \\ Q_{min} + S_f * K_f & \text{if } S_f => S_{f,min} \end{cases} \tag{3.12}$$

Slow storage
The slow storage takes in water through deep percolation (Q_{perc}) and the leftover fraction of the overland outflow of the root zone storage ($Q_{r,s}$). Water leaves the storage through capillary rise (Q_{cap}) and slow discharge (Q_s) as a fraction (K_s) of the amount of water in the storage. The equations for percolation (Q_{perc}) and capillary rise (Q_{cap}) are already described in the unsaturated root zone section.

$$dS_s/dt = Q_{r,s} + Q_{perc} - Q_{cap} - Q_s \tag{3.13}$$

$$Q_{r,s} = Q_r * d_s \tag{3.14}$$

$$Q_s = S_s * K_s \tag{3.15}$$

3.1.3. Minimum discharge

During the analysis of the discharge in the Geul, a pattern started to emerge. The discharge in the Geul was never lower than 1 m^3/s (0.2 mm/d for the Geul catchment area). This is not remarkable in and of itself, but the discharge also never dipped below 0.2 mm/d during dry periods and produced a flatter-than-expected hydrograph during these low flows.

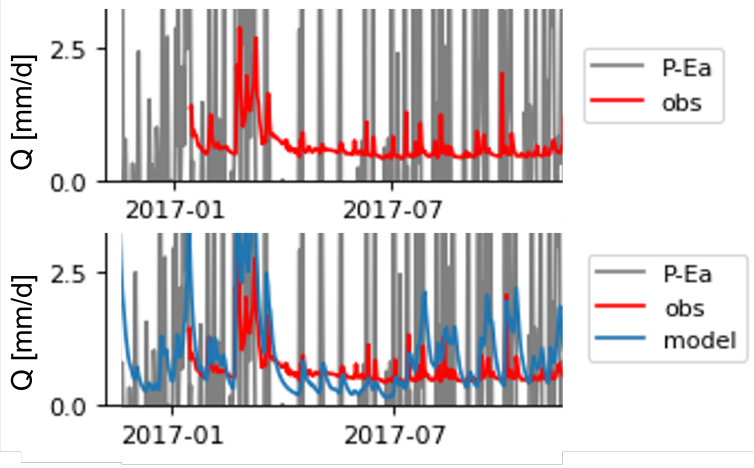


Figure 3.3: Observed and modelled streamflow

In figure 3.3, two graphs are displayed. Both are zoomed-in versions of a larger graph, for a full-size graph, see Appendix D. The grey lines indicate days when the daily precipitation exceeds the evaporation, implying an increase in discharge (and storage). The higher the grey line, the greater the increase in discharge. Although the height of the grey lines is not noticeable in these graph zoom-ins,

the idea is that if a grey line appears on a given day, the hydrograph is expected to rise. The blue line in the lower graph depicts this behaviour. Yet, it is evident that the red line, which represents the observed discharge levels, does not display this behaviour. It is also clear that the model underestimates the discharge during these dry periods.

This minimum observed discharge can be caused by a variety of factors. First and foremost, there can be measurement inaccuracies. As explained in section 2.2.2, the low flows in the Geul are computed using a stage-discharge relationship and not measured by H-ADCP, which is more prone to measurement errors. However, during the monthly manual measurement at Meerssen, there has also never been a discharge recorded below 1.0 m^3/s. A second option is the effect of storage in the catchment. As explained in section 2.1.3, there are numerous small weirs in the catchment, according to the legger of Waterboard Limburg, which has also been spotted in the river during visits to the sight. These weirs cause the water to be held in place during low flows, and through the groundwater, this is discharged slowly over time, resulting in a continuous, minimum discharge in the Geul. It is also the opinion of Waterboard Limburg that the discharge in the Geul is not likely to drop below 0.2 mm/d because of these storages.

To be able to mimic this discharge response, a minimum storage threshold ($S_{f,min}$) must be overcome, which represents the weirs, and a minimum discharge (Q_min) are both added to the fast runoff bucket. Without the addition of these two components, the model was unable to represent the observed runoff, with De_Q scores higher than 0.7, indicating that using the mean of the observed discharge was a better predictor of the discharge than the model. The addition of this process greatly improved the model's performance, as can be seen in the streamflow results in chapter 4.

3.2. Flextopo classes

The difference between a classic HBV model and a flextopo model is the inclusion of landscape classes within the catchment. The inclusion of landscape classes has previously been shown to improve model performance (Gharari et al., 2014), the use of landscapes in this catchment is justified by the distinct landscape pattern that emerges and is represented in figure 3.6a, this is explained further in detail in section 3.2.2. A graphic overview of the three landscape classes, used in the model, can be seen in figure 3.4. Each landscape class has a different dominant runoff process and, therefore, its own HBV model with unique processes and parameter values.

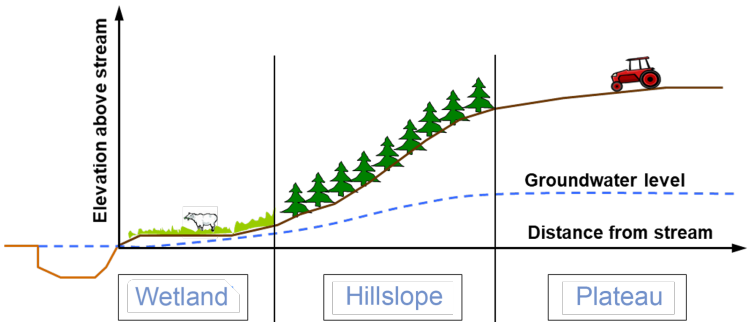


Figure 3.4: Landscape classes (Gao et al., 2014; Gharari et al., 2014; Savenije, 2010)

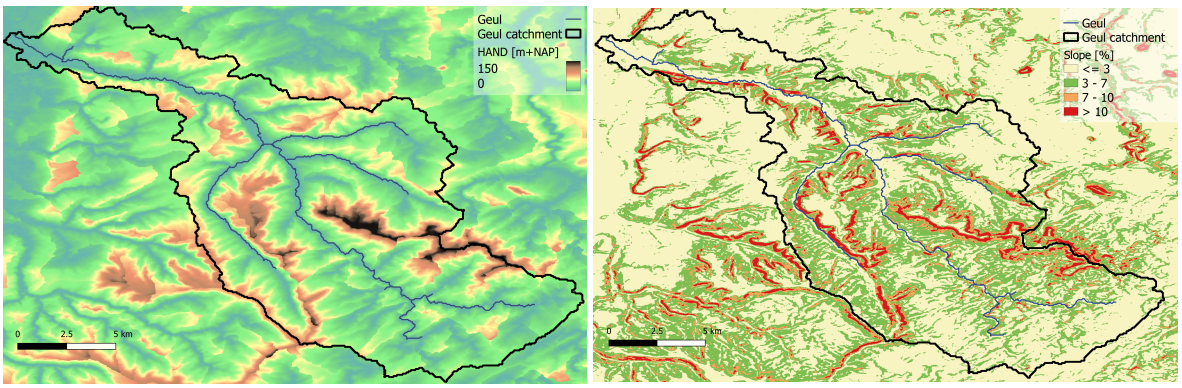
Wetlands are defined as the regions with the lowest elevation above a stream and therefore have less storage space between the groundwater table and the soil surface. This results in less capacity to store water, resulting in the dominant runoff process in wetlands being saturation overland flow (Gharari et al., 2011).

Hillslopes have a deeper groundwater table and, as the name suggests, a higher slope. Therefore, more water can be stored in the root zone, but the slope creates preferential flow paths to the river with quicker runoff, making the dominant runoff process sub surface flow (Gharari et al., 2011).

Plateaus have a more moderate slope and an even deeper groundwater table. This causes the primary flow generation processes on plateaus to be sub-surface storage and groundwater recharge in the form of deep percolation (Savenije, 2010).

3.2.1. Landscape classification

Previous research has shown by which characteristic landscapes can be classified: ”The height above the nearest drainage (HAND) and the surface slope, ..., appear to be the dominant topographical controls for hydrological classification.” (Gharari et al., 2011). In figures 3.5a and 3.5b, the HAND and slope of the Geul catchment are depicted. Already, a pattern emerges, with high HAND values being surrounded by steep slopes.



(a) HAND in m, elevation data from European Environment Agency (2016)

(b) Slope in %, calculated with elevation data from European Environment Agency (2016)

Figure 3.5: The dominant topographical controls for hydrological classification

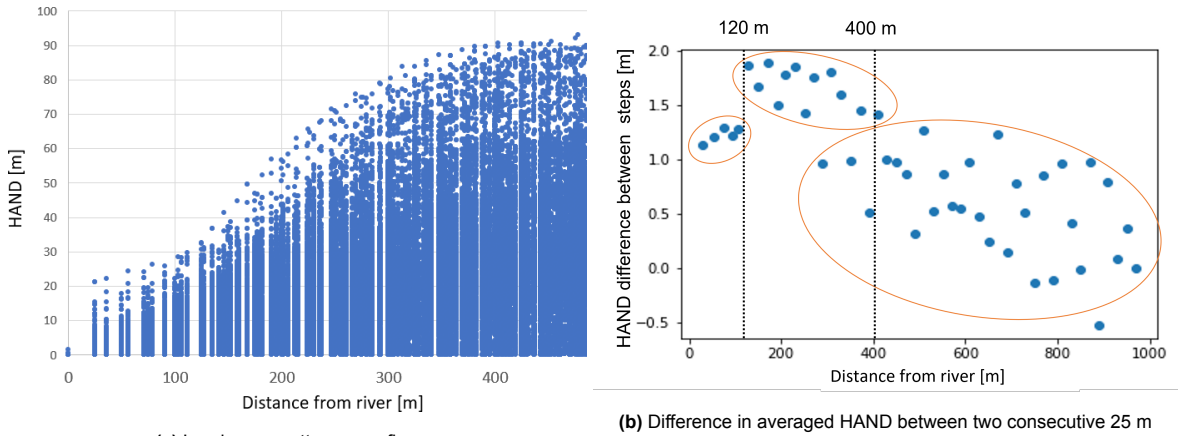
Table 3.2 gives an overview of which landscape corresponds to which combination of HAND and slope values. The specific threshold values for ’high’ and ’low’ are unique for each catchment and therefore need to be calculated.

Table 3.2: Landscape classification with general thresholds

	Low HAND	High HAND
Low slope	Wetlands	Plateau
High slope	Wetlands	Hill slope

3.2.2. Threshold values for HAND and slope

Both HAND and slope threshold is determined by first evaluating the relation between the HAND and the distance from the stream, as seen in Rennó et al. (2008), to determine if a landscape classification is fitting for the catchment. In figure 3.6a every dot represents a location in the Geul catchment with a distance from the river and a corresponding HAND value. This figure shows that the further away a location is from the river, the higher the point is located above the river. With first a more flat area, followed by a more steep area, and finally a flat area again, which is the same pattern as presented in figure 3.4.



(a) Landscape pattern, own figure

(b) Difference in averaged HAND between two consecutive 25 m distance from the river steps

Figure 3.6: Determining distance from the river for each landscape, on average

This pattern in figure 3.6a is used to calculate the distance between each landscape and the river. To determine the specific distance from the river at which each landscape begins, the average HAND per 25-meter increment from the distance to the river is calculated. The difference between every two consecutive steps is calculated and can be seen in figure 3.6b.

In the first 120 m, there is hardly any increase in average HAND between two consecutive steps, the slope is low, and this section is determined to be wetland. After 120 m there is a clear increase in average HAND between two consecutive 25 m steps, meaning a higher slope. The section from 120 m to 400 m is classified as a hillslope. After around 400 m the difference between two consecutive 25 m steps becomes smaller again, meaning that the slope is levelling off and from 400 m onward, the landscape is classified as a plateau.

Since it is now known how far each landscape is from the river, the average HAND and slope values of each landscape section can be calculated, thereby determining the threshold values for each landscape. The thresholds are 3.1 m and 4.8% respectively. Therefore, the landscape classification table can be filled in, and the result can be seen in table 3.3.

Table 3.3: Landscape classification with threshold values

	HAND<=3.1m	HAND >3.1m
Slope <= 4.8%	Wetlands	Plateau
Slope > 4.8%	Wetlands	Hill slope

There has been research on slope values for different dominant runoff processes. The calculated slope is compared with the literature to determine if the found value of 4.8% would actually be able to produce the dominant runoff process of hillslopes, namely subsurface flow. In two different studies (Haggard et al., 2005; Hümann & Müller, 2013) it has been found that subsurface flow becomes the dominant runoff process at a slope of 5% or greater in arable land, grassland, and forests. Which is not far off the 4.8% found in this research

3.2.3. Landscape distribution over the catchment

The threshold values from table 3.3 result in a distribution of classes over the catchment that can be seen in figure 3.7, with the percentages per landscape presented in table 3.4. As the model is lumped, this differentiation over the catchment is not used in the model, however, the percentage per landscape is used to define the landscapes in the model.

Table 3.4: Percentage per landscape class

	Percentage
Plateau	69.70%
Hillslope	23.34%
Wetland	6.96%

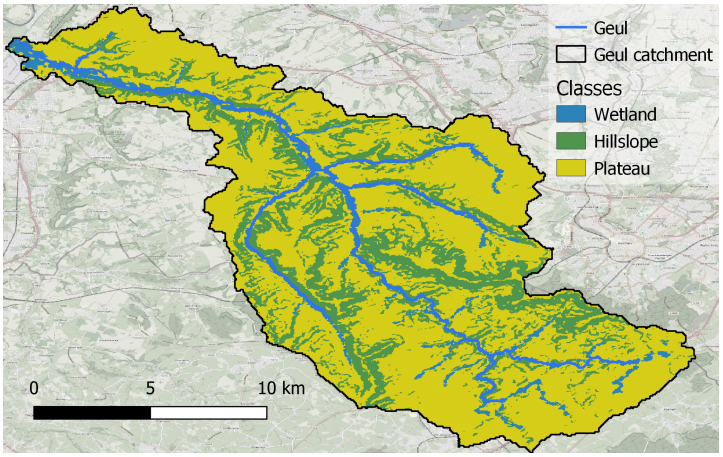


Figure 3.7: Landscape classification, own figure

3.2.4. Conceptual model per landscape class

An overview of the individual conceptual models per landscape class is depicted in figure 3.8. The differences between the models are implemented to assure that the model imitates the response that fits with the dominant runoff process of that landscape. Namely, saturation overland flow for wetlands, subsurface flow for hillslopes and subsurface storage, and deep percolation for plateaus

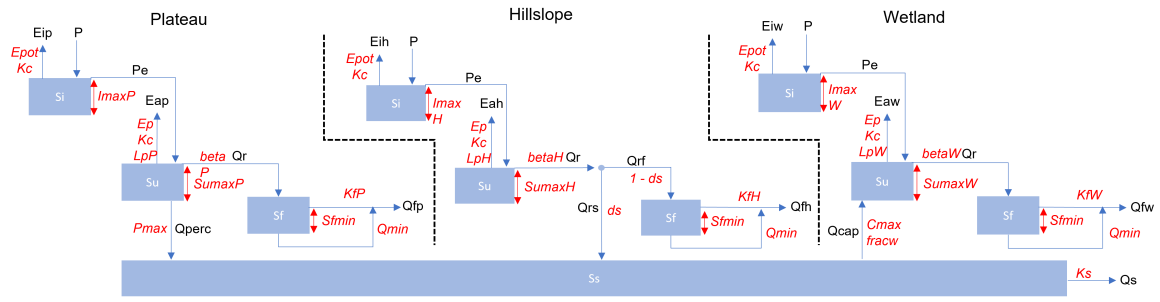


Figure 3.8: Individual model per landscape class, parameter in red are unique value per landscape, parameters in grey have the same value for all three landscapes, an enlargement of this figure can also be found in Appendix C

The discharge and evaporation from each landscape class need to be combined to end up with the total evaporation (E_{tot}) and discharge (Q_{tot} from the model. The groundwater storage is connected between each landscape and only has one outflow.

$$Q_{tot} = Q_{f,tot} + Q_s \tag{3.16}$$

$$Q_{f,tot} = Q_{f,plat} * frac_{plat} + Q_{f,hill} * frac_{hill} + Q_{f,wet} * frac_{wet} \tag{3.17}$$

$$E_{tot} = E_{i,tot} + E_{a,tot} \tag{3.18}$$

$$E_{i,tot} = E_{i,plat} * frac_{plat} + E_{i,hill} * frac_{hill} + E_{i,wet} * frac_{wet} \tag{3.19}$$

$$E_{a,tot} = E_{a,plat} * frac_{plat} + E_{a,hill} * frac_{hill} + E_{a,wet} * frac_{wet} \tag{3.20}$$

3.3. Procedural model

The conceptual model as described above is implemented within Python and should of course run without errors. The most essential test the model needs to pass is the closure of the water balance. All the water that enters the model should also leave the model, or be stored in storage, otherwise there are modelling errors.

$$dS/dt = P - E_{tot} - Q_{tot} - dS_i/dt - dS_u/dt - dS_f/dt - dS_f/dt \tag{3.21}$$

The average change in storage (water balance) during 300.000 runs is -5.20873E-13 mm, with a maximum of 3.52749E-11 mm and a minimum of -3.82752E-11 mm. Because of rounding, the water balance is not equal to zero 11 digits behind the comma, but this is sufficient.

3.4. Calibration

The process of calibrating a model involves selecting appropriate values for each parameter in the model. Generalised Likelihood Uncertainty Estimation (GLUE) is an approach for model calibration and uncertainty assessment based on generalised likelihood measures (Beven & Binley, 1992). The GLUE methodology adopts the concept of equifinality rather than the notion of a single, optimum solution. The calibration portion of the GLUE methodology consists of a number of steps, all of which are described in this section.

1. Each parameter is assigned a minimum and maximum value, as well as a prior distribution within this interval. The process of determining parameter interval values and distribution is described in section 3.4.1.
2. Parameter sets are created with parameter intervals, Monte Carlo random sampling, and parameter constraints. The process of parameter set creation is described in section 3.4.2.
3. Each parameter set from step two is evaluated by means of an objective function, in the case of this research the Nash Sutcliff Efficiency (Nash & Sutcliffe, 1970). The objective function is further described in section 3.4.3.
4. All parameter sets are run during a calibration period, and 'behaviour models' are selected, all other runs are rejected. The selection of these behaviour models is described in section 3.4.4.

The calibration of the model is between 2014 and the end of 2018. The warm-up period is from the first day of 2013 until the end of 2013. Calibration could not start in 2013 as the evaporation data for that year is not correct, as indicated in section 2.2.

3.4.1. Parameter interval

An interval is assigned to each parameter. This interval represents the maximum and minimum values between which the parameter value is expected to fall. The maximum and minimum interval values are assigned in one of three ways:

- Catchment observation-based parameter interval values. Some parameter values, such as maximum root zone storage ($S_{u,max}$), can be determined by calculations from available catchment data.
- Definition-based parameter interval values. Some parameter values, such as the splitting parameter (d_s), are always between 0 and 1.
- Literature-based parameter interval values. Some parameters do not have a direct link with the physical world, meaning that they do not reflect a natural process or storage but are parameters needed to, for instance, determine the degree of non-linearity of the soil runoff (β) or no observations of the parameter are available for the catchment.

In this research, the calibration process is done in two steps, first, all parameter sets which are acquired with the steps above are used. However, as some parameter values come from literature,

they might not be a direct fit with the hydrological processes in the Geul catchment, therefore dotty plots are used to increase or narrow down the intervals as necessary after a first calibration run. After some parameter intervals are improved, the calibration process is repeated. The interpretation and use of dotty plots are explained in section 3.5.1.

In the following section, all parameter interval values that are calculated from catchment observations are presented.

Maximum root zone storage capacity

The root zone storage is the total amount of water stored in the soil of the catchment, which plants and other trees or crops can access through their roots. The fact that these plants can survive for several years suggests that they have created a root system that assures continual access to enough water to bridge dry spells, but not more than is required (Hrachowitz et al., 2021). There are various methods for calculating root zone storage, but with the available data, the chosen method is to estimate it directly from observed water balance data using the maximum yearly water deficit.

The storage deficit is determined by calculating the daily cumulative difference between precipitation and evaporation. With the exception that this storage deficit can never be greater than zero because water cannot be stored in the ground once it is full, see equation 3.22. The minimum value of the cumulative storage deficit is the maximum storage required to bridge the dry periods; see equation 3.23.

$$S_{r,t} = min(0, \sum_{t}(P_t - E_{A,t})) \tag{3.22}$$

$$S_{r,max} = min(S_{r,t}) \tag{3.23}$$

In figure 3.9, the storage deficit over time is plotted next to daily precipitation and evaporation data. The effect of a large precipitation event can clearly be seen in the summers of 2014 and 2021 as the storage deficit is directly filled afterwards. The opposite behaviour can be observed in the dry summer of 2018, with the storage deficit reaching its lowest points. The maximum root zone storage is determined by the storage deficit in 2018, which is 230 mm. This, however, is only the lowest point in a 10-year time series. A longer time series is preferable because it is possible that a larger storage deficit can be calculated in a different year, therefore, an interval is taken between 200 and 260, where the individual landscapes each have different amounts.

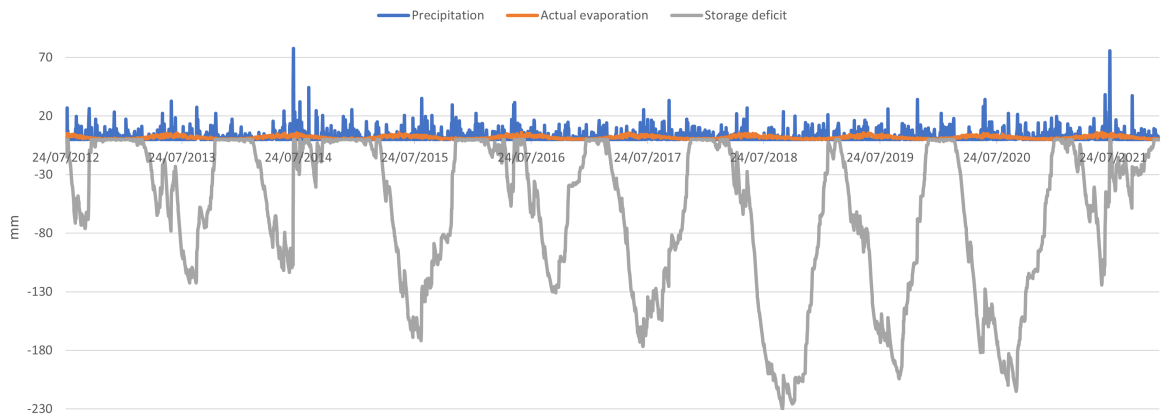


Figure 3.9: Storage deficit

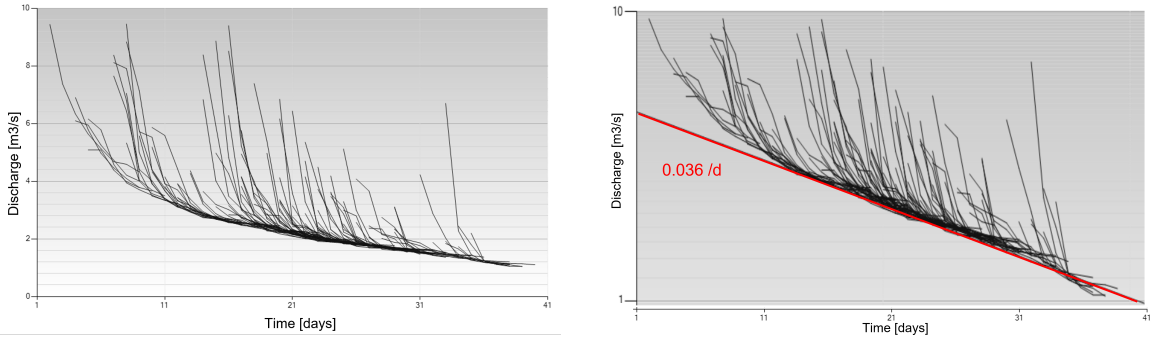
Slow runoff recession constants

The slow runoff recession constant determines how much of groundwater storage is discharged at each time step, as can be seen in equation 3.24.

$$Q_{s,t} = K_s * S_{S,t} \tag{3.24}$$

A large recession constant implies that more water flows out of the storage at each time step, and that water remains in the storage for only a short amount of time. One way of obtaining the recession

constant is to create a Master Recession Curve (MRC). In their publication, Fenicia et al. (2006) outline the technique for developing an MRC. An MRC is a graph that layers all recession curves on top of each other, indicating that all continuously declining regions of the hydrograph are plotted together. Because these recession curves occur during dry conditions, there is no other discharge reaction except the groundwater response. The pattern that emerges, as illustrated in figure 3.10a, indicates that all lines overlap in the lowest part of the graph, and this represents the recharge from the groundwater.



(a) Master recession curve on linear paper

(b) Master recession curve on log paper

Figure 3.10: Master recession curve

This lowest common part of the graph is linear on log paper, as can be seen in figure 3.10b, and the recession coefficient of this line is the same as on the slow runoff recession constant, namely 0.036. This means that a drop of water stays in the groundwater storage for around 1/0.036 = 28 days, which is relatively short for groundwater storage since groundwater recharge has a temporal scale between a couple of days and years (Loon, 2015).

Crop growing stages

All plants and crops evaporate water through their leaves. Because plants have changing leaf surfaces and hence different water needs throughout the year, transpiration is not a constant value throughout the year. To determine transpiration, crop coefficients are used. The transpiration is the potential evaporation times this crop coefficient (Allen et al., 1998), as expressed in equation 3.25.

$$E_T = K_{c,t} * E_{pot} \tag{3.25}$$

The crop coefficient (K_c) changes during the growing stages of the plant, as can be seen in figure 3.11. The MODIS Leaf Area Index (LAI) is used as a proxy to calculate the dates of the growing stages, as previously done by Pierik (2022).

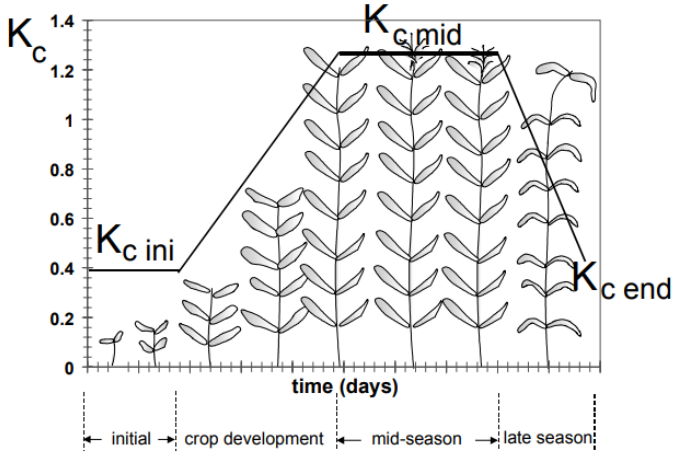


Figure 3.11: Crop coefficient curve (Allen et al., 1998)

figure 3.12 shows the average LAI of the Geul catchment throughout the years. The LAI indicates how many square meters of leaf there are on every square meter of area. A high LAI indicates that, on average, the plants in the catchment area in their mid-season with lots of leaves, and a low LAI indicates that the plants are in their initial season. By fitting a function through the LAI plot, the average starting dates of each season are determined.

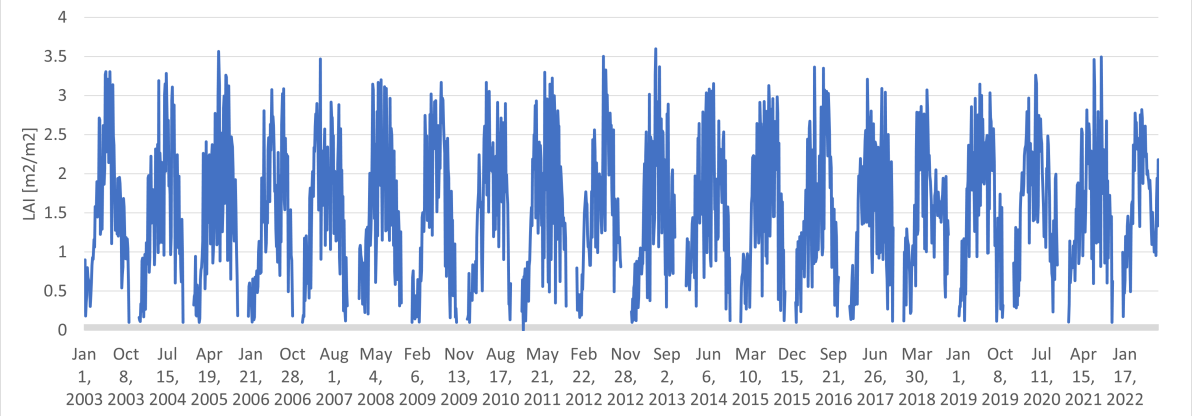


Figure 3.12: Average Leaf Area Index of the Geul catchment

The initial growing season runs from November 7 until March 23. The crop development season runs from the 23rd of March until the 13th of April. The mid-season runs from the 3rd of April until the 30th of September, and the late-season runs from the 30th of September until the 7th of November.

Minimum discharge

To be able to mimic the discharge response of the Geul, an extra storage element has been added to the fast runoff bucket, and a minimal discharge is implemented as explained in section 3.1.3. However, the exact amount of the minimum discharge needs to be calibrated. The lowest minimum discharge value is 0.2 mm/d, as this is the lowest measured discharge between 2014 and 2018. The upper minimal discharge value is 0.5 mm/d. This is due to the fact that the minimum discharge level varies from year to year, with the highest minimum discharge being 0.5 mm/d in the summer of 2017. As a result, the parameter value interval for the minimum discharge is 0.2 to 0.5 mm/d.

Time lag

The time lag is the time between a precipitation event and a resulting increase in discharge, as represented in figure 3.13.

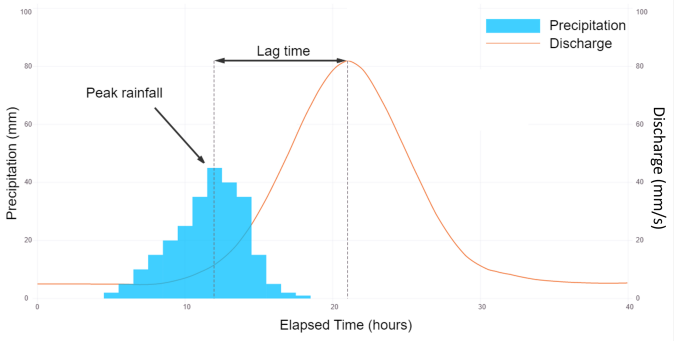


Figure 3.13: Visualisation of lag time (Jackson, 2014)

To calculate the lag time, the 10 largest precipitation events between 2014 and 2018 have been selected, and the time between the peak rainfall and the following peak discharge has been determined. The calculations have been done with daily precipitation and discharge measurements. The average time lag was 1.73 days, with a minimum of 1 and a maximum of 3 days. Therefore, the parameter interval value has been set between 1 and 3 days.

Overview of parameter interval values
An overview of all parameters that are used in the model and which need to be calibrated can be found in table 3.5 and table 3.6. Table 3.5 consists of all parameters that are the same for all landscapes or are only used in one landscape. Table 3.6 includes parameters that differ in value among landscapes. It is specified for each parameter whether the interval values are derived from catchment observations, definitions, or literature. When citing multiple sources, the interval represents the highest and lowest values rather than an average.

Table 3.5: Calibration catchment parameters including interval values

Name	Unit	Interval	Reference
$S_{u,max}$	mm	[200, 260]	Catchment observations
$S_{f,min}$	mm	[0, 500]	Wide interval, as there is no information
P_{max}	mm Δt^{-1}	[0, 6]	Beck et al. (2016), Seibert (1997)
C_{max}	mm Δt^{-1}	[0, 4]	Wawrzyniak et al. (2017)
K_s	Δt^{-1}	0.036	Catchment observations
d_s	-	[0, 1]	By definition between 0 and 1
Kc_{ini}	-	[0.1, 0.6]	Allen et al. (1998), Corbari et al. (2017)
Kc_{mid}	-	[0.4, 1.0]	Allen et al. (1998), Corbari et al. (2017)
Q_{min}	mm Δt^{-1}	[0.2, 0.5]	Catchment observations
T_{lag}	Δt	[1.0, 3.0]	Catchment observations

Hydrological processes differ among landscapes, as mentioned in section 3.2.4. To capture this behaviour, each landscape needs to have its own parameter value for specific parameters. The evaporation reduction threshold (L_p) varies depending on the soil type (Haqiqi et al., 2021), which differs per landscape; however, each landscape has the same parameter interval. The same goes for maximum interception storage (I_{max}), which differs per land use. The fast recession constant (K_f) also differs as water will run off more quickly under a larger slope. Furthermore, the fast recession constant should, per definition, be larger than the slow recession constant.

Table 3.6: Calibration landscape parameters including interval values

Name	Unit	Interval	Reference
$I_{max,p}$	mm	[0, 4]	Verbeiren et al. (2016), Zhong et al. (2022)
$I_{max,h}$	mm	[0, 4]	
$I_{max,w}$	mm	[0, 4]	
$S_{u,max,p}$	mm	[250, 400]	Catchment observation and expert opinion (Hrachowitz, 2022)
$S_{u,max,h}$	mm	[150, 300]	
$S_{u,max,w}$	mm	[5, 75]	
β_p	-	[0.3, 3.0]	de Boer-Euser (2017), Beck et al. (2016), Seibert (1997)
β_h	-	[0.3, 3.0]	
β_w	-	[0.3, 3.0]	
$L_{P,p}$	-	[0,1]	By definition between 0 and 1
$L_{P.h}$	-	[0,1]	
$L_{P,w}$	-	[0,1]	
$K_{f,p}$	Δt^{-1}	[0.036, 0.6]	Beck et al. (2016), Seibert (1997) and K_f > K_s
$K_{f,h}$	Δt^{-1}	[0.036, 0.6]	
$K_{f,w}$	Δt^{-1}	[0.036, 0.6]	

In this respect, maximum root zone storage ($S_{r,max}$) holds a unique position, as also the intervals differ per landscape. The maximum root zone parameter interval values in table 3.5 represent the interval between which the average of all catchments together amounts to, indexed by the percentage of the catchment that has that landscape.

Parameter interval distribution
With the parameter intervals provided in table 3.5 and table 3.6. There could be a previously known distribution within each interval. For example, a normal distribution with the parameter interval mean as the mean value of the distribution. However, there is no previous knowledge about the distribution, and therefore all parameters are given a continuous uniform distribution. This is most commonly done and represents the limited prior understanding of the uncertainties coming from models, parameters, and variables (Beven & Binley, 1992).

3.4.2. Parameter set creation

To form a parameter set, a value is selected for each parameter within it own interval, by its own distribution. As the distribution is assumed to be uniform, the parameter value is selected at random between the minimum and maximum values. This selection process makes sure that the parameter value in itself is correct, but there are also relationships between parameter values, also called constraints, that make sure some parameter values can never be larger or smaller than another value. Parameter sets that do not fulfil both rules are rejected.

Monte Carlo random sampling
The parameter sets are created using Monte Carlo random sampling. As Monte Carlo is a random sampling method, this means that each combination of parameter values is selected at random which can lead to oversampling of some areas and undersampling of others. Therefore, it is important to have a lot of runs, more generally, for n samples per parameter and p parameters the number of required runs is calculated with equation F.34:

$$\text{number of runs} = n^p \tag{3.26}$$

This limitation of Monte Carlo sampling ties in with the third factor namely the large number of parameters in the Wflow model. For 24 parameters with only 10 randomly selected values each, 10^{24} runs are necessary, with a lower limit of at least 100.000 runs (expert opinion Hrachowitz, 2023). The random sampling method is implemented by selecting a random value between 0 and 1 for each parameter set. The value per parameter is calculated with this random value, as can be seen in equation 3.27 with a random value of 0.313412, a minimum parameter value of 0.3 and a maximum parameter value of 3.0

$$\beta_h = R * (\beta_{h,max} - \beta_{h,min}) + \beta_{h,min} \tag{3.27}$$

$$\beta_h = 0.313412 * (3.0 - 0.3) + 0.3 = 1.1462124 \tag{3.28}$$

Parameter constrains
There are parameter constraints in addition to the interval in which a certain parameter value must be. Certain parameters are constrained by external factors in relation to other parameters. A large number of parameters makes it problematic to select 'true' values of parameters, as this leads to equifinality (Beven, 1993). To reduce the degrees of freedom, parameter constraints are used to set up external constraints in relationship to other parameters next to individual parameter intervals. As they reduce the parameter space qualitatively higher parameter sets are tested which lead to better calibrated results (Nijzink et al., 2016)

Parameter constraints	Explanation
200 < $S_{u,max,w}$ * $frac_w$ + $S_{u,max,h}$ * $frac_h$ + $S_{u,max,p}$ * $frac_p$ < 260	Total root zone storage needs to represent catchment observation
$S_{u,max,w}$ < $S_{u,max,h}$ < $S_{u,max,p}$	Proximity of the ground level to the water table
$K_{f,w}$ $K_{f,p}$ < $K_{f,h}$	Higher slopes cause faster flow

3.4.3. Objective function

Each parameter set is used to run the model once, and the resulting discharges and/or evaporation are compared to the observed discharge and/or evaporation. The goal of the objective functions is to find the parameter sets that most closely represent the observed data. The calibration period for this model is from 2014 till the end of 2018.

To further decrease the degree of freedom, to limit equifinality, there are multiple objection functions. Four in total, two for discharge and two for evaporation. The objective of calibration is to minimise the aggregated errors between model output and measured data. For this goal four objective functions are used, the Nash-Sutcliffe efficiency (NSE) and the log NSE for both discharge and evaporation, equation 3.29 and 3.30 for discharge and equation 3.31 and 3.32 for evaporation.

$$NSE_Q = 1 - \frac{\sum_{t=1}^{T}(Q_o^t - Q_m^t)^2}{\sum_{t=1}^{T}(Q_o^t - \overline{Q_m})^2} \tag{3.29}$$

$$NSE_{Q,log} = 1 - \frac{\sum_{t=1}^{T}(log(Q_o^t) - log(Q_m^t))^2}{\sum_{t=1}^{T}(log(Q_o^t) - \overline{log(Q_m)})^2} \tag{3.30}$$

$$NSE_E = 1 - \frac{\sum_{t=1}^{T}(E_o^t - E_m^t)^2}{\sum_{t=1}^{T}(E_o^t - \overline{E_m})^2} \tag{3.31}$$

$$NSE_{E,log} = 1 - \frac{\sum_{t=1}^{T}(log(E_o^t) - log(E_m^t))^2}{\sum_{t=1}^{T}(log(E_o^t) - \overline{log(E_m)})^2} \tag{3.32}$$

Both NSE and NSE log are used, as each objective function has a different focus. Absolute errors are larger during high flow events than low flow events, and as can be seen in equation 3.29 the errors between modelled and observed discharge are squared. This has the effect that errors in high flows are inflated and that the parameter sets that get the high flows correct receive a high NSE value. This comes at the price of poor low flow representation. To be able to get both low and high flows correct, also the log of NSE is used. By taking the log of the squared errors, the absolute errors are balanced by compressing the high flows and stretching the low flow errors. Therefore, combining the two objective functions results in a model which can predict the low flow and high flows equally well, the same is done for evaporation.

These two efficiencies are combined into one number to make the comparison process easier. This is done by using the Euclidean distance, see equation 3.33 for discharge and 3.34 for evaporation. Both these scores can also be combined, as can be seen in equation 3.35. This means that there are three different scores per parameter set:

- $D_{e,Q}$, see equation 3.33, representing how closely the discharge output resembles the observed discharge while taking both the low flows and high flows into account.
- $D_{e,E}$, see equation 3.34, representing how closely the evaporation output resembles the actual evaporation while taking both the low evaporation and high evaporation into account.
- $D_{e,Tot}$, see equation 3.35, representing how closely the discharge and evaporation output observed resembles discharge and actual evaporation while taking both the low flows and evaporation and high flows and evaporation into account.

$$D_{e,Q} = \sqrt{(1 - NSE_Q)^2 + (1 - NSE_{Q,log})^2} \tag{3.33}$$

$$D_{e,E} = \sqrt{(1 - NSE_E)^2 + (1 - NSE_{E,log})^2} \tag{3.34}$$

$$D_{e,Tot} = \sqrt{(1 - NSE_Q)^2 + (1 - NSE_{Q,log})^2 + (1 - NSE_E)^2 + (1 - NSE_{E,log})^2} \tag{3.35}$$

Interpretation
An NSE score of 0 indicates that a model can predict the streamflow or evaporation equally well as the average observed streamflow or evaporation would, and an NSE score of 1 indicates a perfect fit between modelled and observed streamflow or evaporation. However, a model will often score between 0 and 1 and these values also need an interpretation. Moriasis et al. concluded in (1983) that in general, model simulation can be judged as satisfactory if NSE > 0.50. Later in a different research group, Moriasis et al. (2015) gave some more differentiation by indicating different classes within the NSE scores as represented in table 3.7. The same intervals are assumed to be correct for the NSE log. The difference between the two only lies in the meaning of the score. Meaning that a high NSE score indicates a good representation of high flows and a high NSE log score indicates a good representation of low flows.

Table 3.7: Interpretation of NSE values (Moriasi et al., 2015)

Unsatisfactory	Satisfactory	Good	Very good
NSE =< 0.50	0.50 < NSE =< 0.70	0.70 < NSE =< 0.80	NSE > 0.80

As the Euclidean distance score are a combination of an NSE and NSE log score, the values from table 3.7 can be used to calculate the same boundaries for the Euclidean distance scores. These are presented in table 3.8, however, it is important to recognise that a De score > 0.7 does not mean that both NSE and NSE log are below 0.5. A combination of an NSE score of 0.4 and an NSE log score 0.8 also gives a De score of 0.63.

Table 3.8: Interpretation of Euclidean distance values (Moriasi et al., 2015)

Unsatisfactory	Satisfactory	Good	Very good
De => 0.70	0.70 > De => 0.42	0.42 > De => 0.28	De < 0.28

3.4.4. Selecting behaviour model runs
Within the GLUE calibration methodology, not one parameter set is selected as being the optimum solution, but instead, a selection of all behaviour model runs is selected. Behaviour models are all model runs that score above a certain threshold on the objective functions. The selection of the value for this distinction between behavioural and non-behavioural can have a significant impact on posterior distributions (Melching, 1995). From the results of the calibration, it became clear that one threshold value for all objective functions would not work, as it turned out that evaporation was easier to calibrate and therefore the parameter runs scored way higher on $D_{e,E}$ than $D_{e,Q}$ and $D_{e,Tot}$. Therefore, not a threshold value but a threshold amount has been used in this study. From each objective function, the 300 runs with the highest scores have been selected as the behaviour model runs, and these runs will be used to perform the model evaluation.

3.5. Model evaluation

During the evaluation of the model, the selected behaviour models are run once again. However, this time for a new period, between 2019 and 2021, once again with a warm-up period. The warm-up period is from 01/01/2013 until the end of 2018. With the results of these new runs, the model output can be evaluated.

3.5.1. Model parameter evaluation
Parameter uncertainty can be represented with dotty plots, an example of a dotty plot of two different parameters can be seen in figure 3.14. Each dot in a dotty plot represents one parameter set. The x-axis represents the parameter interval value of the parameter in question. On the y-axis, the likelihood score that each parameter set has scored is represented. In this research, the likelihood score will

either be $D_{e,Q}$, $D_{e,E}$ or $D_{e,Tot}$ meaning that each parameter can be plotted against all three objective functions.

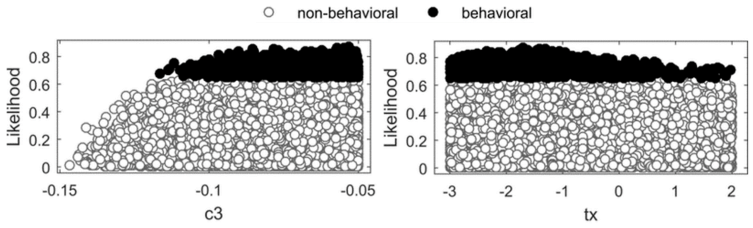


Figure 3.14: Example of two parameter dotty plots including behaviour model threshold (Tweldebrahn et al., 2018)

In figure 3.14, there is also a colour distinction visible. This distinction represents the difference between behaviour model runs, and non-behaviour model runs. In this case, the threshold value was about 0.6 on the likelihood scale.

3.5.2. Model performance evaluation

Each run is once again scored by the objection functions. During the calibration period, the parameters are selected that represent the streamflow and evaporation observations the best between 2014 and 2018. During the evaluation period, these selected parameter sets are put to the test, and it is determined how good the parameter sets are at predicting. Within the traditional evaluation, each model run would be scored and the parameter set which scored the highest on all objective functions is selected as the best predictor. With the GLUE methodology, all parameter sets of behaviour models as selected and thereby give an interval of predicted discharge. With this interval, the uncertainty in input and calibration data as well as the model structure is also represented in the output of the model, where this uncertainty is neglected in traditional evaluation with only one model as output and therefore allows the GLUE procedure the modeller to be realistic about the uncertainties associated with his or her modelling (Beven & Binley, 1992).

streamflow

The creation of these plots, which consist of all behaviour models, has a number of steps which need to be followed.

1. Each behaviour model needs to be re-run for the evaluation period.
2. Per behaviour model, the objective function score is calculated.
3. Each behaviour model is given a weight by dividing the objective score of that model by the sum of the objective scores of all behaviour models.
4. For every time step, the predicted streamflow values are sorted in ascending order and the corresponding weights follow this order.
5. The cumulative weight is determined, starting from the lowest discharge. The cumulative weight of the highest discharge prediction should be 1.
6. The discharges with a cumulative weight lower than 10 or higher than 90 are dropped.
7. The remaining discharge prediction is used as the interval of that time step.

In essence, this methodology looks a lot like calculating a normal confidence interval, but the difference here is that the model runs are scored, and therefore models that were better at predicting the observed values are more dominant in the final interval. Furthermore, this is the interval of the selected behaviour model, therefore the interval could be different if a higher or lower number of models was selected.

Runoff coefficient

The runoff coefficient is described as the fraction of precipitation that turns into discharge. As described in section 2.2.3, the Budyko framework is used to evaluate the long-term water balance of the input data. The distance a point plots below the water limit is equal to the runoff coefficient. The distance between the x-axis and the plotted catchment is the evaporation coefficient. Where the Budyko framework has

previously been used to check data quality, it is also possible to evaluate the models' runoff coefficient with the observed runoff coefficient. If the model can predict the runoff coefficient correctly, is also the model's evaporation correct as these two are bound together by the water balance, see equations below.

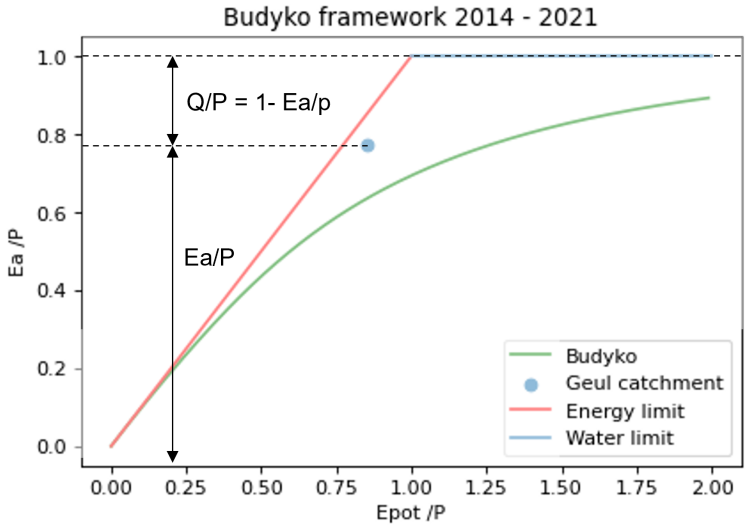


Figure 3.15: Runoff coefficient in Budyko framework

Flow duration curve

The time series of discharge over time, a hydrograph, is only one of the hydrological signatures of a catchment. Because the goal of a hydrological model is to predict streamflow, the hydrograph is most commonly used to determine if a model is working well. However, there are multiple other hydrological signatures which contain complementary information about the response, and can therefore also be used to evaluate the model on. The Flow Duration Curve (FDC) visualises the probability a discharge is exceeded during a time period. A steep slope in the flow duration curve indicates the flashiness of the streamflow response to precipitation inputs, whereas a flatter curve indicates a relatively damped response and higher storage (Yadav et al., 2007).

The FDC of the observed streamflow will be compared with the FDC of the behaviour models of the model calibrated on discharge and the model calibrated on both discharge and evaporation with the NSE log and not NSE, to decrease the influence the high flows have on the outcome.

Monthly streamflow and evaporation averages

Monthly average streamflow and evaporation comparison between the observed values and the modelled outcome can give an indication of how the model performs throughout the year, it can for instance be the case that the model always overestimates evaporation in the winter and underestimates evaporation in the summer. With these average monthly statistics, the model performance can be compared to the observed values for each month.

The monthly average streamflow and evaporation of the observed values will be compared with the monthly average streamflow and evaporation of the behaviour models of the model calibrated on discharge and the model calibrated on both discharge and evaporation with the NSE log and not NSE, to decrease the influence the high flows have on the outcome.

Cumulative streamflow and evaporation

The cumulative streamflow and evaporation comparison between the observed values and the modelled outcome can give an indication of how the model performs throughout the years, it can, for instance, be the case that the model calibrated on discharge, gives good predictions for all years except one, where it grossly overestimates the stream. Furthermore, it becomes visible if the models predict in the right order of magnitudes.

As these plots are cumulative, not the normal average observed values are used, but the accumulated average discharge and evaporation. The average discharge per day is 0.710 mm/d, therefore

equation 3.36 with x as the number of days since the first of January 2019 and y as the cumulative discharge is used as the average value in the NSE log instead of the normal average of 0.710 mm/d.

$$y = 0.710x \tag{3.36}$$

The average evaporation per day is 1.806 mm/d, therefore equation 3.37 with x is the number of days since the first of January 2019 and y is the cumulative evaporation is used as the average value in the NSE log instead of the normal average of 1.806 mm/d. The model calibrated on discharge and evaporation outperforms the model calibrated on discharge only.

$$y = 1.806x \tag{3.37}$$

3.5.3. Significance level

For each performance indicator that is described in the previous sections, the model calibrated on streamflow is compared with the model calibrated on both streamflow and evaporation, and the significance level is calculated. Meaning that the probability is calculated of concluding that a difference exists between the two runs when there is no actual difference.

A significance score of 0.3 indicates for instance that there is a 30.0% chance of concluding that a difference exists between the two models when there is no actual difference. The generally accepted threshold for this value is 5% or a p-value of 0.05, although there has been some debate to lower the threshold to 0.005 most researchers still make use of the classic 0.05 threshold (Leo & Sardanelli, 2020). In conclusion, the average of the behaviour models of the two models only performs significantly differently from each other on a specific performance indicator if the significance level between the results of the two models is lower than 0.05.

4

Results

In this chapter, the results from the model calibration and evaluation are presented. As previously described, the calibration period is between 2014 and the end of 2018, and the evaluation period is between 2019 and the end of 2021. During the evaluation of the model, the parameter sets that scored the best during the calibration period are used for the evaluation. The warm-up period before calibration is one year.

4.1. Calibration sets results

Three different calibration sets are run during the calibration period, from which an overview of their characteristics can be found in table 4.1. The original calibration intervals described in tables 3.5 and 3.6 are used to make 300.000 parameter sets. The dotty plots of these runs can be found in Appendix E. From the dotty plots of these original intervals, it becomes clear that some parameter intervals were too wide while others were too narrow; furthermore, only 4.07% of the parameter sets scored lower than 1.4 on De_Q. Therefore, some changes are made to the parameter interval values, which can be found in table 4.1 under new interval.

Calibration set A, with the new parameter intervals, has been run with 100.000 parameter sets. This lower number of runs has been selected because the percentage of runs that scored De_Q below 1.4 has greatly improved. With the new parameter intervals, 29.9% of the parameter sets scored lower than 1.4. As calibration set A now outperforms the original interval, the original interval will not be used.

Calibration set B uses the same parameter intervals as calibration set A; however, the storages in the model are now filled halfway at the start of the warm-up period. Once again, 100.000 parameters are constructed and run. The addition of initial conditions did not improve the percentage of runs that scored De_Q < 1.4, but the calibration set is used for evaluation to determine the effect of initial conditions on model performance.

Table 4.1: Calibration sets, percentage accepted is the percentage of runs where both NSE Q and NSE Q log were above 0.0.

Name	**Unit**	**Original interval**	**New interval**	**Initial condition**
P_{max}	mm Δt^{-1}	[0, 6]	[0, 2]	[0, 2]
$L_{P,p}$	-	[0,1]	[0, 0.75]	[0, 0.75]
β_P	-	[0.3 , 3]	[1, 3]	[1, 3]
Kc_{mid}	-	[0.4, 1]	[0.6, 1]	[0.6, 1]
T_{lag}	mm Δt^{-1}	[1, 3]	[1.5, 3.5]	[1.5, 3.5]
S_{ini}	mm	0	0	0.5 * S_{max}
Calibration set	-	Not used	A	B
Percentage De_Q < 1.4	-	4.07%	29.9 %	16.2 %

4.1.1. Calibration set A and B

From each calibration set, only the 300 best-performing parameter sets when scoring on discharge are kept and form Model Q. The 300 best scoring parameter sets on both discharge and evaporation are kept and form Model QE. The results of Model Q and Model QE for both calibration sets A and B are visualised in figure 4.1. It becomes clear that within calibration set A and B, the selected parameter sets for Model Q and Model QE score very similarly, with an average of 0.573 for Model Q and 0.576 for Model QE in calibration set A and 0.641 for Model Q and 0.643 for Model QE in calibration set B. But there is a difference between the performance of both models between calibration sets A and B.

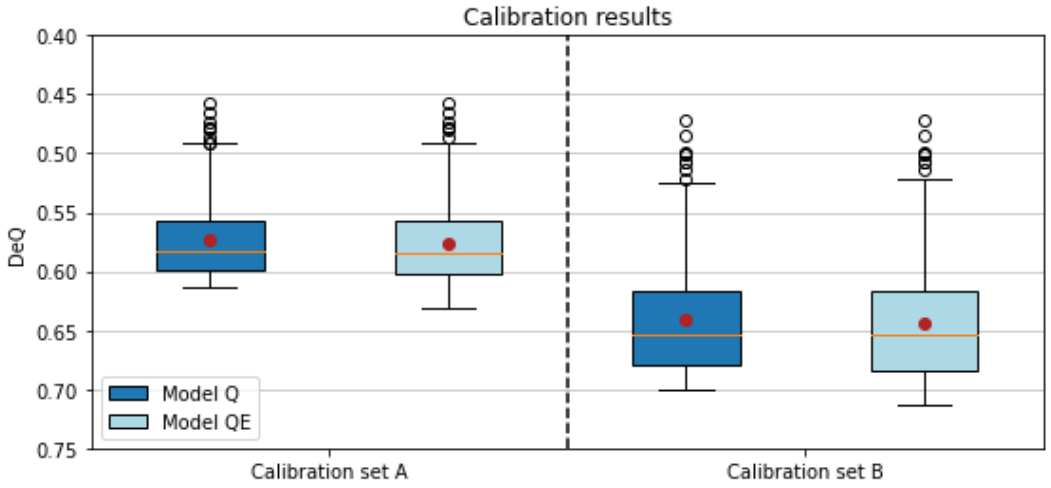


Figure 4.1: Calibration results per calibration set

From the results in figure 4.1, it becomes clear that the parameter sets predict streamflow amounts closer to the observed streamflow, if there is no initial storage in the model, compared to having the storages halfway filled at the start of the warm-up of the calibration period.

Table 4.2: Significance level between models and calibration sets

	Model QE Calibration set A	Model Q Calibrationset B
Model Q Calibration set A	0.424	<0.0001
Model QE Calibration set B	<0.0001	0.5527

From the significance levels in table 4.2, it becomes clear that the visible difference between the results of calibration sets A and B for both Model Q and Model QE are significant, but there is no significant difference between Model Q and Model QE within each calibration set. Therefore it can be concluded that calibration set A has better performance for both Model Q and Model QE, but there is no difference in performance between Model Q and Model QE.

4.1.2. Parameter value uncertainty intervals

As calibration set A, for both models Q and QE, significantly outperforms both models in calibration set B, the dotty plots of calibration set A are plotted in the following figures. In figure 4.2, the parameter values of calibration set A are plotted against De_Q. In figure 4.3, the parameter values are plotted against De_{Tot} values. The dotty plots of the parameter values against the underlaying NSE and NSE log scores of discharge and evaporation can be found in Appendix E, as well as the dotty plots of calibration set B. From the dotty plots, it becomes clear that many of the parameters are not clearly defined, except for some such as Pmax, LpP, kfH and KCmid. This indicates that some parameters do not seem to have a large impact on the model's performance.

Calibration set A, model Q

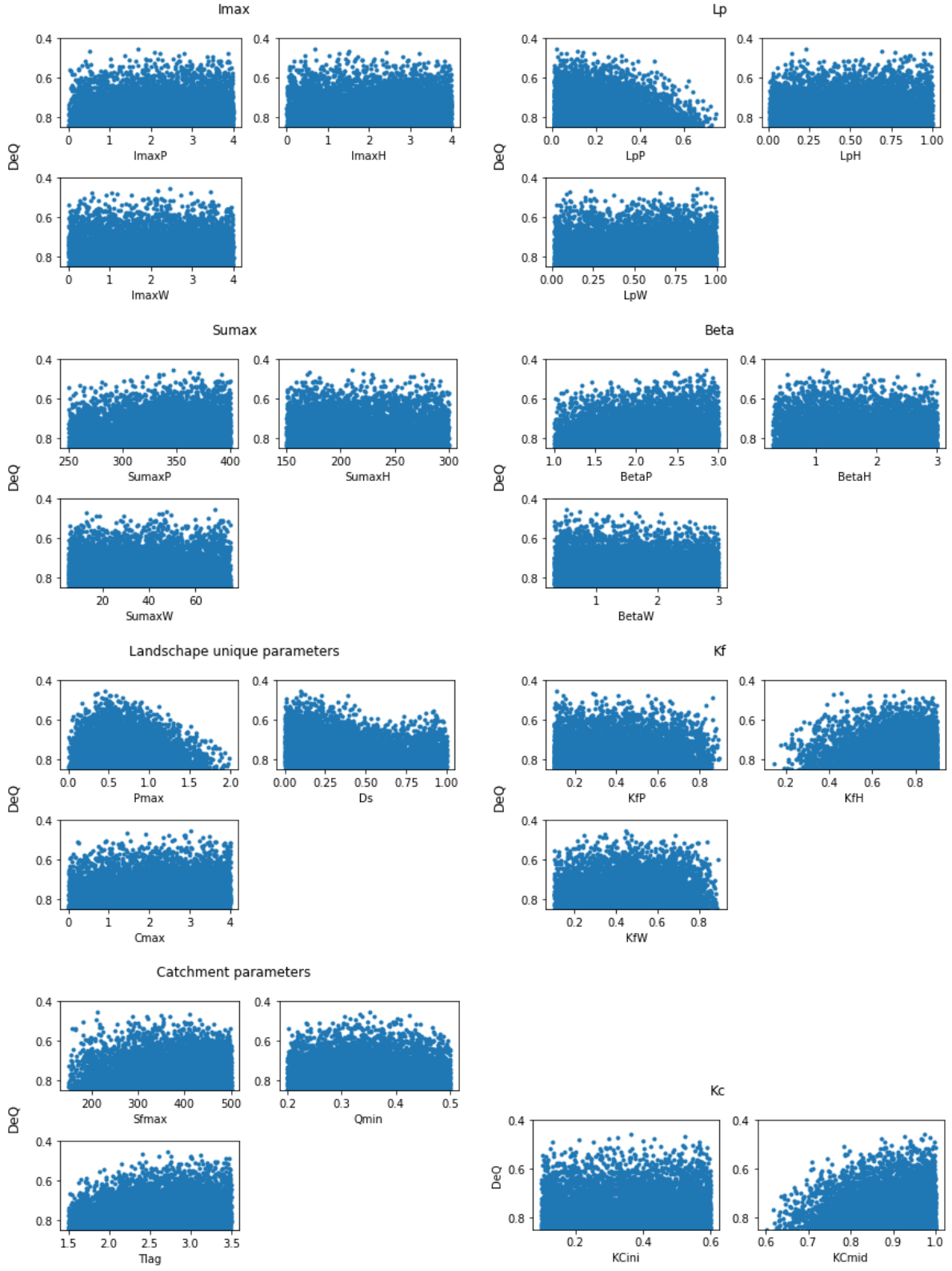


Figure 4.2: Dotty plot DeQ of calibration set A, Model Q

Calibration set A, model QE

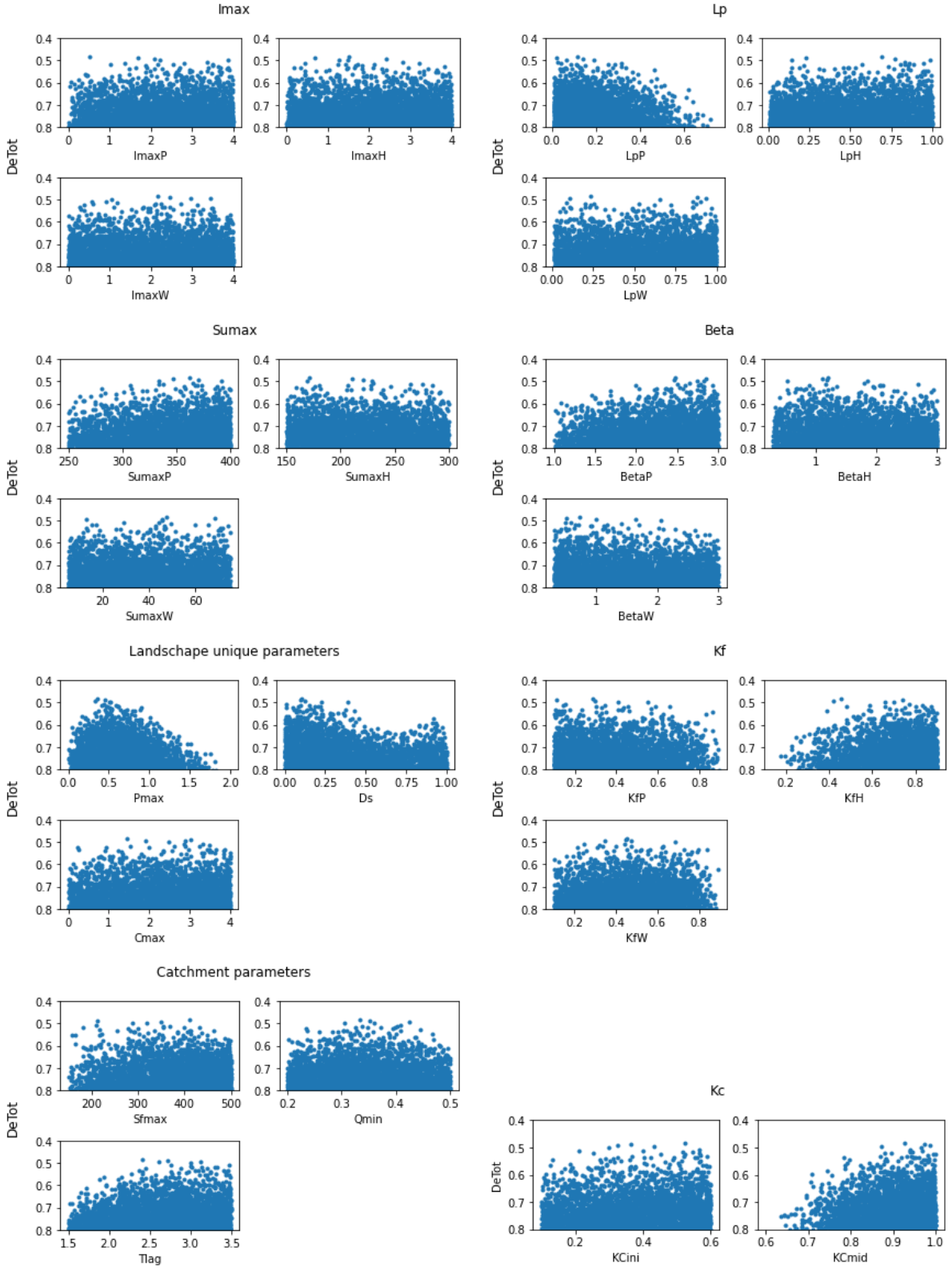


Figure 4.3: Dotty plot DeTot of calibration set A, Model QE

4.2. Evaluation runs results

During the evaluation period, the 300 best scoring parameter sets from Model Q and Model QE from calibration sets A and B are used to evaluate model performance from 2019 until 2021. In table 4.3, the different evaluation runs are depicted. Evaluation run 1 uses calibration set A, meaning no initial storage for the warm-up period, and evaluation run 2 uses calibration set B, with initial storage.

After evaluation run 2, it became clear that the inclusion of initial conditions did not have a positive effect on the evaluation results. Therefore, evaluation run 3 has been done without initial conditions. The difference between evaluation runs 1 and 3, which are both done with calibration set A, is that the minimum discharge method that had been designed to capture the catchment observed hydrograph, as described in section 3.1.3 has been turned off since February 2021.

This has been done as the river hydrograph from evaluation run 2 changed following a high flow event in the Geul in February 2021. As the high flow event could have influenced the conditions in the catchment, the model is adapted to switch off the minimal discharge as it is likely that some weirs in the catchment have been damaged and no longer hold the water back. By comparing evaluation runs two and three, the effect of excluding the minimal discharge process from February 2021 onward becomes clear.

Table 4.3: Evaluation runs

Name	Run 1	Run 2	Run 3
Calibration set number	A	B	A
Calibration set with initial storages?	No	Yes	No
Minimal Q turned off	No	No	Yes

Each model is scored on streamflow and the monthly runoff coefficient. As streamflow is scored on De_Q and the monthly runoff coefficient on NSE, it is important to notice that a lower score on De_Q and a higher score on NSE indicated better model predictions compared to the observed values.

4.2.1. Stream flow results

In this section, the observed stream flow is compared to modelled streamflow. The models are scored on De_Q. The results are visualised in a box plot per run per model, as can be seen in figure 4.4.

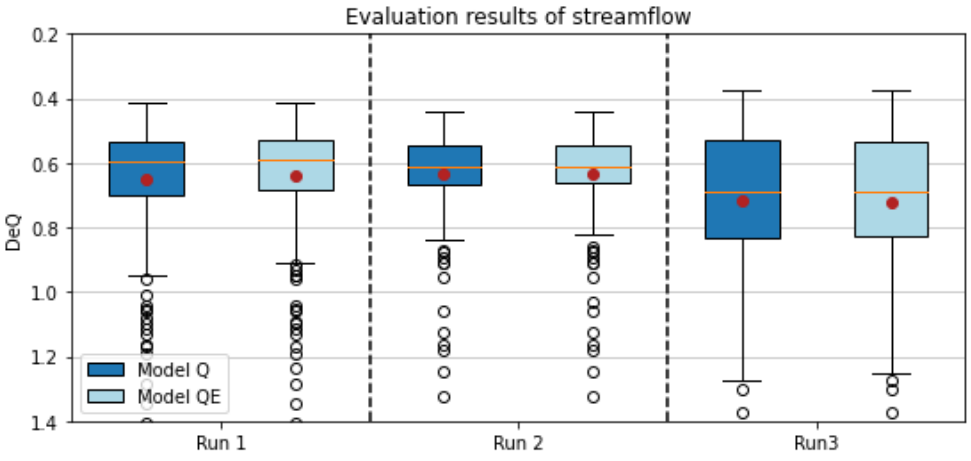


Figure 4.4: Evaluation results of streamflow

In figure 4.4, the streamflow results of both models for each run are plotted. Which showed that Model Q and Model QE, scored very similarly during each run, and the averages of Model Q and Model QE of each run also lay close together. The average scores per run per model can be seen in table 4.4.

Table 4.4: Average model performance per evaluation run for streamflow [DeQ]

		Run 1	Run 2	Run 3
Stream flow [DeQ]	Model Q	0.649	0.633	0.643
	Model QE	0.642	0.632	0.642

In table 4.5, the significance levels between Model Q and Model QE are presented, as the averages form table 4.4 and the spread in figure 4.4 also indicated, there is no significant difference in performance between Model Q and Model QE in any of the runs.

In table 4.6, the significance levels between each evaluation run for both Model Q as Model QE are presented. As no significance score is lower than 0.05 is must be concluded that there is no difference in performance between evaluation runs for both Model Q as Model QE.

Table 4.5: Significance level between Model Q and Model QE

	Significance level
Run 1	0.6394
Run 2	0.9289
Run 3	0.9438

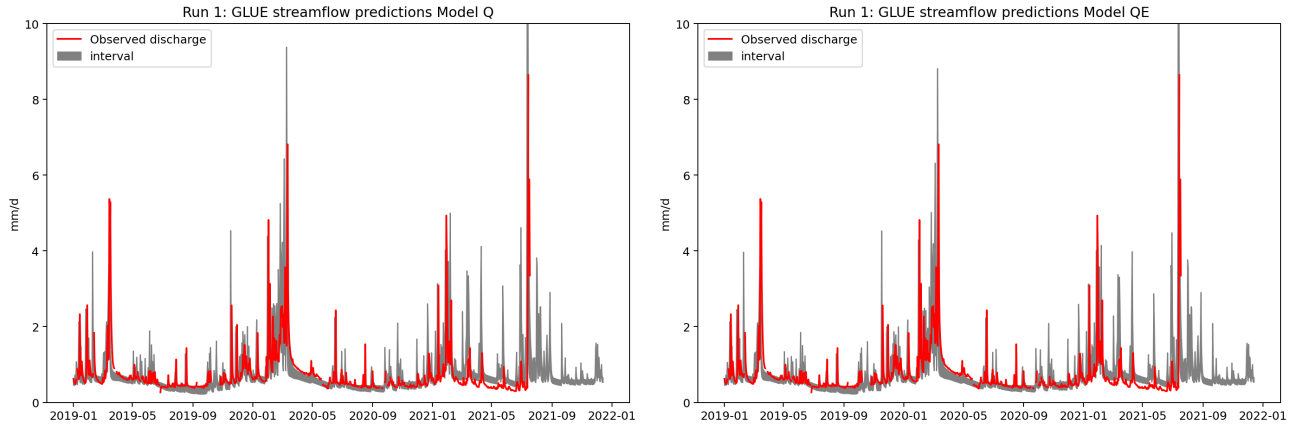
Table 4.6: Significance level between each of the three runs

		Significance level
Run 1 vs Run 2	Model Q	0.2579
	Model QE	0.4683
Run 2 vs Run 3	Model Q	0.5398
	Model QE	0.5576
Run 3 vs Run 1	Model Q	0.7394
	Model QE	0.9919

In the following sections, the streamflow predictions per model and per evaluation run are plotted All streamflow plots in the following sections are also available in full size in Appendix E.

Run 1

In figure 4.5a and figure 4.5b, the streamflow predictions for Model Q and Model QE are plotted. The difference between the two models is mostly visible in the peak discharges. The peak in 2020-03 in Model Q is higher than in Model QE and the same goes for the smaller peaks around 2021-02 and the July 2021 peak. Overall, Model QE seems to predict a lower streamflow than Model Q. To see if this could have something to do with the initial storage amount, the same run is done with initial storage in run 2.



(a) Model Q

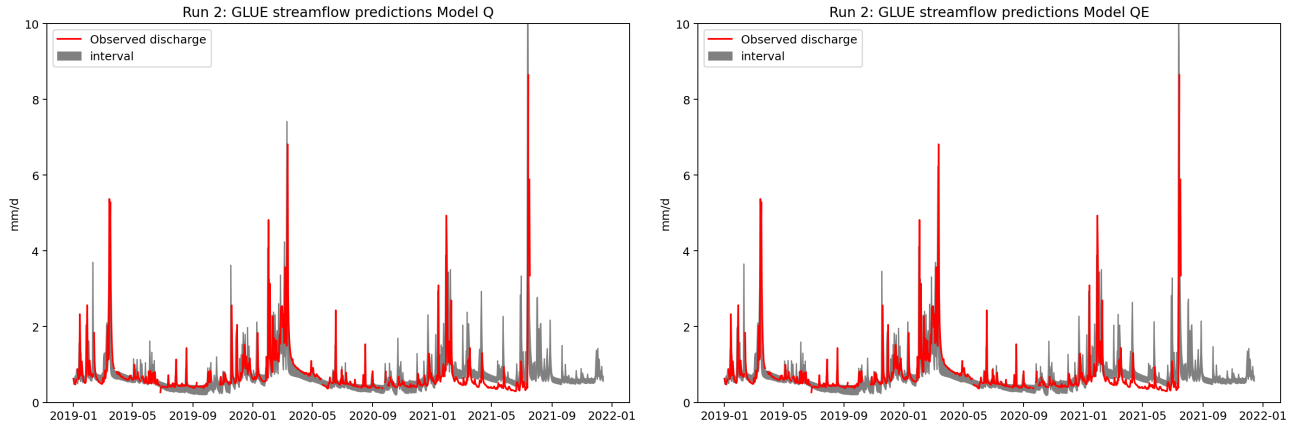
(b) Model QE

Figure 4.5: Run 1

Run 2

In figure 4.6a and figure 4.6b, the streamflow predictions for Model Q and Model QE are plotted, respectively. Compared to the results from run 1, both Model Q and Model QE predict lower discharges,

with Model QE predicting discharges even lower than Model Q. This can once again be most clearly seen in the peaks of 2020–03, around 2021–02, and the July 2021 peak. A second point of interest is the area between 2021-02 and 2021-07. Here, the predicted interval lies entirely above the observed values for both Model Q and Model QE. This was also the case during run 1, but during run 2, this difference is more apparent, which is interesting as the other parts of the prediction interval lay lower in run 2.

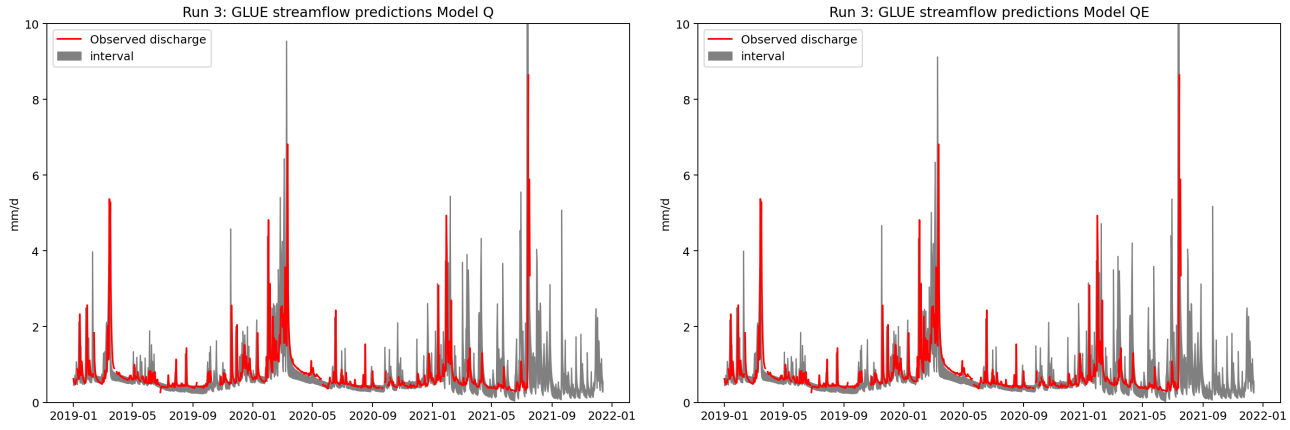


(a) Model Q **(b)** Model QE

Figure 4.6: Run 2

Run 3

In figure 4.7a and figure 4.7b, the streamflow predictions for Model Q and Model QE are plotted, respectively. The effect of switching the Q_{min} after the high flow event in 2021-02, is clearly visible in the plots of both models. After 2021-02, the prediction interval is more volatile and has quicker changes. Model Q once again, predicts larger peak flows, such as in 2020–03, than Model QE.



(a) Model Q **(b)** Model QE

Figure 4.7: Run 3

4.2.2. Runoff coefficient results

In this section, the monthly runoff coefficient from the observations is compared to the modelled monthly runoff coefficients of the three calibrated models during all three runs. The models are scored on NSE between each modelled run and the observed. These results are visualised in a box plot per run. Furthermore, the observed monthly runoff coefficient is plotted against the monthly runoff coefficient of all behaviour models.

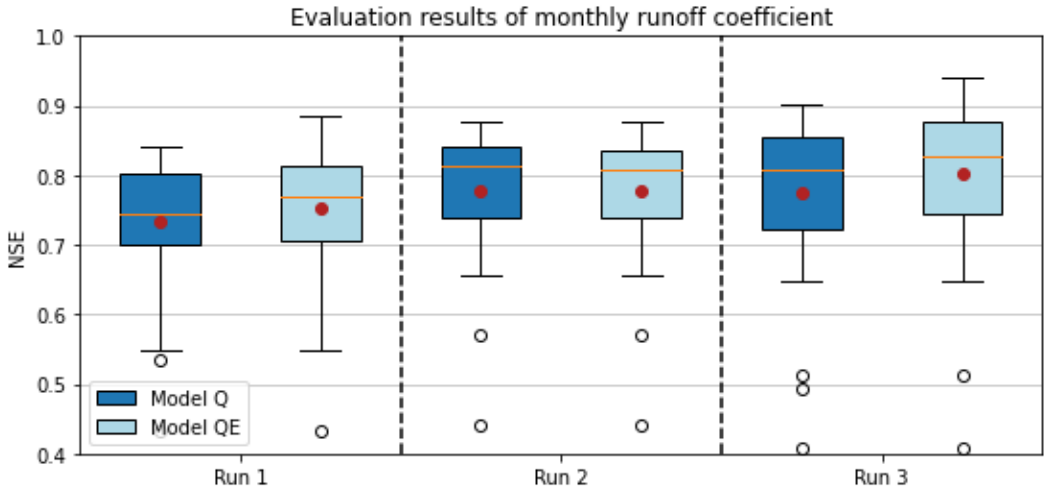


Figure 4.8: Evaluation results of the monthly runoff coefficient

In figure 4.8, the runoff coefficient results of both models for each run are plotted. In contrast to the streamflow results, there is a visible difference between Model Q and Model QE for runs 1 and 3, and between the evaluation runs, there seems to be an increase in NSE at each run. The average scores per run per model can be seen in table 4.7.

Table 4.7: Average model performance per evaluation run for monthly runoff coefficient [NSE]

		Run 1	Run 2	Run 3
Runoff coefficient [NSE]	Model Q	0.733	0.779	0.774
	Model QE	0.753	0.777	0.801

In table 4.8, the significance levels between Model Q and Model QE are presented. And in runs 1 and 3, it can be concluded that there is a significant difference between Model Q and Model QE. When looking at the average score in table 4.7, it can be concluded that Model QE outperforms Model Q in both runs 1 and 3. In run three, there is no significant difference between Model Q and Model Qe.

In table 4.9, the significance levels between each evaluation run for both Model Q and Model QE are presented. It becomes clear from the significance levels that almost every comparison is significant except for the difference in results between evaluation runs 2 and 3 for Model Q.

Table 4.8: Significance level between Model Q and Model QE

	Significance level
Run 1	0.008
Run 2	0.762
Run 3	0.003

Table 4.9: Significance level between each of the three runs

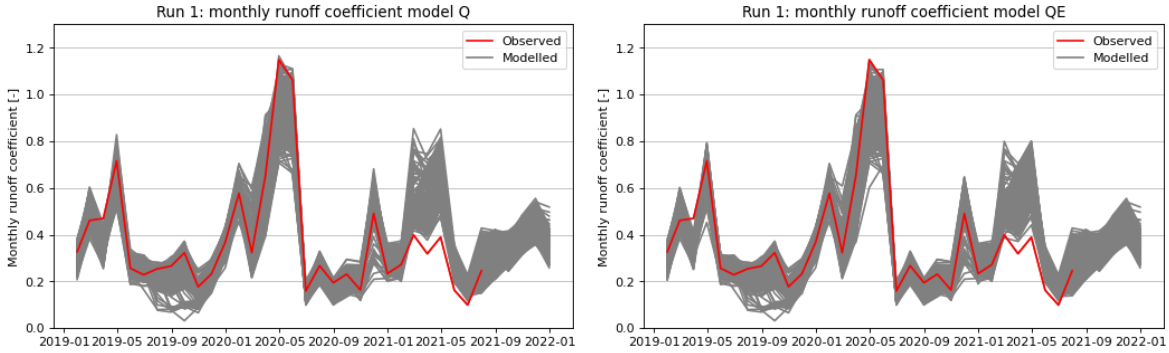
		Significance level
Run 1 vs Run 2	Model Q	<0.0001
	Model QE	0.001
Run 2 vs Run 3	Model Q	0.552
	Model QE	0.003
Run 3 vs Run 1	Model Q	<0.0001
	Model QE	<0.0001

From the results above, it becomes clear that for each run, Model QE outperforms Model Q, indicating that the model that is not only calibrated on discharge but also on evaporation is better at producing the observed monthly runoff coefficient compared to the model that is only calibrated on discharge. Furthermore, the inclusion of initial storage did improve model performance for both Model Q and Model QE. The inclusion of the swith off of the minimal discharge from February 2021 on improved the model

performance of Model Q and Model QE when compared with run 1, but only for Model Q when compared with evaluation run 2. In the following sections, the monthly runoff coefficients per model and per evaluation run are plotted against the observed runoff coefficient.

Run 1

In figure 4.9a and figure 4.9b, the runoff coefficient predictions for Model Q and Model QE are plotted, respectively. Model Q predicts higher monthly runoff coefficients than Model QE. This is, for instance, visible in 2019-05 and from 2021-03 until 2021-07. It also becomes clear that from 2021-02 onwards, the modeled interval does not follow the observed runoff coefficient as well as it did before 2021-02.

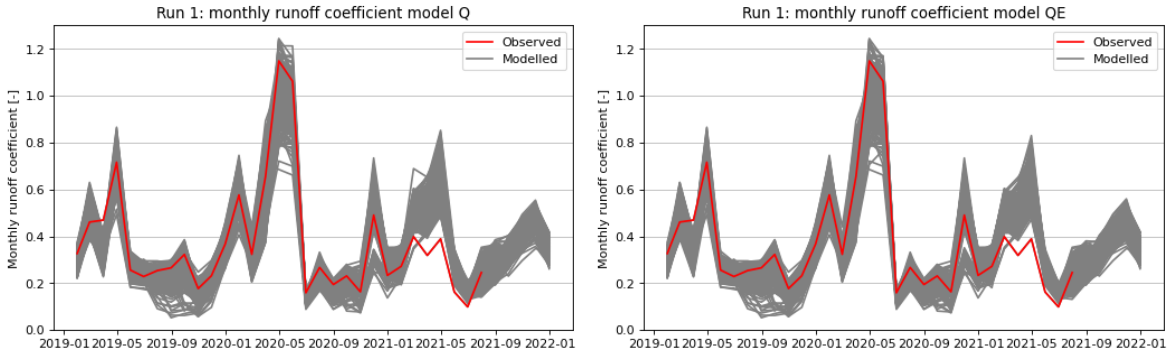


(a) Model Q **(b)** Model QE

Figure 4.9: Run 1

Run 2

In figure 4.11a and figure 4.11b, the runoff coefficient predictions for Model Q and Model QE are plotted, respectively. With the addition of initial values, the runoff coefficient prediction is larger. This is most visible in the peak during 2020–05, where now the interval wraps around the observed values, whereas in run 1, almost the entire interval predicted lower runoff coefficients than the observations.

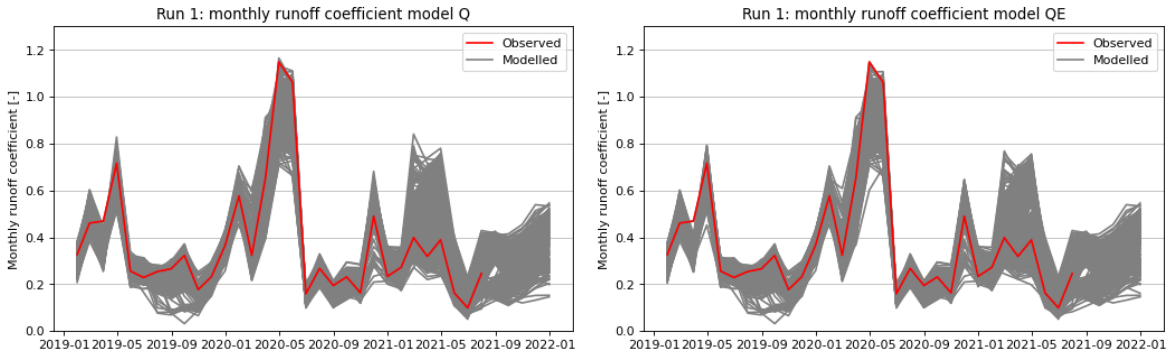


(a) Model Q **(b)** Model QE

Figure 4.10: Run 2

Run 3

In figure 4.11a and figure 4.11b, the runoff coefficient predictions for Model Q and Model QE are plotted, respectively. The influence of switching off Q_min is very visible, from 2021–02, both models are able to make predictions that more closely represent the observed runoff coefficients„ and the observed runoff coefficient is now part of the modeled interval for both models. Model Q remains higher in runoff coefficients than Model QE, as is still visible in sections from 2021–02 onwards as well.



(a) Model Q

(b) Model QE

Figure 4.11: Run 3

4.3. Model performance of best evaluation run

In the previous section, it became clear that there was no significant difference between any of the runs or between Model Q and Model QE for the streamflow performance indicator De_Q. For the monthly runoff coefficient, there was a significant difference between runs and between the models, with evaluation run 3 outperforming the other two. Therefore, evaluation run 3 is selected to evaluate the model's performance on different metrics in this section. To gain more insight into the model's performance. During these evaluations, it became clear that the switch off of the minimum flow was not always beneficial; therefore, evaluation run 1 is also used to depict the difference.

4.3.1. Flow duration curve

In figure 4.12, the clearest deviation between the modelled and the observed FDC is between the exceedance probabilities of 0.8 and 1.0. For some of the runs from Model Q and Model QE, the storages have probably been empty for some time, which resulted in flows lower than the minimum flow. However, the average FDC of both Model Q and Model QE follows the observed FDC very well, with NSE log scores of 0.908 and 0.909, respectively, as depicted in table 4.10. In the same table, it is stated that the significance level is 0.953, which indicates that there is no significant difference in performance between the two models.

Table 4.10: NSE log score per model

	NSE log
Model Q	0.908
Model QE	0.909
Significance level	0.953

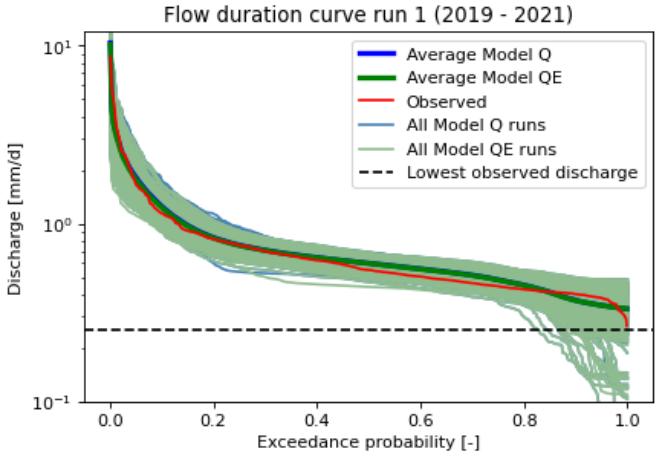


Figure 4.12: Flow duration curves from evaluation run 1

In figure 4.13, the FDC of evaluation run 3 is presented. The effect of switching off the minimal discharge is clearly visible in the section with an exceedance probability between 0.8 and 1.0. The switch-off gave the opportunity for Model Q and Model QE to predict lower discharges. In table 4.11, the average NSE logs of both models are presented, and both Model Q and Model QE score lower than in run 1, with a score of 0.757 for Model Q and a score of 0.740 for Model QE. Once again, the difference between the two models is not significant, with a significance level of 0.524.

Table 4.11: NSE log score per model

	NSE log
Model Q	0.757
Model QE	0.740
Significance level	0.524

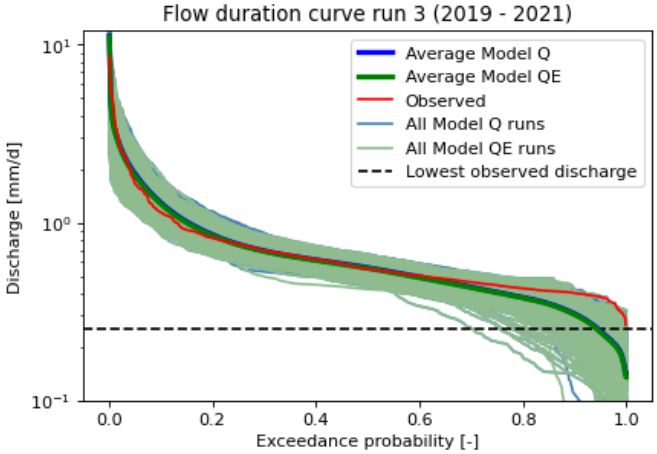


Figure 4.13: Flow duration curves from evaluation run 3

4.3.2. Monthly average evaporation and discharge

In figure 4.14, the peak discharges in February and July 2021 are clearly visible, the model also has a larger spread in predictions during peak discharges than during normal flow conditions. The individual lines of the two models are hard to distinguish from each other, but from the first month until the 8th month, the average of Model Q lays slightly above Model QE, and the average of Model QE lays slightly above Model Q from the 8th month onward. In table 4.12, the NSE log scores are presented, with Model QE having a slightly better score of 0.698 compared to 0.693 for Model Q. However, the difference between the two models is not significant, with a significance level of 0.7319.

Table 4.12: NSE log score per model

	NSE log
Model Q	0.693
Model QE	0.698
Significance level	0.7319

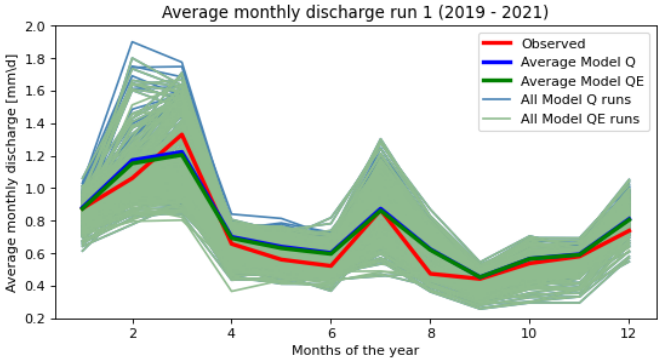


Figure 4.14: Monthly average discharge run 1

In figure 4.13, the average monthly discharge is presented from evaluation run 3, which shows lower monthly discharges in the first two months compared to evaluation run 1, as expected. This can clearly be seen when comparing the location of the average Model Q and Model QE monthly discharge in the first two months; the lines lay closer to the observed values and continued on to be closer towards the observed value in months 4 until 6, even dipping under the observed line between months 9 and 11. This closer position towards the observed line results in slightly higher NSE log scores for both Model Q and Model QE, than in evaluation run 1, with scores of 0.775 and 0.786, respectively. As can be seen in table 4.13, the difference between the two models is not significant, with a significance level of 0.238.

Table 4.13: NSE log score per model

	NSE log
Model Q	0.775
Model QE	0.786
Significance level	0.238

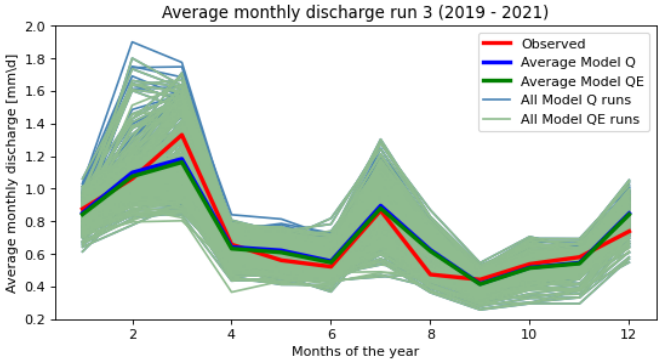


Figure 4.15: Monthly average discharge run 3

In figure 4.16, a clear evaporation pattern can be seen throughout the year, with a peak in the summer months. The figure is for both evaluation runs 1 and 3, as the switchoff of the minimum discharge is later in the model than all evaporation fluxes and thus has no influence on the evaporation amounts. Both Model Q and Model QE follow the seasonal pattern very good with NSE log values of 0.884 and 0.891, respectively. However, it is clear that during the first three months of the year and during the last two months, the models are less capable of producing evaporation rates close to the observed evaporation. As is presented in table 4.14, the models do differ significantly, with a significance level of 0.006. Meaning that Model QE significantly outperforms Model Q, in predicting monthly evaporation compared to the observed evaporation.

Table 4.14: NSE log score per model

	NSE log
Model Q	0.884
Model QE	0.891
Significance level	0.006

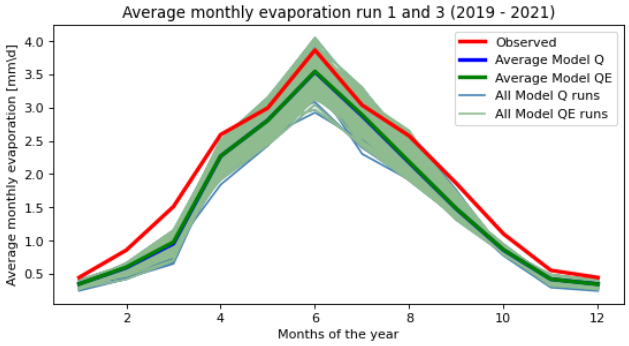


Figure 4.16: Monthly average evaporation run 1 and 3

4.3.3. Cumulative evaporation and discharge

In figure 4.17 the cumulative discharge from run 1 is plotted for both Model Q and Model QE. The averages of both models follow the observed line well in general, with NSE log scores of 0.793 and 0.759, respectively. As can be seen in table 4.15, the difference between the two models is not significant, with a significance level of 0.592. A deviation from the observed line is visible from 2021–02 onward, where both models overestimate the cumulative discharge.

Table 4.15: NSE log score per model

	NSE log
Model Q	0.793
Model QE	0.759
Significance level	0.592

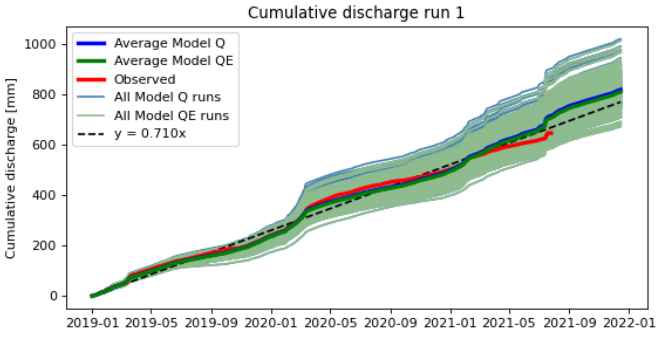


Figure 4.17: Cumulative discharge run 1

In figure 4.18, the effect of switching off the minimum discharge in 2021-02 is visible, the averages of both models follow the observed line more closely. This also comes forward in the NSE log scores in table 4.16, where Model Q has now increased from 0.793 to 0.823 and Model QE has increased from 0.759 to 0.787. The difference between the two models in evaluation run 3 is not significant, with a significance level of 0.533.

Table 4.16: NSE log score per model

	NSE log
Model Q	0.823
Model QE	0.787
Significance level	0.533

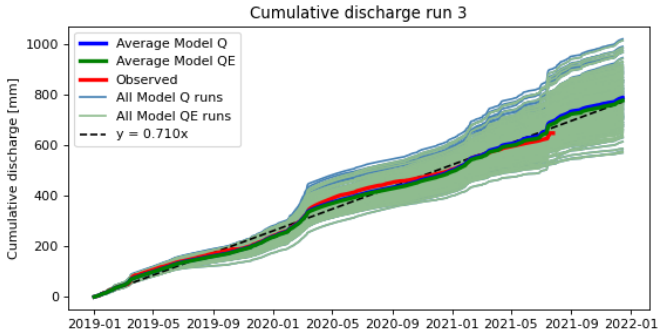


Figure 4.18: Cumulative discharge run 3

In figure 4.19 the cumulative evaporation for Model Q and Model QE is plotted. Both models predict cumulative evaporation amounts lower than the observed values for the entire timeserie, indicating that there is structurally not enough water evaporating in the catchment. Both models are able to represent the observed values good with NSE log scores of 0.760 for Model Q and 0.774 for Model QE. With the limited spread within each model, the difference between the two models is significant, as can be seen in table 4.3.3, with a significance score of 0.021. This means that Model QE can significantly better predict the cumulative evaporation compared to the observed evaporation than Model Q.

Table 4.17: NSE log score per model

	NSE log
Model Q	0.760
Model QE	0.774
Significance level	0.021

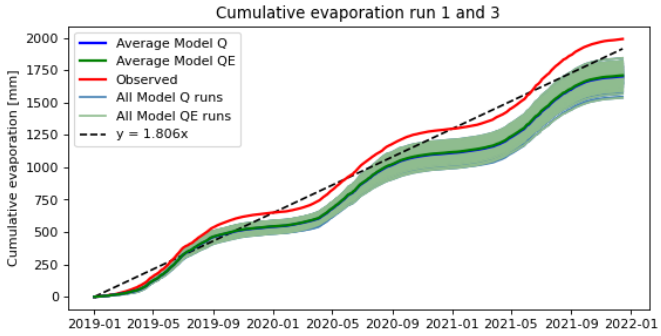


Figure 4.19: Cumulative evaporation run 1 and 3

4.3.4. Minimum discharge switch off

The inclusion of the switch-off in February 2021, gave varying results. This switch-off increased performance for the runoff coefficient, average monthly discharge and cumulative discharge, but a decrease in performance of the FDC, when compared to evaluation run 1. Indicating that there is a difference in the streamflow pattern from February 2021 onward, however a complete switch-off of the minimum discharge process might have been to aggressive, and as the model was able to predict very low discharge, decreased its performance on FDC.

5

Discussion

The decisions and assumptions made throughout this research have had an impact on the study's results. Unknown values or processes were thoroughly investigated wherever possible, however, this was not always the case. The sections below describe these assumptions and decisions, first pertaining to the data collection and quality control in section 5.1, followed by the model choices and assumption in section 5.2 and those of the calibration process in section 5.3. These are followed by a discussion of the evaluation process in section 5.4 and the results in section 5.5.

5.1. Data

Thiessen polygons were used to interpolate the precipitation data. Thiessen polygons are formed by drawing perpendicular bisectors on the lines joining two neighbouring measurement stations, and values are assigned by closest neighbour. For two measurement stations in Belgium, the perpendicular bisectors drawn on the line between themselves and the nearest measurement station within the catchment (Gemmenich) overlapped with the catchment border, implying that within the catchment, all area would be assigned the value of Gemmenich, therefore these measurement stations were not added. However, using a different interpolation method like inverse distance, these measurement stations would have had an effect on the area within the catchment, as inverse distance interpolates between the values of the measurement stations. In this research, with a lumped model, the impact of this change might be limited, but with a distributed model, this change in interpolation technique might improve model performance.

Priestly Taylor potential evaporation has been used within this study, but as shown in figure 2.5, Makkink potential evaporation predicted substantially lower potential evaporation rates. Makkink potential evaporation has not been utilised because there are no actual evaporation datasets that use Makkink. However, because Makkink projected much lower evaporation rates, the evaporation in the model would be lower as well, increasing storage and discharge. This could have had an effect on the model's predictive power, but it is unknown to what extent.

Within the Budyko framework, groundwater fluxes are not taken into account. Measuring errors aren't the only reason a catchment can exceed its water and energy limits. The inflow or outflow of groundwater beneath catchment borders can also cause a catchment to plot beyond the limits. When a catchment plots outside the framework, it is obvious that something is going on, but plotting within the framework can also be a false positive error. Meaning that the groundwater may already play a role, but because the catchment plots within the margins, the groundwater flux can be wrongfully neglected. Especially in a limestone karst region like the south of Limburg, transboundary groundwater fluxes may play a role.

The elevation data used is from EUDEM (2016). This is a widely used elevation data set, however, it is worth the mention that this data set does not differentiate between a tree, a building, and the ground. Meaning that sometime the elevation levels are higher in an urban area, or a forested area. As the EUDEM spatial resolution is averaged out over a 25 by 25 m square, these objects could be averaged out, but specifically with the landscape classification with is dependent on HAND and Slope with are both derived from elevation data, this could have had an effect.

5.2. Model

The hydrological model uses a daily time step, which has multiple implication for the model itself and its output. To begin with, a daily time step yields fewer exact forecasts, and if modelled discharge could be on an hourly time step, there would be more data points to compare the modelled output to the observed output. The same is true for modelled evaporation data, except that evaporation has a daily rhythm in and of itself, which is currently not captured in the model. The increase of temporal resolution to an hourly timescale would improve the calibration process and thereby possibly improve the prediction, but it would mostly make the model more useful. If a hydrological model is used to predict high stream flow values, a daily time step is simply too rough. A daily time step would however also increase the computational time of the model, which would mean that to be able to do the same amount of runs, an increase in computer power would be needed. This means that the predictive power might decrease if an hourly time step is used, but the same computing power is available.

The model is represented by one cell, which means that all input values are assumed to be homogeneous for the entire catchment. The data is available at a higher resolution, from which it is also clear that the forcing data has a heterogeneous pattern over the catchment, with consistently higher precipitation and evaporation rates in the South East part of the catchment. But, as with increasing temporal resolution, it would lengthen the run time.

A snow process has not been added to the hydrological model of the Geul, even though snow does occur in the catchment, with on average of 18 snow days in the year between 1991 and 2020, (Huirne, 2020). The addition of a snow process to the hydrological model could improve model predictions, however, this also adds a lot of parameters to the model, making the model more difficult to calibrate. However, the addition of a snow process could have increased model performance.

The landscape classification is in essence fully made from elevation data, as both the slope and HAND are derived from elevation data. Because the elevation is used to derive the HAND, the location of the river is not determined by actual location in the catchment but determined at the location where the elevation is the lowest locally. There is not a large difference between these two locations in the Belgium part of the catchment, as the Geul is more defined by elevation levels in that region. However, in the lower and flatter parts of the catchment, the river meanders throughout the plane and the location of the river differs from that of the location of the river in the HAND map, up to a hundred of meters.

The Geul's land use does not correspond to the defined landscapes. As explained in Gharari et al. (2014): "Hill slopes are generally characterised by forest, while plateaus and valley bottoms are mostly used as cropland and pastures, respectively.". This distinction is not entirely correct in the Geul, as some plateaus are completely covered by forest and some hill slopes are being used for agriculture. A land use map was utilised in the WFlow model, as well as a map of the three landscapes, see appendix F. This was not possible in the FLEX HBV model, since the model could only differentiate between landscape or land use, not both at the same time.

The first choice of model for this research was the WFLOW Flextopo model from Deltares. This model is a distributed model, with an hourly time step. As explained in appendix F. The model's run time was too long to be able to have a meaningful calibration, meaning that within the duration of this thesis, not enough calibration runs would have been run. Even when downscaling the spatial and temporal resolution to a lumped model on a daily time step, the run time was still too long. Therefore, the choice had been made to switch to the HBV FLEX model, which is a lumped model on a daily time step, and this improved the run time significantly.

If the distributed hydrological model would have been used, the heterogeneity in both precipitation and evaporation throughout the catchment would have been able to be captured better in the model. Not only would the input precipitation data more accurately depict the precipitation patterns in the catchment, but the more interesting aspect would be the possibility to calibrate each grid cell of the model with corresponding actual evaporation data. This distributed calibration could have a positive impact on the effect that calibration on evaporation data has on model performance (Dembélé et al., 2020). The hourly time step of the WFlow model would still have to be downscaled to a daily time step, as the evaporation data is currently still only available on a daily time step.

5.3. Calibration

The root zone storage parameter interval is determined by calculating the yearly water deficit with observed catchment data. This water deficit should be calculated by using transpiration instead of actual evaporation since transpiration is the amount of water that evaporates from the unsaturated zone. However, because no transpiration data was available, actual evaporation has been used. This means that there is a slight overestimation of evaporation, as interception evaporation is now also taken into account. therefore, a slight overestimation of the water deficit. However, the amount is only very limited, as interception evaporation can be no more than a couple of millimetres per day because interception storage is very limited on trees and plants. Furthermore, the assumption is made that there is no interaction between the groundwater and the unsaturated zone. This probably in reality is not the case, with a net flux from the unsaturated zone towards the groundwater storage. As a larger area of the catchment has deep percolation processes than capillary rise. The influence of this assumption would be that the water deficit is larger in reality, with water leaving the unsaturated zone and entering the groundwater storage.

The slow recession constant has been determined by way of an multiple recession curve (MRC). In this curve, all recession curves that extend over a longer period of time are plotted on top of each other. However, during the creation of this plot, it became clear that during larger precipitation events, the discharge data was often missing. The recession curve in the MRC is mostly the curves from medium precipitation events. However, this will have a limited effect on the actual recession coefficient as there were still multiple curves to use, and the slow recession coefficient only looks at the lowest part of the curves.

Monte Carlo random sampling has been used to create parameter sets. This sampling methodology takes, as the name indicated, random samples. This means that some parameter value spaces can be selected more than others and some less, leading to clustering. Therefore, a large number of runs need to be used to calibrate the model on, to make sure that the parameter space is sufficiently sampled. However, with computing limitations, not all parameter sets can be run. This can have an effect on the parameter sets selection since there can always be better parameter combination sets, which have not been selected. Therefore, Monte Carlo random sampling should not be used to compute a single best run, as the chance that the perfect combination of parameter values is in the parameter sets that are used is very limited. However, using the parameter sets in a GLUE methodology where multiple parameter sets are used, decreases this risk. Nonetheless, there is still an opportunity that once a way larger amount of parameter sets is selected, the outcomes will differ significantly from the results found in this study.

5.4. Evaluation

In this study, the NSE has been used as the objective function. An NSE value of 0 indicates that the aggregated errors of the modelled hydrograph are equal to the aggregated errors of the average observed discharge with respect to the observed discharge. This means that it is more difficult to get a high NSE value if the average observed discharge is very close to the actual value for long periods of time. This is the case in the Geul catchment, which has limited peak flows and a consistent base flow throughout the year, as the flow duration curve in figure 4.12 also indicates. Therefore, the NSE values of this study might be underestimating the predictive value of the model. In contrast, the average evaporation is a much worse predictor with the evaporation following a sin-like trend throughout the year, making it more easily to acquire high NSE and NSE log scores.

The option of evaluating evaporation on monthly anomalies instead of NSE or NSE log time series has been explored, however, this also comes with its challenges. The usable evaporation data only started in 2014 meaning that the monthly average evaporation amount would be calculated over the same period as the calibration and evaluation period and therefore this average would be very close to the observed value. Now causing extremely low NSE and NSE log values. The goal would still be to find an objective function that would put calibration on evaporation and calibration on discharge in the same range. However, for further research this methodology would probably be preferable over NSE and NSE log, to take the seasonality out of the evaporation data, especially if longer time series become available.

As discussed in section 3.4.4, GLUE is not an undisputed methodology. This is mostly because of the selection of a threshold value. This value is defined by the model maker and is therefore subjective to his or her opinion of what a behaviour model is. The interval that is created with the selected behaviour model should therefore be carefully interpreted. This means that this interval will change if a different threshold value is selected, and therefore the evaluation results of this study could differ if a different number of behaviour models were selected.

5.5. Results

The dotty plots of the parameters are not very well-defined, indicating that a quite some parameters could be any of the value in its own interval and still get a high objective score. This indicated that maybe the parameter intervals of some parameters were too narrow, to begin with, or that parameter has a limited effect on model performance. The latter is probably the reason the specific parameters of the wetland and hillslopes are less defined than the parameter of the plateaus, as there is more plateau area in the catchment, meaning that parameters of the plateau landscape have a larger impact on the outcome of the model.

When comparing the streamflow results from Model Q and Model QE, there is no significant difference in performance between the two models in any of the evaluation runs. Even though it was the hypothesis of this thesis that additional calibration on evaporation would increase streamflow predictive power, this is not always the case, as is shown in the literature. Jiang et al. (2020) researched 28 natural river basins and concluded that models calibrated on both discharge and evaporation produce better or similar streamflow simulations in 29% of the basins compared to models only calibrated on discharge. Dembélé et al. (2020) researched 12 different remote sensing evaporation datasets with four distinct multivariate calibration strategies and concluded that evaporation datasets have a good potential for improving model calibration, but this is dependent on the calibration strategy, with distributed calibration strategies outperforming catchment average strategies. As the model used in this study is lumped, this could have limited the potential of the evaporation data calibration.

The monthly runoff coefficient score did show a significant difference in the performance of Model Q and Model QE for evaluation runs 1 and 3. This difference with the streamflow results can be attributed to the more direct link between the runoff coefficient and evaporation, as precipitation is divided into streamflow and evaporation, and these two factors therefore directly influence the runoff coefficient (Li, Niu, He, & Wang, 2022).

The evaluation of the flow duration curve, monthly average streamflow and evaporation, and cumulative streamflow and evaporation only show a significant difference in model performance by Model Q and Model QE for monthly average evaporation and cumulative evaporation. This result is what was to be expected as almost all of the studies from Jiang et al. (2020) concluded that, compared to the Qobs calibration scheme, the discharge and evaporation calibration was able to improve the evaporation simulation while maintaining a suitable model streamflow performance.

6

Conclusion

The aim of this study was to develop a model that could accurately replicate the Geul catchment's response to precipitation, with subsequent calibration on discharge (Model Q) and on both discharge and evaporation (Model QE). The models were evaluated based on their ability to represent observed streamflow, monthly runoff coefficient, flow duration curve (FDC), average monthly streamflow and evaporation and cumulative streamflow and evaporation, with the goal of identifying if there was an improvement in model performance in Model QE compared to Model Q.

6.1. Does evaporation data make a difference?

Through three evaluation runs, it became clear that both Model Q and Model QE were able to predict streamflow satisfactorily, with De_Q scores between 0.632 and 0.649 across all runs. There was no significant difference between the evaluation runs, indicating that the addition of initial storage or the switch off of the minimum discharge from February 2021 onward, did not significantly change the models' ability to predict streamflow. There is no significant difference between the performance of Model Q and Model QE in any of the evaluation runs.

Further analysis showed that both Model Q and Model QE scored good on the monthly runoff coefficient, achieving NSE scores between 0.733 and 0.801. In this case, the difference between the evaluation runs was significant with the highest scores in evaluation run 3, when there were no initial storages and a minimum discharge switch-off was implemented in February 2021. Both in evaluation run 1 and 3 there was a significant difference between Model Q and Model QE, with Model QE outperforming Model Q.

With evaluation run 3 giving the best results for the monthly runoff coefficient and indifference between evaluation runs for the streamflow predictions, this evaluation run is used to compare the performance of Model Q and Model QE on different hydrological signatures. When comparing the FDC, monthly average evaporation and discharge, and cumulative evaporation and discharge, the difference between the performance of Model Q and Model QE was only significant for the monthly average evaporation and cumulative evaporation. With the different performance of the two models in some areas but equal in others, it is now possible to answer the main research question of this thesis:

Is the predictive power of a discharge calibrated hydrological model of the Geul catchment in the Netherlands, greater than the predictive power of an evaporation-calibrated model, in addition to discharge?

Model QE outperformed Model Q on monthly runoff coefficient, monthly average evaporation and cumulative evaporation, increasing the NSE score of Model Q of the monthly runoff coefficient from 0.774 to 0.801, the NSE log score of the monthly average evaporation from 0.884 to 0.891 and the cumulative evaporation from 0.760 to 0.774. However, Model QE did not outperform Model Q on streamflow. Meaning that the strict answer to the research question is no. However, Model QE outperformed Model Q on some signatures, without scoring significantly worse than Model Q on any other, including streamflow.

This indicates that calibrating on both discharge and evaporation instead of only calibrating on discharge increases the model's capability to divide the incoming precipitation into discharge and evaporation and better predict monthly and cumulative evaporation.

6.2. Implications of this study

The study suggests that incorporating evaporation as a calibration variable can enhance hydrological models' performance, however not on streamflow predictions in this case. Even though no significant difference in streamflow predictions is observed between Model Q and Model QE, the latter performs better on evaporation signatures. These findings underscore the significance of using additional calibration variables to limit the degrees of freedom a hydrological model has.

Even though the model's streamflow predictions do not improve with additional calibration on evaporation data, it should not be concluded that calibration on evaporation data is not beneficial to overall model performance. Models that predict evaporation amounts that more closely represent observed evaporation data can be assumed to have internal processes that more closely represent the hydrological processes in a catchment.

A number of studies have been performed with calibration on evaporation data (Jiang et al., 2020), and within these studies, most hydrological models are distributed and in larger catchments. This is no coincidence, with a gridded model, the full potential of gridded evaporation data can be used. Where discharge measurements automatically integrate over the entire upstream area, evaporation data can be used on a local scale. Therefore, especially with the 100m by 100m evaporation datasets available in the Netherlands, the impact of the use of calibration on evaporation data should be researched further with preferably a distributed model.

6.3. Recommendation for future research

The addition of more spatial resolution in the hydrological model would give the opportunity to differentiate between precipitation and evaporation values throughout the catchment, and as indicated in section 2.1.1, there is a substantial difference in precipitation and evaporation within the catchment. With a distributed grid of the model, it would also be possible to have distributed output. This means that the evaporation data would not have to be averaged out over the catchment but each grid cell could be calibrated against its own evaporation data, which could improve the calibration of the model.

Next to a higher spatial resolution of the hydrological model would also the increase of temporal resolution is a great addition. The catchment is relatively small meaning that hydrological processes happen more quickly. With the daily time steps some processes like Horton runoff can be missed in the current model. Furthermore, an hourly time step would also give the opportunity to calibrate streamflow to hourly observed discharges instead of daily which could improve the overall model performance and the calibration precision. Calibration of evaporation data on an hourly timescale is not possible, as the data is for now only available on a daily timescale.

Furthermore, if more calibration runs could be run, more parameter combinations can be tested which would possibly improve the calibration results. Moreover, it may give the opportunity to try out even broader parameter ranges. It could namely be possible that the model could not move freely enough in its parameter space to improve its predictive powers.

Bibliography

Allen, R., Pereira, L., Raes, D., & Smith, M. (1998). Crop evapotranspiration. guidelines for computing crop water requirements. *FAO Irrigation and Drainage Paper (FAO)*, *56*. Retrieved from https://agris.fao.org/agris-search/search.do?recordID=XF1999085851 doi: 10.3/JQUERY-UI.JS

Baldassarre, G. D., & Montanari, A. (2009). Hydrology and earth system sciences uncertainty in river discharge observations: a quantitative analysis. *Hydrol. Earth Syst. Sci*, *13*, 913-921. Retrieved from www.hydrol-earth-syst-sci.net/13/913/2009/

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., … Richardson, F. (2001, 11). Fluxnet: A new tool to study the temporal and spatial variability of ecosystem–scale carbon dioxide, water vapor, and energy flux densities. *©2001 American Meteorological Society*, *82*. doi: 10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2

Beck, H. E., de Jeu, R. A., Bruijnzeel, L. A., Schellekens, J., & van Dijk, A. I. (2009). Improving curve number based storm runoff estimates using soil moisture proxies. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *2*, 250-259. doi: 10.1109/JSTARS.2009.2031227

Beck, H. E., van Dijk, A. I., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., & Bruijnzeel, L. A. (2016, 5). Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, *52*, 3599-3622. doi: 10.1002/2015WR018247

Becker, R., Koppa, A., Schulz, S., Usman, M., aus der Beek, T., & Schüth, C. (2019, 10). Spatially distributed model calibration of a highly managed hydrological system using remote sensing-derived et data. *Journal of Hydrology*, *577*, 123944. doi: 10.1016/J.JHYDROL.2019.123944

Bergström, S., & Forsman, A. (1973). Development of a conceptual deterministic rainfall-runoff model. *NORDIC HYDROL.*, *4*, 147-170. doi: 10.2166/NH.1973.0012

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

Beven, K., & Binley, A. (1992, 7). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*, 279-298. Retrieved from https://onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/full/10.1002/hyp.3360060305 doi: 10.1002/HYP.3360060305

Bouaziz, L. (2021). Internal processes in hydrological models a glance at the meuse basin from space. Retrieved from https://doi.org/10.4233/uuid:09d84cc1-27e2-4327-a8c7-207a75952061 doi: 10.4233/uuid:09d84cc1-27e2-4327-a8c7-207a75952061

Bouaziz, L. (2022). *Flextopo · wflow.jl.* Retrieved 2022-19-09, from https://deltares.github.io/Wflow.jl/stable/model_docs/vertical/flextopo/#vert_flextopo

Budyko, M. I. (1948). *Evaporation under natural conditions*. Gidrometeorizdat, Leningrad. (English translation by IPST, Jerusalem)

Copernicus. (2018). *Clc 2018 — copernicus land monitoring service.* Retrieved from https://land.copernicus.eu/pan-european/corine-land-cover/clc2018

Corbari, C., Ravazzani, G., Galvagno, M., Cremonese, E., & Mancini, M. (2017, 11). Assessing crop coefficients for natural vegetated areas using satellite data and eddy covariance stations. *Sensors (Basel, Switzerland)*, *17*. Retrieved from /pmc/articles/PMC5713072//pmc/articles/PMC5713072/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5713072/ doi: 10.3390/S17112664

de Boer-Euser, T. (2017). Added value of distribution in rainfall-runoff models for the meuse basin. Retrieved from https://repository.tudelft.nl/islandora/object/uuid%3A89a78ae9-7ffb-4260-b25d-698854210fa8 doi: 10.4233/UUID:89A78AE9-7FFB-4260-B25D-698854210FA8

Deltares. (2022). *About wflow · wflow.jl.* Retrieved 2022-07-24, from https://deltares.github.io/Wflow.jl/stable/

Dembélé, M., Ceperley, N., Zwart, S. J., Salvadore, E., Mariethoz, G., & Schaefli, B. (2020, 9). Potential of satellite and reanalysis evaporation datasets for hydrological modelling under various model calibration strategies. *Advances in Water Resources*, *143*. doi: 10.1016/j.advwatres.2020.103667

de Moor, J. J., Kasse, C., van Balen, R., Vandenberghe, J., & Wallinga, J. (2008, 6). Human and climate impact on catchment development during the holocene — geul river, the netherlands. *Geomorphology*, *98*, 316-339. doi: 10.1016/J.GEOMORPH.2006.12.033

de Moor, J. J., & Verstraeten, G. (2008, 3). Alluvial and colluvial sediment storage in the geul river catchment (the netherlands) — combining field and modelling data to construct a late holocene sediment budget. *Geomorphology*, *95*, 487-503. doi: 10.1016/J.GEOMORPH.2007.07.012

Direction généraleopérationnelle de la Mobilité et des Voies hydrauliques. (2022). *Directoraat-generaal mobiliteit en waterwegen - directory's en statistieken.* Retrieved 2022-11-06, from http://voies-hydrauliques.wallonie.be/opencms/opencms/fr/hydro/Archive/annuaires/index.html

Duethmann, D., Bloschl, G., & Parajka, J. (2020, 7). Why does a conceptual hydrological model fail to correctly predict discharge changes in response to climate change? *Hydrology and Earth System Sciences*, *24*, 3493-3511. doi: 10.5194/HESS-24-3493-2020

European Environment Agency. (2016). *Eu-dem v1.0 — copernicus land monitoring service.* Retrieved 2022-07-09, from https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1-0-and-derived-products/eu-dem-v1.0?tab=download

Expertise Netwerk Waterveiligheid. (2021). *Hoogwater 2021 feiten en duiding.* Retrieved from https://klimaatadaptatienederland.nl/publish/pages/192998/hoogwater-2021-feiten-en-duiding.pdf

Fenicia, F., Savenije, H. H. G., Matgen, P., & Pfister, L. (2006). Hydrology and earth system sciences is the groundwater reservoir linear? learning from data in hydrological modelling. *Hydrol. Earth Syst. Sci*, *10*, 139-150. Retrieved from www.hydrol-earth-syst-sci.net/10/139/2006/

Forzieri, G., Alkama, R., Miralles, D. G., & Cescatti, A. (2017, 6). Satellites reveal contrasting responses of regional climate to the widespread greening of earth. *Science*, *356*, 1180-1184. Retrieved from https://www.science.org/doi/10.1126/science.aal1727 doi: 10.1126/SCIENCE.AAL1727/SUPPL_FILE/AAL1727-FORZIERI-SM.PDF

Frenken, H. (2021). *Waterschap: Valkenburg beschermen tegen wateroverlast is onmogelijk.* Retrieved from https://nos.nl/artikel/2391449-waterschap-valkenburg-beschermen-tegen-wateroverlast-is-onmogelijk

GADM. (2022). *Download gadm data.* Retrieved 2022-07-09, from https://gadm.org/download_country.html

Gao, H., Hrachowitz, M., Fenicia, F., Gharari, S., & Savenije, H. H. (2014, 5). Testing the realism of a topography-driven model (flex-topo) in the nested catchments of the upper heihe, china. *Hydrology and Earth System Sciences*, *18*, 1895-1915. doi: 10.5194/HESS-18-1895-2014

Gash, J. H. (1979). An analytical model of rainfall interception by forests. *Quarterly Journal of the Royal Meteorological Society*, *105*, 43-55. doi: 10.1002/QJ.49710544304

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., & Savenije, H. H. (2014, 12). Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration. *Hydrology and Earth System Sciences*, *18*, 4839-4859. doi: 10.5194/hess-18-4839-2014

Gharari, S., Hrachowitz, M., Fenicia, F., & Savenije, H. H. (2011). Hydrological landscape classification: Investigating the performance of hand based landscape classifications in a central european meso-scale catchment. *Hydrology and Earth System Sciences*, *15*, 3275-3291. doi: 10.5194/HESS-15-3275-2011

Gyasi-Agyei, Y. (2020, 7). Identification of the optimum rain gauge network density for hydrological modelling based on radar rainfall analysis. *Water 2020, Vol. 12, Page 1906*, *12*, 1906. Retrieved from https://www.mdpi.com/2073-4441/12/7/1906/htmhttps://www.mdpi.com/2073-4441/12/7/1906 doi: 10.3390/W12071906

Haggard, B., Moore, P., & Brye, K. (2005, 01). Effect of slope on runoff from a small variable slope box-plot. *Journal of Environmental Hydrology*, *13*.

Haqiqi, I., Grogan, D. S., Hertel, T. W., & Schlenker, W. (2021, 2). Quantifying the impacts of compound extremes on agriculture. *Hydrology and Earth System Sciences*, *25*, 551-564. doi: 10.5194/HESS-25-551-2021

Hersbach, H., Bell, B., Biavati, G., Berrisford, P., Horányi, A., Sabater, J. M., … Thépaut, J.-

N. (2018). *Era5 hourly data on single levels from 1959 to present.* Retrieved 2022-07-09, from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview

Hoitink, A. J., Buschman, F. A., & Vermeulen, B. (2009, 11). Continuous measurements of discharge from a horizontal acoustic doppler current profiler in a tidal river. *Water Resources Research*, *45*. doi: 10.1029/2009WR007791

Horton, P., Schaefli, B., & Kauzlaric, M. (2022, 1). Why do we have so many different hydrological models? a review based on the case of switzerland. *Wiley Interdisciplinary Reviews: Water*, *9*, e1574. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/wat2.1574 doi: 10.1002/WAT2.1574

Hrachowitz, M., Stockinger, M., Coenders-Gerrits, M., Ent, R. V. D., Bogena, H., Lücke, A., & Stumpp, C. (2021). Reduction of vegetation-accessible water storage capacity after deforestation affects catchment travel time distributions and increases young water fractions in a headwater catchment. *Hydrol. Earth Syst. Sci*, *25*, 4887-4915. Retrieved from https://doi.org/10.5194/hess-25-4887-2021 doi: 10.5194/hess-25-4887-2021

Huang, S., Shah, H., Naz, B. S., Shrestha, N., Mishra, V., Daggupati, P., … Vetter, T. (2020, 12). Impacts of hydrological model calibration on projected hydrological changes under climate change—a multi-model assessment in three large river basins. *Climatic Change*, *163*, 1143-1164. Retrieved from https://link.springer.com/article/10.1007/s10584-020-02872-6 doi: 10.1007/S10584-020-02872-6/FIGURES/3

Huirne, J. (2020). *Aantal witte dagen gehalveerd - weer.nl.* Retrieved from https://www.weer.nl/nieuws/2020/aantal-witte-dagen-gehalveerd

Humanitarian OpenStreetMap Team. (2021). *Hotosm belgium waterways (openstreetmap export) - humanitarian data exchange.* Retrieved 2022-19-10, from https://data.humdata.org/dataset/hotosm_bel_waterways

Hümann, M., & Müller, C. (2013, 3). Improving the gis-drp approach by means of delineating runoff characteristics with new discharge relevant parameters. *ISPRS International Journal of Geo-Information*, *2*, 27-49. doi: 10.3390/IJGI2010027

Immerzeel, W. W., & Droogers, P. (2008, 2). Calibration of a distributed hydrological model based on satellite evapotranspiration. *Journal of Hydrology*, *349*, 411-424. doi: 10.1016/J.JHYDROL.2007.11.017

Jackson, A. (2014). *Discharge hydrographs.* Retrieved from https://geographyas.info/rivers/discharge-and-hydrographs/

Jain, S. K., & Singh, V. (2003, 1). Acquisition and processing of water resources data. *Developments in Water Science*, *51*, 47-121. doi: 10.1016/S0167-5648(03)80056-4

Jarosz, W., Enayet, A., Kensler, A., Kilpatrick, C., & Christensen, P. (2019). *Orthogonal array sampling for monte carlo rendering* (Vol. 38).

Jennings, K. S., Winchell, T. S., Livneh, B., & Molotch, N. P. (2018, 3). Spatial variation of the rain–snow temperature threshold across the northern hemisphere. *Nature Communications 2018 9:1*, *9*, 1-9. Retrieved from https://www.nature.com/articles/s41467-018-03629-7 doi: 10.1038/s41467-018-03629-7

Jiang, & Wang. (2019, 8). The role of satellite-based remote sensing in improving simulated streamflow: A review. *Water 2019, Vol. 11, Page 1615*, *11*, 1615. Retrieved from https://www.mdpi.com/2073-4441/11/8/1615/htmhttps://www.mdpi.com/2073-4441/11/8/1615 doi: 10.3390/W11081615

Jiang, Wu, H., Tao, J., Kimball, J. S., Alfieri, L., & Chen, X. (2020, 1). Satellite-based evapotranspiration in hydrological model calibration. *Remote Sensing 2020, Vol. 12, Page 428*, *12*, 428. Retrieved from https://www.mdpi.com/2072-4292/12/3/428/htmhttps://www.mdpi.com/2072-4292/12/3/428 doi: 10.3390/RS12030428

Khatami, S., Peel, M. C., Peterson, T. J., & Western, A. W. (2019, 11). Equifinality and flux mapping: A new approach to model evaluation and process representation under uncertainty. *Water Resources Research*, *55*, 8922-8941. Retrieved from https://onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/full/10.1029/2018WR023750 doi: 10.1029/2018WR023750

Kirchner, J. W. (2006, 3). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*. Retrieved from https://onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/full/

10.1029/2005WR004362https://onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/abs/10.1029/2005WR004362https://agupubs-onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/10.1029/2005WR004362 doi: 10.1029/2005WR004362

Klein, A. C. (2022). Hydrological response of the geul catchment to the rainfall in july 2021. Retrieved from https://repository.tudelft.nl/islandora/object/uuid%3Aee25d687-70af-4aca-ae41-78e3f83943bf

KNMI. (2020). *Klimaatviewer.* Retrieved from https://www.knmi.nl/klimaat-viewer/kaarten/temperatuur/gemiddelde-temperatuur/winter/Periode_1991-2020

KNMI. (2022a). *Dagwaarden van weerstations.* Retrieved 2022-07-09, from https://daggegevens.knmi.nl/klimatologie/daggegevens

KNMI. (2022b). *Uurwaarden van weerstations.* Retrieved 2022-07-09, from https://daggegevens.knmi.nl/klimatologie/uurgegevens

Kuczera, G., Kavetski, D., Franks, S., & Thyer, M. (2006, 11). Towards a bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*, *331*, 161-177. doi: 10.1016/J.JHYDROL.2006.05.010

Leo, G. D., & Sardanelli, F. (2020, 12). Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *European Radiology Experimental*, *4*, 1-8. Retrieved from https://eurradiolexp.springeropen.com/articles/10.1186/s41747-020-0145-y doi: 10.1186/S41747-020-0145-Y/METRICS

Li, X., Niu, Y., He, Q., & Wang, H. (2022, 1). Identifying driving factors of the runoff coefficient based on the geographic detector model in the upper reaches of huaihe river basin. *Open Geosciences*, *14*, 1421-1433. doi: 10.1515/geo-2022-0438

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997, 12). Development and test of the distributed hbv-96 hydrological model. *Journal of Hydrology*, *201*, 272-288. doi: 10.1016/S0022-1694(97)00041-3

Liu, Z., Wang, Y., Xu, Z., Duan, Q., Liu, Z., Wang, Y., … Xu, Z. (2017). Conceptual hydrological models. *Handbook of Hydrometeorological Ensemble Forecasting*, 1-23. Retrieved from https://link-springer-com.tudelft.idm.oclc.org/referenceworkentry/10.1007/978-3-642-40457-3_22-1 doi: 10.1007/978-3-642-40457-3_22-1

Loon, A. F. V. (2015, 7). Hydrological drought explained. *Wiley Interdisciplinary Reviews: Water*, *2*, 359-392. doi: 10.1002/WAT2.1085

Lu, J., Wang, G., Chen, T., Li, S., Fiifi, D., Hagan, T., … Su, B. (2021). A harmonized global land evaporation dataset from model-based products covering 1980-2017. *Earth Syst. Sci. Data*, *13*, 5879-5898. Retrieved from https://doi.org/10.5194/essd-13-5879-2021 doi: 10.5194/essd-13-5879-2021

López, P. L., Sutanudjaja, E. H., Schellekens, J., Sterk, G., & Bierkens, M. F. (2017, 6). Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products. *Hydrology and Earth System Sciences*, *21*, 3125-3144. doi: 10.5194/HESS-21-3125-2017

López-Ballesteros, A., Senent-Aparicio, J., Srinivasan, R., & Pérez-Sánchez, J. (2019, 9). Assessing the impact of best management practices in a highly anthropogenic and ungauged watershed using the swat model: A case study in the el beal watershed (southeast spain). *Agronomy 2019, Vol. 9, Page 576*, *9*, 576. Retrieved from https://www.mdpi.com/2073-4395/9/10/576/htmhttps://www.mdpi.com/2073-4395/9/10/576 doi: 10.3390/AGRONOMY9100576

Martens, B., Miralles, D. G., Lievens, H., Schalie, R. V. D., Jeu, R. A. D., Fernández-Prieto, D., … Verhoest, N. E. (2017, 5). Gleam v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, *10*, 1903-1925. doi: 10.5194/GMD-10-1903-2017

Melching, C. S. (1995). Reliability estimation. *Computer Modelsof Watershed Hydrology*, 69-118.

Miralles, D. G., Gash, J. H., Holmes, T. R., Jeu, R. A. D., & Dolman, A. J. (2010, 8). Global canopy interception from satellite observations. *Journal of Geophysical Research: Atmospheres*, *115*, 16122. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1029/2009JD013530https://onlinelibrary.wiley.com/doi/abs/10.1029/2009JD013530https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2009JD013530 doi: 10.1029/2009JD013530

Miralles, D. G., Holmes, T. R., Jeu, R. A. D., Gash, J. H., Meesters, A. G., & Dolman, A. J. (2011). Global land-surface evaporation estimated from satellite-based observations. *Hydrology and Earth Sys-*

tem Sciences, *15*, 453-469. doi: 10.5194/HESS-15-453-2011

Moges, E., Demissie, Y., Larsen, L., & Yassin, F. (2021, 1). Review: Sources of hydrological model uncertainties and advances in their analysis. *Water (Switzerland)*, *13*. doi: 10.3390/W13010028

Montanari, A., & Brath, A. (2004, 1). A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research*, *40*, 1106. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1029/2003WR002540 doi: 10.1029/2003WR002540

Moriasi, D. N., Arnold, J. G., Liew, M. W. V., Bingner, R. L., Harmel, R. D., & Veith, T. L. (1983). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*.

Moriasi, D. N., Gitau, M. W., Pai, N., & Daggupati, P. (2015). Hydrologic and water quality models: Performance measures and evaluation criteria. *Transactions of the ASABE*, *58*, 1763-1785. Retrieved from https://elibrary.asabe.org/azdez.asp?JID=3&AID=46548&CID=t2015&v=58&i=6&T=1http://elibrary.asabe.org/abstract.asp?aid=46548&confalias=&t=1&redir=&redirType=https://doi.org/10.13031/trans.58.10715 doi: 10.13031/TRANS.58.10715

Nash, J. E., & Sutcliffe, J. V. (1970, 4). River flow forecasting through conceptual models part i — a discussion of principles. *Journal of Hydrology*, *10*, 282-290. doi: 10.1016/0022-1694(70)90255-6

Nijzink, R. C., Almeida, S., Pechlivanidis, I. G., Capell, R., Gustafssons, D., Arheimer, B., … Hrachowitz, M. (2018, 10). Constraining conceptual hydrological models with multiple information sources. *Water Resources Research*, *54*, 8332-8362. Retrieved from https://onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/full/10.1029/2017WR02189 doi: 10.1029/2017WR021895

Nijzink, R. C., Samaniego, L., Mai, J., Kumar, R., Thober, S., Zink, M., … Hrachowitz, M. (2016, 3). The importance of topography-controlled sub-grid process heterogeneity and semi-quantitative prior constraints in distributed hydrological models. *Hydrology and Earth System Sciences*, *20*, 1151-1176. doi: 10.5194/HESS-20-1151-2016

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., & Loumagne, C. (2005, 3). Which potential evapotranspiration input for a lumped rainfall-runoff model? part 2 - towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling. *Journal of Hydrology*, *303*, 290-306. doi: 10.1016/j.jhydrol.2004.08.026

Oudin, L., Michel, C., & Anctil, F. (2005, 3). Which potential evapotranspiration input for a lumped rainfall-runoff model? part 1 - can rainfall-runoff models effectively handle detailed potential evapotranspiration inputs? *Journal of Hydrology*, *303*, 275-289. doi: 10.1016/j.jhydrol.2004.08.025

Pierik, C. (2022). *The impact of an additional phenology model on the performance of conceptual hydrological models.* Retrieved from https://repository.tudelft.nl/islandora/object/uuid%3Ad065bbed-1eb0-4bb6-84b8-ae2370b84e61

Planet. (2022). *Producten — planet planetary variables 3.0-champaign documentation.* Retrieved 2022-20-10, from https://docs.vandersat.com/projects/satdata/producten.html#actuele-verdamping

Priestley, C. H. B., & Taylor, R. J. (1972). *On the assessment of surface heat flux and evaporation using large-scale parameters* (Vol. 100).

Provincie Limburg. (2021). *Ontwerp-kernrapport geuldal (157) 2021-2027.* Retrieved from https://www.bij12.nl/wp-content/uploads/2021/11/Natura-2000-Kernrapport-Beheerplan-157-Geuldal.pdf

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010, 5). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*, 5521. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1029/2009WR008328 doi: 10.1029/2009WR008328

Rennó, C. D., Nobre, A. D., Cuartas, L. A., Soares, J. V., Hodnett, M. G., Tomasella, J., & Waterloo, M. J. (2008, 9). Hand, a new terrain descriptor using srtm-dem: Mapping terra-firme rainforest environments in amazonia. *Remote Sensing of Environment*, *112*, 3469-3481. doi: 10.1016/J.RSE.2008.03.018

Savenije, H. H. G. (2010). Hydrology and earth system sciences hess opinions "topography driven conceptual modelling (flex-topo)". *Hydrol. Earth Syst. Sci*, *14*, 2681-2692. Retrieved from www.hydrol-earth-syst-sci.net/14/2681/2010/ doi: 10.5194/hess-14-2681-2010

Schumann, A. H. (1998, 2). Thiessen polygon. *Hydrology and Lakes*, 648-649. Retrieved from https://link.springer.com/referenceworkentry/10.1007/1-4020-4513-1_220 doi:

10.1007/1-4020-4513-1_220

Seibert, J. (1997). Estimation of parameter uncertainty in the hbv model. *Nordic Hydrology*, *28*, 247-262. doi: 10.2166/nh.1998.15

Senese, A., Maugeri, M., Vuillermoz, E., Smiraglia, C., & Diolaiuti, G. (2014). Air temperature thresholds to evaluate snow melting air temperature thresholds to evaluate snow melting at the surface of alpine glaciers by t-index models: the case study of forni glacier (italy) air temperature thresholds to evaluate snow melting. *TCD*, *8*, 1563-1587. Retrieved from www.the-cryosphere-discuss.net/8/1563/2014/ doi: 10.5194/tcd-8-1563-2014

Shokri, A., Walker, J. P., van Dijk, A. I., Wright, A. J., & Pauwels, V. R. (2018, 4). Application of the patient rule induction method to detect hydrologic model behavioural parameters and quantify uncertainty. *Hydrological Processes*, *32*, 1005-1025. Retrieved from https://onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/full/10.1002/hyp.11464 doi: 10.1002/HYP.11464

Sirisena, T. A. G., Maskey, S., & Ranasinghe, R. (2020, 11). Hydrological model calibration with streamflow and remote sensing based evapotranspiration data in a data poor basin. *Remote Sensing*, *12*, 1-24. doi: 10.3390/RS12223768

STOWA. (2010). Stowa 2010-36. Retrieved from https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.stowa.nl%2Fsites%2Fdefault%2Ffiles%2Fassets%2FPUBLICATIES%2FPublicaties%25202010%2FSTOWA%25202010-36.pdf&psig=AOvVaw0pvWVfASNd-pGPZusH_QE6&ust=1682761723845000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCJjam4emzP4CFQAAAAAdAAAAABAJ

Tweldebrahn, A. T., Burkhart, J. F., & Schuler, T. V. (2018). Parameter uncertainty analysis for an operational hydrological model using residual based and limits of acceptability approaches. *Hydrology and Earth System Sciences Discussion*, *15*. Retrieved from https://doi.org/10.5194/hess-2018-158 doi: 10.5194/hess-2018-158

Vandenberghe, J., de Moor, J. J., & Spanjaard, G. (2012, 7). Natural change and human impact in a present-day fluvial catchment: The geul river, southern netherlands. *Geomorphology*, *159-160*, 1-14. doi: 10.1016/J.GEOMORPH.2011.12.034

van Kraalingen, D., & Stol, W. (1997). *Evapotranspiration modules for crop growth simulation : implementation of the algorithms from penman, makkink and priestley-taylor*. DLO Research Institute for Agrobiology and Soil Fertility. Retrieved from https://edepot.wur.nl/4413#:~:text=In%20the%20Priestley%2DTaylor%20equation,vapour%20pressure%20and%20wind%20speed.

Verbeiren, B., Nguyen, H. K., Wirion, C., & Batelaan, O. (2016, 6). An earth observation based method to assess the influence of seasonal dynamics of canopy interception storage on the urban water balance. *http://journals.openedition.org/belgeo*. Retrieved from http://journals.openedition.org/belgeo/17806 doi: 10.4000/BELGEO.17806

Villarini, G., Mandapaka, P. V., Krajewski, W. F., & Moore, R. J. (2008, 6). Rainfall and sampling uncertainties: A rain gauge perspective. *Journal of Geophysical Research: Atmospheres*, *113*, 11102. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1029/2007JD009214 doi: 10.1029/2007JD009214

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

Waterschap Limburg. (2019). *Leggerkaart - waterschap limburg.* Retrieved from https://www.waterschaplimburg.nl/uwbuurt/kaarten-meetgegevens/leggerkaart/

Wawrzyniak, T., Osuch, M., Nawrot, A., & Napiorkowski, J. J. (2017, 7). Run-off modelling in an arctic unglaciated catchment (fuglebekken, spitsbergen). *Annals of Glaciology*, *58*, 36-46. doi: 10.1017/aog.2017.8

Whelan, M. J., Kim, J., Suganuma, N., & MacKay, D. (2019, 7). Uncertainty and equifinality in environmental modelling of organic pollutants with specific focus on cyclic volatile methyl siloxanes. *Environmental Science: Processes Impacts*, *21*, 1085-1098. Retrieved from https://pubs-rsc-org.tudelft.idm.oclc.org/en/content/articlehtml/2019/em/c9em00099bhttps://pubs-rsc-org.tudelft.idm.oclc.org/en/content/articlelanding/2019/em/c9em00099b doi: 10.1039/C9EM00099B

Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, *30*, 1756-1774. Retrieved from www.elsevier.com/locate/advwatres doi: 10.1016/j.advwatres

.2007.01.005

Zektser, I. S., & Loaiciga, H. A. (1993). Groundwater fluxes in the global hydrologic cycle: past, present and future. *Journal of Hydrology*, *144*, 405-427.

Zhang, Y., Chiew, F. H., Zhang, L., & Li, H. (2009, 8). Use of remotely sensed actual evapotranspiration to improve rainfall–runoff modeling in southeast australia. *Journal of Hydrometeorology*, *10*, 969-980. Retrieved from `https://journals.ametsoc.org/view/journals/hydr/10/4/2009jhm1061_1.xml` doi: 10.1175/2009JHM1061.1

Zhao, L., Xia, J., yu Xu, C., Wang, Z., Sobkowiak, L., & Long, C. (2013, 4). Evapotranspiration estimation methods in hydrological models. *Journal of Geographical Sciences*, *23*, 359-369. doi: 10.1007/s11442-013-1015-9

Zhong, F., Jiang, S., van Dijk, A. I. J. M., Ren, L., Schellekens, J., & Miralles, D. G. (2022, 11). Revisiting large-scale interception patterns constrained by a synthesis of global experimental data. *Hydrology and Earth System Sciences*, *26*, 5647-5667. doi: 10.5194/HESS-26-5647-2022

A

Measurement locations

Table A.1: Precipitation measurement location, including measurement periods

Station	Location X	Location Y	Start	End	Source
18.P.37	201510	321755	1993-03-16 20:00:00	2021-08-25 06:00:00	WL
10.P.36	177314	320266	1994-09-30 01:00:00	2021-08-25 06:00:00	WL
3.P.29	193962	345513	1996-03-02 01:00:00	2021-10-20 00:00:00	WL
6.P.21	189335	337212	1994-06-20 01:00:00	2021-11-12 13:00:00	WL
15.P.41	185950	308371	2010-12-15 08:00:00	2021-11-12 13:00:00	WL
10.P.30	191269	318617	1998-03-27 01:00:00	2021-11-12 13:00:00	WL
1.P.2	196887	359107	1993-12-02 06:00:00	2021-11-12 13:00:00	WL
6.P.40	186546	326311	2010-12-07 14:00:00	2021-11-12 13:00:00	WL
8.P.18	180697	331683	1993-03-16 19:00:00	2021-11-12 13:00:00	WL
12.P.25	198559	308595	1998-09-03 01:00:00	2021-11-12 13:00:00	WL
Maastricht	5.761872	50.905348	1906-01-01 00:00:00	2021-12-31 00:00:00	KNMI
Battice	253200	149235	2002-01-01 00:00:00	2021-12-31 00:00:00	SPW
Gemmenich	263361	161066	2002-01-01 00:00:00	2021-12-31 00:00:00	SPW
Jalhay	266002	138638	2002-01-01 00:00:00	2018-12-31 00:00:00	SPW
Ternell	276725	141812	2002-01-01 00:00:00	2018-12-31 00:00:00	SPW

Table A.2: Discharge measurement location, including measurement periods

Station	Location X	Location Y	Start	End	Source
Meerssen	178825	32243	1993-03-16 20:00:00	2021-08-25 06:00:00	WL
Schin op Geul	188938	318437	1994-09-30 01:00:00	2021-08-25 06:00:00	WL
Azijnfabriek	193962	345513	1996-03-02 01:00:00	2021-10-20 00:00:00	WL
Eys	189335	337212	1994-06-20 01:00:00	2021-11-12 13:00:00	WL
Partij	185950	308371	2010-12-15 08:00:00	2021-11-12 13:00:00	WL
Hommericht	191269	318617	1998-03-27 01:00:00	2021-11-12 13:00:00	WL
Cottessen	196887	359107	1993-12-02 06:00:00	2021-11-12 13:00:00	WL
Sippenaeken	186546	326311	2010-12-07 14:00:00	2021-11-12 13:00:00	SPW
Kelmis	180697	331683	1993-03-16 19:00:00	2021-11-12 13:00:00	SPW

B

Data sources

B.1. Forcing

- Hourly radiation and temperature data from KNMI measurement point Maastricht (KNMI, 2022b)
- Daily precipitation, Makkink potential evaporation and temperature data from KNMI measurement point Maastricht (KNMI, 2022a)
- Hourly precipitation data Waterschap Limburg (received by Delteres, 2022)
- Hourly precipitation data Beglium from SPW measurement points (Direction généraleopérationnelle de la Mobilité et des Voies hydrauliques, 2022)
- ERA5 hourly actual evaporation gridded (Hersbach et al., 2018)

B.2. Geul catchment

- Land borders are from GADM (GADM, 2022)
- Waterways in The Netherlands from Waterschap Limburg (received by Delteres, 2022)
- Waterways in Belgium from the Humanitarian OpenStreetMap Team (2021)
- Elevation data from the European Environment Agency (2016)
- Slope data is calculated with elevation data

C

Conceptual model per landscape

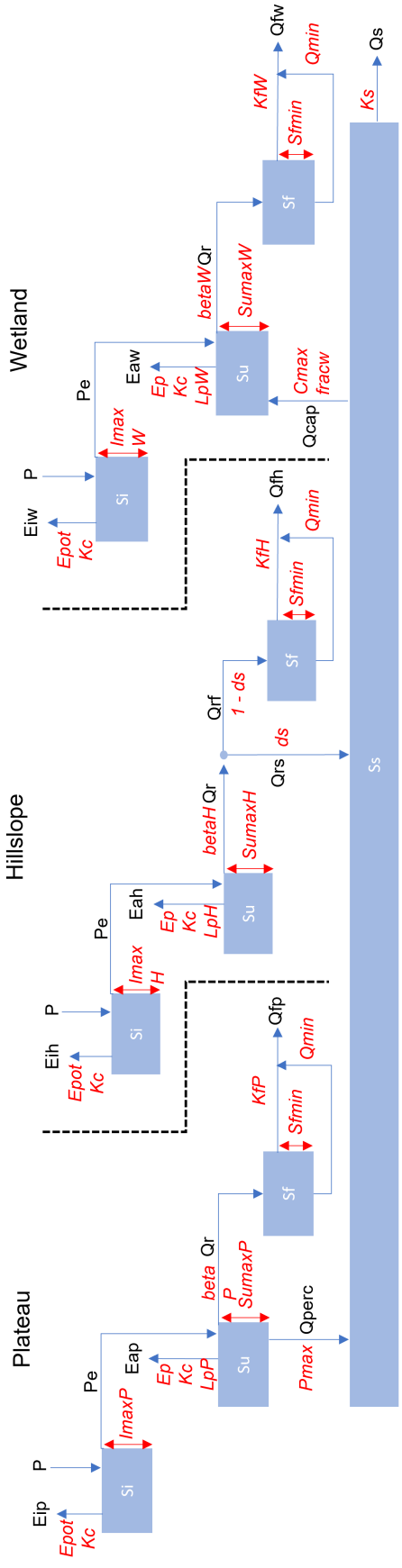


Figure C.1: Individual model per landscape class

D

Minimum discharge

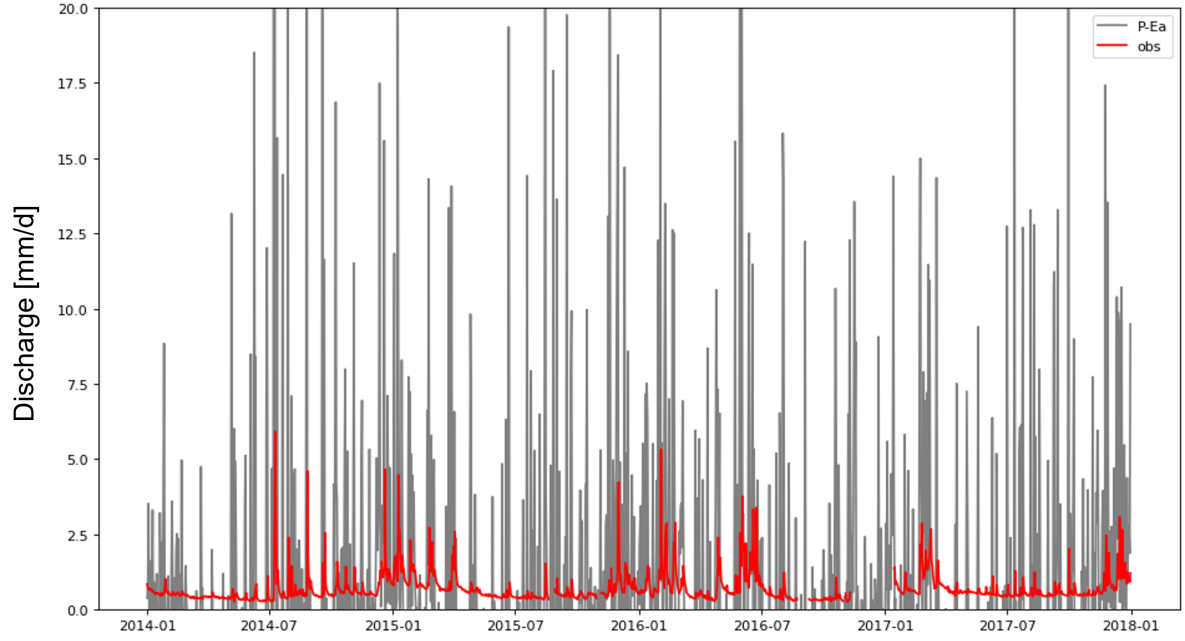


Figure D.1: Minimum discharge plot 1

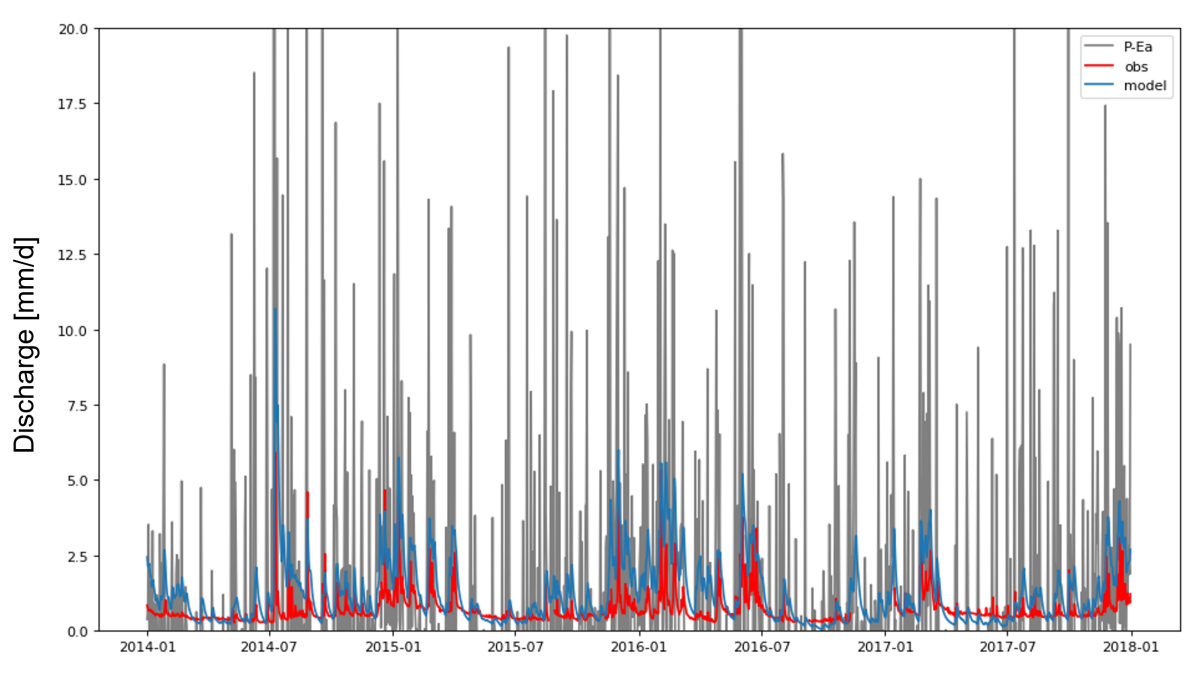
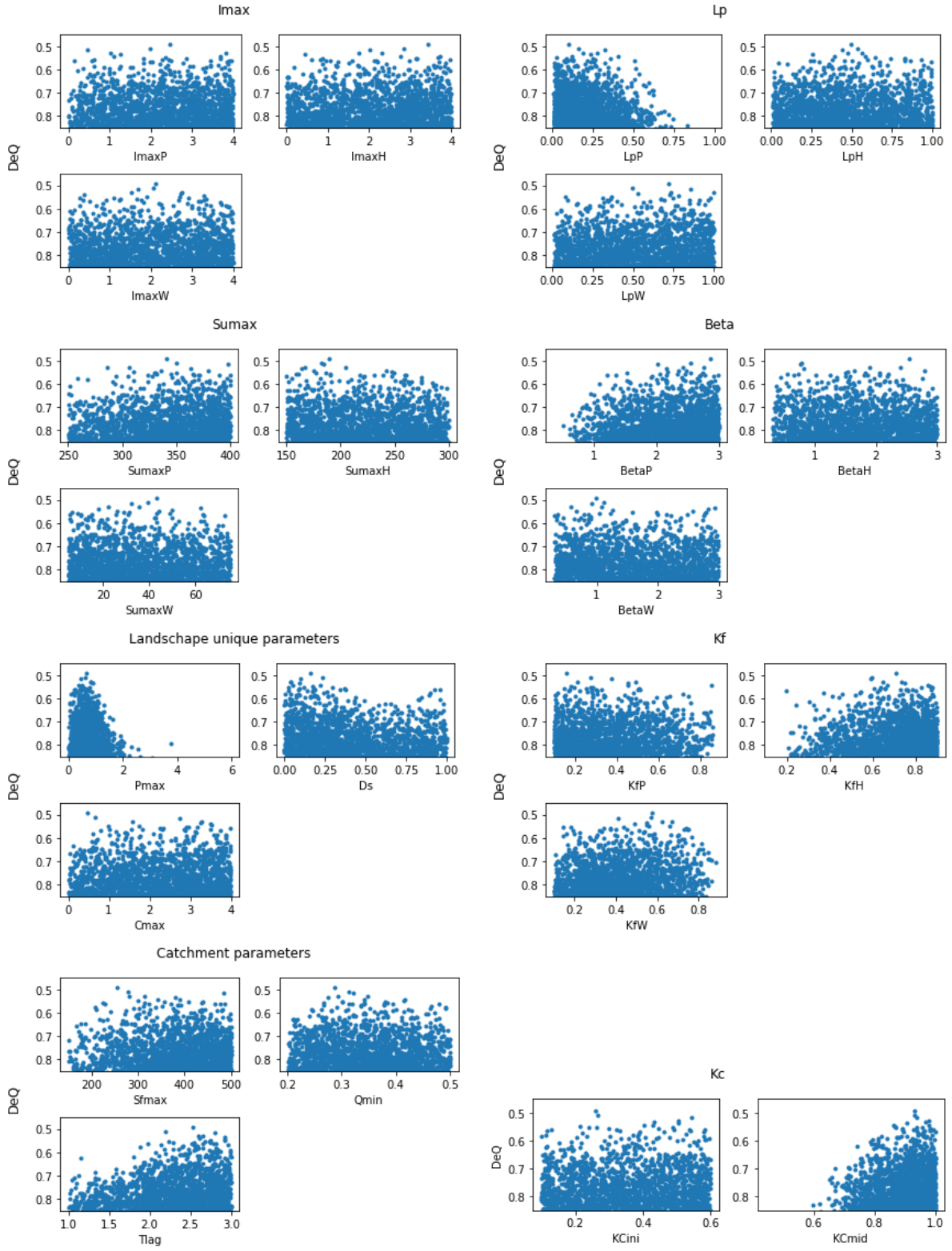


Figure D.2: Minimum discharge plot 2

E

Results

E.1. Dotty plots original parameter intervals



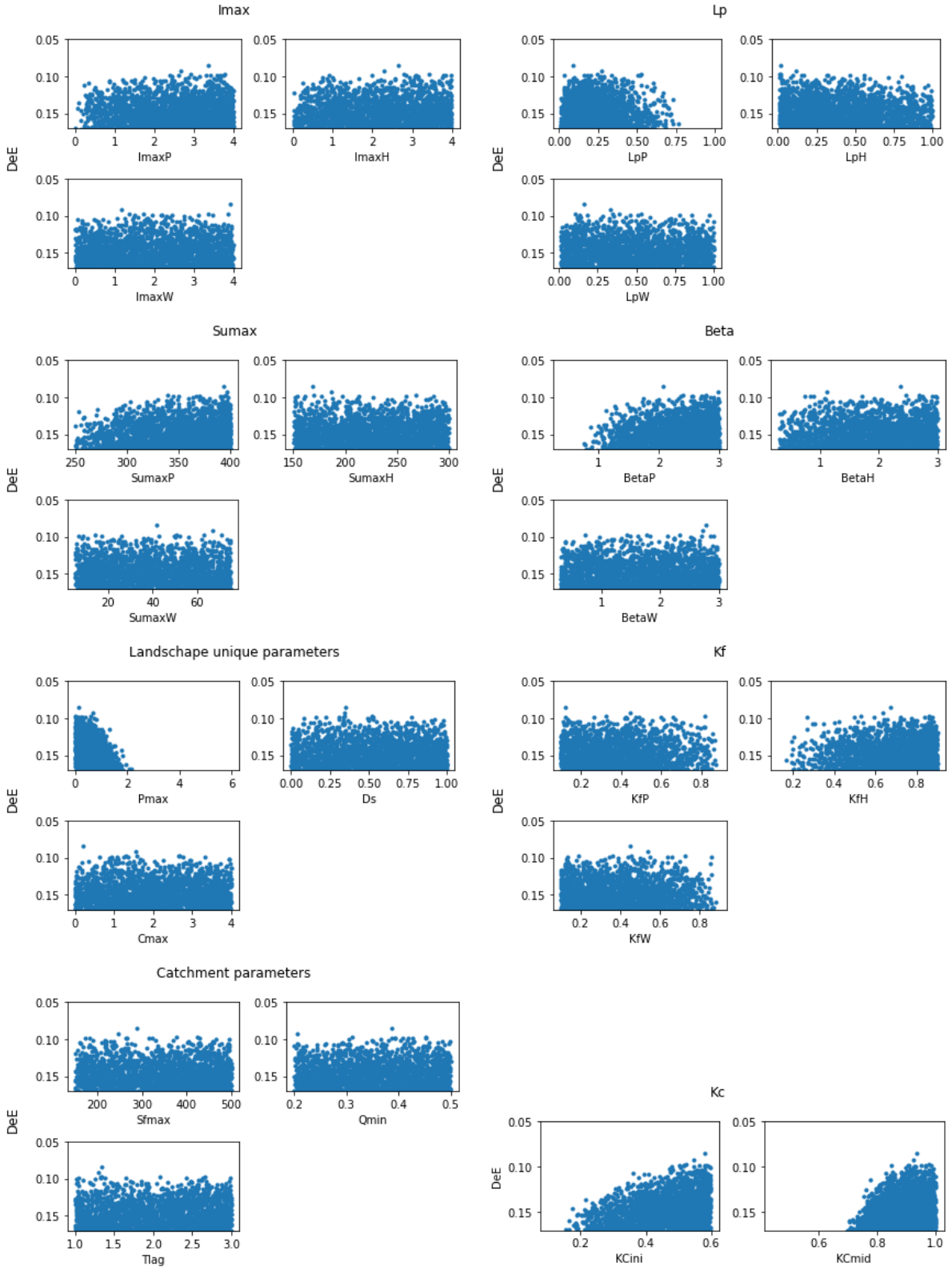


Figure E.2: Dotty plot DeE

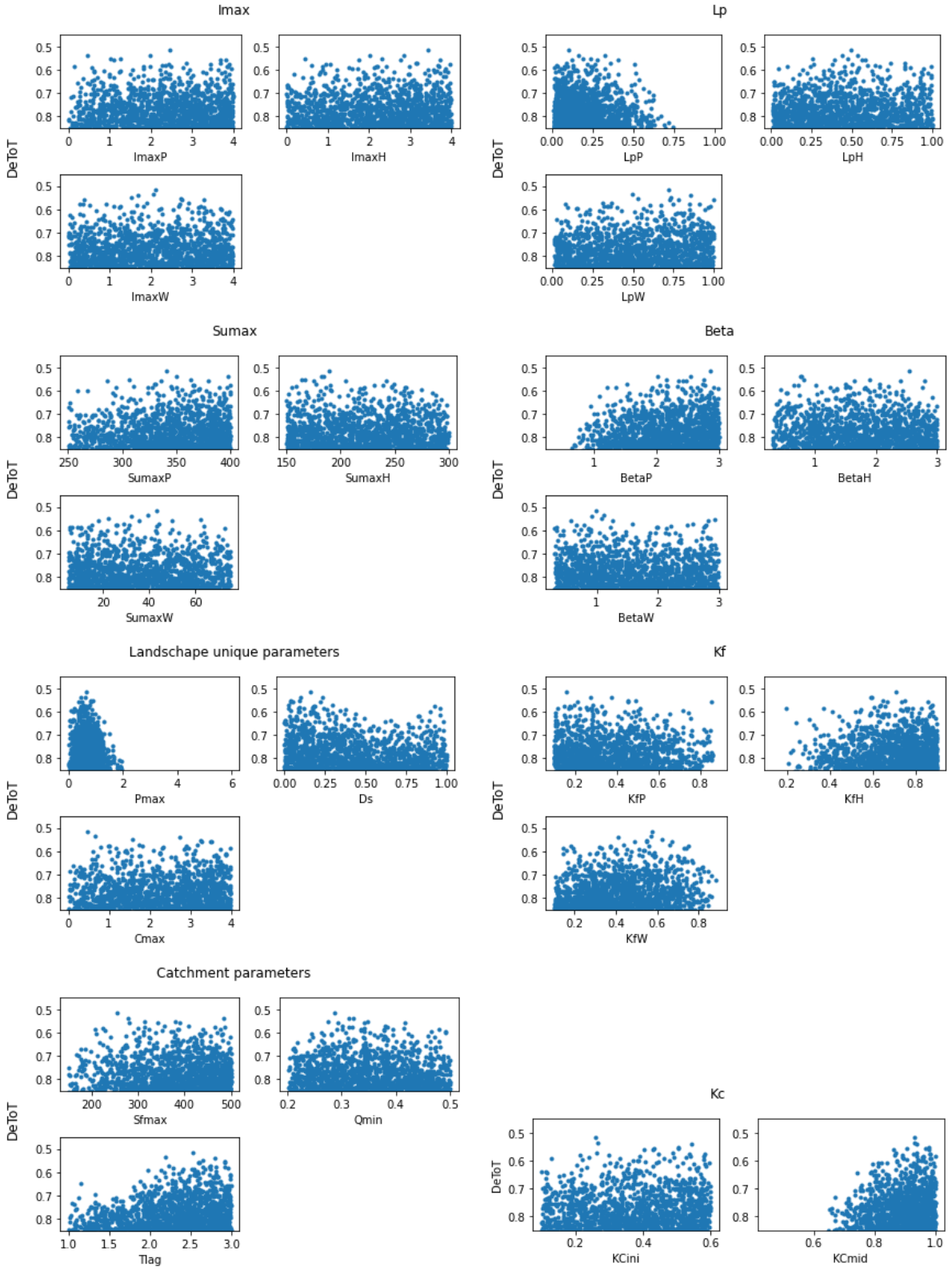


Figure E.3: Dotty plot DeTot

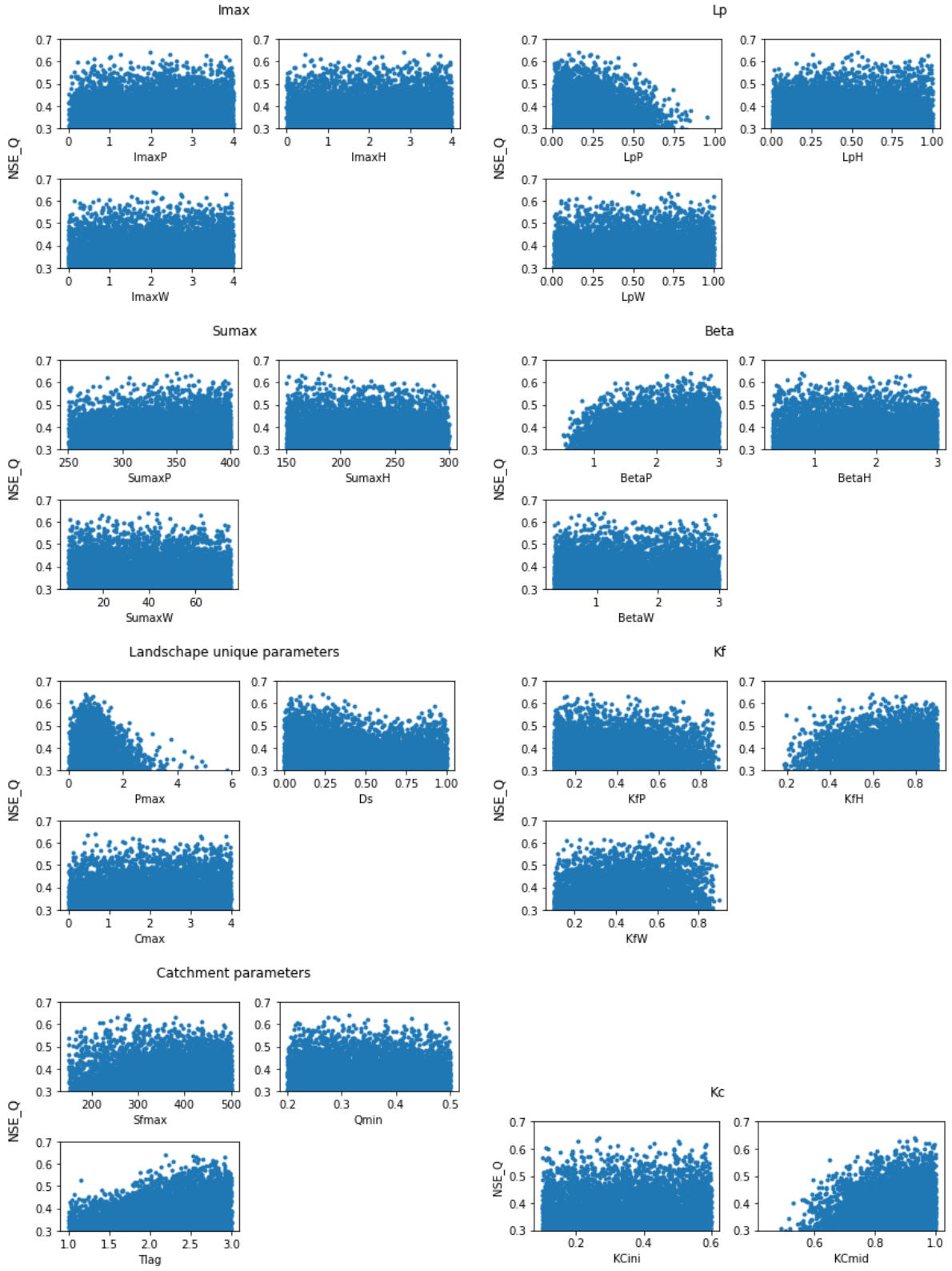


Figure E.4: Dotty plot NSE Q

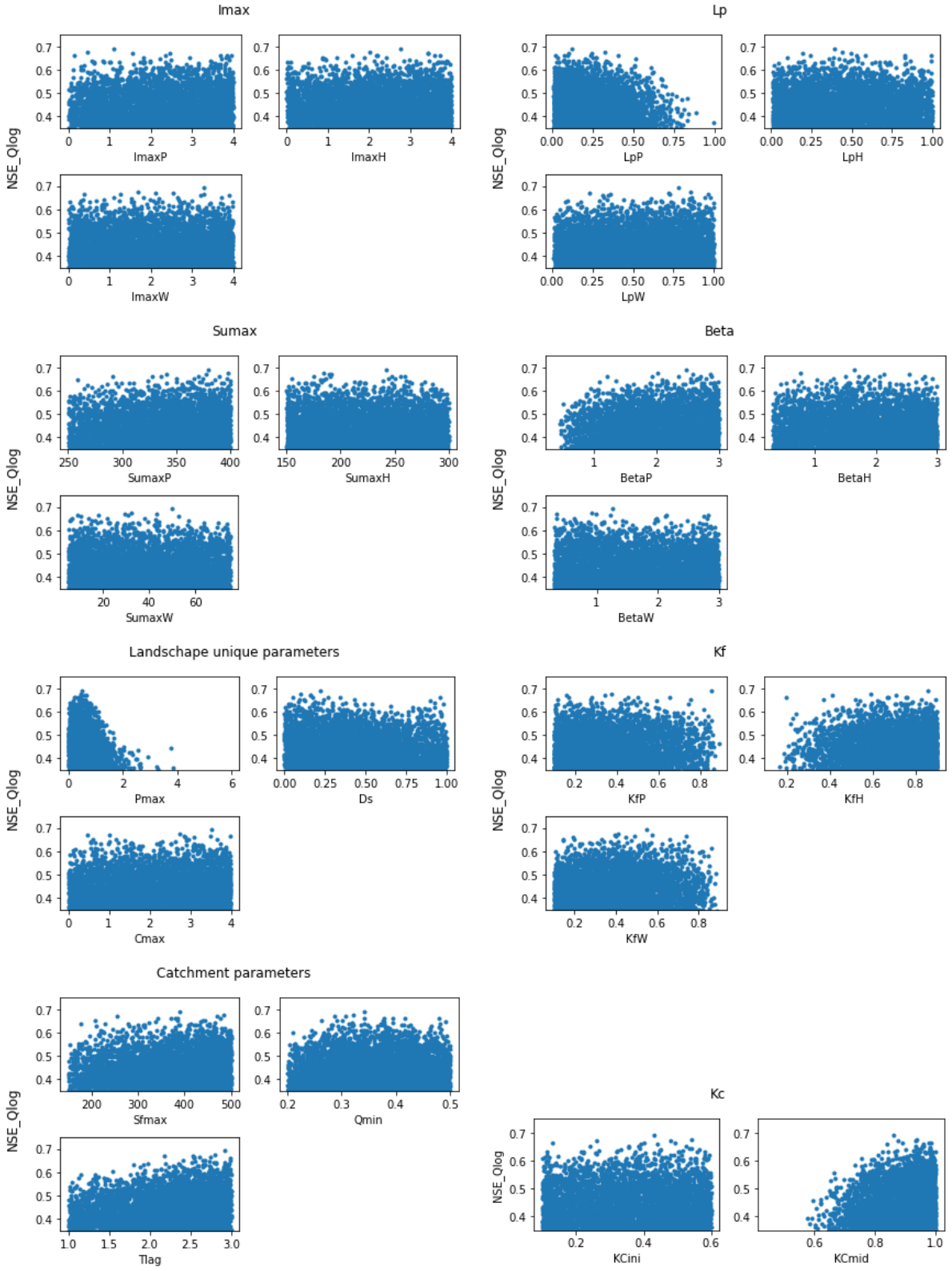


Figure E.5: Dotty plot NSE Qlog

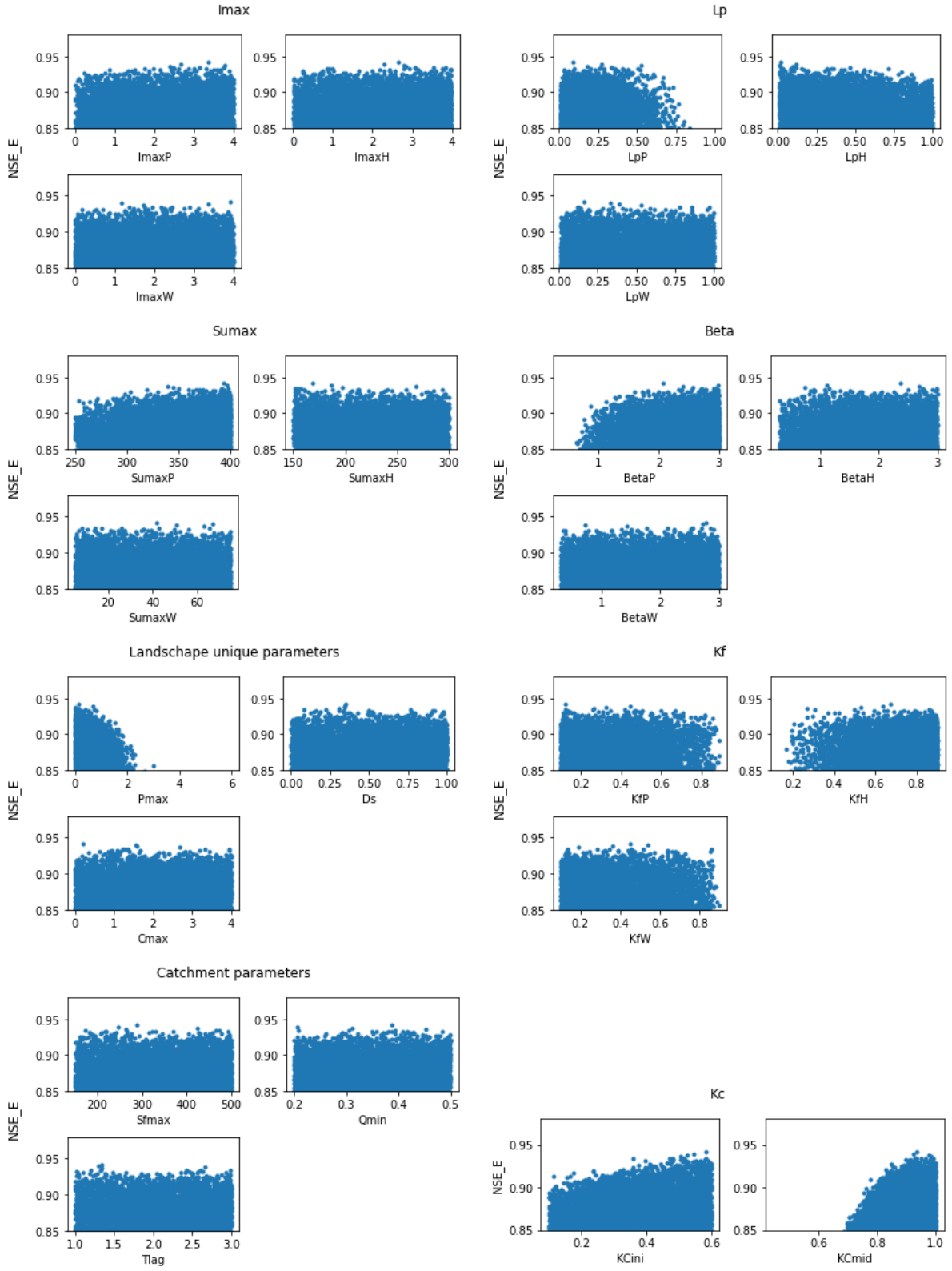


Figure E.6: Dotty plot NSE E

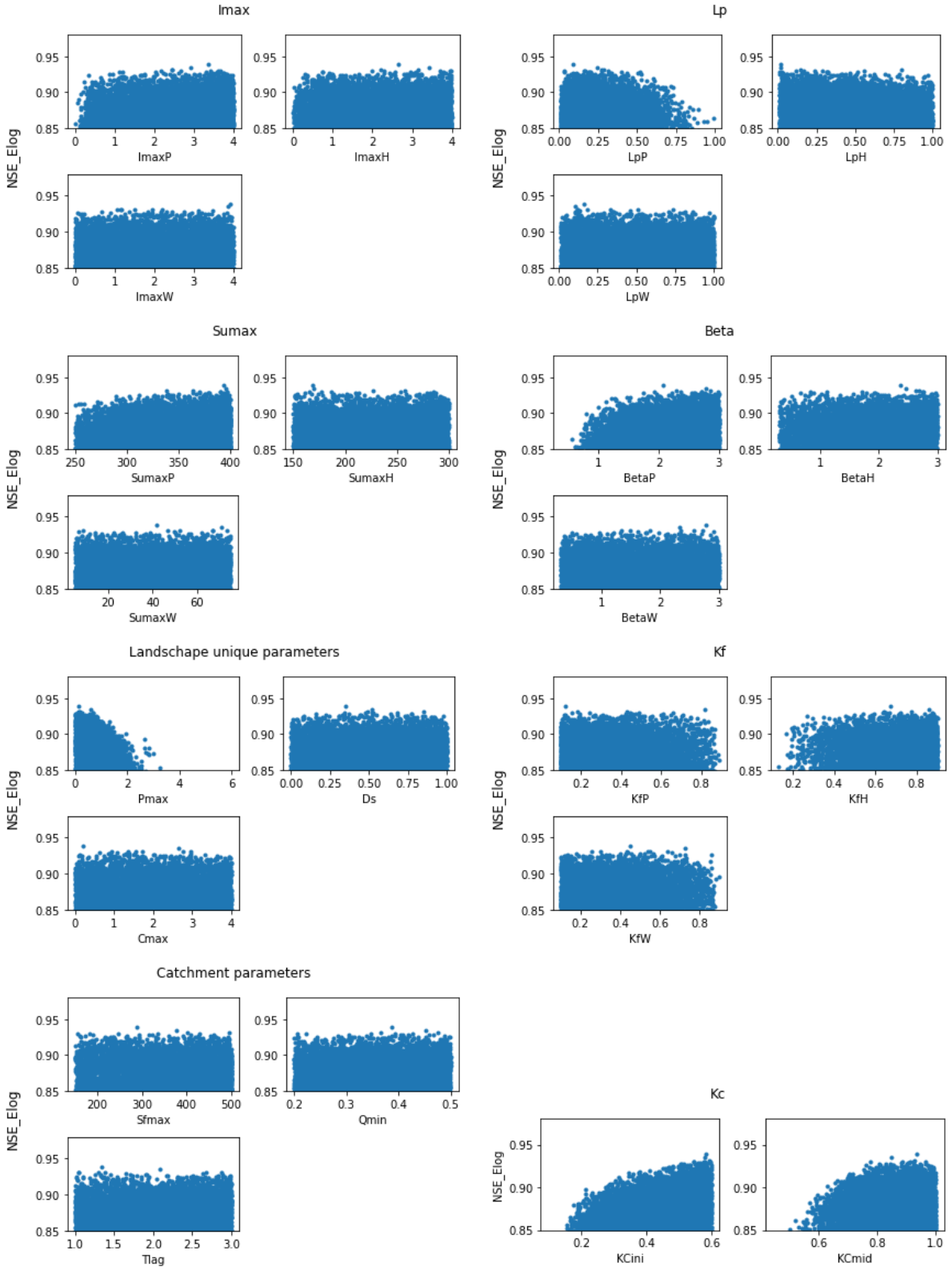


Figure E.7: Dotty plot NSE Elog

E.2. Dotty plots revised parameter intervals

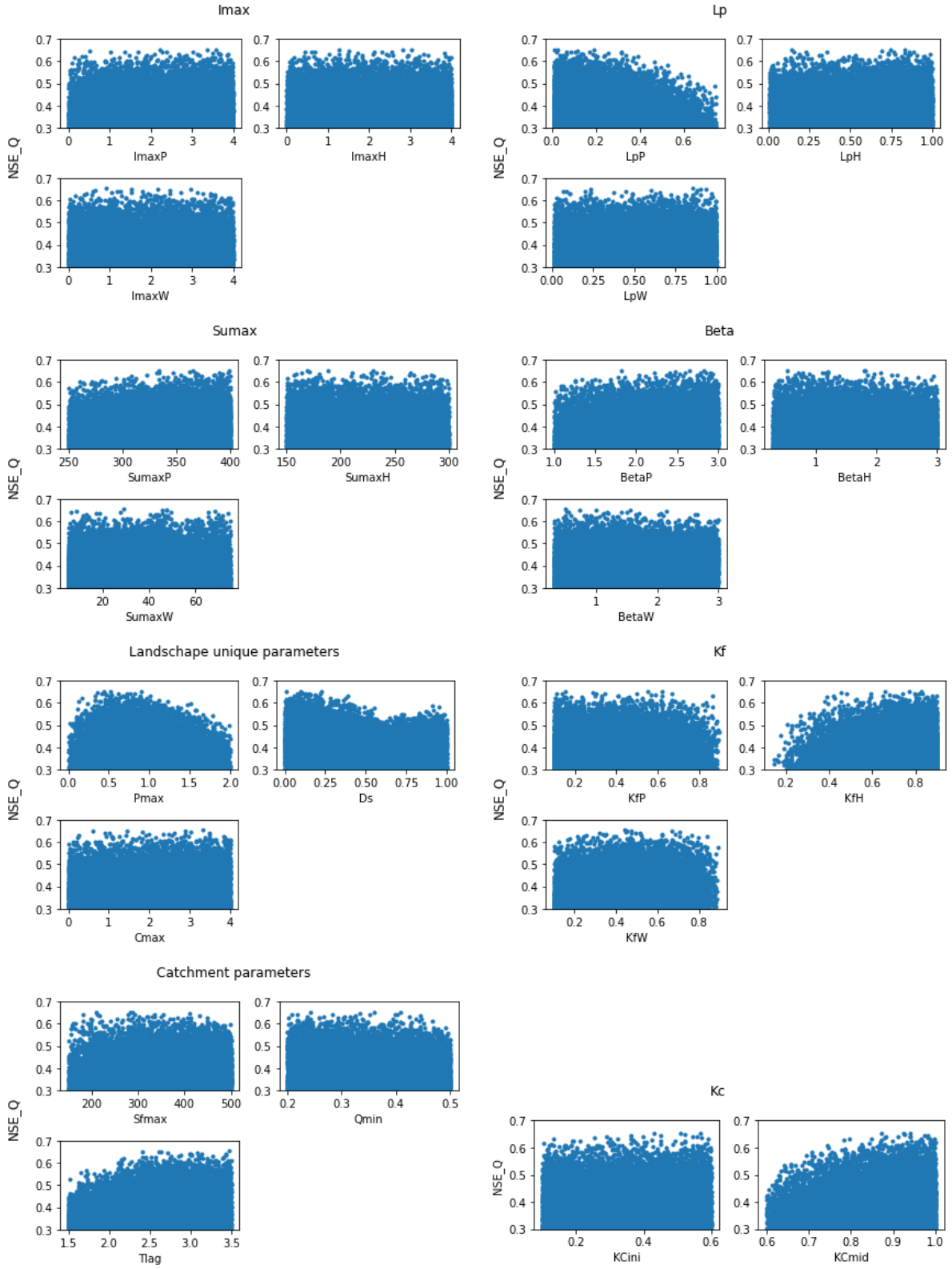


Figure E.8: Dotty plot NSE Q

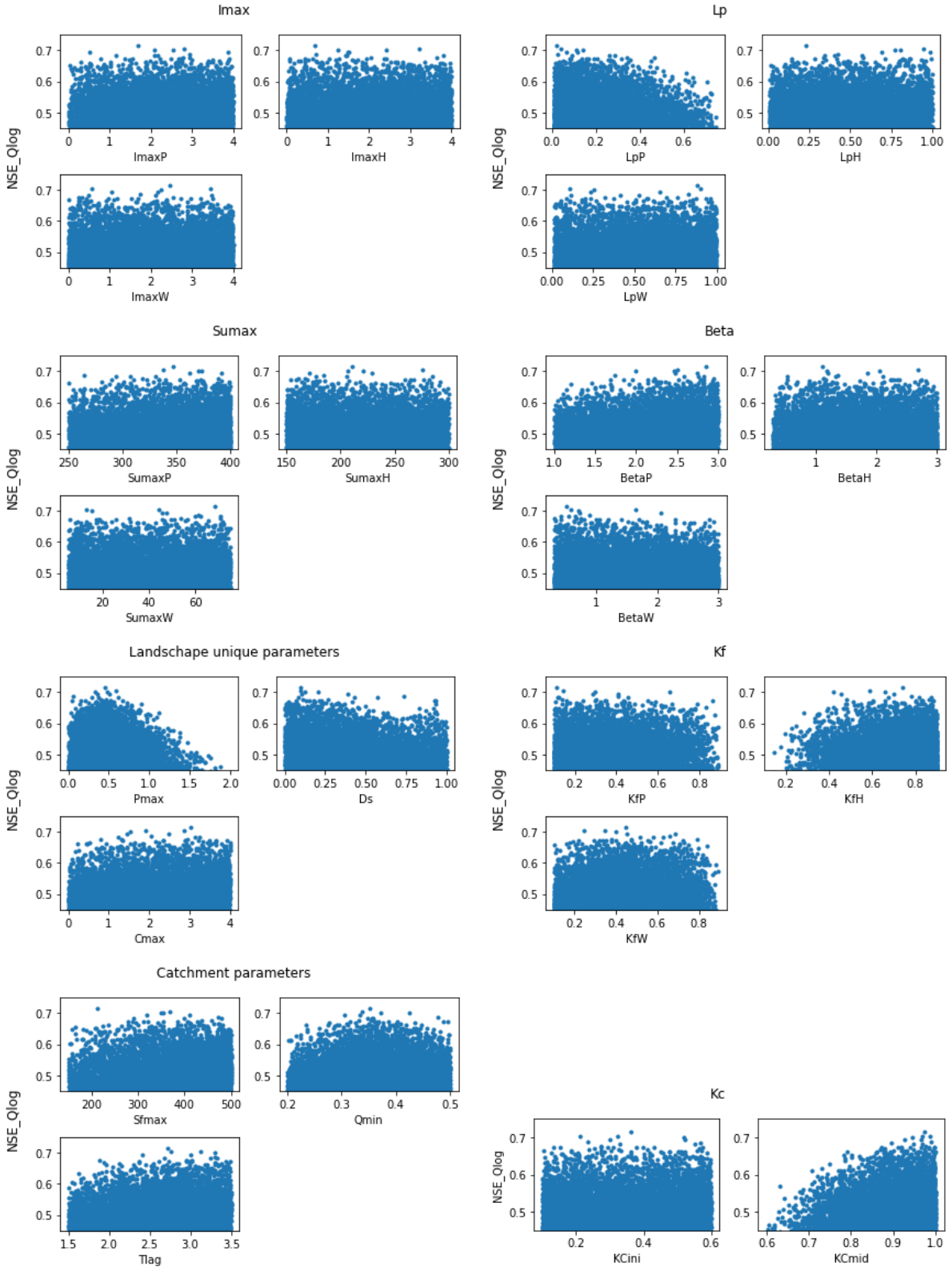


Figure E.9: Dotty plot NSE Qlog

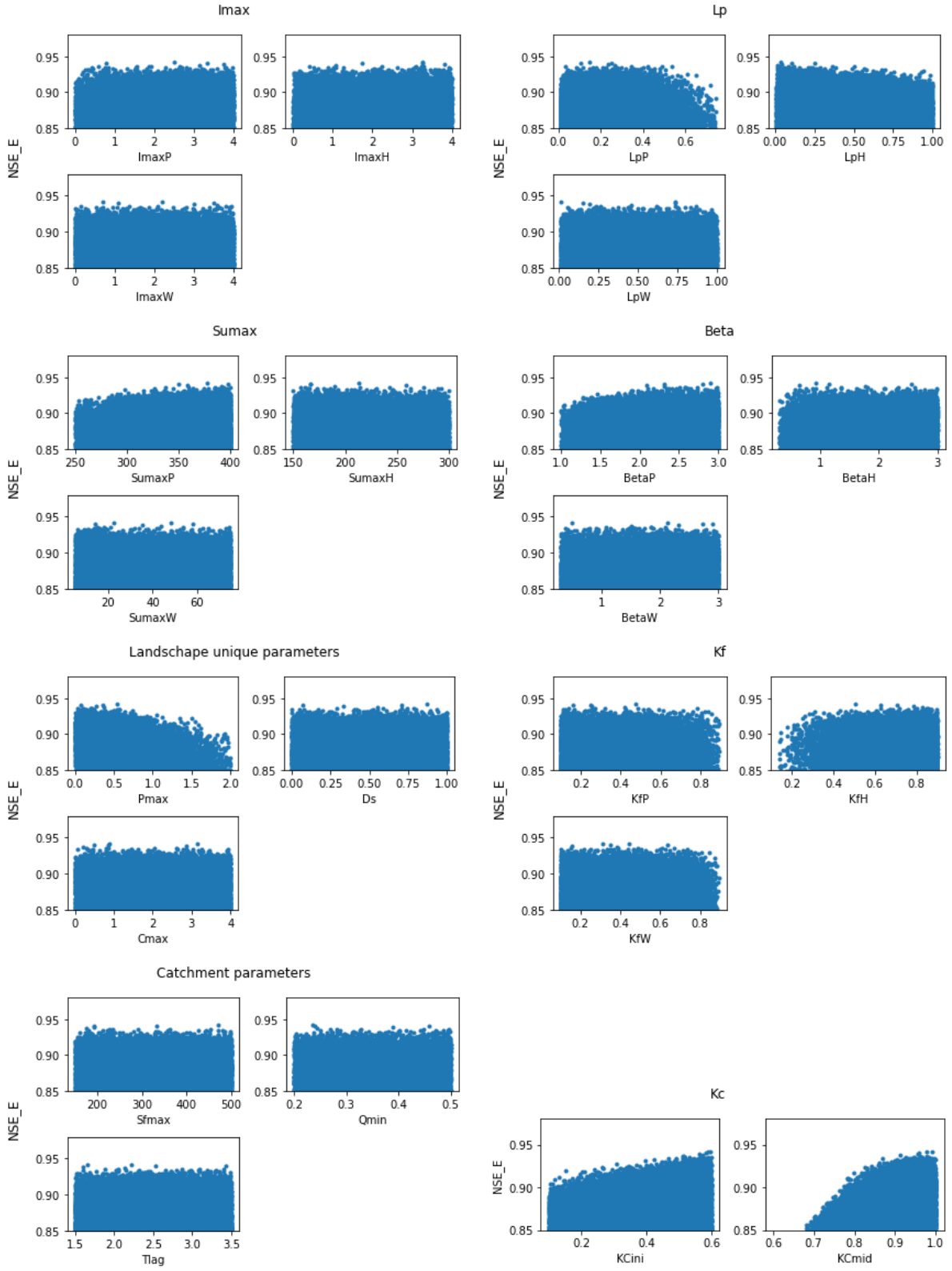


Figure E.10: Dotty plot NSE E

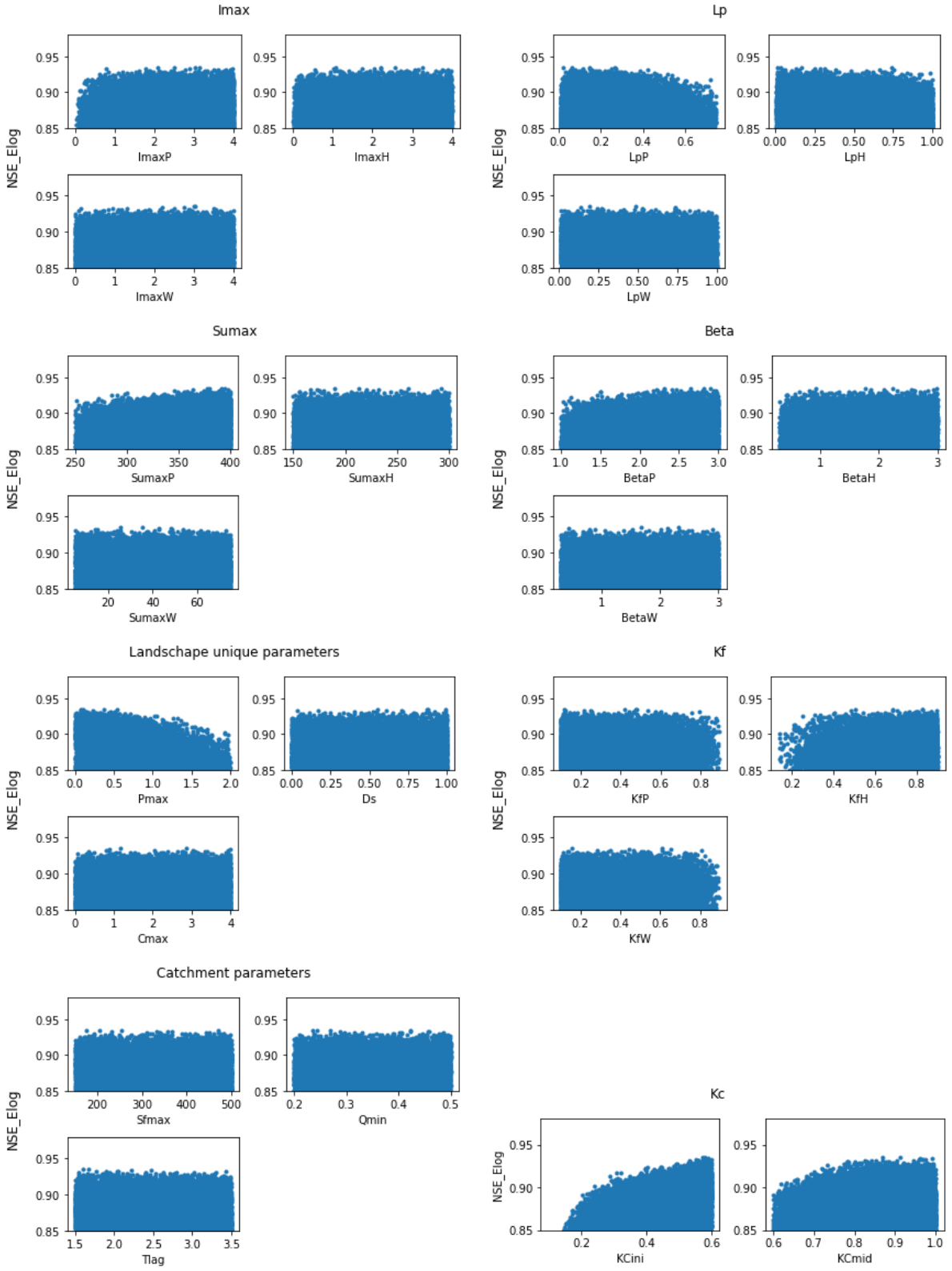


Figure E.11: Dotty plot NSE Elog

E.3. Stream flow results

E.3.1. Run 1

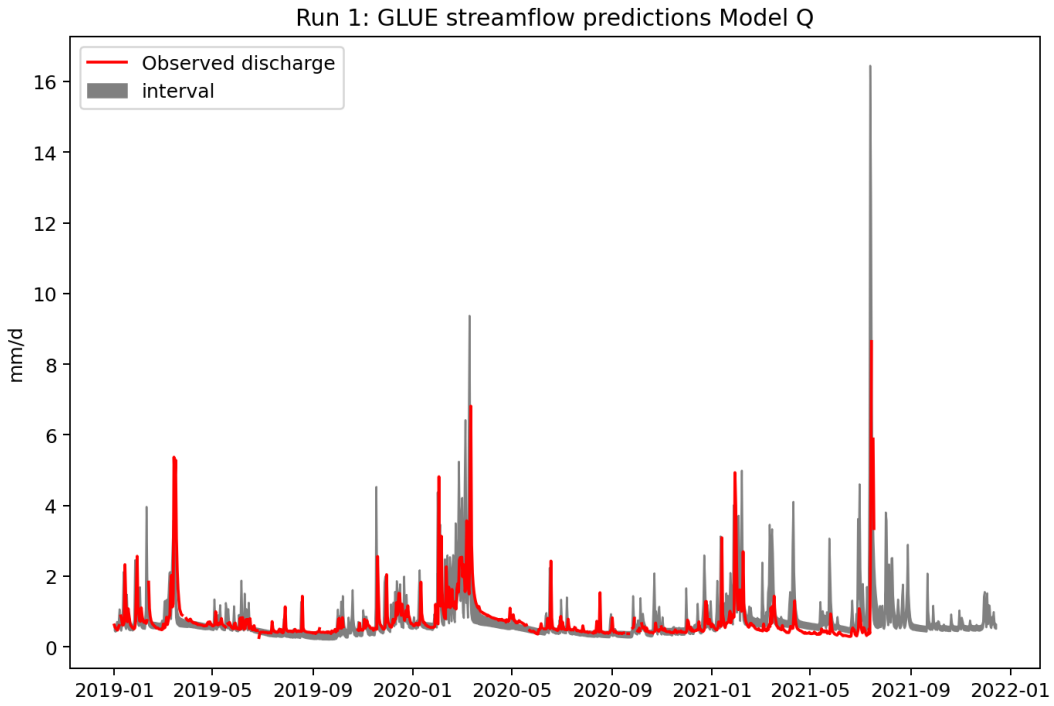


Figure E.12: Run 1:Stream flow predictions from model calibrated on Q 300 best calibration runs using GLUE

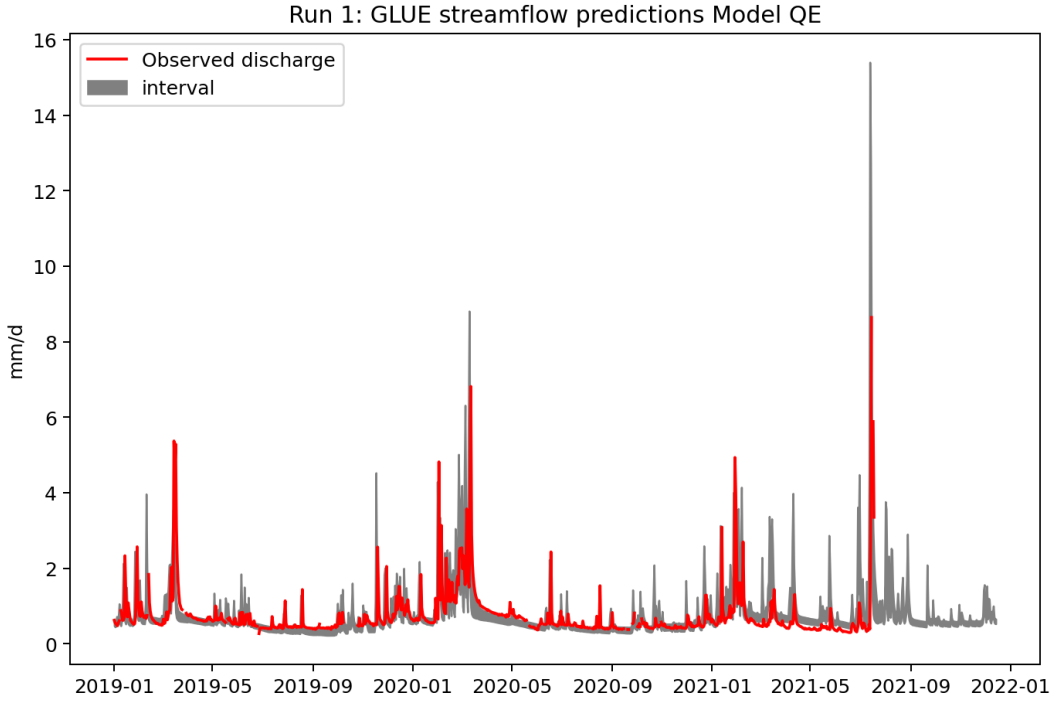


Figure E.13: Run 1:Stream flow predictions from model calibrated on Q and E of 300 best calibration runs using GLUE

E.3.2. Run 2

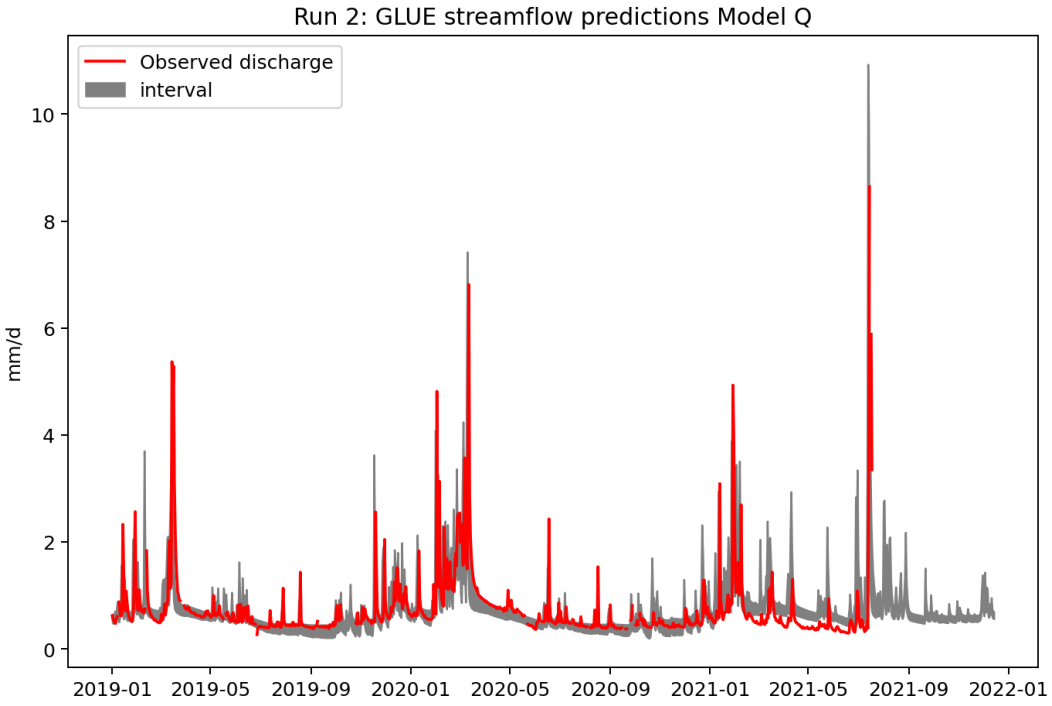


Figure E.14: Run 2:Stream flow predictions from model calibrated on Q 300 best calibration runs using GLUE

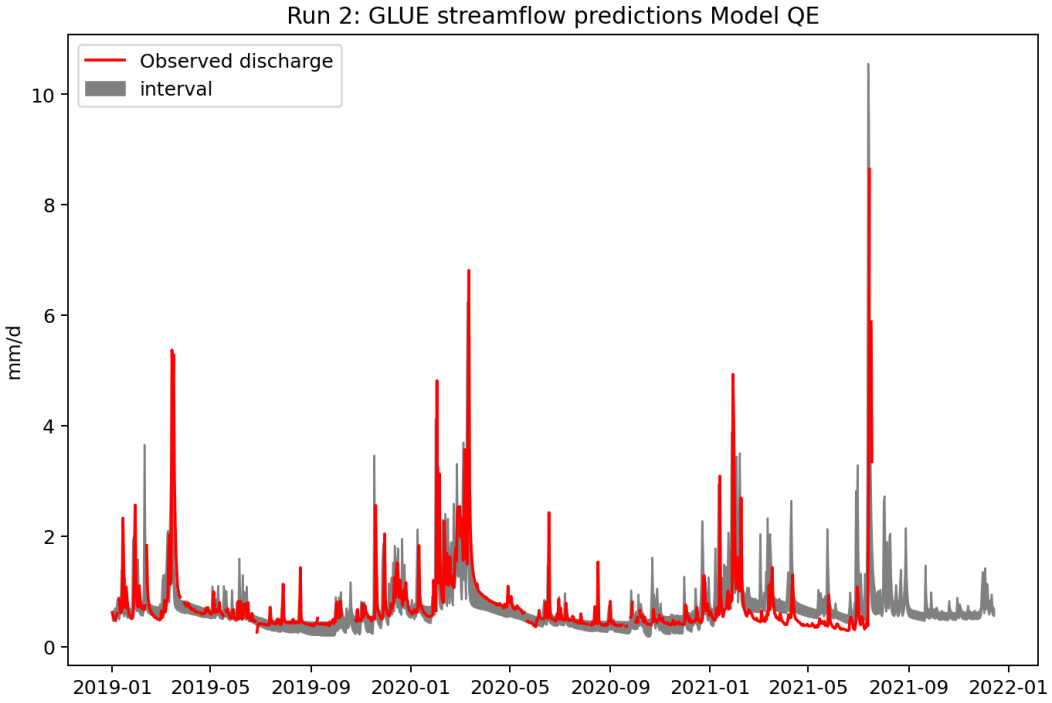


Figure E.15: Run 2:Stream flow predictions from model calibrated on Q and E of 300 best calibration runs using GLUE

E.3.3. Run 3

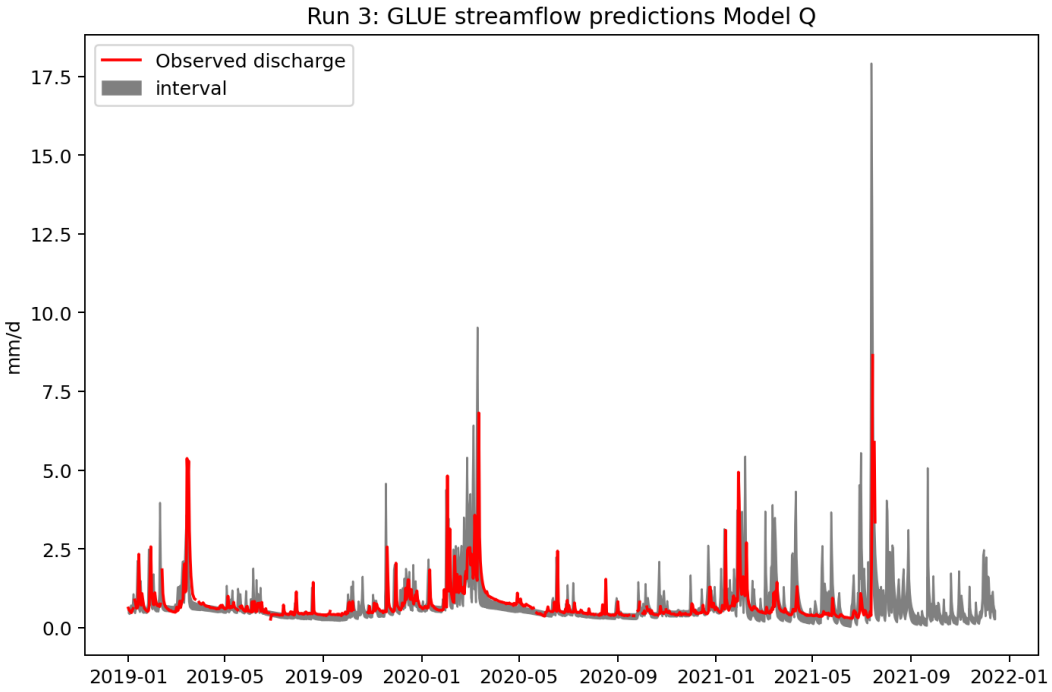


Figure E.16: Run 3:Stream flow predictions from model calibrated on Q 300 best calibration runs using GLUE

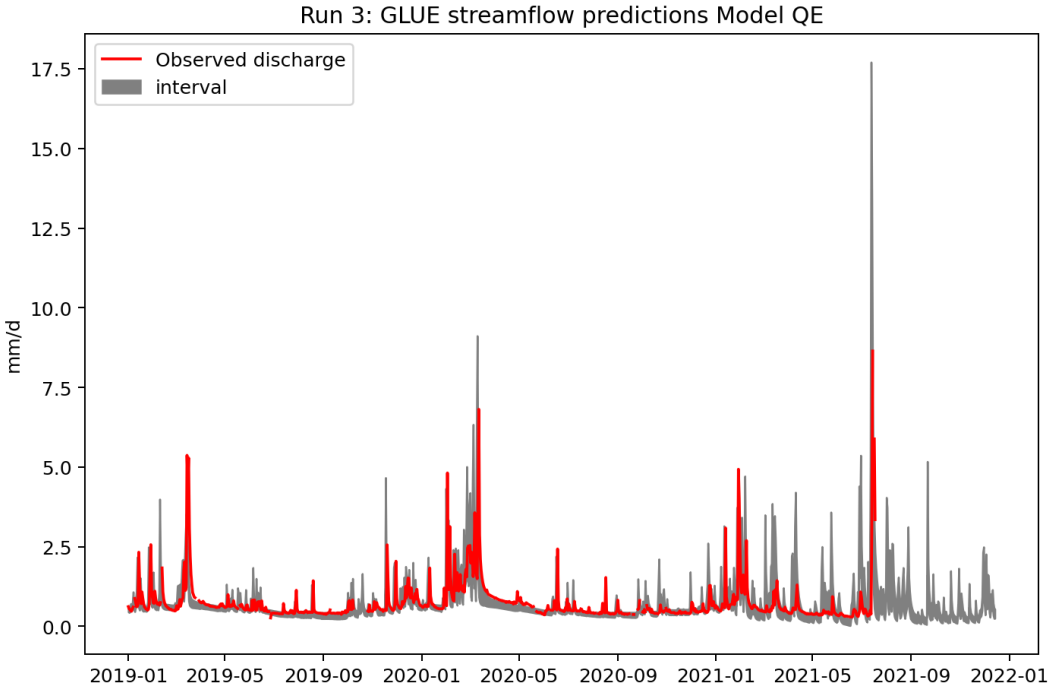


Figure E.17: Run 3:Stream flow predictions from model calibrated on Q and E of 300 best calibration runs using GLUE

E.4. Runoff coefficient results

E.4.1. Run 1

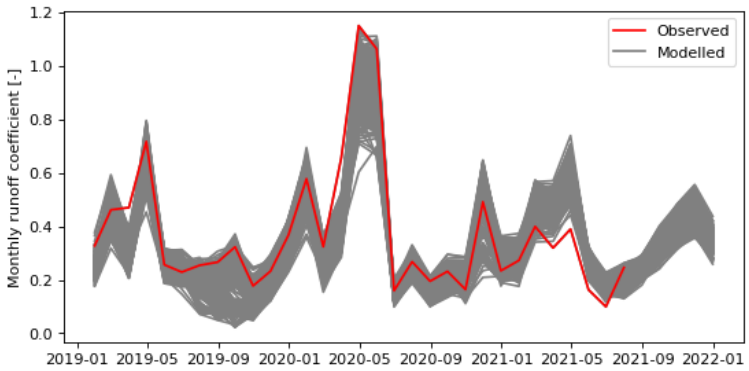


Figure E.18: Run 1: Monthly runoff coefficient from model calibrated on Q and E of 300 best calibration runs

E.4.2. Run 2

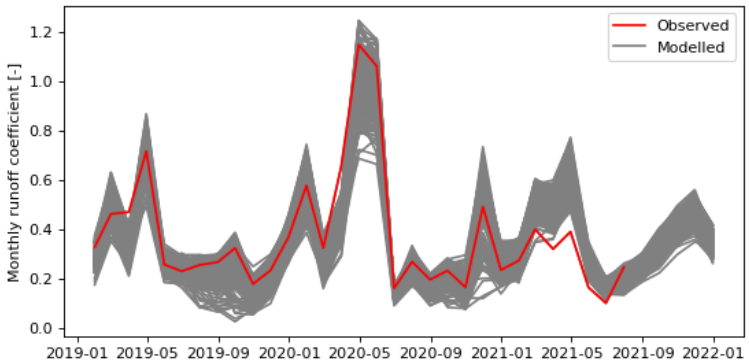


Figure E.19: Run 2: Monthly runoff coefficient from model calibrated on Q and E of 300 best calibration runs

E.4.3. Run 3

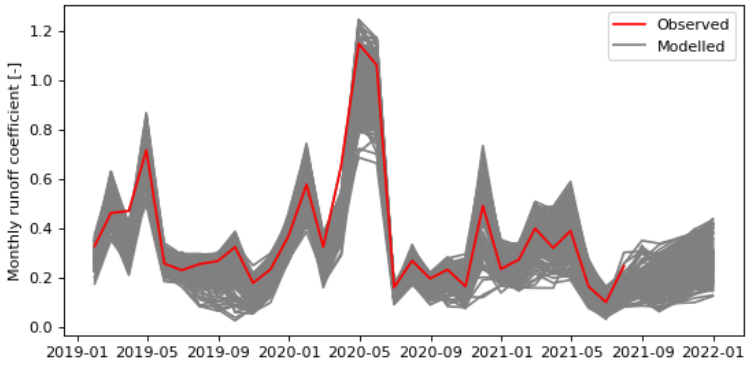


Figure E.20: Run 3: Monthly runoff coefficient from model calibrated on Q and E of 300 best calibration runs

F

Wflow Flextopo Model

”Wflow is Deltares’ solution for modelling hydrological processes, allowing users to account for precipitation, interception, snow accumulation and melt, evapotranspiration, soil water, surface water and groundwater recharge in a fully distributed environment.” (Deltares, 2022). There are different concepts of the Wflow model and in this research Wflow flextopo is used. This model has been selected because of the ease with which changes can be made to the processes, and the possibility to differentiate between different types of landscapes. The underlaying modelling methodology is a distributed Hydrologiska Byråns Vattenbalansavdelning (HBV) model. The HBV model is originally developed by Bergström (1973) and later updated from a lumped to a distributed HBV by Lindström (1997) and is to date one of the most used hydrological models in the world. The rainfall runoff model uses a bucket approach to imitate the different hydrological processes between rainfall and runoff in the river. This bucket approach is visible in the schematic overview of the Wflow flextopo model in figure F.1. The flextopo aspect of the model is explained in section 3.2.

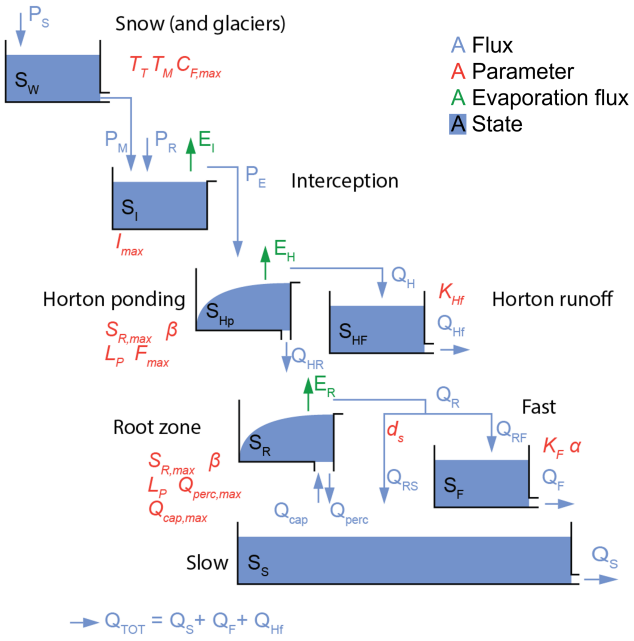


Figure F.1: Wflow flextopo schematic representation (Bouaziz, 2022)

Name	Description	Unit	Type
cfmax	degree-day factor ($C_{F,max}$)	mm °C^{-1} Δt $^{-1}$	parameter
tt	threshold temperature for snowfall (T_T)	°C	parameter
tti	threshold temperature interval length	°C	parameter
ttm	threshold temperature for snowmelt (T_M)	°C	parameter
whc	water holding capacity as fraction of current snow pack	-	parameter
cfr	refreezing efficiency constant in refreezing of freewater in snow	-	parameter
ecorr	evaporation correction	-	parameter
pcorr	correction factor for precipitation	-	parameter
rfcf	correction factor for rainfall	-	parameter
sfcf	correction factor for snowfall	-	parameter
imax	maximum interception storage (I_{max})	mm	parameter
shmax	maximum horton ponding storage capacity ($S_{H,max}$)	mm	parameter
srmax	maximum root zone storage capacity ($S_{R,max}$)	mm	parameter
beta	exponent in soil runoff generation equation (β)	-	parameter
lp	fraction of root zone capacity below which actual evaporation is equal to potential evaporation (L_P)	-	parameter
ks	recession constant slow groundwater storage (K_S)	Δt^{-1}	parameter
kf	recession constant fast storage (K_F)	Δt^{-1}	parameter
khf	recession constant horton runoff storage (K_{Hf})	Δt^{-1}	parameter
alfa	measure of non-linearity of upper reservoir (α)	-	parameter
perc	maximum percolation flux from root zone to slow storage ($Q_{perc,max}$)	mm Δt^{-1}	parameter
cap	maximum capillary rise from slow storage to root zone ($Q_{cap,max}$)	mm Δt^{-1}	parameter
ds	splitter parameter determining fraction of root zone outflow to slow storage (d_s)	-	parameter
fdec	exponent for the decline of infiltration capacity (F_{dec})	-	parameter
fmax	maximum infiltration capacity from horton ponding (F_{max})	mm Δt^{-1}	parameter
hrufrac	fraction of class within cell ($F_{hrufrac}$)	-	parameter
precipitation	precipitation (P)	mm Δt $^{-1}$	forcing
temperature	air temperature (T_a)	°C	forcing
precipcorr	corrected precipitation	mm Δt^{-1}	flux
epotcorr	corrected potential evaporation	mm Δt^{-1}	-
potential_evaporation	potential evapotranspiration (E_{PET})	mm Δt^{-1}	forcing
potsoilevap	potential soil evaporation (E_P)	mm Δt^{-1}	-
snow	snow storage (S_{Sm})	mm	state
snowwater	available free water in snow (S_{Sw})	mm	state

interceptionstorage	interception storage (S_I)	mm	state
hortonpondingstorage	horton ponding storage (S_{HP})	mm	state
hortonrunoffstorage	horton runoff storage (S_F)	mm	state
rootzonestorage	root zone storage (S_R)	mm	state
faststorage	fast storage (S_F)	mm	state
slowstorage	slow storage (S_S)	mm	state
soilevap	soil evaporation	mm Δt^{-1}	flux
intevap	evaporation from interception storage (E_I)	mm Δt^{-1}	flux
hortonevap	evaporation from horton ponding storage (E_H)	mm Δt^{-1}	flux
rootevap	evaporation from root zone storage (E_R)	mm Δt^{-1}	flux
actevap	actual evapotranspiration (intevap + hortonevap + rootevap) (E_A)	mm Δt^{-1}	flux
precipeffective	Effective precipitation (P_E)	mm Δt^{-1}	flux
rainfallplusmelt	snow melt + precipitation as rainfall (P_M + P_R)	mm Δt^{-1}	flux
snowmelt	snowmelt (P_M)	mm Δt^{-1}	flux
snowfall	snowfall (P_S)	mm Δt^{-1}	flux
facc	modeled accumulated frost	°C Δt	-
qhortonpond	Flux from the hortonian ponding storage to the hortonian runoff storage (Q_H)	mm Δt^{-1}	flux
qhortonrootzone	Flux from the hortonian ponding storage to the root zone storage (Q_{HR})	mm Δt^{-1}	flux
qhortonrun	Flux from the hortonian runoff storage (Q_{Hf})	mm Δt^{-1}	flux
qrootzone	Flux from the root zone storage (Q_R)	mm Δt^{-1}	flux
qrootzonefast	Pref. recharge to fast storage (Q_{RF})	mm Δt^{-1}	flux
qcapillary	Capillary flux from the slow to the root-zone storage (Q_{cap})	mm Δt^{-1}	flux
qfast	runoff from fast storage (Q_F)	mm Δt^{-1}	flux
qslow	runoff from slow storage (Q_S)	mm Δt^{-1}	flux
runoff	total specific runoff per cell (qslow + qfast_tot) (Q)	mm Δt^{-1}	flux

F.1. Hydrological processes

The hydrologic cycle on earth consist of different processes, which occur simultaneously and continuously as long as the conditions for the process are satisfied. The hydrological cycle is not fully represented in this model, as condensation is out of scope for a catchment hydrological model.

Snow (and glaciers)

Glaciers are left out of the model, as these do not occur in the Netherlands. Snow on the other hand does occur if the air temperature (T_a) is below the threshold for snowfall (tt). The snowfall enters the storage and is added to the refrozen snow water. Water leaves the storage through snow melt and if the liquid water concentration (whc) is exceeded.

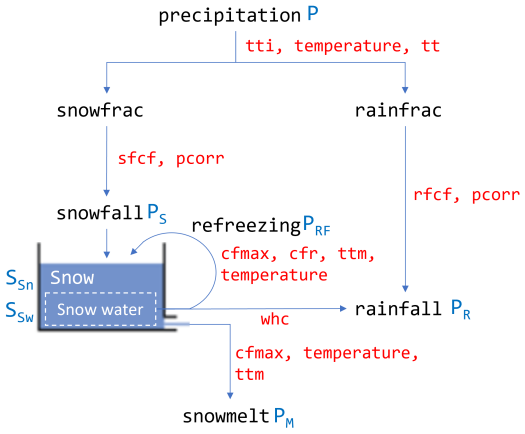


Figure F.2: Snow process, own figure

$$dS_Sn/dt = P_S + P_{Fr} - P_M - P_R \tag{F.1}$$

$$P_S = snowfrac * sfcf * P * pcorr \tag{F.2}$$

$$snowfrac = 1.0 - min((T_a - (tt - tti/2.0))/tti, 1) \tag{F.3}$$

$$P_{Fr} = C_{F,max} * C_{Fr}(ttm - T_a); T_a < ttm \tag{F.4}$$

$$P_M = C_{F,max}(T_a - ttm); T_a > ttm \tag{F.5}$$

$$P_R = max(S_{Sn} - S_{Sw} * whc, 0.0) + rainfrac * rfcf * P * pcorr \tag{F.6}$$

$$rainfrac = 1.0 - snowfrac \tag{F.7}$$

Interception

If the air temperature (T_a) is above the threshold for snowfall (tt), rainfall occurs. This is added to the snow melt, and these together are the inflows of the interception storage. Water leaves the storage through evaporation and outflow into Horton ponding.

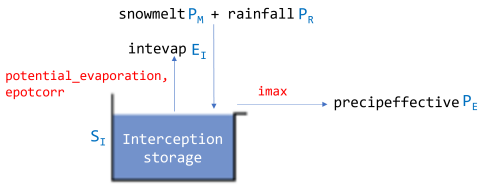


Figure F.3: Interception process, own figure

$$dS_I/dt = (P_R + P_M) - E_I - P_E \tag{F.8}$$

$$P_E = max(0, S_I - I_{max}) \tag{F.9}$$

$$E_I = minE_{PET}, S_I) \tag{F.10}$$

Horton ponding

When the water leaves the inception storage, it falls upon the ground. Here it either stays for a while (Horton ponding) or it runs off (Horton runoff). This depends on how full the storage capacity (S_H) already is. First, Horton ponding is described.

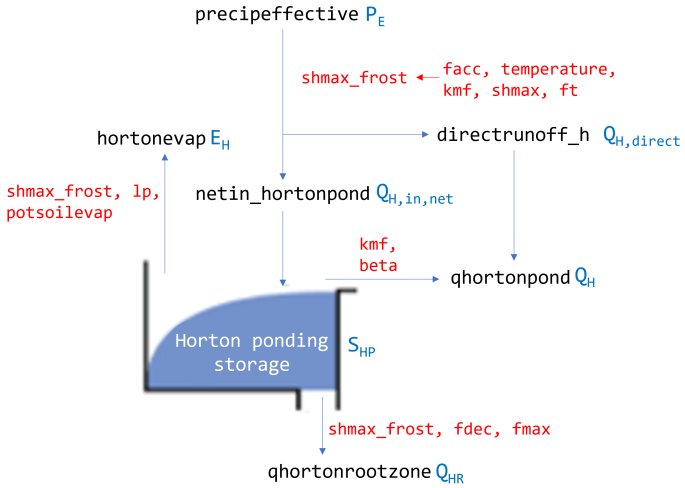


Figure F.4: Horton ponding process, own figure

$$dS_{HP}/dt = P_E - E_H - Q_H - Q_{HR} \tag{F.11}$$

$$E_H = \begin{cases} min(S_{HP}, E_P) & \text{if } S_{HP} > S_{H,max} * L_P \\ min(E_P * (S_{HP}/(S_{H,max} * L_P)), 0.0) & \text{if } S_{HP} <= S_{H,max} * L_P \end{cases} \tag{F.12}$$

$$E_P = max(0.0, epotcorr * E_{PET} - E_i) \tag{F.13}$$

$$Q_H = Q_{H,in,net} * (1 - (1 - min(\overline{S_{Hp}}, 1)^{\beta}) + Q_{H,direct} \tag{F.14}$$

$$Q_{H,in,net} = P_E - Q_{H,direct} \tag{F.15}$$

$$\overline{S_{Hp}} = S_{Hp}/S_{H,max} \tag{F.16}$$

$$Q_{H,direct} = max(S_{HP} + P_E - S_{H,max}, 0) \tag{F.17}$$

$$Q_{HR} = \begin{cases} min(F_{max} * exp(-F_{dec}(1 - min(\overline{S_{Hp}}, 1))), S_{HP}) & \text{if } min(\overline{S_{Hp}}, 1) > 0 \\ 0 & \text{if } min(\overline{S_{Hp}}, 1) <= 0 \end{cases} \tag{F.18}$$

Horton runoff

All the water that cannot immediately be stored in Horton ponding goes into Horton runoff (Q_H) and enters Horton runoff storage. Water leaves the storage with a certain recession constant (K_{HF}) and flows into the stream.

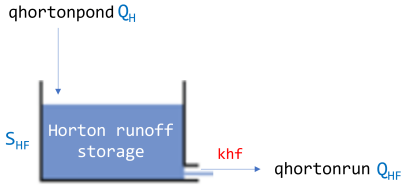


Figure F.5: Horton runoff process, own figure

$$dS_{HF}/dt = Q_H - Q_{HF} \tag{F.19}$$

$$Q_{HF} = min(K_{HF} * S_{HF}, S_{HF}) \tag{F.20}$$

Root zone storage

The outgoing water from the Horton ponding storage infiltrates and enters the root zone storage. Another source of the root zone storage is capillary rise from the slow storage. Water leaves the storage through evaporation, percolation into the slow storage and outflow which can end up in both the fast and slow storage.

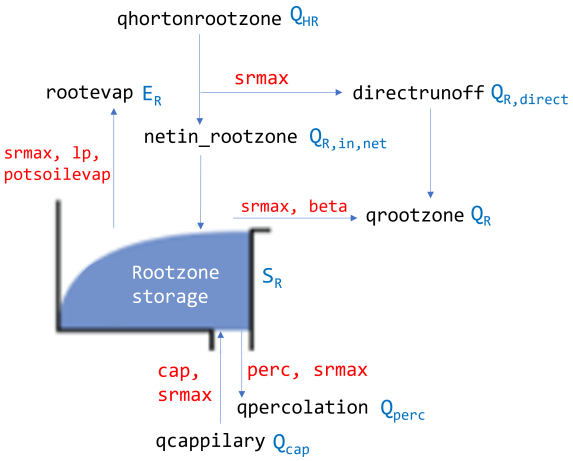


Figure F.6: Root zone process, own figure

$$dS_R/dt = Q_{HR} - E_R - Q_R - Q_{perc} + Q_{cap} \tag{F.21}$$

$$E_R = \begin{cases} min(S_R, E_P) & \text{if } S_R > S_{R,max} * L_P \\ min(E_P * (S_R/(S_{R,max} * L_P)), 0) & \text{if } S_R <= S_{R,max} * L_P \end{cases} \tag{F.22}$$

$$E_P = max(0.0, epotcorr * E_{PET} - E_i - E_H) \tag{F.23}$$

$$Q_R = Q_{R,in,net} * (1 - (1 - S_{R,max})^{\beta}) + Q_{R,direct} \tag{F.24}$$

$$Q_{R,in,net} = Q_{HR} - Q_{R,direct} \tag{F.25}$$

$$Q_{R,direct} = max((S_R + Q_{HR} - S_{R,max}), 0.0) \tag{F.26}$$

$$Q_{perc} = Q_{perc,max} * S_{R,max} \tag{F.27}$$

$$Q_{cap} = Q_{cap,max} * (1 - S_{R,max}) \tag{F.28}$$

Fast storage and runoff

The fast storage is the fast aspect of the groundwater storage in the catchment, it collects water from recharge through the root zone. Water leaves the storage trough outflow into the stream.

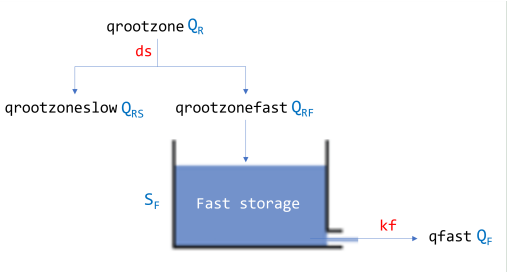


Figure F.7: Fast storage and runoff process, own figure

$$dS_F/dt = Q_{RF} - Q_F \tag{F.29}$$

$$Q_{RF} = Q_R * (1 - ds) \tag{F.30}$$

$$Q_F = min(S_F, K_F * S_F^{\alpha}) \tag{F.31}$$

Slow storage and runoff

The slow storage is seen as the slow groundwater storage in the catchment, it collects water from percolation and recharge through the root zone. Water leaves the storage by capillary rise to the root zone and linear outflow into the stream. Percolation and capillary rise are described under root zone storage.

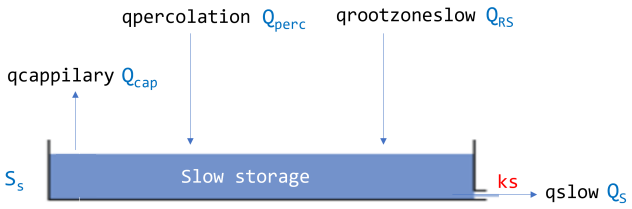


Figure F.8: Slow storage and runoff process, own figure

$$dS_S/dt = Q_{RS} + Q_{perc} - Q_S - Q_{cap} \tag{F.32}$$

$$Q_S = min(S_S, S_S * K_S) \tag{F.33}$$

Interception capacity

The distributed nature of the model, opens up the opportunity to differentiate parameter values throughout the catchment as is done with the landscape classification. Nonetheless, this differentiation based on HAND and slope does not fully capture the hydrological landscape of the Geul. As there is a land use bias under laying the landscape classification. The catchment in Gharari et al. (2014) had a land use differentiation as follows: ”Hill slopes are generally characterised by forest, while plateaus and valley bottoms are mostly used as crop land and pastures, respectively.”. This meant that the land use distinction followed the landscape distinction, this is not the case in the Geul catchment where land use is more heterogeneously distributed. To be able to capture these land-uses. The interception capacity of a location is not linked to the landscape classification of that location but to land use. The land use of the Geul catchment can be seen in figure F.9.

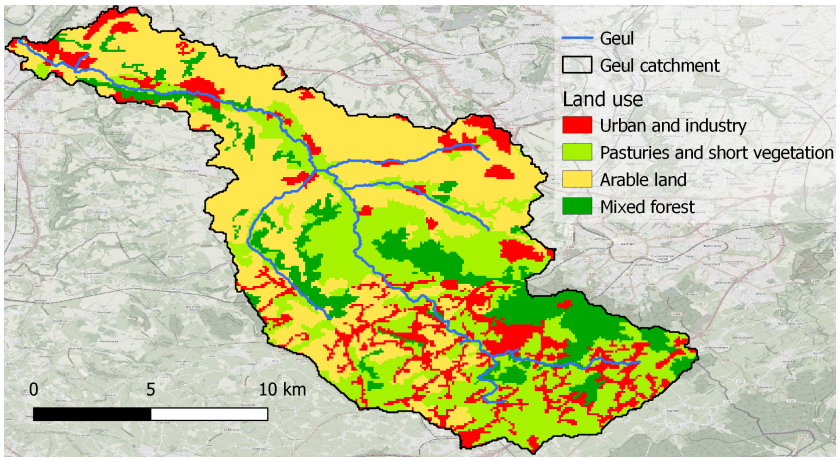


Figure F.9: Land use in the Geul catchment (Copernicus, 2018)

F.2. Calibration runtime Wflow Flextopo

The Wflow Flextopo model is not suitable for meaningful calibration. There are three main factors that contributed to this. Firstly, the run time of the model is on average 3.5 minutes per run. On the available equipment 4 runs can run in parallel, decreasing the average run time to 2.5 minutes. Secondly, Monte Carlo sampling is a random sampling method, this means that each combination of parameter values is selected at random which can lead to oversampling of some area's and under sampling of others. Therefore, it is important to have a lot of runs, more generally, for n samples per parameter and P parameters the number of required runs is calculated with equation F.34:

$$\text{number of runs} = n^P \tag{F.34}$$

This limitation of Monte Carlo sampling ties in with the third factor namely the large number of parameters in the Wflow model. For 15 parameters with only 10 randomly selected values each, 10^{15} runs are necessary, with an lower limit of at least 100.000 run (consultation with expert M. Hrachowitz, 2023). However there are methods to decrease these three problems namely:

1. Decrease runtime by decreasing the space and time resolution of the model.
2. Change to a sampling method, which requires less runs because of more effective parameter set selection.
3. Decrease parameters to be calibrated.

Each of these solutions has been explored. The decrease of the spatial resolution from 1 km by 1 km to both 4 km by 4 km as 5 km by 5 km is implemented into the model. However, this only gave a run time decrease of 30 seconds per run on average (without running in parallel). The decrease of spatial resolution from an hourly time step to a daily time step gave a decrease in run time of 3 minutes per run on average (without running in parallel). With this decrease, the total run time, while running in parallel, would be almost 25 days.

Changing the sampling method from Monte Carlo towards a stratified sampling method, which divides the sampling domain into regular, even-sized strata and places one sample in each. This is easily done for two parameters, but extending such sampling approaches to simultaneously stratify successively in higher dimensions as well remains a challenge (Jarosz et al., 2019). Because of time constrains, and limited chance of successfully creating a 15 dimensional stratified parameter set, this method is not explored further.

The number of parameter to calibrate is decreased by taking the average values for all parameters concerning the snow process in the model, as these value do not differ much between catchments.

Table F.1: Parameters which values are not calibrated, but average interval value is used

Name	Description	Unit	Average
tt	threshold temperature for snowfall	°C	1.0
tti	threshold temperature interval length	°C	3.5
ttm	threshold temperature for snowmelt	°C	-2.05
whc	water holding capacity as fraction of current snow pack	-	0.1
cfmax	degree day factor	mm °C $^{-1}$ Δt $^{-1}$	0.22
cfr	refreezing efficiency constant in refreezing of freewater in snow	-	0.005

Table F.2: Calibration parameters which are used extra in Wflow Flextopo

Name	Description	Unit	Interval	Reference
tt	threshold temperature for snowfall	°C	[–0.4, 2.4]	Jennings et al. (2018)
tti	threshold temperature interval length	°C	[0, 7]	Wawrzyniak et al. (2017)
ttm	threshold temperature for snowmelt	°C	[-4.6, 0.5]	Senese et al. (2014)
whc	water holding capacity as fraction of current snow pack	-	[0, 0.2]	Beck et al. (2016), Seibert (1997)
cfmax	degree day factor	mm $°C^{-1}$ Δt^{-1}	[0.02, 0.42]	Beck et al. (2016), Seibert (1997)
cfr	refreezing efficiency constant in refreezing of freewater in snow	-	[0,0.1]	Beck et al. (2016), Seibert (1997)
fdec	exponent for the decline of infiltration capacity	-	[0.01 , 0.5]	Euser (2017)
fmax	maximum infiltration capacity from horton ponding	mm Δt^{-1}	[2.5, 37.5]	Klein (2022)
khf	recession constant horton runoff storage	Δt^{-1}	[0.036, 0.6]	Beck et al. (2016), Seibert (1997)
shmax	maximum horton ponding storage capacity	mm	[0, 50]	Euser (2017)

F.3. Implications of lumped vs distributed model

The swith from a distributed to a lumped hydrological model can have multiple implications. As distributed models are able to provide a more detailed and spatially explicit representing of the hydrological processes in the basin, they are more suited for catchments with large regional differences in for instance precipitation and evaporation, like the Geul has, as described in section 2.1.1.

If the computational power or time had been available to use the distributed Wflow model, the impact of calibration on evaporation might have been greater than in the current model. As the re-gridded evaporation data from Planet (2022) is available on 100m by 100m grid, it would have been possible to calibrate each gridcel of the model (1km by 1km) with the average actual evaporation of that section. Which

Furthermore, the precipitation data of the region would not have had to be averaged out over the catchment, but could have stayed distributed, giving a more accurate representation of the precipitation over the catchment.