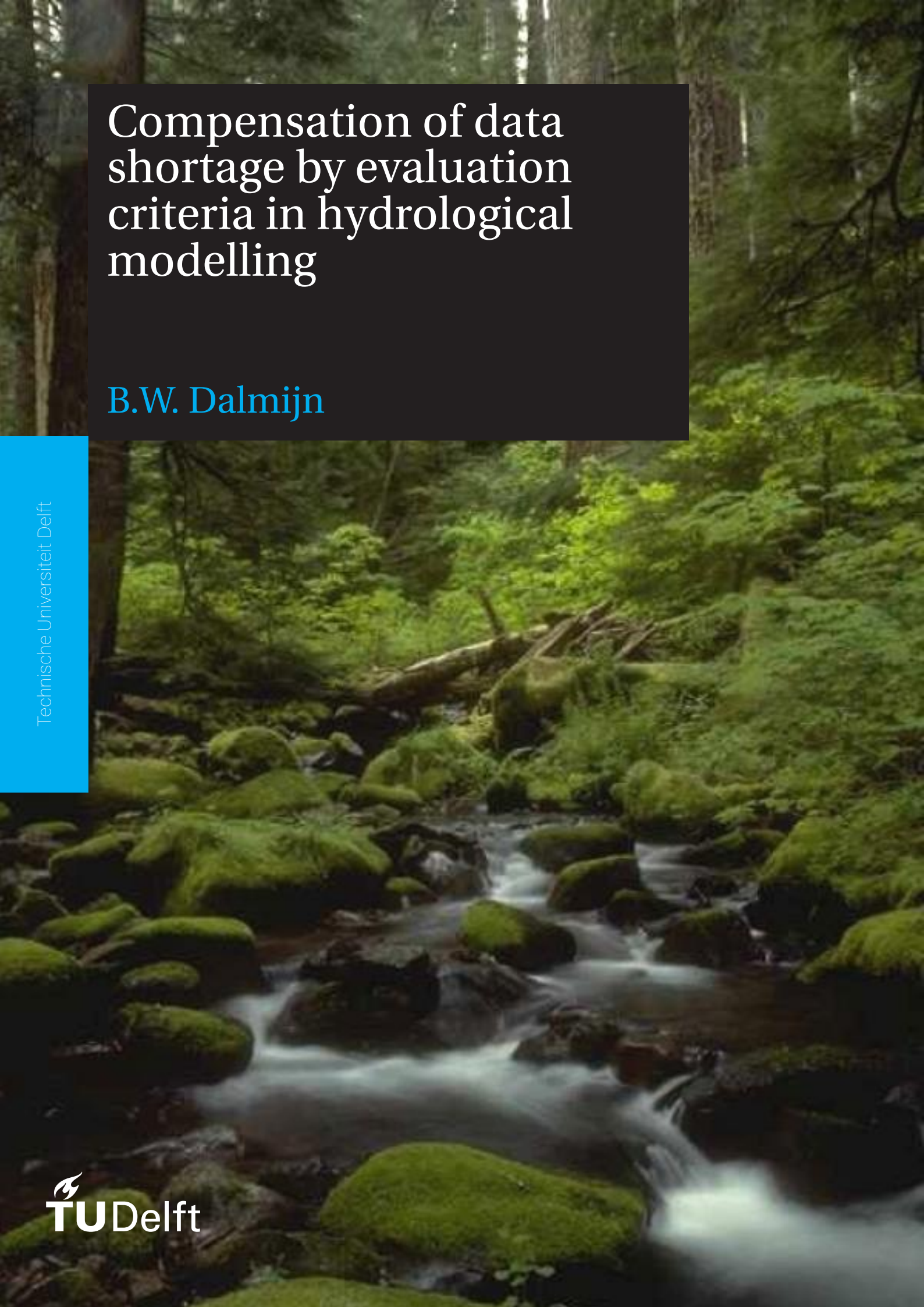
Compensation of data shortage by evaluation criteria in hydrological modelling

B.W. Dalmijn

Technische Universiteit Delft



TU Delft

Compensation of data shortage by evaluation criteria in hydrological modelling

by

B.W. Dalmijn

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on October 10th, 2019

Student number:	4247027	
Project duration:	March, 2018 – April, 2019	
Thesis committee:	Dr. ir. M. Hrachowitz,	TU Delft, supervisor
	Prof. dr. ir. M. Bakker,	TU Delft
	Prof. dr. ir. J.P. Van der Hoek,	TU Delft

TU Delft

Abstract

This thesis proposes an approach for tackling the problem of data-shortage in hydrological modelling. A hydrological model, in the context of this thesis, translates meteorological data to stream-flow of a river, for a specific catchment. A model incorporates parameters that describe the dynamics of a chosen model, in this case a FLEX (Flux-Exchange) based model. All parameter together form a set. These parameters are unknown and therefore need to be determined, which is done via a calibration process. In the calibration process, the modelled flow will be evaluated against the observed stream-flow over a period of time for a certain hydrological signature. A hydrological signature is defined as the quantification of specific information concerning the rainfall-runoff dynamics of a catchment, e.g. the mean stream-flow. The evaluation for a certain hydrological signature is known as an evaluation criterion. Literature deems the use of one or two criteria in combination with multiple years of data enough to ensure a good performance from the parameter-sets that exit the calibration process. As this amount of data is not available everywhere, this thesis proposes to still ensure good model performance for less observed data. For this purpose, a selection was made of evaluation criteria to be applied in the calibration process. These criteria were selected in such a way that various different characteristics of the hydrological response were covered. A benchmark was created for comparison purposes in which 10 years of data was used during calibration and one evaluation criterion. After the benchmark, the observed data was shortened and evaluation criteria were added. The results showed that 6 months worth of data in combination with all criteria would create benchmark-level performances without extra information. The same level of performance can be reached with 3 months of data but this would require extra information. This information would be e.g. during which period of time observed data should be collected.

Preface

This is the report of my master thesis that concludes my master Water Management followed at the Technical University of Delft.

I would like to thank my supervisor and also the chairman of my graduation-committee M. Hrachowitz for giving me the opportunity and time to graduate under him. I would also like to thank him for exploring different types of graduation-subjects with me and eventually handing me the subject this thesis-report is about. Besides the chairman, I would like to thank my remaining two committee-members, M. Bakker and J.P. van der Hoek, for their time, their willingness and enthusiasm to take place in my committee, and, most importantly, their critique and suggestions concerning my thesis.

From my fellow students I thank M.S. van Esch and J.G.V. van Ramshorst for their critical remarks and advice concerning the code for the calibration. Last I would like to thank my friends and family for their support and advice, while writing my thesis.

B.W. Dalmijn
October, 2019
Delft

Contents

Abstract **i**

Preface **iii**

List of Symbols **vii**

1 Introduction **1**
1.1 Relevance . . . 1
1.2 Background Information . . . 3
1.3 Research Question(s) . . . 5
1.4 Proposed Approach . . . 5

2 Model and input **7**
2.1 The catchment . . . 8
2.2 The model . . . 10
2.2.1 Constructing the model . . . 10
2.2.2 Data . . . 17

3 Method **19**
3.1 Evaluation Criteria . . . 20
3.2 Calibration . . . 24
3.2.1 First run & Benchmark . . . 24
3.2.2 Moving window . . . 27
3.3 Validation . . . 28

4 Results **29**
4.1 Calibration . . . 29
4.1.1 First observations . . . 29
4.1.2 Further observations . . . 32
4.1.3 Prediction . . . 37
4.2 Validation . . . 39

5 Discussion **43**
5.1 Criteria thresholds . . . 43
5.2 Runoff coefficient criterion . . . 43
5.3 Identification . . . 45
5.4 Parameter sensitivity . . . 46
5.5 Starting month . . . 48

6 Conslusion & Recommendations **49**
6.1 Conclusion 49
6.2 Recommendations 50

References **51**

A Flex_{nd} model **57**
A.1 Constitutive Equations 57
A.2 Lag-function 58
A.3 Matlab code of the model 59

B Evalutation criteria & calibration **63**
B.1 Criteria 63
B.1.1 Matlab code 63
B.1.2 Figure(s) 66
B.2 Calibration 67
B.2.1 Matlab Code 67
B.2.2 Figures 71

C Results supplement **73**

D Discussion supplement **79**

List of Symbols

Symbol	Description	Units
AC	Auto correlation function	–
β	Shape factor soil moisture function	–
C_e	Limit potential evaporation	–
C_o	Melting factor	$\frac{mm}{°C \times day}$
D	Splitting factor for run-off and preferential percolation	–
ε	Evaluation criterion value	–
E_i	Evaporation from interception	mm/d
E_{pot}	Potential evaporation	mm/d
E_t	Transpiration	mm/d
$FlowDur$	Flow Duration Curve	–
I_{max}	Interception capacity	mm
K_f	Recession coefficient fast reservoir	d^{-1}
K_s	Recession coefficient slow reservoir	d^{-1}
$LogNSE$	Logarithmic Nash-Sutcliffe	–
NSE	Nash-Sutcliffe efficiency	–
P	Precipitation	mm/d
$PeakDis$	Peak Distribution	–
P_{eff}	Effective precipitation	mm/d
P_s	Remaining precipitation after interception	mm/d
Q_f	Flow from fast reservoir	mm/d
Q_m	Modelled stream-flow	mm/d
Q_{obs}	Observed stream-flow	mm/d
Q_s	Flow from slow reservoir	mm/d
Q_{tot}	Outcome of the model	mm/d
R_f	Fast run-off	mm/d
R_i	Infiltration / percolation	mm/d
$R_{i,max}$	Maximum infiltration per unit time	mm/d
RLD	Rising limb density	–
RO	Runoff coefficient	–
R_p	Preferential percolation	mm/d
$R_{Q_m,d}$	Correlation coefficient of sample Q_m and a delayed variant of Q_m	–

R_u	Overland run-off	mm/d
S_f	Fast reservoir	mm
S_i	Interception reservoir	mm
S_n	Snow reservoir	mm
S_s	Slow reservoir	mm
S_u	Unsaturated reservoir	mm
$S_{u,m}$		mm
$S_{u,max}$	Unsaturated Capacity	mm
S_x	Standard deviation of sample x	–
T_A	Actual temperature	°C
T_{lag}	Lag-time for output transfer function	$days$
T_{thresh}	Threshold temperature for snow melt	°C

1

Introduction

1.1. Relevance

L'vovich and White (1990) have shown that since the industrial revolution, what could be considered the starting point of human intervention on an unprecedented scale on the environment as mentioned by Savenije et al. (2014), the distribution of fresh water around the globe changed as a result of human action. Humanity made efforts to manage water for its own cause and altered the urban and rural landscape to influence the flow and storage of water. The hydrological response, in the catchments where these actions were conducted, changed as a result.

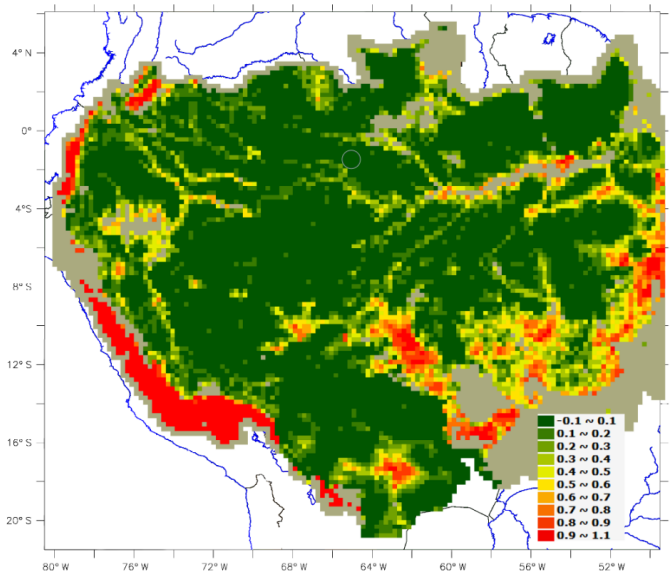


Figure 1.1: The percentage of deforestation is shown over the entire Amazone basin. The pixels in this figure represent an area of 25 by 25 km^2.
Source: Guimberteau et al. (2017)

Some activities that had their resonance in the hydrological response of a catchment were, according to Savenije et al. (2014): **a)** direct diversion of water flows, including inter-basin transfers for water supplies to cities, industries and agriculture, **b)** transformation of the stream network, for example through the construction of dams and reservoirs or the canalisation of rivers, **c)** changing drainage basin characteristics, for example through deforestation, urbanisation, drainage of wetlands and agricultural practices and **d)** activities altering the regional or global climate (Savenije et al., 2014, p. 320-321).

An example of deforestation is presented in figure 1.1 where the percentage of deforestation is shown over the Amazon basin. Guimberteau et al. (2017) concluded an increase in both evaporation and runoff in areas which would not be affected by deforestation as a result of climate change, 5.0 and 14% respectively. The southeast of the Amazon receives 10% less precipitation at the end of the dry season but Guimberteau et al. (2017) concluded a smaller drop in evaporation. This lead to a river stream-flow reduced by 31%. The effects of deforestation, ranging from 7 to 34%, in the south-east region were examined. The results showed a much greater decrease in evaporation as a consequence of less vegetation. This would entail an increase in runoff that could counter balance the decrease in stream-flow, in case of extreme deforestation. Deforestation has a significant influences on the hydrological response of an area but still is just one result of many human activities. Therefore it is not difficult to grasp the influence of mankind on the altering hydrological response.

Blair and Buytaert (2016) state that the human interventions/activities have been largely made to accommodate the requirements of a population that has grown from 0.9 billion, at the start of the industrial revolution, to 7 billion. As a result of trying to accommodate for this tremendous increase in population, human intervention has required such control that in many location water flows as mankind dictates. These far-reaching anthropogenic activities are leading to a coupling of the human and the hydrological systems (Wagener et al., 2010). This entails that the decisions made by humans that impact the hydrological system also impacts the human society. A logical deduction from this information is: mankind needs an understanding of the hydrological system in order to understand the gravity and impact of its decisions regarding these hydrological systems.

A reasonable way to acquire this understanding is trough hydrological modelling, by breaking down a hydrological system into bite-sized pieces. These pieces roughly represent the dominant processes of a hydrological system. By representing the dominant processes in a simple manner, insight can be obtained rather quickly concerning how a hydrological system, from now on referred to as a catchment, responses to changes in terms of e.g. climate and landscape. Thereby is a model able to present insight in the hydrological dynamics of a catchment. How useful this insight is, depends on the type of model that is used and how it is used.

Thereby comes the predictive capabilities of a hydrological model. Climate scenarios could be fed to the model in order to get an idea what the impact of climate change would be on the hydrological response.

1.2. Background Information

Hydrological models use parameters made for describing the different processes in the chosen model. These parameters together form a set which is optimized during a calibration process with the desire to obtain the "best" performing model. The calibration could take the form of simply evaluating the modelled outcome, for a variety of parameter-sets, against the observed data. Calibration therefore does not necessarily mean tweaking in order to acquire the absolute best parameter-set but the definition leans to filtering out the "bad" parameter-sets for the acquisition of the best. The evaluation of modelled versus observed will result in a single value; this value quantifies how similar the modelled outcome is to the observed data. This similarity between modelled and observed concerns a hydrological signature. A hydrological signature is defined as the quantification of specific information concerning the rainfall-runoff dynamics of a catchment (Westerberg and McMillan, 2015). A simple example would be the mean stream-flow. Hydrological signatures provide insight into the dominant hydrological processes of a catchment. The evaluation of modelled flow versus observed data for a specific hydrological signature is called an evaluation criterion. Parameter-sets producing an outcome that passed the pre-determined threshold of a criterion are deemed "good" performing parameter-sets, those that do not are deemed "bad".

Two essential terms concerning hydrological modelling that must be mentioned are: performance and consistency. Performance is described as the ability of a model to mimic a specific hydrological signature of a certain catchment. Consistency is defined as the ability of a model to adequately reproduce hydrological signatures, while using the same parameter-set. A deduction from this is that a model can suffer from consistency problems when only a single evaluation criterion is used during the calibration process for the determination of the best performing parameter-set. To improve both performance and consistency of a hydrological model, research has been done and published. For instance, a framework was designed to assess the realism of model structures (Euser et al., 2013). This framework tests for both performance and consistency using a principal component analysis on a range of evaluation criteria. The test included eight hydrological signatures and eleven model structures, i.e. multiple models that are each built differently. The results showed that some structures could have the same performance for some evaluation criteria but could wildly differ in consistency.

It was found that implementing expert knowledge helped to increase the model's consistency (Hrachowitz et al., 2014). This expert knowledge was implemented into the models in the form of prior-constraints. Prior-constraints can take the form of parameter constraints, which are meant to ensure that the parameter combinations lie within the realm of what is considered realistic. Another form of prior-contraints are process-constraints, which have the goal of ensuring that e.g. individual fluxes must remain within an expected interval. These constraints were implanted to counter-balance the increasing complexity of the model and ensure the model behaves with respect to the modeller's perception of the hydrological system. More complex models with higher quantities of incorporated expert-knowledge showed better consistency than simple well-calibrated models. In Gharari et al. (2014) a semi-distributed model was used. This

entails that the model that was used, consisted of three parallel model structures, each representing a hydrological response unit and together producing one single output, while sharing one component. These hydrological response units represented the wetland, hill-slope and the plateau. The increased complexity of the model, as a result of the parallel model structures, was conditioned by expert-knowledge driven parameter- and process-constraints. The outcome showed that an increasing model complexity in combination with expert knowledge based constraints improved the performance even in an uncalibrated state.

Another reason for incorporating expert knowledge in a hydrological model is to reduce equifinality of parameters, that could be the result of an increasing model complexity. Equifinality entails that the model is able to mimic the observed data using a variety of parameter-sets with different values. This entails that a variety of different internal model dynamics could mimic the observed data, which could signify a lack of correspondence with reality by the model (Kleissen et al., 1990). In Kelleher et al. (2017) a hierarchical approach is presented to reduce the number of parameters sets by usage of regional, observation-driven and expert knowledge-based constraints. The hierarchy is based on availability and cost of obtaining this information. Out of 10000 parameter-sets only nine remained which met all criteria[1].

Trying to model the stream-flow of a catchment where very limited data is available, is not an easy task. The best approach might be to do a few stream-flow measurements. In Seibert and Beven (2009) a number of monitored catchments were used to test this approach. The starting point was a simple model with no stream-flow data implemented. Hereafter different sub-sets of available data, i.e. individual stream-flow measurements, were implemented to slowly constrain the values of the parameter-sets. A simple method was used to calibrate the model. A weighted ensemble mean of simulations, using the parameter-sets left after the calibration process, was used and outperformed the use of the single best performing parameter-set. While showing that not many measurements were needed to obtain roughly the same information, the significance of the few measurements could differ per day and per catchments. Thus it concludes that an intelligent choice is required for the moment the stream-flow measurements should be sampled[2].

While much research has been done in terms of the model's performance and the best way to gather and implement data and even determine how much data is needed (e.g. Bergström, 1976; Fenicia et al., 2006; Moussa and Chahinian, 2009; Berghuijs et al., 2014; Seibert and McDonnell, 2002; Hrachowitz and Clark, 2017; He et al., 2015; Tian et al., 2016), no further research has been done to determine whether a lack of data could be compensated for. Although it was to a certain extent touched upon in Fenicia et al. (2008), the focus was more on achieving this by improving the model in order to identify the hydrological behaviour instead of adding information to the calibration process.

[1] A note made in this article was about the lack of considerations for the internal behaviour of the model and whether this would even be within the realm of possibilities. Extra evaluation of internal dynamics was recommended.

[2] Along those lines an article was written on when measurements should be taken to be most the informative. The most informative moments for model calibration appeared to be at the falling limp of the stream-flow curve (Wang et al., 2017). The falling limb follows the peak-flow of a rainfall-runoff event.

There are many catchments around the globe where limited data is available, whether it is meteorological or stream-flow. This limitation in data makes it difficult to calibrate a model that is consistent due to the probably limited information sheltered in the limited data. Making predictions in these catchments with limited available data by compensating for this lack of available data would save a lot of effort and costs in collecting data from areas that are not easy to travel to. Razavi and Tolson (2013) states that in general large periods of data for calibration are deemed more robust and reliable for identification of the "best" parameter-set (Perrin et al., 2007). These large periods are mentioned to span multiple years. Usually only one or two evaluation criteria are used in the calibration process when using these large periods of calibration data (e.g. Perrin et al., 2007; Juston et al., 2009)

1.3. Research Question(s)

- To what extend is it possible to shorten the period-length of calibration data by using multiple evaluation criteria, for the model to perform similar to being calibrated over a period of 10 years using one evaluation criterion?
  - Is the addition of multiple evaluation criteria enough to see the model perform similarly, while calibrated over a smaller period, to the same model that has been calibrated over a period of 10 years?

1.4. Proposed Approach

First of all, a model has to be constructed to facilitate the research of this thesis. This model will include a certain set of parameters, which are determined from processes, described in the form of equation, that are mentioned in examined literature and subsequently deemed useful. The model and subsequent calibration steps will be programmed in either the program-language Python or Matlab. The model itself is to be designed for a predetermined catchment from which sufficient daily data (meteorological and stream-flow) is available. After this, a calibration method is chosen and a set of evaluation criteria will be reviewed in order to be applied in the calibration process. After the calibration process, the model will be validated for another period, in regards to stream-flow data, than which was used during the calibration-period. First, a set number of randomly generated parameter-sets will be evaluated against one or two common criteria for a set number of years of data (a benchmark period; 10 years). The purpose of this process is to iron out some of the development flaws of the model and the calibration-process. An additional purpose of this process is to set a benchmark, i.e. the outcomes that passed the pre-determined thresholds of said criteria. This benchmark serves as a comparison-tool for the outcomes of shorter-period calibration. The second step of the calibration-process will consist of shortening the calibration-period combined with the step-wise addition of the chosen evaluation criteria. It is likely that the performance and, even more likely, the consistency of the model will decrease as a result of this, which can be viewed after the validation step. The goal is to counter this reduction in performance and consistency by adding new evaluation criteria to the calibration step. This calibration data will be shortened and more evaluation criteria will be added.

2

Model and input

First of all, at the beginning of this chapter information will be given concerning the catchment for which the stream-flow was modelled. This included geography, meteorology etcetera. Hereafter will be the coverage of the type of model that has been used during this thesis. An overview of the model's structure will be presented. The model that is used as a starting point will thereby be discussed and the changes made to this model will be explained, thereby will the reasons why these changes were made be presented. This create a stepwise description of how the model used in this thesis came into existence. Entangled with both the catchment and the model are the data collected from this specific catchment. An overview will be presented of the data required by the model and the calibration process. Besides an overview, information is given about: the modifications made concerning the data, how these are justified, some background information in form of e.g. from which station the data were taken and the implications from this choice.

2.1. The catchment

The catchment that will be modelled for the purpose of this thesis is the HJ Andrews Experimental forest which lies in the United States of America. The choice of catchment was based upon the available meteorological- and stream-flow data. The HJ Andrews catchment has 21 years of continuous daily data available, which suits the needs of this thesis.

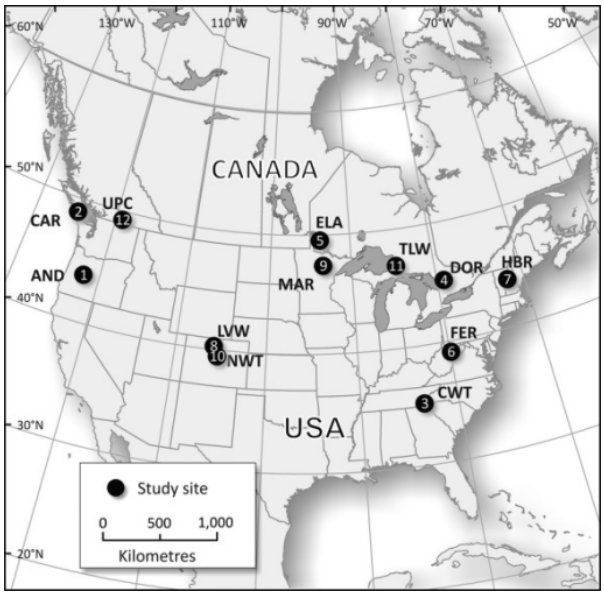


Figure 2.1: Locations of long-term monitoring catchments in the USA (Creed et al., 2014).

The HJ Andrews Experimental forest was established in 1948 by the U.S. forest service. It lies in the Cascade mountain-range in the state of Oregon. The HJ Andrews Experimental forest is marked as 1 in figure 2.1 above.

As stated on the website of the HJ Andrews Experimental forest (USFS, 2018), the forest, since its inception, has witnessed a diverse and impactful research history. Starting from the 1950s, USFS scientists initiated three sets of experimental watersheds[1] to study the effects of logging on the hydrology, sediment yield and nutrient losses. This basically came down to deforestation of areas within the catchment. One of the watersheds was left in its natural state as a reference point (Nijzink et al., 2016). The hydrological response of this catchment has most likely been altered by the deforestation from the 50s and 60s, which entails that the interception capacity of the catchment will have decreased.

In 1980 the Andrews Forest became a charter member of the NSF-sponsored Long-term Ecological Research (LTER) network. The LTER program supported the continuation of long-term studies and environmental monitoring (e.g., climate, streamflow, water quality, population dynamics of sentinel terrestrial and aquatic species, vegetation succession, and disturbance events) (USFS, 2018).[2] This meant the continued availability of daily data,

[1] A watershed is an area that contributes to the larger catchment and functions as a smaller catchment within the larger catchment.

[2] This information is taken directly from the HJ Andrews experimental forest website.

which is a vital necessity for this thesis.

As stated on the website USFS (2018) the HJ Andrews catchment has a surface area of around 6400 hectares. The Landscape is marked by steep hills and deep valleys. Elevations within the catchment can differ from circa 400 metres to around 1600 metres.

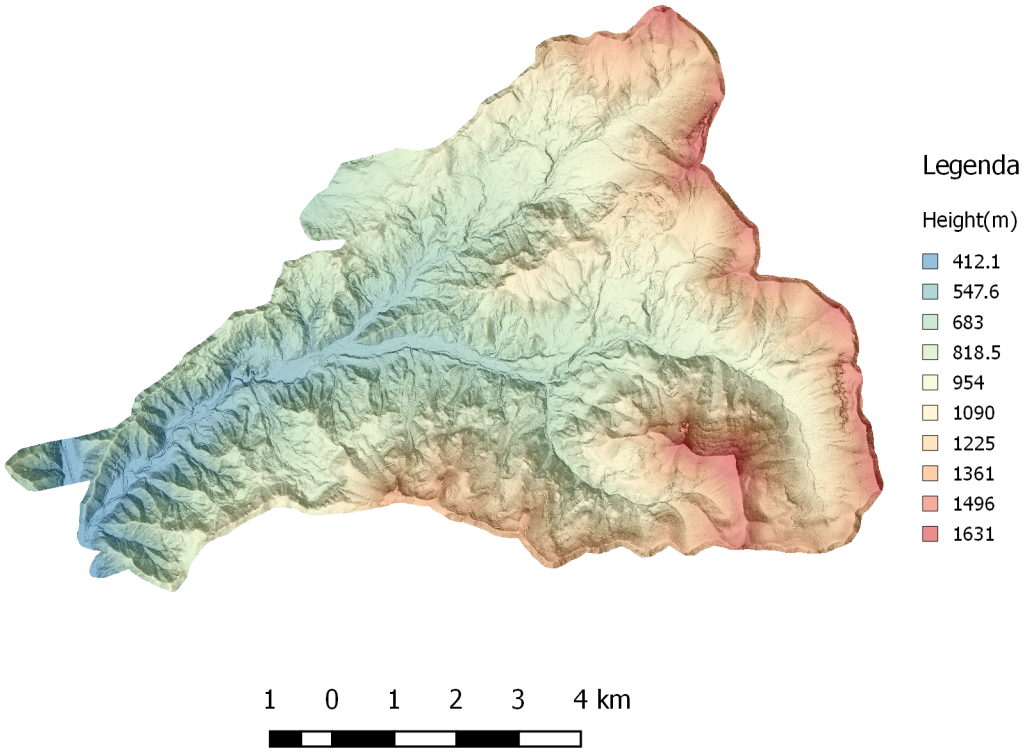


Figure 2.2: Relief of the HJ Andrews catchment.

Figure 2.2 shows that there is not only a big absolute difference between the minimum and the maximum height within the catchment but height is somewhat evenly distributed across the catchment.

In terms of meteorology, the HJ Andrews experimental forest enjoys wet, mild winters and dry, cool summers (USFS, 2018). From the data collected at the central meteorological station, which is one of the benchmark meteorology stations, it shows that the mean temperature lies around a low 1.89 °C in December and a high 17.34 °C in July, for the past 20 years. Most of the precipitation in the HJ Andrews Catchment falls from the month of November to April with a peak daily average of around 12.85 mm in the month of December. The average yearly precipitation measured at this station for the past 20 years lies around 226 cm, which is much higher than the average precipitation taken over the whole of the United States.

2.2. The model

2.2.1. Constructing the model

The type of model used in this thesis is a model based on what is referred to as the FLEX (FLux-EXchange)-Model. The FLEX-model first made its introduction in a scientific paper (Fenicia et al., 2006). In this paper the model is presented as a lumped conceptual model claiming to represent the relevant hydrological processes[3]. This first iteration (referred to as FLEX_b) of the FLEX-model has four reservoirs as its building blocks. These reservoirs are: an interception reservoir (S_i), an unsaturated reservoir (S_u), a fast reservoir (S_f) and a slow reservoir (S_s). The structure of this first iteration is the same as the structure of the mentioned components in figure 2.3. However some of the constitutive equation's differ from those that are used in this thesis. This model has, besides data, an intake of 10 parameters.

The basic principle of this FLEX_b is: precipitation enters the model in the interception reservoir where it fills up to a certain value (I_{max}) and evaporates as long as there is water in the reservoir. Once the amount surpasses I_{max} the surplus continues into two directions, one being a stream that infiltrates into the unsaturated reservoir and one that is an excess flux. From the unsaturated reservoir water can transpire through vegetation and percolate towards the deeper grounds, which in this case is the slow reservoir. The excess flux will be divided in two separate fluxes. The first one chooses the path of preferential percolation into the slow reservoir, the second becomes surface runoff heading for the fast reservoir. Lastly, the slow- and fast reservoir combined produces the output of this model. The state of each reservoir is updated each time-interval (Δt) by simply adding and subtracting the ingoing and outgoing fluxes. In this thesis the time-interval (Δt) has been set at 1 day which is predominately a result of the time-resolution of the input-data (see sub-section 2.2.2).

Parameter	Definition	Units
I_c	Evaporation Coefficient	–
I_{max}	Interception Capacity	mm
S_{fc}	Unsaturated Capacity	mm
L_p	Limit Potential Evaporation	–
β	Shape parameter for runoff	–
D	Runoff partition coefficient	–
P_{max}	Maximum percolation rate	mm/d
N_{lagf}	Lag-time for S_f transfer function	d
N_{lags}	Lag-time for S_s transfer function	d
K_f	Recession coefficient fast reservoir	d

Table 2.1: Parameters of the FLEX_b-model (Fenicia et al., 2006).

In Fenicia et al. (2008), multiple updates and improvements were made to the FLEX-model. The FLEX_b-model was stripped from its interception reservoir for the purpose of testing a few hypotheses of which adding the interception reservoir was one.

[3]This includes: interception processes, evaporation, transpiration, storage capacity of the soil, unsaturated soil drainage, preferential recharge, percolation, runoff, groundwater flow.

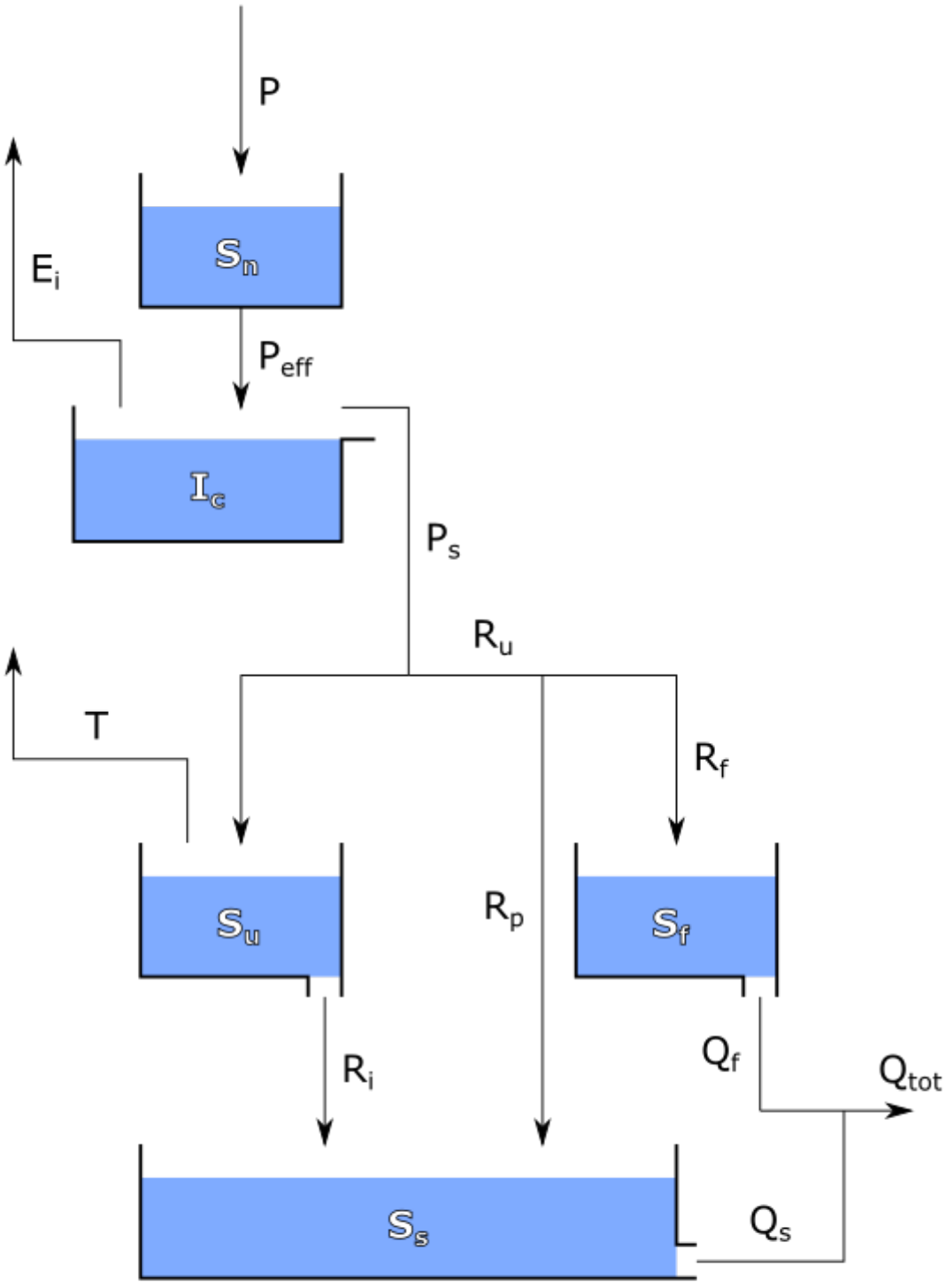


Figure 2.3: Schematic representation of the structure of the $FLEX_{nd}$-model.

By reducing the model and lumping the evaporation and the transpiration together into one term[4], the model's performance worsened, in regards to the used evaluation criteria, even with the proposed modifications, as opposed to the addition of the interception reservoir. Hereafter the parameter I_c was dropped in favour of letting the interception reservoir evaporate at the rate of the potential evaporation[5] during dry periods.

Another attempt at improving the FLEX_b-model was made by taking into account the special heterogeneity of the catchment (Fenicia et al., 2008). Here the catchment was divided in n areas, to be regarded as model units, and a distributed-description[6] of the model components was introduced. When distributing a model component, the parameters corresponding to that component will likely be distributed as well. This would mean that e.g. there would be an I_{max} parameter for every interception reservoir to be described. The consequence is that extra effort has to be made to calibrate the model but also that the problem of equifinality[7] could arise. Equifinality could however be reduced by applying constraints to the model which could stem from expert-knowledge (Kelleher et al., 2017). Such a constraint could take the simple form of confining the values a parameter could take during the calibration process.

The emergence of distributed parameters can be bypassed by using the same parameter over the n number of model units and therefore the corresponding distributed model component. When using the same parameter for the entire model component, the output of each distributed component is weighted, according to the area of the corresponding model unit, and combined to produce a single output from the entire model component. In Fenicia et al. (2008) improvements were achieved in the performance of the model in regard to the objective functions used in the paper when applying this variant of a distributed description of a model component (Fenicia et al., 2008, Figure 3). Especially great improvements were noted when this principle was applied to both the interception reservoir and the unsaturated reservoir noted as the $\text{FLEX}_{\text{Id,Urd}}$-model.

Precipitation does not always take the form of a liquid but can also appear as snow in sub-zero conditions. When in a frozen state, precipitation does not necessary enter the system, but, in a way, just stays on top of it until it melts. This principle urges the consideration of implementing something in the model to take this into account. In Nijzink et al. (2016) a snow reservoir was incorporated in the model in a similar manner discribed in Bergström (1976) to cover this principle. In Bergström (1976), where it is called a snow routine, it is described by the equation below. This is a simple way to describe whether or not precipitation takes the form of snow, and if the routine produces water from snow that melts (snowmelt).

$$\text{snowmelt} = C_o \times (T - T_{thresh}) \tag{2.1}$$

[4]At the time this paper (Fenicia et al., 2008) was written it is stated that these two terms got lumped quite a lot in the field of hydrological modelling.

[5]The potential evaporation is the amount of water that can physically evaporate per unit time given the meteorological conditions at that moment.

[6]A distributed-description of a model component entails that there is to be a model component corresponding to each model unit e.g. an interception reservoir for each area.

[7]Equifinality is the principle of multiple parameters sets producing an outcome that fits the observed data, while having the inherent possibility of getting the internal dynamics of the catchment wrong.

In equation (2.1) T represents the ground surface temperature on that day. If the temperature of that day does not exceed the threshold temperature of T_{thresh} than the precipitation just lies on top the of system in the form of snow. If however the temperature does exceed the threshold temperature, the build-up snow start to melt at the rate C_o. Naturally, it is a possibility that no physical snow is present in the snow routine combined with a temperature higher than the threshold temperature. In this case the precipitation passes through the snow routine into the next component of the model.

For the model used in this thesis the snow reservoir was implemented as described above and is mathematically written down as seen in equation (A.2) (tabel A.1; appendix A.1). While the previously stated possibility, of no present physical snow and a higher temperature than the threshold, is not represented in this equation; it naturally is implemented in the model. The snowmelt here is noted as the effective precipitation P_{eff}.

The principle of a distributed-description of a model-component (Fenicia et al., 2008), as it was described on page 12, was applied to the snow reservoir[8]. This distribution is ought to counter the lack of representation, by a single reservoir, of the height differences' influence on the snow build-up . In order to make a distributed description of the snow reservoir only two inputs are needed: one that is derived from the Digital Elevation Model[9](DEM) of the catchment and the other that is the environmental lapse rate[10].

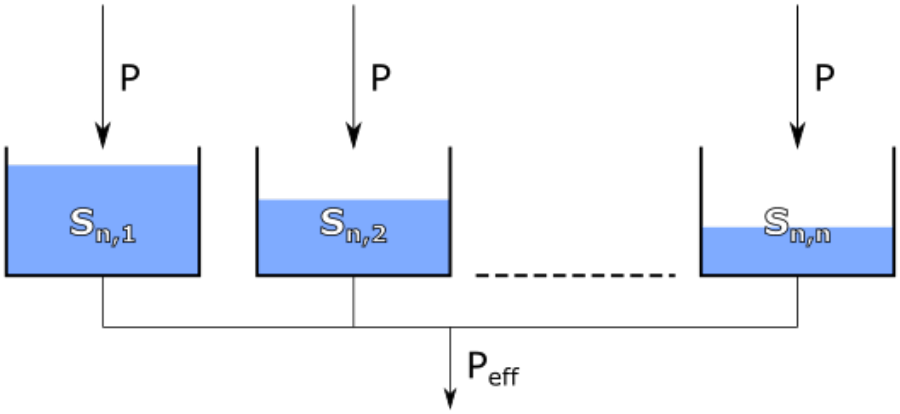


Figure 2.4: A schematic representation of the distributed description of the snow reservoir.

[8]In this thesis this principle was applied to neither the interception reservoir nor to the unsaturated reservoir, but instead it was applied solely to the snow reservoir. The reason behind this is primarily the lack of available distributed, e.g. precipitation, data, which will be touched upon in the next subsection.

[9]A Digital Elevation model, or short DEM, is a digital representation of the relief of an area. The DEM-file consists of, more or less, an array where each element is a pixel containing the height corresponding to the location of that pixel. A visual representation of the DEM-file used in this thesis is visible in figure 2.2. The DEM-file of the HJ Andrews catchment was found on the website of the catchment (USFS, 2018), where a DEM-file with a pixel-grid resolution of 10 metres by 10 metres was acquired.

[10]The environmental lapse rate is defined as the decrease in temperature as the result of an increase in altitude. For the first 10-11 kilometres of the Atmosphere the environmental lapse rate is around 6-6.5 °C/km (Thayyen and Dimri, 2014).

For the distribution of the snow reservoir, the catchment was divided in model units. Each model unit corresponds to a certain height interval[11], therefore every distributed component corresponds with a height interval as well. The next step was to determine, from the DEM-file, the amount of pixels within each height interval. The following table was created as a result:

Height interval(m)	**Pixels(%)**	**Temperature shift(°C)**
250-500	2.46	+4.24
500-750	22.05	+2.62
750-1000	30.76	+0.99
1000-1250	24.82	-0.63
1250-1500	17.86	-2.25
1500-1750	2.04	-3.87
1750-2000	0	-5.50

Table 2.2: Percentage of contribution to the total amount of pixels by each height interval.

One single distributed snow reservoir behaves the same as the snow routine as it was described in Bergström (1976) and adheres to equation (A.2) (table A.1; appendix A). The output of each distributed snow reservoir is then weighted according to the percentage of contribution to the total amount of pixels[12](table 2.2). All the weighted outputs from the distributed snow reservoirs are then summed up to produce one single output from the entire snow component. This output should logically have the same order of magnitude as the output of one single snow routine. In table 2.2 a third column, containing the temperature shift, can be observed. This temperature shift is the result of the snow reservoir taking temperature as an input, as can be seen in equation (2.1). However, since a distributed description is made of the snow component, the temperature input differs from one distributed snow reservoir to another as a result of the height difference. The collected temperature data were therefore shifted according to the environmental lapse rate with respect to the height of the meteorological station the data were collected from. For example, if the meteorological station is located at an altitude of 1028 metres, this means that for the interval of 750 -1000 meters the temperature shift becomes:

$$\text{Temperature Shift(°C)} = \left(\frac{1028 - \frac{750+1000}{2}}{1000} \right) \times 6.49 = 0.99 \tag{2.2}$$

After the implementation of a distributed description of the snow reservoir, this iteration of this FLEX-model was dubbed the FLEX_{nd}-model as a small nod to Fenicia et al. (2008). The constitutive equations of the FLEX_{nd}-model, as seen in table A.1 from appendix A, are heavily based on the constitutive equation used in the paper Nijzink et al. (2016). These most notably differ from those mentioned in Fenicia et al. (2006) in regard to the distribution of runoff[13].

[11] These height intervals were chosen in such a way that not too much information was lost, as a result of choosing a too big of an interval. In contrast, an interval too small could cause an increase in run-time of the model.

[12] The percentage of contribution to the total amount of pixels is another way of formulating the percentage of the surface-area that lies within that specific height interval.

[13] The equations differ from one another, but achieve the same principle when applied. For more information it is recommended to read the supplement of Nijzink et al. (2016) and the paper Fenicia et al. (2006).

A difference from both Nijzink et al. (2016) and Fenicia et al. (2006) is the description of the constitutive equations for both the fast reservoir and the slow reservoir. In Fenicia et al. (2006) the constitutive equation of the slow reservoir is described as a relationship between storage and discharge that needed to be determined. For the fast reservoir in this paper it is formulated as follows:

$$Q_f = K_f \times S_f \tag{2.3}$$

This is an equation gained from the fact that the fast reservoir is stated to be linear in its storage-discharge relationship and is defined by the time-scale K_f (which is better known as the recession coefficient of the fast reservoir). A part of the research of Fenicia et al. (2006) was reviewing whether the slow reservoir could be represented by a linear storage-discharge[14] relation. The conclusion was that this is mostly correct, though articles before and since refuted this conclusion (Tallaksen, 1995; Moore, 1997; Chapman, 1999; Kirchner, 2009), which itself is stated by Fenicia et al. (2006). However, as it is partly correct[15], the slow reservoir will be represented by a linear storage-discharge relationship and this should therefore result in the following equation for the slow reservoir:

$$Q_s = K_s \times S_s \tag{2.4}$$

When applying the principle of (ordinary) operator-splitting[16] on the first-order ordinary differential equation (A.13), the flux Q_s can be isolated in the split equation:

$$\frac{dS_s}{dt} = -Q_s \tag{2.5}$$

Here S_s is preliminary updated by solving the preceding split equations concerning the fluxes $(D) \times R_u$ and R_i. Combining equations 2.4 and 2.5 by substituting Q_s creates the following first-order linear ordinary differential equation:

$$\frac{dS_s}{dt} = -K_s \times S_s$$

Which is easily solved and results in:

$$S_s = c_1 e^{-K_s \times t}$$

[14]The storage-discharge relation is a description for the outflow from groundwater relative to the amount of groundwater present. Instead of using an a-priori assumption for this relation, it is derived from a constillation of recession-segments, known as the Master Recession Curve (MRC). Recession comes after the period of direct runoff, resulting from precipitation, and markes the period of flow dominated by the output of groundwater. In the case of this thesis, this is primarily the output of the slow reservoir.

[15]Tallaksen (1995) shows via numerical analyses that a linear reservoir only represents a small portion of the recession period, especially when taking into account the time variability found in recessions and the variability encountered in the recession behaviour of individual segments. However, at the time this paper was written, it concluded, in a nutshell, that further research was required. Kirchner (2009) states that the recession is closer to a power relationship in his search for a single non-linear storage-discharge relationship to characterize a catchment. Moore (1997), however, found that two linear reservoirs delivered a better performance than one single non-linear reservoir, which is in line with the finding of the variability in recession behaviour (Tallaksen, 1995). Chapman (1999) concluded, at least for recession of relative small duration, that a linear storage-discharge relation is sufficient while reviewing alternatives.

[16]The principle of operator-splitting states that one could split a partial differential equation and solve the split equations over the same interval in order to obtain an estimated overall solution. These split equations are solved in a pre-determined sequence, where one split equation takes input from the previous. In case of a first-order linear differential equation like equation (A.11), it would mean: solving for the fluxes separately (Fenicia et al., 2011).

When going one time step (Δt) forwards from t_n to t_{n+1}, it will result in:

$$S_{s,n+1} = S_{s,n} * e^{-K_s \times \Delta t} \tag{2.6}$$

Equation (2.6) shows that the state of the slow reservoir degrades with each time step by a factor $e^{-K_S \times \Delta t}$. From split equation (2.5) can be concluded the outgoing flux Q_s is equal to the negative change of the slow reservoir. A logical deduction from this in combination with equation (2.6) will lead to the following equation:

$$Q_s = S_{s,n} - S_{s,n} \times e^{-K_s \times \Delta t}$$

Here n is *pro forma* as it indicates the current time-step. When left out, one is left with the constitutive equation as is shown in appendix A. Applying the set of proceedings to both the fast and slow reservoir will result in the equations (A.12) and (A.14).

Lastly, there is the implementation of a lag-function. The lag-function's primary use boils down to offsetting fluxes to simulate the time-lag[17] caused by e.g. travelling of water through preferential pathways or recharging ground water[18]. In Fenicia et al. (2006) two lag-functions were applied to the FLEX_b-model (their representative parameters were already mentioned in table 2.1). These two lag-functions offset the fluxes that enter the fast and slow reservoir. In this thesis however, only one lag function is used[19]. The purpose of this transfer function is to offset the outflow of the model in order to account for the routing in the channel until it reaches the outlet of the catchment. Hereby the two parameters mentioned in table 2.1 are substituted for the parameter T_{lag}, which is the lag-time for the output transfer function. As stated above, the transfer function is characterised by a triangular distribution defined by the parameter. The basic principle is: the larger the parameter is, the "flatter" the triangular function becomes. This makes sense, as the triangular distribution and the output of the model, i.e. the modelled stream-flow, are used as input afterwards in a process known as convolution[20] to produce one single output where lag is accounted for. As stated before: as the lag-time (T_{lag}) becomes larger, the triangular distribution becomes flatter. This then leads to a more spread-out output of the model, where also the peaks are shifted, after convolution, which coincides with the theory about the precipitation-stream flow[21] relationship in a catchment.

[17] A lag-function is, in Fenicia et al. (2006), stated to be characterized by a triangular distribution of linearly increasing weights defined by its respective parameter.

[18] E.g. recharging ground water deals with the permeability of the soil/ground. The model only "thinks" in time intervals in which everything happens instantaneously. This logically is incorrect. Lag-functions help to make the model "less wrong" in this regard.

[19] One lag-function offsetting the outflow accomplishes the same goal as the two mentioned above. However, this comes at far lesser computational costs, as this function is only run once per model-run.

[20] Convolution of two functions, e.g. $f(t)$ and $g(t)$, represents the amount of overlap there is between the two functions. If both functions range over a finite set of values, which could be represented as a vector, then the solution becomes $C_t = \sum_u a_u \times b_{t-u+1}$. Here vector a represents function $f(t)$, vector b represents function $g(t)$ and C_t is the result; the output of the model in this thesis and the triangular distribution both are finite sets of values and therefore could be seen as vectors. A deduction from the equation above is that each element of C_t is the sum of the overlapping part of the two functions up until that point.

[21] E.g. if a river is surrounded by non-porous and impermeable rocks then a high peak-flow will be recorded. The cause is water which cannot infiltrate and therefore can only flow overland, which is the fastest route. However if the catchment is larger with a more porous and permeable ground-type, i.e. water can infiltrate, the lag-time increases as a partition of the water will take a slower path towards the river. Interception also adds to the lag as water is first temporarily stored before continuing its path.

With everything discussed in terms of model-attributes, one is left with the following set of parameters:

Parameter	Definition	Units	Range
C_o	Melting factor	$\frac{mm}{°C \times day}$	1 to 5
T_{thresh}	Temperature threshold	°C	−1 to 1
I_{max}	Interception capacity	mm	0 to 5
$S_{u,max}$	Unsaturated capacity	mm	1 to 1000
C_e	Limit potential evaporation	–	0.5 to 0.5
β	Shape parameter for runoff	–	0.1 to 4
$R_{i,max}$	Maximum percolation rate	mm/d	1 to 10
D	Runoff partition coefficient	–	0 to 1.0
K_f	Recession coefficient fast reservoir	d^{-1}	0.1 to 3
K_s	Recession coefficient slow reservoir	d^{-1}	0.0424 to 0.0707
T_{lag}	Lag-time for output transfer function	d	0.1 to 3

Table 2.3: The parameters of the $FLEX_{nd}$-model. A couple of parameters are denoted differently when compared to table 2.1. $S_{u,max}$ represents the same parameter as S_{fc}, C_e as L_p and $R_{i,max}$ as P_{max}.

2.2.2. Data

Naturally a model cannot be run without providing it with data. For the $FLEX_{nd}$-model the following types of data were required:

- Meteorological data
  - Precipitation data
  - Potential evaporation data[22]
  - Temperature data
- Stream-flow data[23]

The meteorological data were taken from the Central Meteorological Station (CENMET), which is located at an altitude of 1028 metres, as was mentioned earlier in this section (page 14; equation (2.2)). This station was one of the few, if not the only station, where daily data could be collected for over a period of more than 20 years, without too many hiatus in the data. The daily data was collected for a timespan ranging from the 1st of January 1994 to the 31st of December 2014.

The gaps that did exist in meteorological data, whatever set of data it was, were inferred by data from other stations using the simple AA-method (Sattari et al., 2016). This method is defined as follows:

$$V_0 = \frac{\sum_{i=1}^{n} V_i}{N} \tag{2.7}$$

Here V_0 is denoted as the to be estimated value. V_i is the value taken from station i and N are the number of stations. This method works particularly well when the stations are uniformly distributed across the area (Sattari et al., 2016).

[22] This is inferred by incoming short-wave radiation and temperature data.
[23] This is used during the calibration process.

However, as mentioned before, CENMET is one of the few stations that has collected daily data for more than 20 years. This means that the simple AA-method is not always applicable, because of possible lack of data from the other stations. In this case the principle of linear interpolation was applied in order to fill in the gaps. It must be said that most hiatus were not larger than a few days.

Potential evaporation
All of the meteorological data were taken directly from the data-catalogue from the HJ Andrews Experimental Forest website. The potential evaporation however, needed to be determined by inferring it by other data. The first thought was to estimate the potential evaporation by using the equation of Penman-Monteith, although this approach is disputed for its biases on a leaf-scale (Schymanski and Or, 2017), for which it originally was derived, and its poor performance at the scale of an ecosystem (Maes et al., 2019).

$$\rho_w \lambda E = \frac{\Delta(R_n - G) + \frac{\rho_a c_p}{r_a}(e_s(T_a) - e(z_r))}{\Delta + \gamma(1 + \frac{r_s}{r_a})} \tag{2.8}$$

Here E represents the potential evaporation, which is to be estimated. However, without covering every symbol in equation (2.8), it is easy to determine that there are no sufficient data to make use of this equation. E.g. R_n, the net radiation, is only available for the last year in regard to the data-range to be used. Thereby comes a complete absence of data in regard to G, the ground heat flux. Limited availability of data regarding the net radiation also ruled out methods like the Bowen-ratio; although in reality the potential evaporation is needed, the Bowen-ratio provides the actual in-situ evaporation. This disputed the use of e.g. the Bowen-ratio to begin with.

A decent alternative, according to Aguilar and Polo (2011), comes in the form of the Hargreaves equation (Hargreaves and Richard, 2003, Equation 8):

$$ET_0 = 0.0023 R_a (T_{mean} + 17.8) \times \sqrt{T_{max} - T_{min}} \tag{2.9}$$

ET_0 is denoted as the potential evaporation, where R_a represents the incoming short-wave radiation. Besides the requirement of daily mean temperature, as input for the model, the maximum and the minimum temperature are needed in order to estimate the potential evaporation according to equation (2.9). Although a modified version was presented (Aguilar and Polo, 2011) and the estimated values were higher than expected, the overall signal was deemed reasonable and was therefore used.

Stream-flow
Daily stream-flow, which is needed for the calibration process, was taken from the Andrews Lookout Creek Gaging Station (GSLOOK). Hiatus in the dataset could not be filled by inferring it by data from other gauging stations, as GSLOOK is the only station where the gauged watershed area covers the entire catchment. Therefore the only option available was: interpolation.

3

Method

This chapter covers the evaluation criteria that were chosen for the calibration process. For every evaluation criterion an explanation will be given of what the criterion entails, what is does, how it came to be and which hydrological signature from the hydro-graph it will cover. This chapter will also cover what type of calibration method is used and what the outcomes of this calibration process will be. As this is a thesis about shortening calibration data, an overview of the calibration period lengths will be presented here. Lastly, there is a small section dedicated to the validation process. Here it is explained why a validation process is needed in the determination of parameters.

3.1. Evaluation Criteria

All evaluation criteria used will be cover in the order in which they are applied in the calibration process. Each criterion value is denoted as ε, which represents the goodness of fit for its respective evaluation criterion. These values range from -∞ to 1, where 1 is a perfect fit and around a value of 0 already implies a "bad" fit.

The Nash-Sutcliffe efficiency coefficient (NSE)

$$\varepsilon = 1 - \frac{\sum(Q_m - Q_{obs})^2}{\sum\left(Q_{obs} - \overline{Q_{obs}}\right)^2} \tag{3.1}$$

The Nash-Sutcliffe efficiency coefficient was first introduced in Nash and Sutcliffe (1970). In this paper a search was made for a preconceived rule to measure/evaluate and take into account the results of an optimization step, for the purpose of applying it in an automated optimization process of the parameters. This preconceived rule was determined to be composed of a linear regression analysis between the modelled and the observed and a initial variance component concerning the observed. The written-out version of this formulation is viewed as equation (3.1). The Nash-Sutcliffe efficient has, since its inception, been a regular for evaluating the performance of models in hydrology (e.g. Pachepsky et al., 2016; Tarawneh et al., 2016; Seibert and Beven, 2009; Pool et al., 2017; Jain and Sudheer, 2008). However, a model has been cited to insufficiently reproduce hydrological signatures when solely making use of the NSE (Tian et al., 2016; Pool et al., 2017; He et al., 2015; Jain and Sudheer, 2008). Therefore it is often used in conjunction with other types of information (e.g. expert-knowledge based constraint) (e.g. Muleta, 2011; Euser et al., 2013; Hrachowitz et al., 2014; Moussa and Chahinian, 2009; Gharari et al., 2014; Nijzink et al., 2016), and evaluation criteria, possibly for the purpose of multi-objective(evaluation criteria) calibration. In conclusion, the Nash-Sutcliffe efficiency coefficient is chosen as the stepping stone for the calibration process.

The log Nash-Sutcliffe (LogNSE)

$$\varepsilon = 1 - \frac{\sum(log_{10}(Q_m) - log_{10}(Q_{obs}))^2}{\sum\left(log_{10}(Q_{obs}) - \overline{log_{10}(Q_{obs})}\right)^2} \tag{3.2}$$

Like stated in paragraph above, solely using the NSE in the calibration process will most likely not ensure the model's consistency. The squared terms above and below the bar tent to put a heavy emphasis on the peaks of the hydro-graph[1] and therefore "neglect" to some extent the lower flows. However, taking the logaritm of flows places a heavier emphasis on lower flows (Gharari et al., 2014; He et al., 2015). The result is the log Nash-Sutcliffe as formulated in equation (3.2). Santos et al. (2018) made a comment about using the logarithm of flows and concluded it to be flawed as it could lead to misinterpretation in the estimation of the water balance. This on an intuitive level is logical. Another comment is that the evaluation criterion, where the logarithm is inserted, loses some of its physical meaning as it does not handle zero flow recordings very well. The LogNSE, in this thesis,

[1] A hydro-graph is a graphical representation of stream-flow over time.

is meant to complement the NSE and not to be used solely. Muleta (2011) used a variant, called MNS, formulated as:

$$\varepsilon = 1 - \frac{\sum |Q_m - Q_{obs}|}{\sum \left| Q_{obs} - \overline{Q_{obs}} \right|}$$

Muleta (2011) used the MNS for the same reason, which is the over-sensitivity of the NSE to peak-flows. However after some testing, the addition of the MNS was deemed unnecessary as it had too much common ground with the LogNSE[2] in its applicability.

The flow duration Curve (FlowDur)

$$\varepsilon = 1 - \frac{\sum (Y_m - Y_{obs})^2}{\sum (Y_{obs} - \overline{Y_{obs}})^2} \tag{3.3}$$

where

$$Y_m = Sorted(Q_m): \quad \text{descending}$$
$$Y_{obs} = Sorted(Q_{obs}): \quad \text{descending}$$

The flow duration curve is a hydrological signature constructed from stream-flow data (modelled or observed). The flow duration curve represents the frequency distribution of the stream-flow defined over a specific time-step (Jothityangkoon et al., 2001; Berghuijs et al., 2014). This basically represents the probability that a stream-flow measurement will exceed a specific magnitude (Jothityangkoon et al., 2001; Sawicz et al., 2011). By taking the NSE of the flow duration curve instead of just the NSE of the flows, more emphasis is played on the magnitude of the flows instead of the timing of the peaks (Euser et al., 2013).

The autocorrelation function (AC)

$$\varepsilon = NSE(\,[R_{Q_m,1}, R_{Q_m,2}, \ldots R_{Q_m,i}]\,,\,[R_{obs,1}, R_{obs,2}, \ldots R_{obs,i}]\,) \tag{3.4}$$

Autocorrelation is the correlation of a signal (a hydro-graph is a physical signal) with a delayed version of itself, which in essence is the extent to which a delayed signal stays similar to its original counterpart. In modelling this could be used as an evaluation criterion (Winsemius et al., 2009; Euser et al., 2013) to measure the smoothness[3] of a hydro-graph. To evaluate how the autocorrelation of the modelled hydro-graph compares itself to the observed hydro-graph, it can be opted to divide the modelled value by the observed. This is then modified so it is usable as a quantitative evaluation criterion.

$$\varepsilon = 1 - \left| 1 - \frac{AC(Q_m)}{AC(Q_{obs})} \right| \tag{3.5}$$

This however is just one calculated autocorrelation making use of a 1 day delay (Euser et al., 2013). In Hrachowitz et al. (2014) usage has been made of a range of delays increasing from

[2] A pleasant remark, for this thesis, from Muleta (2011) was that the MNS and the NSE complemented each other rather well. This to some extent further justifies the use of the LogNSE besides the NSE.

[3] Smoothness of a hydro-graph is in essence how sharp or smooth the peaks of a hydro-graph arise. If the peaks are sharp then the smoothness is low. A low smoothness combined with a increasing lag will result in a faster decline of the correlation coefficient when compared to a high smoothness.

1 to e.g. i. This creates a vector covering the spectral properties[4] (Montanari and Toth, 2007; Hrachowitz et al., 2014) of a signal, e.g. a hydro-graph; thereby taking the form of a series of autocorrelations. In equation (3.4) every element of either one of the two vectors is the correlation between its respective original series of data, e.g. Q_m and a delayed variant, e.g. $Q_{m,d}$. The delayed variant is constructed by shifting the series of data i number of days forwards. If the original series start at April 1st then, in order to acquire $R_{Q_{m,1}}$, the delayed variant starts at April 2nd. The number of days, over which the correlation coefficient is calculated, has been set at 45 days. This at least would allow the creation of a large enough signal in regard to the smallest calibration period (section 3.2).

$$R_{Q_m,d} = \frac{\sum_{i=1}^{n} \left(Q_m - \overline{Q_m}\right)\left(Q_{m,d} - \overline{Q_{m,d}}\right)}{(n-1) \times S_{Q_m} \times S_{Q_{m,d}}}$$

where

$$S_x = \sqrt{\frac{1}{1-n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

For the same reason stated behind the formulation of equation (3.5), equation (3.4) is formulated as the NSE of the vectors containing the autocorrelations; primarily to acquire a quantitative evaluation criterion.

The runoff coefficient (RO):

$$\varepsilon = NSE(\,[Ru_{m,1}, Ru_{m,2}, \ldots Ru_{m,i}]\,,\,[Ru_{obs,1}, Ru_{obs,2}, \ldots Ru_{obs,i}]\,) \tag{3.6}$$

where

$$Ru_x = \frac{Q_x}{P} \tag{3.7}$$

The runoff coefficient (or ratio) Ru_x is defined as the relationship between the stream-flow Q_x and precipitation P over an interval, as is formulated in equation (3.7). The runoff coefficient represents the water balance between water exiting as stream-flow and water exiting as evaporation and transpiration, when the used data spans a long period of time (long-term) (Yadav et al., 2007; Sawicz et al., 2011). As being classified as a hydrological signature, it can be used as an evaluation criterion as has been done frequently in the past (e.g. Sawicz et al., 2011; Euser et al., 2013; Gharari et al., 2014). In this thesis the weekly runoff coefficient has been used. However, instead of determining the mean of those weekly coefficients over the respective calibration period, the coefficients were used as a signal represented by a vector. This vector, much like the autocorrelation function, contains the coefficients in chronological order. Therefore the vector representing the

[4] In short, spectral properties are revealed after decomposition (spectral density function) of a signal into a sum of sinusoidal components (frequency content). This in essence is a Fourier analysis. When a signal is analysed by its frequency content it is called a spectrum, hence the term 'spectral properties'. Autocorrelation, spanning a range of lag-time, represents the same properties as the spectral density function but in the time-domain opposed to the frequency-domain of the spectral density function.

runoff coefficients inferred by the modelled stream-flow is evaluated against its observed counterpart using the NSE.

The rising limp density (RLD):

$$\varepsilon = 1 - \left| 1 - \frac{L_{Q_m}}{L_{Q_{obs}}} \right| \tag{3.8}$$

where

$$L_x = \frac{\sum T_r}{n} \tag{3.9}$$

The rising limb density, first described as a hydrological signature in Shamir et al. (2005), is determined by the ratio between the sum of the time the hydro-graph is rising $\sum T_r$, to reach its peaks, and the number of peaks n. This is the inverse of the peak density (Morin et al., 2002). By dividing the time of rising by the number of peaks, one is left with the average time to reach a peak. This, like the autocorrelation, is a way to measure the smoothness of the hydro-graph. However, unlike the autocorrelation, it is averaged over the calibration period and independent of the flow volume (Shamir et al., 2005; Euser et al., 2013), as it does not matter what the magnitude of the flow is concerning the RLD. However, opposite to the other evaluation criteria, the peaks and their subsequent rising limbs are a product of analytical determination. What is considered a peak and a rising limb therefore becomes somewhat subjective. Some thresholds were built in the algorithm to filter out some of the "noise" of the hydro-graph[5].

The peak distribution (PeakDis):

$$\varepsilon = 1 - \left| 1 - \frac{PD_{Q_m}}{PD_{Q_{obs}}} \right| \tag{3.10}$$

where

$$PD_x = \frac{Q_{10} - Q_{50}}{0.9 - 0.5} \tag{3.11}$$

Peak distribution, as the name already alludes to, is a signature that represents the distribution of peak magnitudes over a given period. Unlike the RLD, the peak-flows used for the peak distribution are defined/determined by a lower recorded stream-flow on both the previous and the following time step (Euser et al., 2013). A flow duration curve is constructed from all the gathered peak-flows. The slope (PD_x) between the 10^{th} (Q_{10}) and the 50^{th} (Q_{50}) percentile is calculated from this peak duration curve, which, like the RLD, only focusses on higher, although not on extreme, peaks. Those are stated to be the most interesting for this analysis (Euser et al., 2013; Sawicz et al., 2011). By using the slope of the peak distribution curve, emphasis is placed on the relative peak magnitudes. This, to some extent, bypasses errors in e.g measurements, observed data, which affect the absolute peak magnitudes.

[5]The algorithm for determining the peaks and the subsequent rising limbs is presented in listing B.2 in appendix B section B.1. An example of the determination comes in form of figure B.1 in appendix B.

3.2. Calibration

Before the calibration process, a method of calibration had to be selected. The choice fell on the Monte-Carlo method, which can be seen inside the for-loop in listing B.4. The Monte-Carlo method, within the context of this thesis, can be considered a 'brute-force' method[6].

3.2.1. First run & Benchmark

First of all, an amount of 100,000 parameter-sets are generated. The sets of parameters assume values randomly sampled between the intervals as viewed in table 2.3[7]. These randomly generated parameter-sets will be used for the entirety of this thesis, i.e. no other parameter-sets will be used. For testing purposes, such as ironing out design-flaws, the 100.000 parameter-sets are forced through the loop and thereby only evaluated against the NSE over the entire benchmark period (10 years).

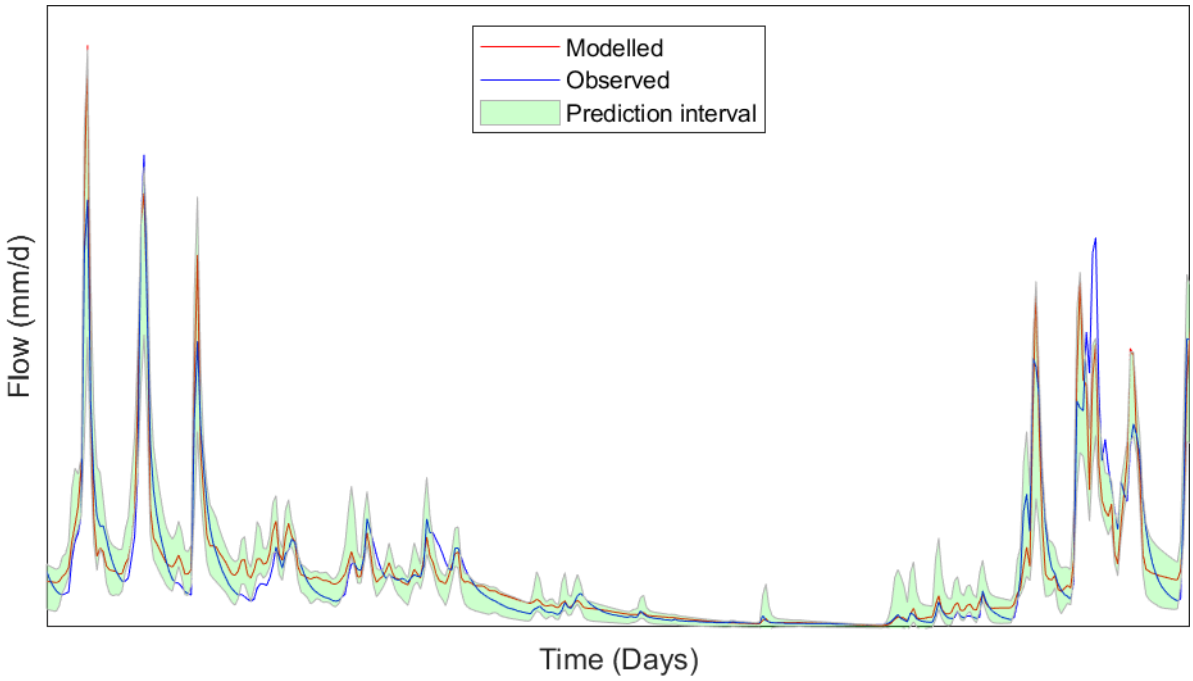


Figure 3.1: The modelled flow versus the observed flow evaluated solely by the NSE for a 10 year benchmark period. The modelled flow depicted used the parameter-set that had the highest NSE value. The prediction interval (lightgreen beam) represents the boudaries of the values the modelled flow can assume with a confidence level of 95 percent, when passing the NSE threshold.

[6] A method is considered brute-force when no use is made of a more efficient algorithm. It basically comes down to generating and testing, which is the purpose of the Monte-Carlo method in this thesis (listing B.4, appendix B.2).

[7] One thing to address is the small interval for the recession coefficient of the slow reservoir (K_s). A linear storage-discharge relationship was ought to be sufficient (Fenicia et al., 2006; Chapman, 1999), though be it for a limited part of the recession period (Tallaksen, 1995). This limited part is at the tail-end of the recession period. This "tail" could be viewed solely as a flux exiting the slow reservoir. As these tails display exponential behaviour, it is possible to fit the outflow of the slow reservoir to these tails. This then would lead to a K_s that is constant. A margin of error has been added to the range of K_s in order to let it fluctuate a little, as these tails are unique as a result of different circumstances and errors in measurements are a possibility.

The calibration process uses 1 year preceding the 10-year evaluation period to "warm-up" the model[8], as the initial conditions, e.g. the soil moisture, are unknown at the moment the model starts.

Next, all 100.000 parameter-sets are evaluated against all remaining evaluation criteria individually over the benchmark period, as was done with the NSE. The purpose of this process is to establish the thresholds of the individual criteria. The threshold of an evaluation criterion is defined as the 95^{th} percentile[9] of all the values acquired from evaluating all parameter-sets against said criterion. The individual criteria thresholds are presented in table 3.1.

Evaluation criterion	**Threshold value**
NSE	0.7859
LogNSE	0.8833
FlowDur	0.9677
AC	0.8698
RO	0.9554
RLD	0.9664
Peakdis	0.8209

Table 3.1: The thresholds of the individual evaluation criteria.

From here it is possible to create the first benchmark. The benchmark serves as a measurement to evaluate the equivalent performance[10] of parameter-sets that were deemed "good" by the shorter-period calibration. The first benchmark consists of parameter-sets that solely passed the NSE-threshold. However, when a parameter-set is chosen solely based on the NSE value, the hydro-graph tents to show misjudgement of the recession period, as is seen in figure 3.1. This is mainly the result of the NSE's tendency to lay too much emphasis on the peak-flows (section 3.1), which leads to a model "trying" to get the peak-flows right.

Thereby comes the fact that eventually more than one evaluation criterion will be used, it would therefore be biased to base the equivalent performance solely on the Nash-Sutcliffe efficiency, while using multiple evaluation criteria during the shorter-period calibration. A more desirable benchmark would be based on all of the mentioned evaluation criteria. The most practical approach would be the creation of one single quantitative value representing all criteria. The basic principle of this approach is displayed in figure 3.2.

[8] It is advisable to let the model run for a period of time to sort out the initial conditions itself. Otherwise the model calibration period would start on arbitrary initial conditions that could result in a great bias in regard to the evaluation criteria. Yu et al. (2018) states that the the warm-up time is dependent on factors like the soil texture, soil profile length etc., which made the determination of the warm up in this thesis an educated guess.

[9] It has to be noted however, that the determination of the threshold is somewhat subjective. An other possibility would be a threshold based on e.g. export-knowledge.

[10] Equivalent performance is defined as the perfomance the model will have during the benchmark period (in this case 10 years), when using the same parameter-sets that passed the thresholds of their respective evaluation criteria during the shorter-period calibration. A model logically will have a different performance for different periods, especially when the length of period differs greatly.

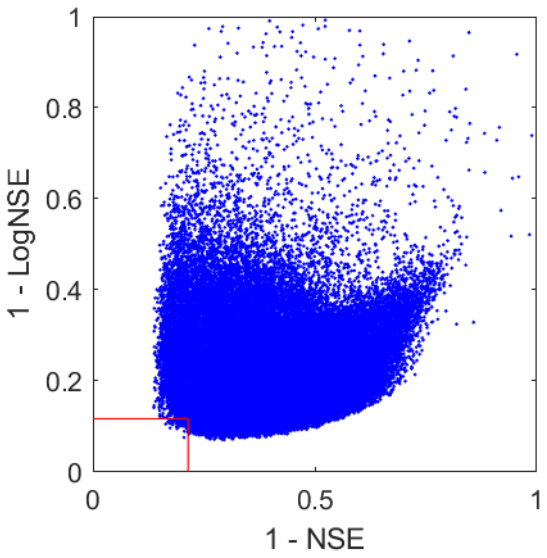


Figure 3.2: A depiction of the relationship between the complements of two evaluation criteria. The red lines indicate the thresholds of the respective evaluation criteria. The outer rim of the dots within the red square is defined as the Pareto-front. The dot that is closest to the origin is defined as the Pareto optimal value and determined by the euclidean distance to the origin. This dot should represent the best performing parameter-set when both evaluation criteria have equal weight.

This single value will be the euclidean distance to the origin of the complements[11] of all the evaluation criteria (figure 3.2). The euclidean distance is a multi-dimensional use of the Pythagoras-formula with the origin as its starting point. All the evaluation criteria values have equal weight in the determination of the euclidean distance.

$$\varepsilon = \sqrt{(1-NSE)^2 + (1-LogNSE)^2 \quad \ldots \quad + (1-PeakDis)^2} \tag{3.12}$$

By looking at figure 3.2 and how equation (3.12) is formulated, it can be concluded that a lower euclidean distance value indicates a better performance of a parameter-set, which is the opposite of an individual criterion. A decision was made not to base the benchmark on parameter-sets passing all the individual criterion thresholds (dots within the red box; figure 3.2), as no parameter-set can pass all the thresholds while evaluated over a period as long as 10 years. This entails that the benchmark would be empty. Therefore a single threshold value was determined; this value is formulated as the 5^{th} percentile of the euclidean distance values of all parameter-sets evaluated over the benchmark-period. This value was set at: 0.7671. All parameter-sets with a euclidean distance value lower than the threshold were included in this benchmark.

[11] The complement of a normalized quantity is described as one minus said quantity.

3.2.2. Moving window

Five period lengths were chosen for the shorter-period calibration process:

- 5 years
- 2 years
- 1 year
- 6 months
- 3 months

Each of these period-lengths is moved across the 10-year period, that was used to create the benchmark, one month at a time. This creates a "moving window", that allows each period-length to start at every month. This entails that a calibration period-length of 3 months will have 118 calibration periods in a span of 10 years. The last two months, for a period length of 3 months, can not be used as starting points as the calibration period would extent beyond the benchmark period of 10 years.

Evaluation criteria will be added during the calibration-process for every period-length, starting with the NSE. This implies that e.g. for the 3 months period-length it is done for all 118 periods. After the NSE, the LogNSE will be added. Hereafter the FlowDur will be added, etcetera. The criteria are added in the order in which they are mentioned in section 3.1. After the addition of each criterion, an evaluation is carried out for all the criteria that are implemented at that moment. The parameter-sets that pass the thresholds (table 3.1) of the implemented criteria are saved while the others are rejected. This process continues until all the criteria are implemented in the calibration process[12].

The parameter-sets that are saved after each criterion addition are saved with the equivalent performance values of both the NSE and the euclidean distance for an easy comparison with both benchmarks. When the same parameter-set is used for multiple periods with the same period length, it creates the possibility of saving duplicates per period length[13]. However, this is completely fine as it would only matter less in which month the calibration starts for that specific parameter-set.

Two last things need to be addressed concerning the calibration process:

- The fixed C_e parameter
- Exclusion of extreme peak-flows in the observed data

The fixation of the C_e parameter is the result of C_e being a remnant of an earlier built of the model. In e.g. Fenicia et al. (2006) it was an active parameter while in e.g. Nijzink et al. (2016) it was fixed on a value of 0.5. In equation (A.9) from appendix A, C_e would take the spot of the 0.5 in the determination of the transpiration E_t.

The reason behind the exclusion of extreme peak-flows is that the model can not generate enough flow to reach those peaks. If the outcome is evaluated solely by the NSE, then the

[12] A part of the moving window calibration can be found in listing B.5 in appendix B.2.2.

[13] Although the period length of 3 months is a constellation of 118 periods, the results are still saved under the single banner "3 months".

model, using the parameter-set with the best NSE value, will most likely emphasize the faster processes in order to reach those peaks. However, the model is not able to produce enough outflow to reach these extreme peak-flows but will "sacrifice" the reproduction of hydrological signatures in its attempt. Therefore these peaks were excluded from the observed data for the NSE, LogNSE, FlowDur, AutoCor and the RO. As the RLD and the PeakDis do no suffer punishment for leaving these peak-flows in, the peak-flows were left in for these criteria.

3.3. Validation

The manner in which the model, with the parameter-sets acquired from the calibration process, is able to reproduce hydrological signatures, for a period outside the calibration period, is not clear. This is the result of not knowing much about the consistency of the model using these parameter-sets.

Therefore a validation period is introduced. The purpose is to evaluate the consistency of the model utilising these parameter-sets. The parameter-sets are evaluated over the entire validation period, also spanning 10 years, for every single evaluation criteria, in order to review the performance of the parameter-sets over a different period of time. This will give insight into the consistency of the model utilizing these parameter-sets. The euclidean distance value will also be determined over the entire validation period for comparison with the benchmark.

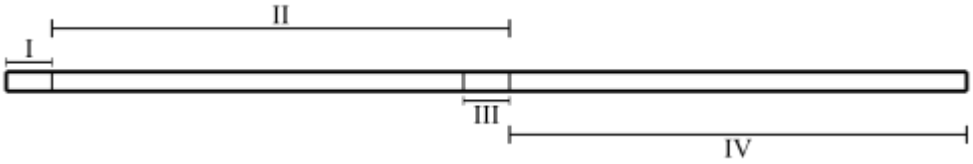


Figure 3.3: Timeline that represents the use of data of the entire period. I is the warm-up period for the calibration process. II is the calibration (benchmark) period. III is the warm-up period for the validation process. IV is the validation period.

The last year of the calibration period (III) is used as a warm-up of the model for the validation period (IV) as is seen in figure 3.3 above.

4

Results

4.1. Calibration

This section covers the results of the Monte-Carlo moving-window calibration (subsection 3.2.2). The equivalent performance is determined for every parameter-set that came through the calibration process given any period-length. The equivalent performance used in this context is defined in terms of the euclidean distance value. The equivalent performance is used for the purpose of an easy comparison with the benchmark. Figure 4.2 is a graphical representation of the equivalent performances of the parameter-sets from every shorter-period calibration. The first box plot of figure 4.2 (highlighted in green; mean euclidean distance value of: 0.6526) is the benchmark performance, as was described in subsection 3.2.1.

4.1.1. First observations

The first group of box plots to the right of the benchmark represent the 5 year calibration period-length. The performance of the model using the parameter-sets that passed the threshold of the NSE solely (far left) appears to be worse than the benchmark performance (mean equivalent value of: 1.2606 to 0.6526), as one would expect[1]. The stepwise addition of evaluation criteria seems to greatly improve the performance, until the performance of the 5 year calibration period-length even exceeds(!) (far right) the performance of the benchmark. This observation entails that utilizing the first 4 evaluation criteria promises a performance better than the threshold of the benchmark, when 5 years worth of daily data is available.

No box plots are shown after the addition of the RO, RLD and the PeakDis for this calibration

[1] It becomes "easier" to pass the threshold when the calibration period-length is made shorter, as there are less data to get right. This results in parameter-sets passing the calibration process, while quite possibly having a worse equivalent performance. This becomes increasingly apparent when one looks at e.g. the 3 months period calibration utilizing only the NSE (figure 4.2).

period length. This is very likely the result of strict/high thresholds. In general, the evaluation criteria values tend to be lower, when a larger period is used to evaluate the modelled flow against the observed. This is a logical result as it gets harder for the model to get everything 'right'. In conclusion, for larger periods it becomes increasingly difficult for parameter-sets to pass all the thresholds of the respective evaluation criteria.

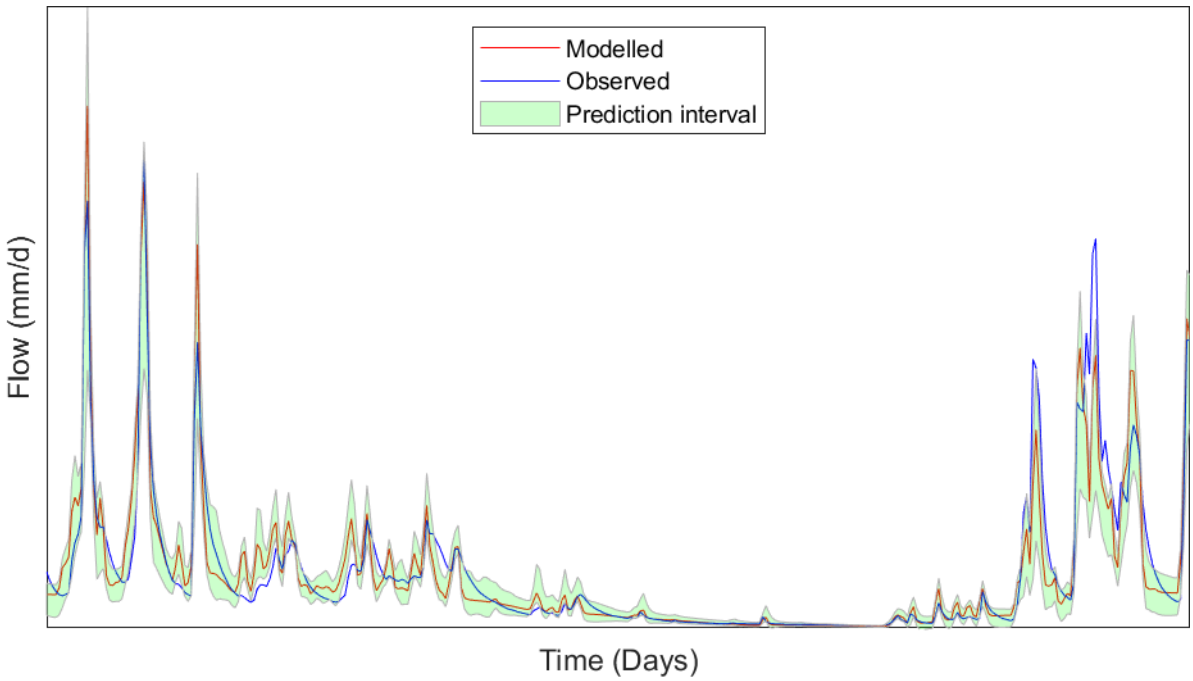


Figure 4.1: A depiction of the modelled flow versus the observed. The model used, out of the remaining parameter-sets, the parameter-set with the highest equivalent NSE-value after the addition of the last criterion. The calibration period-length is 6 months.

The performance of the model decreases as the calibration period-length shortens. The stepwise addition of calibration criteria mostly seems to improve the performance of the model for any given calibration period-length (figure 4.2) and the importance of more evaluation criteria appears to increase, as one shortens the calibration period progressively. The performance improves to benchmark-level values for a period-length up to 6 months, after the addition of the last criterion that still secures remaining parameters. A parameter-set (Par_{6m}) produced by the 6 months period-length calibration

Evaluation criterion	**Individual value**
NSE	0.8333
LogNSE	0.8476
FlowDur	0.9577
AC	0.9125
RO	0.5175
RLD	0.9088
PeakDis	0.8304

Table 4.1: The equivalent individual criteria values of the model using the parameter-set which had the highest equivalent NSE-value. Calibration period length is 6 months and all criteria were added in the process.

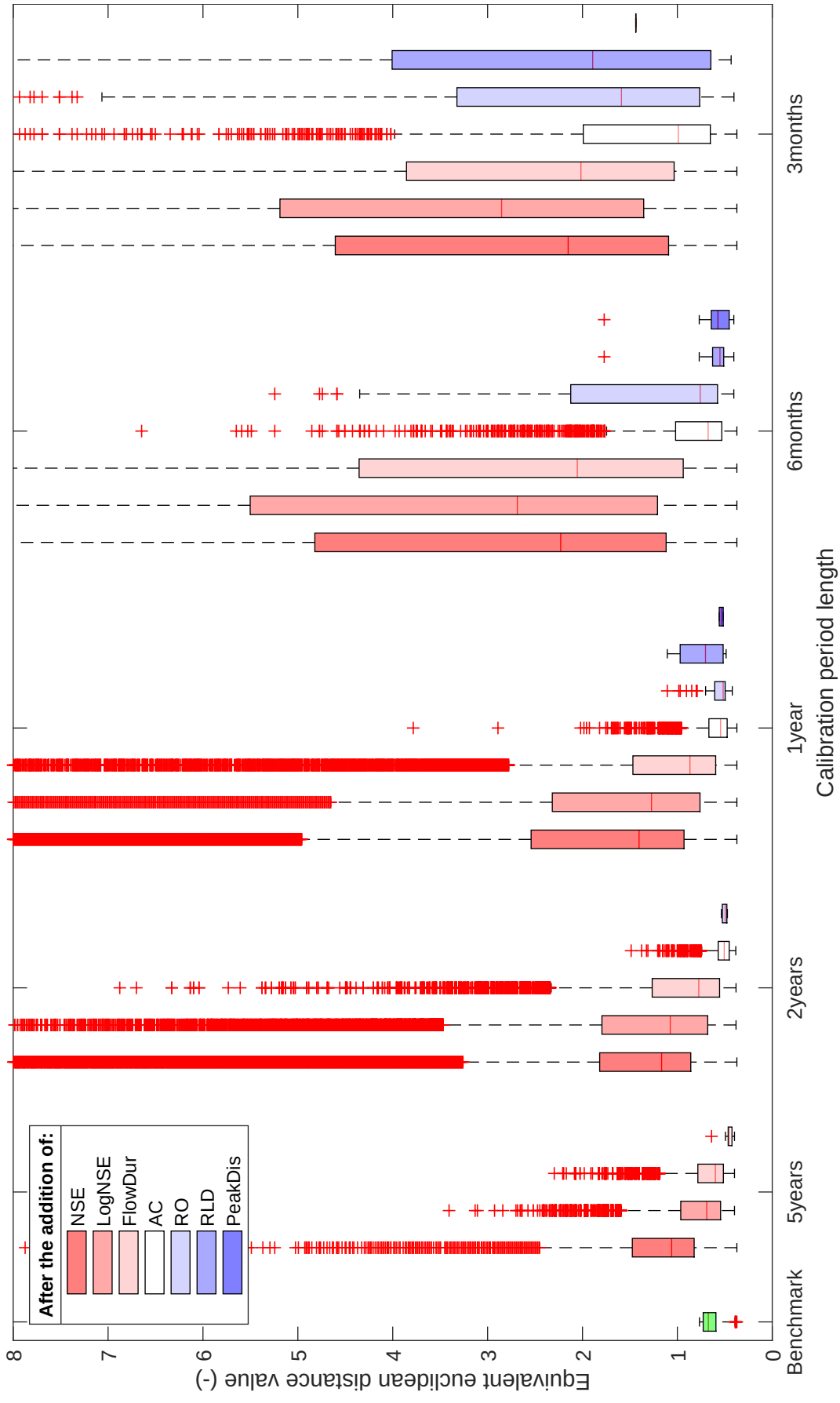


Figure 4.2: The equivalent performance, in terms of euclidean distance over the entire benchmark calibration period, of the model using the parameter-sets that made it through the calibration process, i.e. passing the thresholds, for the different period lengths. LogNSE stands for the NSE+LogNSE, FlowDur for NSE+LogNSE+FlowDur, etcetera.

shows a better ability to mimic the hydro-graph (figure 4.1) than the parameter-set evaluated solely against the NSE over a period of 10 years (figure 3.1). Thereby is the prediction interval narrower than shown in figure 3.1, which entails less "randomness" in performance (subsection 4.1.3). The individual equivalent criterion values of Par_{6m} (table 4.1) are comparable to the threshold values that were determined over a period of 10 years (table 3.1). This would entail a "good" performance according to the criteria.

4.1.2. Further observations

While looking at figure 4.2 and moving passed the 6 months period-length towards the 3 months period length, one is able to make a few eye-catching observations:

I Only one parameter-set is left after the addition of the last evaluation criterion for the 3 month period length

II For the shorter calibration periods, especially the 3 month period length, it appears that the performance improvement stagnates (cluster to the far right)

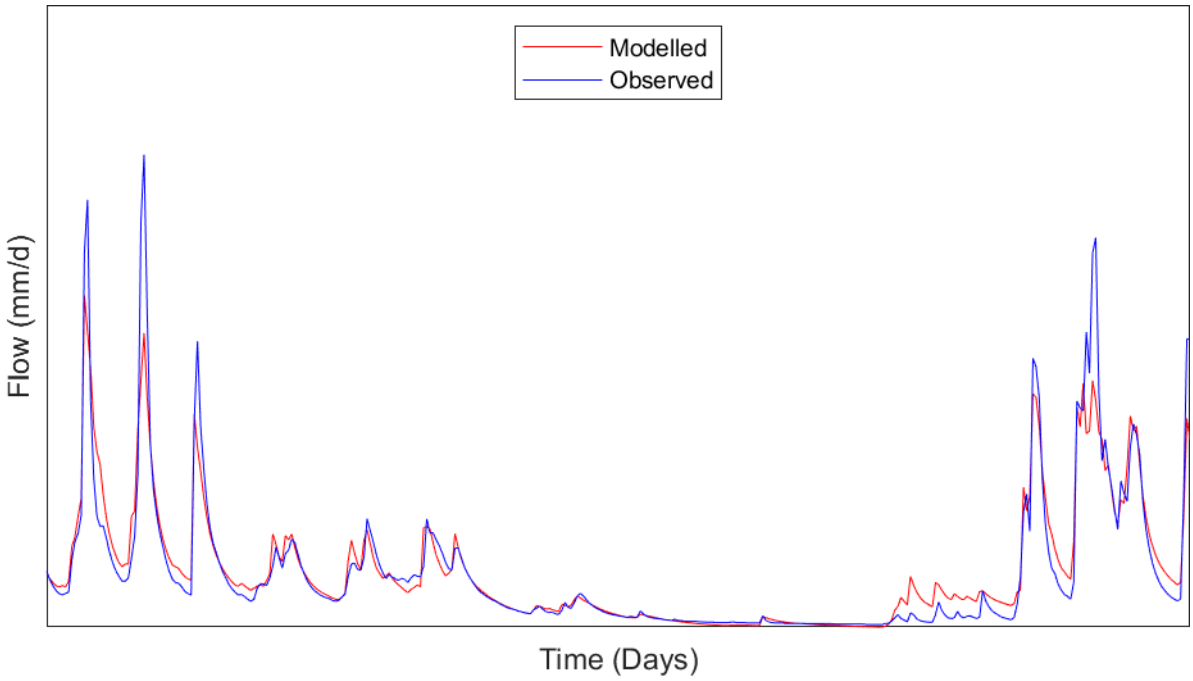


Figure 4.3: A depiction of the modelled flow versus the observed. The model used the last remaining parameter-set that remained after the addition of the last evaluation criterion over a calibration period of 3 months.

From observed point I it can be concluded that this sole remaining parameter-set should lead to the best performing model, especially as the equivalent performance value is reasonably close to the set of benchmark values. The fact that the visual modelled outcome, shown in figure 4.3, seemingly is an improvement in comparison to figure 3.1 further strengthens this belief[2].

[2]Figure C.1 from appendix C shows the outcome of the model using the parameter-set that had the highest NSE-value after the addition of the NSE over a calibration period of 3 months. The visual performance shown in figure 4.3 is significantly better than in figure C.1; at least implying it is a significant improvement.

This however could be very deceiving. By looking at the modelled flow versus the observed, as shown in figure C.2 from appendix C, the visual performance of the model using this parameter-set raises some eyebrows when reviewing the outcome over a larger period. Over the entire benchmark period, the model tends to overshoot some of the smaller peak-flows and does not "even attempt" to produce a few of those peak-flows, while utilising this sole remaining parameter-set.

Evaluation criterion	**Individual value**
NSE	0.7792
LogNSE	0.6475
FlowDur	0.8689
AC	0.0962
RO	2.26×10^{-4}
RLD	0.9750
PeakDis	0.7527

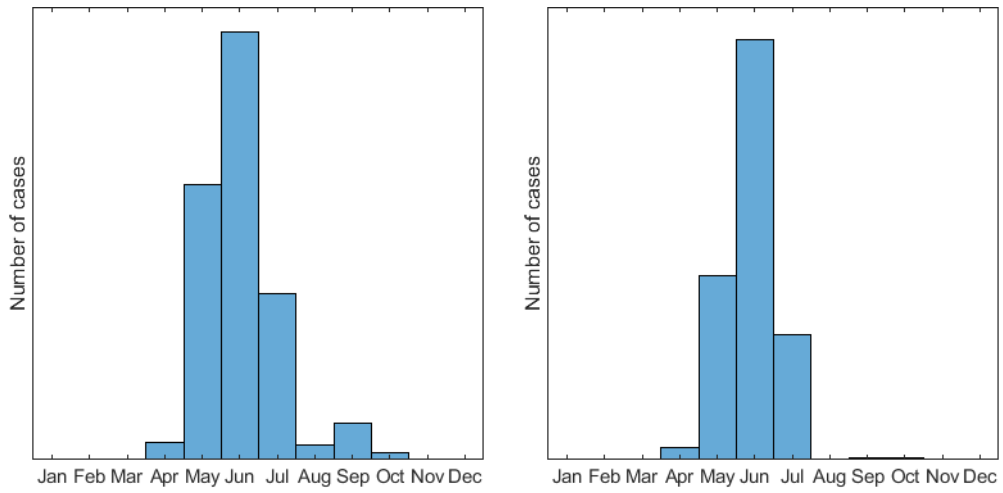
Table 4.2: The values of the evaluation of the model using the sole remaining parameter-set, over the benchmark period, for every individual criterion.

A low value for the runoff coefficient criterion is not disastrous as this criterion value decreases rather quickly when the magnitudes differ a bit from the observed flow. More alarming is the low value of the autocorrelation criterion, as this basically "checks" the shape of the modelled hydro-graph. Figure C.2 from appendix C arguably depicts this low autocorrelation value.

Observed point II is most likely the result of a change in rainfall-runoff dynamics, starting around the month of May. Figure C.4 from appendix C shows a fast decrease in daily precipitation during the summer months. However, as shown in figure C.5 from appendix C, the monthly mean runoff coefficient increases by a great margin in those months while being mostly constant for most of the year. This implies that there still is a significant enough stream-flow without much precipitation. A deduction from this information is that the contribution to the stream-flow is dominated by various types of groundwater flows[3] during those summer months.

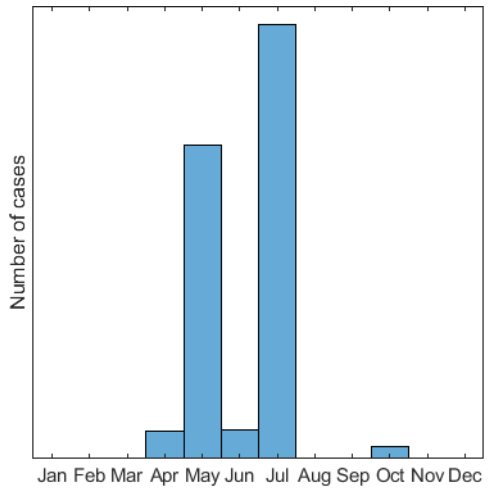
A large emphasis would therefore be placed on the slow reservoir and the unsaturated/soil moisture reservoir within the FLEX_{nd}-model and, as a result, on their respective parameters during the calibration process. The period from the decline of precipitation until the end of significant recorded stream-flows (high runoff coefficient) spans around 4 months, from May until August. If the entire 3 month calibration period falls within these months, then performance issues will arise in terms of equivalent performance. In essence, to obtain a good performance during these months for a 3 month period-length, only the parameters associated with the slow reservoir, to a limited degree the unsaturated reservoir and their respective fluxes, have to be calibrated "correctly". This entails that parameters that are associated with e.g. the interception reservoir could have a random value and the model would still have a decent performance for those 3 months.

[3] As there is little to no precipitation, the groundwater reservoirs tend to empty themselves as water still wishes to flow in the direction of decreasing pressure, i.e. the main stream.

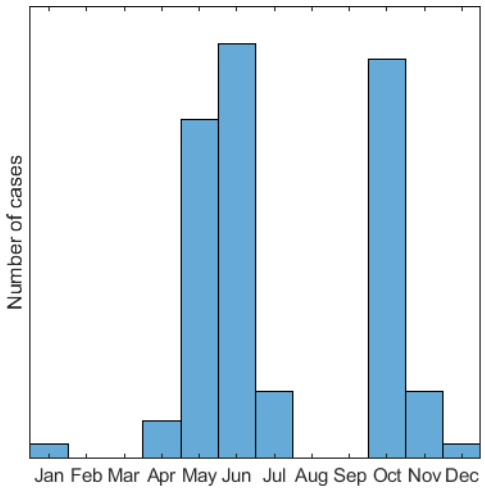


(a) NSE

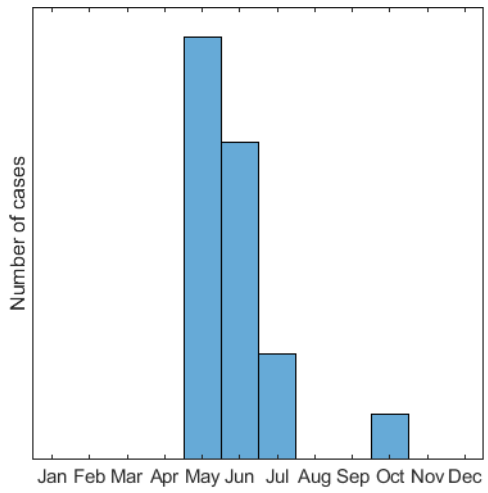
(b) LogNSE



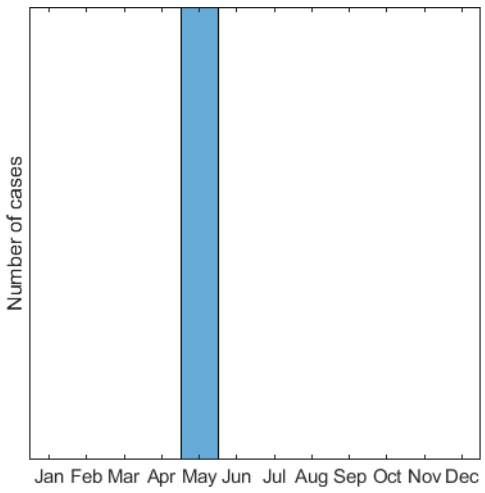
(c) FlowDur



(d) AC



(e) RO



(f) RLD

Figure 4.4: Worst 10 percent of the equivalent euclidean distance values after each evaluation criteria addition for the 3 month period-length. The Peak Distribution is left out, as there was only one parameter-set left.

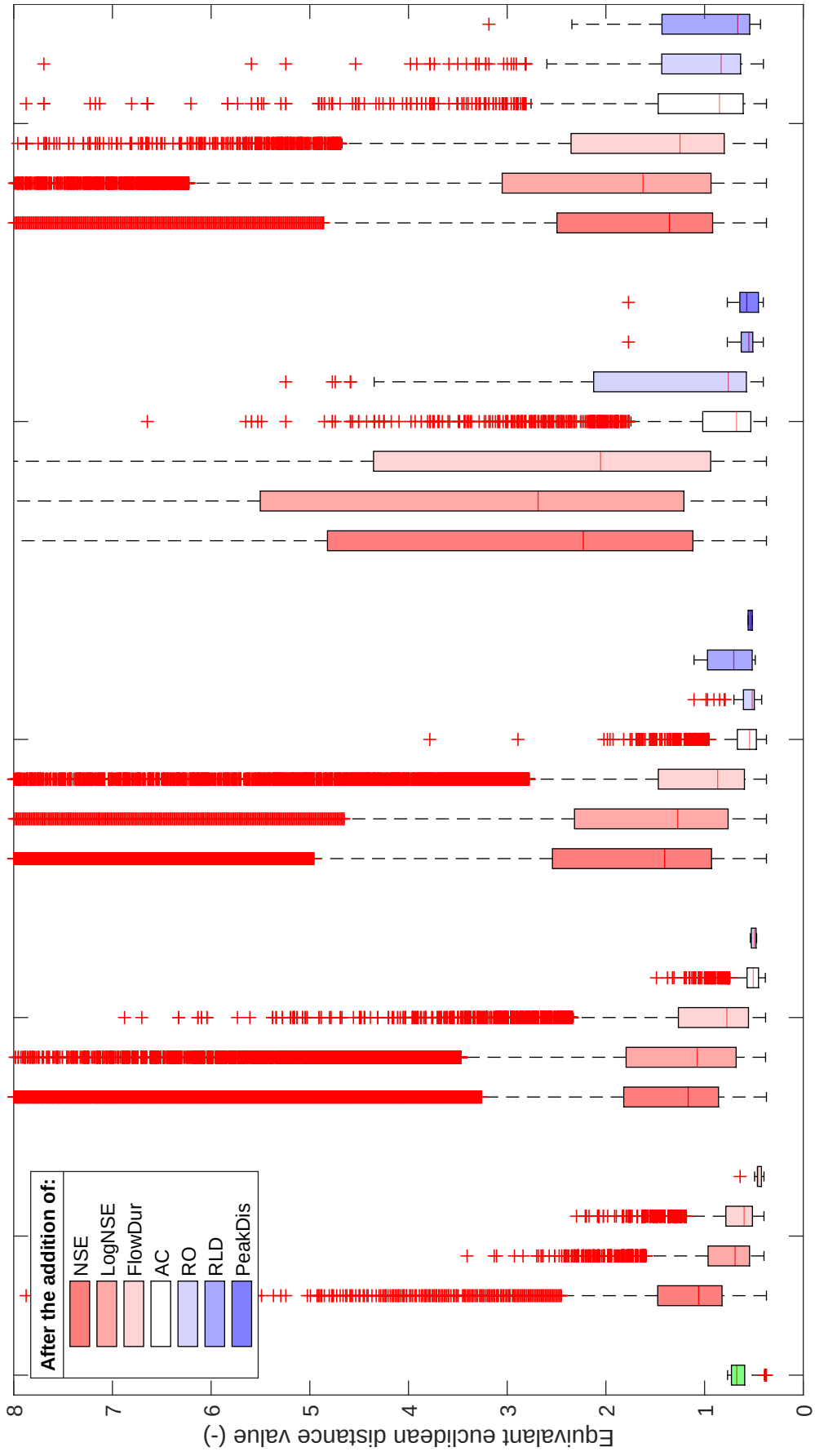


Figure 4.5: The equivalent performance, in terms of euclidean distance over the entire benchmark calibration period, of the model using the parameter-sets that made it through the calibration process. For a calibration period of 3 months, the periods starting with the months May, June and July were excluded in accordance with the results observed in figure 4.4.

Especially the parameters concerning the snow reservoir will suffer as the temperature in most of these summer months (figure C.6; appendix C) do not reach sub-zero values, which basically means that the snow reservoir is bypassed and therefore deemed irrelevant in this calibration process. In conclusion, a model calibrated during the summer months over a period of 3 months will most likely not ensure consistency.

A step in the right direction would therefore be the exclusion of calibrations, for the 3 month period length, that would start in the months May, June and July. Although a period would end with September when it would start in July; July is still excluded as the amount of precipitation in September is still low on average (figure C.4; appendix C), which entails that the model would still not "learn" how to utilise e.g. its interception reservoir. Plotting the worst 10 percent of the equivalent individual evaluation criteria values (figure 4.4) reveals that indeed the calibration periods starting with the months May, June and July should be excluded. Figure 4.5 shows a much better average equivalent performance of the remaining parameter-sets in comparison to figure 4.2 (mean equivalent euclidean distance value of 1.0442 compared to 2.6773; far right box-plot of the far right cluster). The exclusion of these starting months neatly solves the problem that arose with observed point I. The starting month of that sole remaining parameter-set was May, which implies that it now is filtered out. This paragraph also serves as an explanation why the performance of that sole remaining parameter set was lacking. For the purpose of this thesis a parameter-set is determined to replace the previously mentioned parameter-set in order to present a graphical representation of the "best performing" parameter-set that comes out of the 3 month period-length calibration. The determination is based on the highest equivalent NSE value after the addition of the RLD evaluation criterion. This parameter-set will be denoted as Par_x.

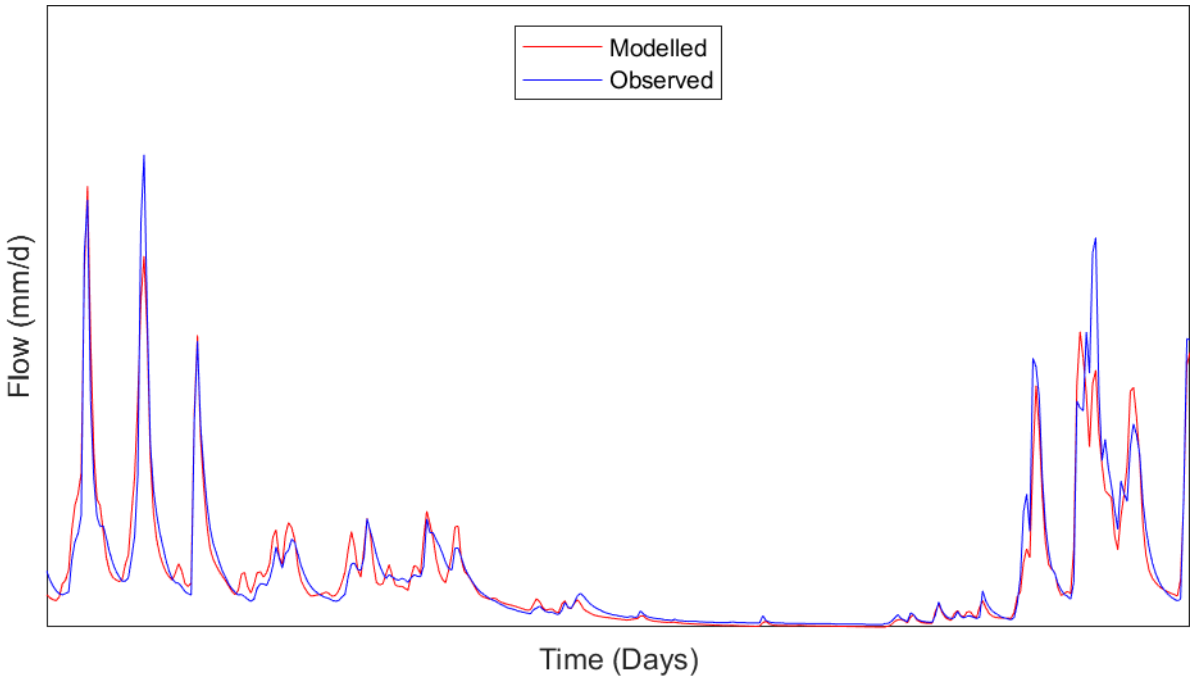


Figure 4.6: Modelled versus Observed, where the model utilises the parameter-set (Par_x) with the highest equivalent NSE values after the addition of the RLD.

Although the euclidean distance, with equal weight for each evaluation criterion, is a good quantity to be used as tool for comparison, it is not adequate for determining the best parameter-set. This is shown by figure B.2 from appendix B. In this figure the benchmark period is used for evaluation utilising every criterion, after which the parameter-set was chosen with the lowest euclidean distance value. Visually, the performance of the model is rather poor using this parameter-set.

Evaluation criterion	Individual value
NSE	0.8617
LogNSE	0.7410
FlowDur	0.8371
AC	0.8595
RO	−2.1618
RLD	0.9354
PeakDis	0.8948

Table 4.3: The values of the evaluation of the model using Par_x, over the benchmark period, for every individual criterion.

The visual performance of the model using Par_x is already an improvement as is seen by comparing figure 4.6 with figure 4.3. By looking at table 4.3 the first thing that stands out is the negative value for the runoff coefficient criterion. However, as mentioned before, this does not entail poor performance as this tends to happen rather quickly. The biggest difference between table 4.3 and table 4.2 is the significantly higher value of the autocorrelation criterion when utilising Par_x. This leads to the conclusion that Par_x is far more capable of producing an outcome that has a similar shape to the observed hydro-graph.

4.1.3. Prediction

The light green beam as seen in figure 3.1 represents the interval, with a confidence level of 95 percent, in which future modelled flows are predicted to lie in. For figure 3.1 this prediction interval is determined according to the outcomes produced by the model using the benchmark parameter-sets, i.e. the parameter-sets that passed the NSE threshold while being evaluated over the benchmark period.

One would expect the prediction interval to be wider when utilizing the parameter-sets that passed the threshold(s) during shorter period calibration, especially after the earlier additions of evaluation criteria. The step-wise addition of evaluation criteria in the calibration process should in theory narrow the prediction interval, as the step-wise addition of evaluation criteria improves the equivalent performance of the set of remaining parameters. Figure 4.7 indeed shows a wider prediction interval than shown in figure 3.1. However, the step-wise addition of criteria do not appear to significantly narrow this interval for the most part, when looking at a period-length of 3 months. This largely is the result of the yet to be excluded summer months, which also caused the stagnation of the performance improvement (observation point II; subsequent explanation on page 32 & 35). The exclusion of the periods, with the summer months as their starting months, sees a narrowing of the prediction intervals (figure 4.8). Figure 4.8 also shows an increased impact on the interval by the step-wise addition of criteria when compared to figure 4.7.

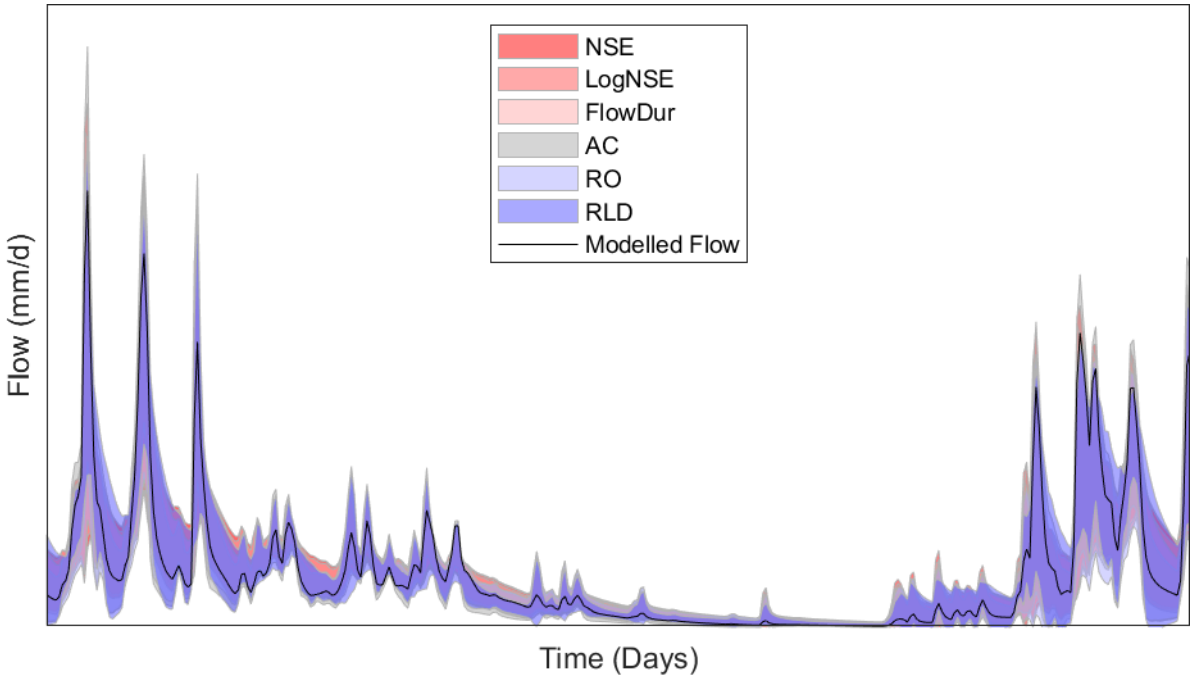


Figure 4.7: The prediction intervals after the step-wise addition of evaluation criteria for the parameters-sets that come through the calibration process for a period-length of 3 months. The modelled flow in this graph is produced by the model utilizing Par_x.

The reason why a more narrow prediction interval is desirable, is to ensure less "randomness" in the models performance, while utilizing one of the parameter-sets deemed useful by the calibration process.

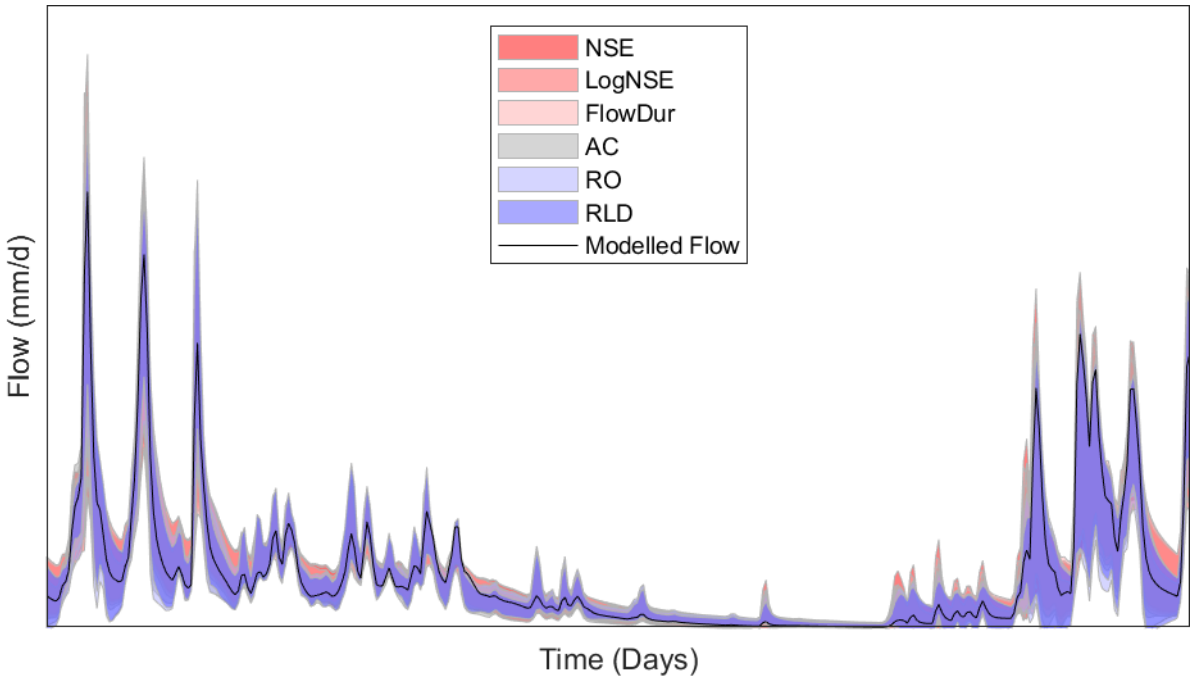


Figure 4.8: The prediction intervals after each criterion addition. The parameter-sets were evaluated over a 3 month period from which the results were excluded that began with the months May, June or July.

The prediction interval after the addition of the RLD appears, for the most part, to be fairly close to the prediction interval shown in figure 3.1. This entails that the predictability of the model calibrated over a period of 3 months and evaluated against all the mentioned criteria is on par with that of a model calibrated over a period of 10 years and evaluated solely against the NSE. Logically, the prediction interval becomes smaller as the calibration period-length becomes larger, which is viewed in the figures C.7 (6 months) & C.8 (1 year) from appendix C.

4.2. Validation

Figure 4.9 shows that each addition of an evaluation criterion during the calibration process leads to a better euclidean distance value while evaluating for the validation period. The criterion values, after the addition of the RLD, associated with the validation period, are very close to the values of the calibration period, which itself is fairly close to the values of the calibration benchmark. This would lead one to suspect that these parameter-sets are able to reproduce the hydrological signatures and thereby ensure model performance consistency.

Evaluation criterion	Individual value
NSE	0.8403
LogNSE	0.7560
FlowDur	0.8378
AC	0.8361
RO	−1.6463
RLD	0.9927
PeakDis	0.8274

Table 4.4: The values of the evaluation of the model using Par_x, over the validation period, for every individual criterion.

Figure 4.11 seems to strengthen this believe as the individual criteria all show improvement after each addition in the calibration process, when evaluating over the validation period. The individual values of the evaluation criteria for the validation period seem to be very close to the values for the benchmark calibration period, when the model utilises parameter-set Par_x (table 4.3 and table 4.4). This entails that Par_x will most likely ensure consistency for this specific catchment.

Although most individual criteria show improvement after each addition in the calibration process (figure 4.4), it does not seem to be the case for the runoff coefficient criterion and to some extent the autocorrelation criterion. As stated before, the marginal improvement of the runoff coefficient criterion value is not a big concern when it is used for judging a parameter-set's performance outside the calibration period for a period with a larger duration. The autocorrelation criterion values show a small improvement. However, visually well performing parameter-sets, over a large period, mostly perform well in terms of the autocorrelation criterion and visa versa.

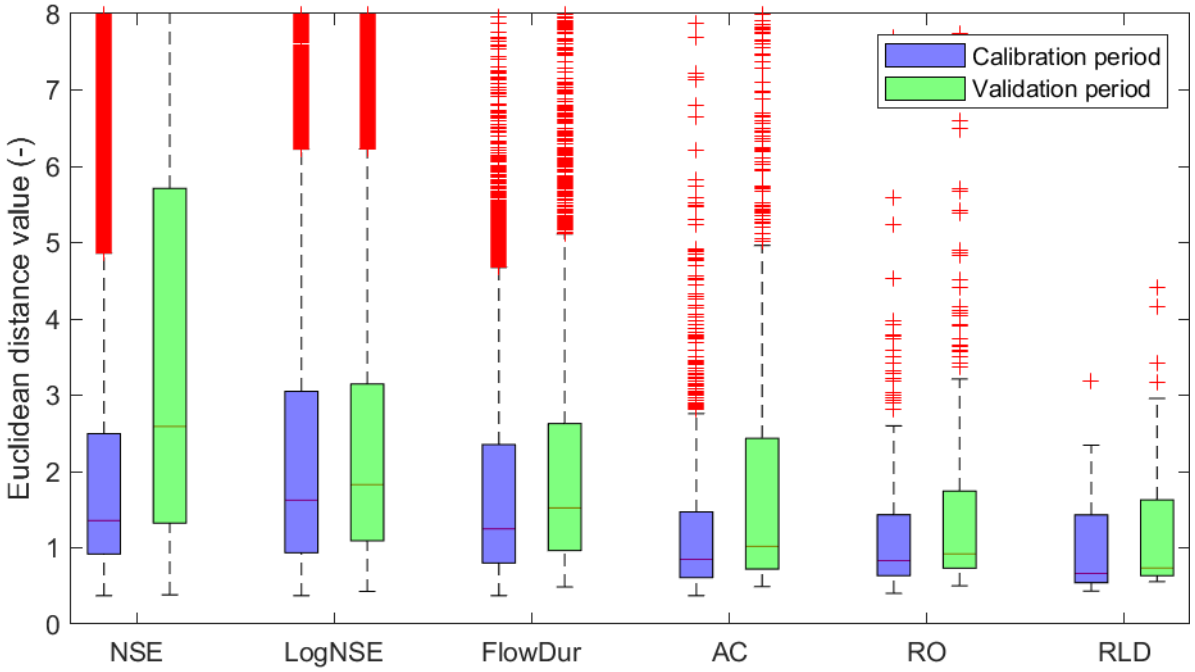


Figure 4.9: Equivalent euclidean distance values after every criterion addition versus the euclidean distance values of the validation period, also using the parameter-sets that passed the thresholds after each criterion addition. Only the outcomes of the 3 month period-length calibrations are depicted in this graph. The exclusion of the aforementiod starting months leads to no-data after the addition of the PeakDis.

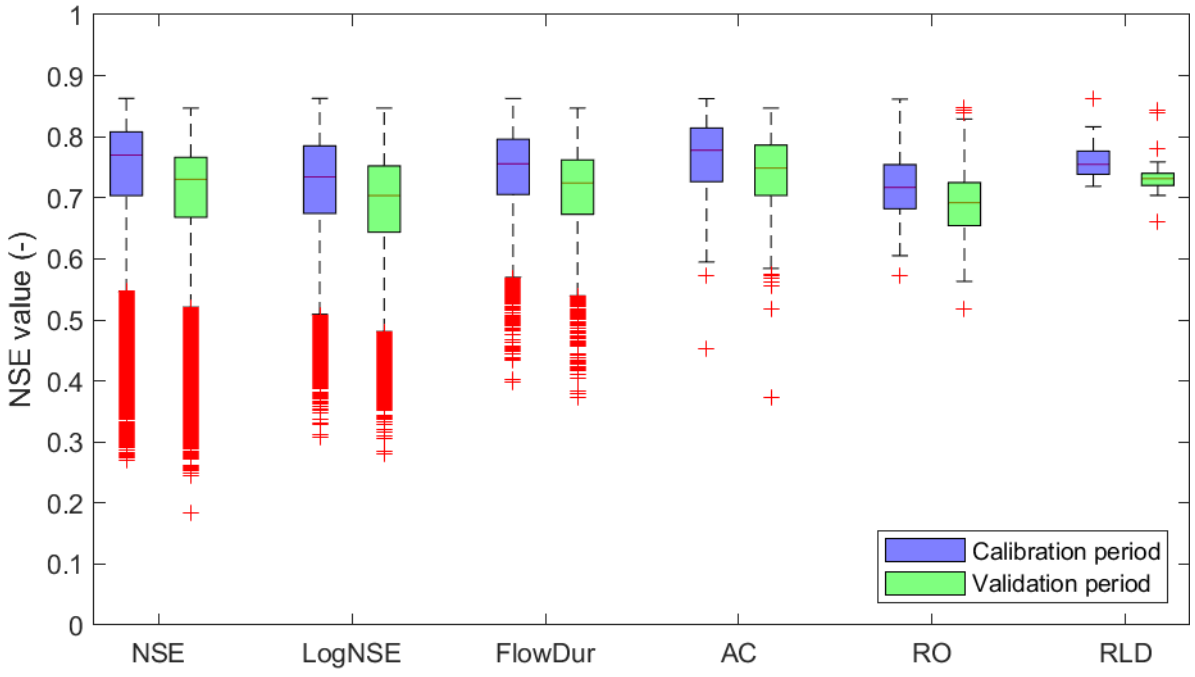


Figure 4.10: Equivalent NSE values after every criterion addition versus the NSE values of the validation period. Only the outcomes of the 3 month period-length calibrations are depicted in this graph. The exclusion of the aforementioned starting months leads to no-data after the addition of the PeakDis.

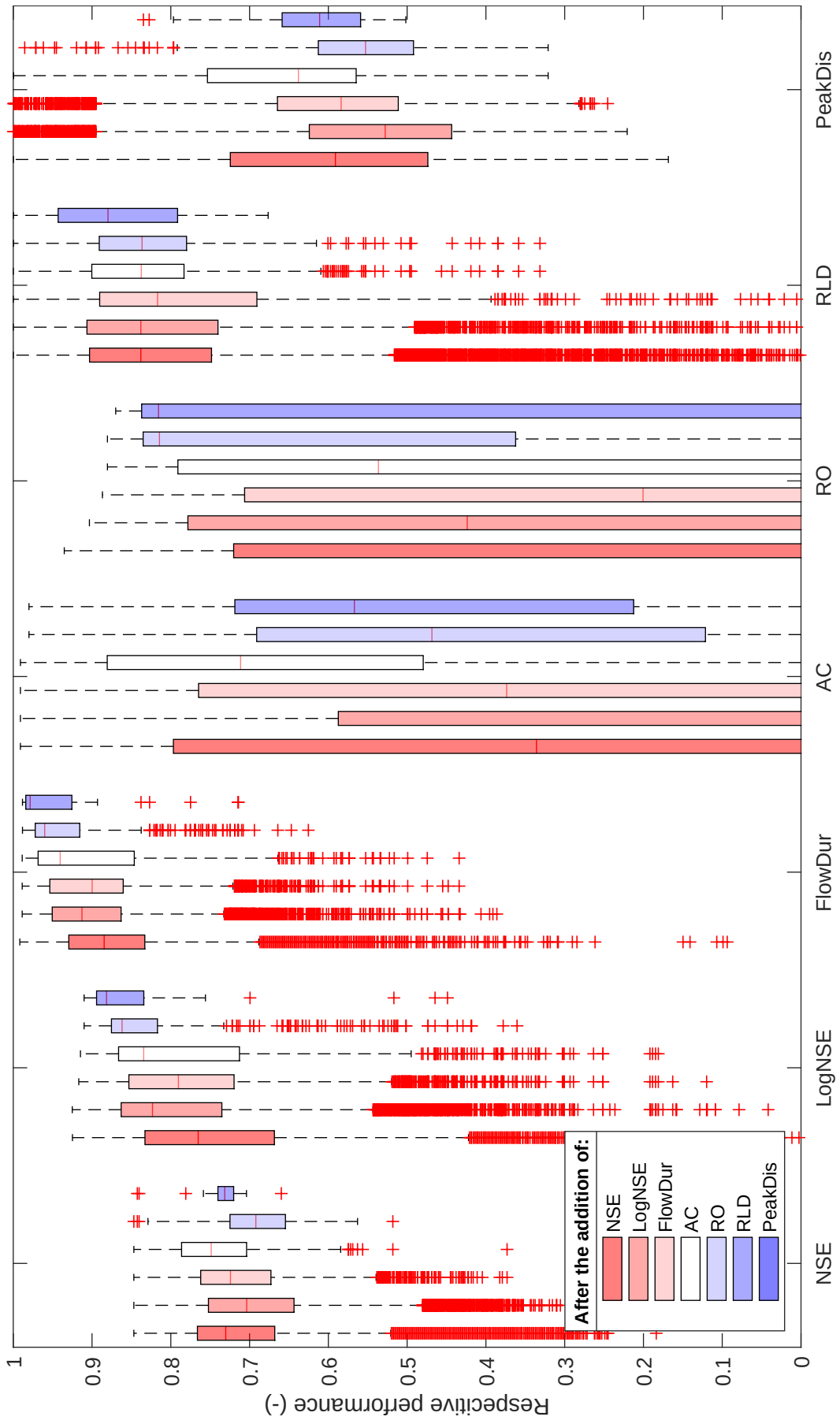


Figure 4.11: A depiction of the individual criterion values using the parameter-sets that pass the threshold after each criterion addition during the 3 month period-length calibration. The parameter-sets are validated against the entire validation period of observed stream-flow.

5

Discussion

5.1. Criteria thresholds

First point of attention would be the relaxation of the thresholds. As could be observed in figures 4.2 & 4.5 and be read in the subsequent explanation in section 4.1, the absence of remaining parameter-sets, evaluated over the larger periods, can be attributed to the strict thresholds of the evaluation criteria. In addition, the sole remaining parameter set (subsection 4.1.2), from the 3 months period-length calibration after the implementation of every criterion, is most likely also the product of strict thresholds. The incorporation of the knowledge concerning the starting months of calibration for the 3 month periods left the outcome with no remaining parameter-sets that passed all the thresholds of their respective criteria. However, e.g. Par_x visually outperformed the sole remaining parameter-set, though it did not pass all the thresholds (PeakDis). It would therefore be desirable that parameter-sets like and similar to Par_x would remain after the addition of every criterion during the calibration process. This could lead to more helpful way of determining the best performing parameter-set.

5.2. Runoff coefficient criterion

The equivalent performance of the remaining parameters-sets after the addition of the runoff coefficient criterion is questionable for most of the calibration period-lengths, in particular those of 1 year and shorter. As mentioned in subsection 4.1.2, a low equivalent value of the runoff coefficient (e.g. table 4.3) does not necessarily indicate poor performance; these parameter-sets still are able to mimic the hydro-graph rather well. However, the low equivalent RO values do affect the equivalent euclidean distance values; this would make the visual representation of the performance, as shown in figures 4.2 and 4.5, biased but still very much useful. More alarming is the apparent inconsistency of the RO value during calibration and the equivalent RO value. A parameter-set that passes the RO threshold (e.g. Par_x) could have a subzero equivalent RO value. One would expect

a lower equivalent value in comparison to the value encountered during the calibration process. However, the equivalent value is still expected to indicate some level of good performance in terms of RO over the benchmark[1] period. The inconsistency is most likely the result of taking the Nash-Sutcliffe coefficient of the runoff coefficient signal, as is displayed in equation (3.6) and (3.7). The Nash-Sutcliffe coefficient lays a large emphasis on the larger differences (section 3.1; page 20) and a difference between the modelled and observed weekly runoff coefficient occurs rather quickly as a result of e.g. one lower peak-flow. These two factors combined will lead a criterion that is highly sensitive for possibly the wrong reasons, as this sensitivity could result in low equivalent RO values for reasonable performing parameter-sets. An easy solution would be making use of the runoff coefficient over the entire calibration period instead of creating a signal of the weekly runoff coefficients. This "global" runoff coefficient sacrifices the trend that the signal created but thereby bypasses the problems mentioned above.

$$E = 1 - \left| 1 - \frac{Ru_m}{Ru_{obs}} \right| \tag{5.1}$$

where

$$Ru_x = \frac{\sum Q_x}{\sum P} \tag{5.2}$$

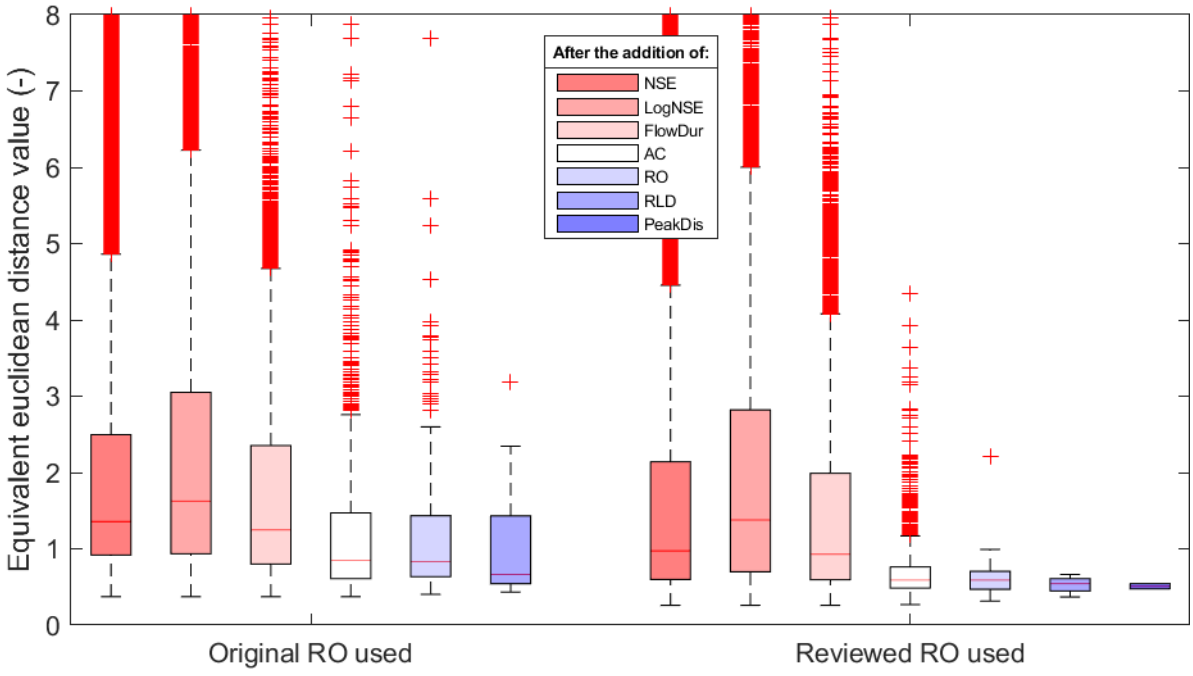


Figure 5.1: Equivalent perfomance of the parameter-sets that passed each addition of evaluation criteria over a 3 month period. The cluster of boxplots on the left used the RO criterion according to equations (3.6) & (3.7) and the cluster to the right according to equations (5.1) & (5.2).

[1]The inconsistency of the RO also persists throughout the validation period as is seen in figure 4.11.

The necessity of a trend in this criterion is somewhat dubious as this already is covered by other evaluation criteria (e.g. AC). This global runoff coefficient would then represent the long term water balance (Yadav et al., 2007; Sawicz et al., 2011) as was described in section 3.1.

When looking at the comparison between the original RO and the reviewed RO (figure 5.1), in terms of equivalent performance (euclidean distance) after each criterion addition, the improvement of performance is significant both in general and after the addition of the runoff coefficient criterion. Also notable is the reintroduction of remaining parameter-sets after the addition of the PeakDis, as figure 5.1 took into account the information gained from figure 4.4. The improvement also, by making use of the reviewed RO, is significant for larger calibration period-lengths (e.g. figure D.1; appendix D), even more so in comparison to the benchmark (figure D.2; appendix D).

A necessary reiteration is: it is an inconsistency between the values during calibration and the equivalent values/performance. However, when one would try to calibrate a model on a poorly gauged catchment, one would not have access to enough data to look at the equivalent performance. The equivalent performance only serves as a measurement of what one could expect from a calibration over a shorter period. The original runoff coefficient criterion did perform in terms of filtering "faulty" parameter-sets. Therefore the results, while using the original RO, are deemed useful, especially for the smaller calibration periods.

5.3. Identification

As mentioned in section 5.2, the availability of data, when modelling a poorly gauged catchment, is not sufficient for the creation of equivalent performance. However, as mentioned in subsection 4.1.2, the determination of Par_x was based on the highest equivalent NSE value. This entails the search of another method of determining/identifying the "best" parameter-set. A first idea could have to do something with the evaluation criteria values obtained during the calibration over a certain period, though the method of choosing a parameter-set based on the smallest euclidean distance has already been refuted (section 4.1). However, with neither expert knowledge about the criterion weights nor about parameter constraints, the options become limited.

	Calibration value		**Equivalent value**	
Evaluation Criterion	**Par_x**	**Par_z**	**Par_x**	**Par_z**
NSE	0.8007	0.8474	0.8617	0.7597
LogNSE	0.9626	0.9436	0.7410	0.8883
FLowDur	0.9777	0.9830	0.8371	0.9790
AC	0.9008	0.9143	0.8595	0.8141
RO	0.9818	0.9588	−2.1618	0.9082
RLD	0.9767	0.9969	0.9354	0.9950
PeakDis	0.4285	0.7241	0.8948	0.7278

Table 5.1: A comparison between Par_x and Par_z

An other possible option is the summation of the criterion values. This bypasses the penalty on individual criteria differences but yields the same outcome as taking the euclidean distance. This outcome is dubbed Par_z.

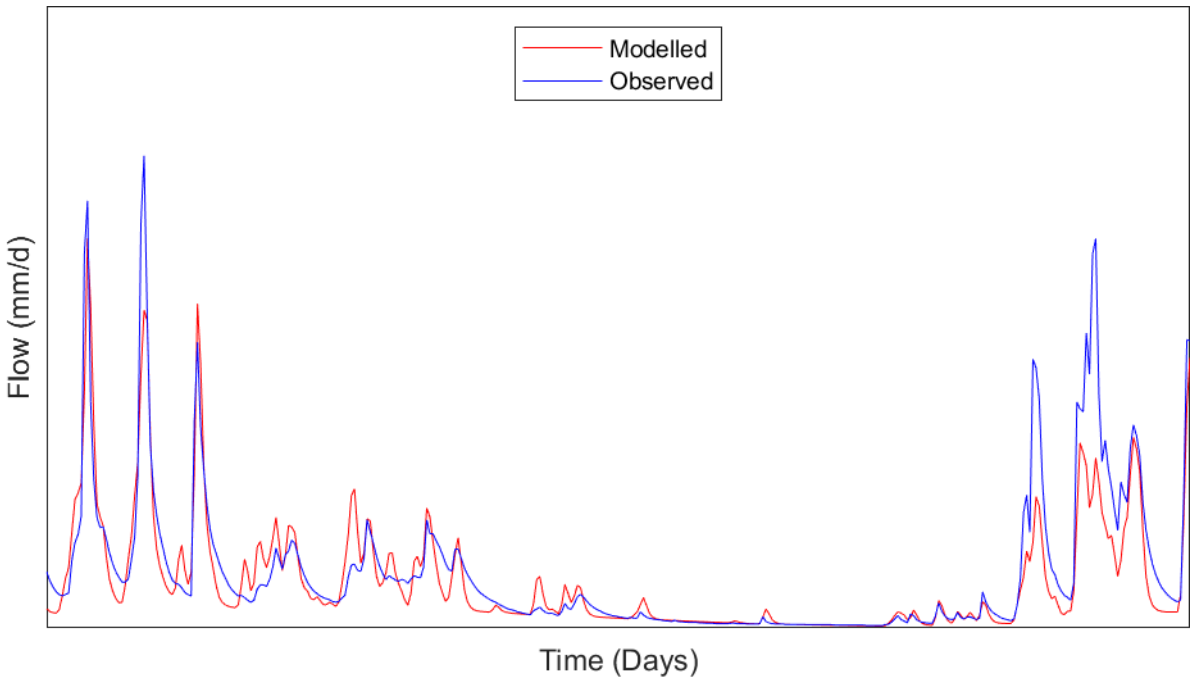


Figure 5.2: Modelled versus Observed, where the model utilises the parameter-set Par_z.

Figure 5.2 shows a visually worse performance compared to the one seen in figure 4.6 (Par_x). However, when looking at table 5.1, it becomes apparent that the model using Par_z is up to par or even performing better than the model using Par_x, most notably seen in regards to the original equivalent runoff coefficient value. A critical notion is: none of the evaluation criteria are weighted, which infers that not much is known about the relative importance of those criteria. Thereby, three parameters[2] differ greatly as can be seen in table D.1 from appendix D. A higher β and C_o tend to lead to the spiky behaviour observed in figure 5.2, which can be concluded from the equations of table A.1; appendix A. This entails that the problem of equifinality could lurk in the remaining parameter-sets.

5.4. Parameter sensitivity

Parameter sensitivity signifies how well a parameter can be identified[3] within the parameter space (Fenicia et al., 2008). If a parameter cannot be well identified then its corresponding constitutive equation or even the model holds little correspondence with reality (Kleissen et al., 1990; Fenicia et al., 2008). This thesis follows the same approach of identifying the parameter-sensitivity as is explained in Fenicia et al. (2008), which is an approach described by Freer et al. (2004) and based on the Regional Sensitivity Analysis (RSA) (Spear and Hornberger, 1980).

[2]The remainder of the parameters were of the same order of magnitude.
[3]A well identified parameter is one that will hover around a certain value within the parameter space when it has a high corresponding criterion value.

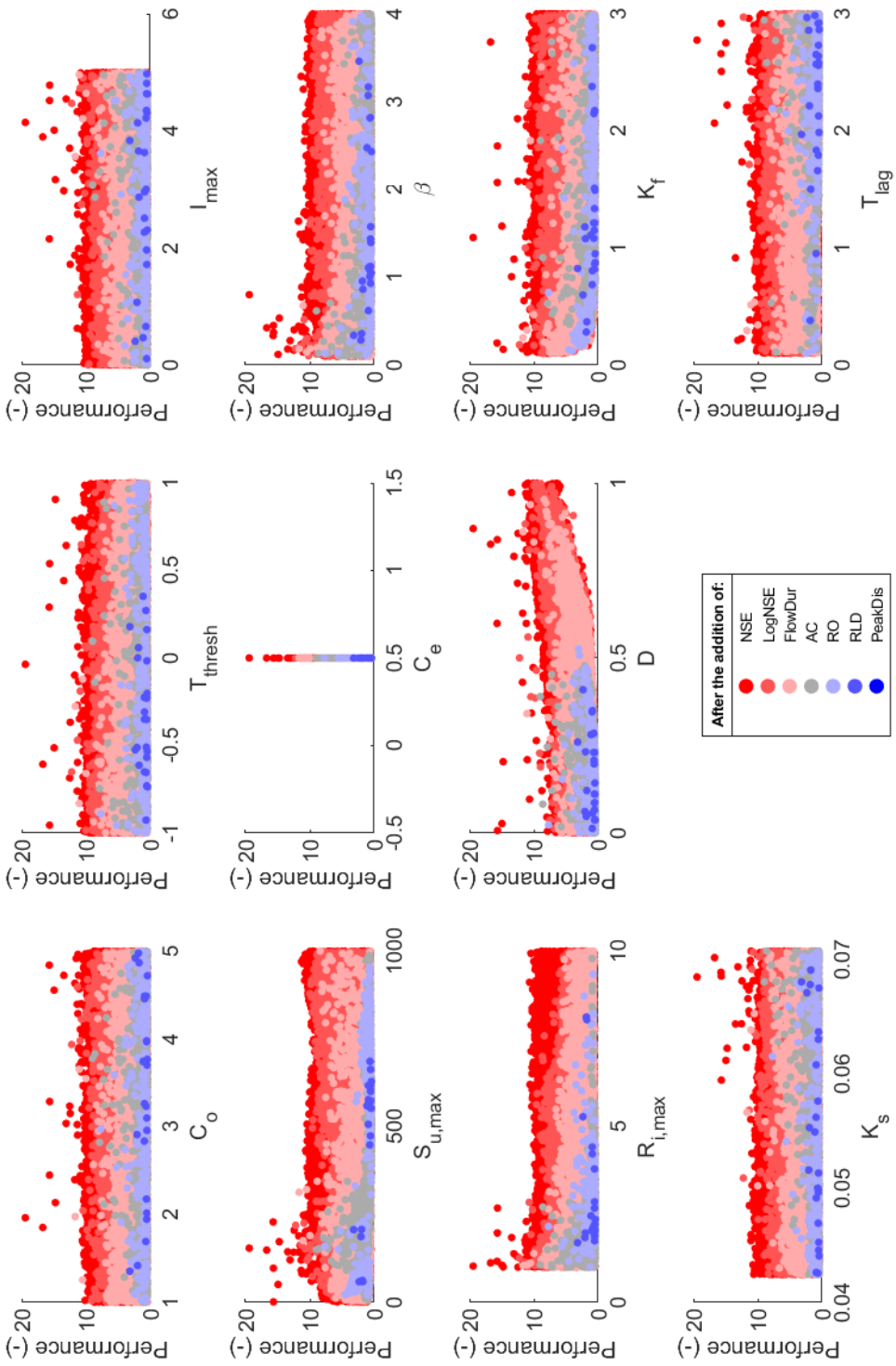


Figure 5.3: Sensitivity of each parameter is displayed after each criterion addition. The performance shown in this figure is the equivalent euclidean distance. The parameters used in this figure are from the remaining parameter-sets of the 3 months period-length calibration. Knowledge of starting months has been assimilated.

As mentioned in Fenicia et al. (2008), the RSA is based on random sampling. This, however, is fairly inefficient in terms of calculation-time. So, in similar fashion to Fenicia et al. (2008), the same parameter-sets are used that were generated for the Monte-Carlo method calibration. The approach is visually represented by figure 5.3. Figure 5.3 shows a great influence of the evaluation criteria on the sensitivity of the parameters. E.g. after the addition of only the NSE the parameters are highly insensitive while after the addition of the RLD most parameters show some sensitivity and can be identified. Parameters like I_{max} and C_o show no sensitivity, which could signify a deficiency in realism of those model components. $S_{u,max}$ shows two clusters after the addition of the RLD, which could be the equifinality mentioned in section 5.3. Taking into account the information regarding the starting month of calibration significantly improves the sensitivity (figure D.3; appendix D).

5.5. Starting month

As explained in subsection 4.1.2, there appears to be a trend in performance in regards to the starting month of calibration. This has thus far been shown for the 3 months calibration period-length. The larger period-lengths also show some form of trend in this regard (figures D.6 & D.8; appendix D). However, from a period larger than 6 months onwards, it becomes less as to what causes this trend, which would limit the applicability of this knowledge. Thereby lies the focus on creating meaningful outcomes with less data, placing an emphasis on e.g. the 3 months period length. An other point of interest is the apparent isolation of a “best” starting month for a period-length of 3 months (figure D.5; appendix D), which appears to be the month direcly after the period of change (subsection 4.1.2), which in this case would be October. Furthermore, from a period-length of 6 months and larger, there does not appear to be a noticeable trend in regards to a best starting month (figures D.7 & D.9; appendix D). These trends are either insignificant/slightly random (figure D.7; appendix D) or coincide with the worst performing starting months (figure D.8 combined with figure D.9; appendix D). Although it seems a very useful piece of information, the lack of data available in the area for which this thesis is intended highly limits the applicability of this finding.

6

Conslusion & Recommendations

6.1. Conclusion

The main research question of this thesis read:

> *"To what extend is it possible to shorten the period-length of calibration data by using multiple evaluation criteria, for the model to perform similar to being calibrated over a period of 10 years using one evaluation criterion?"*

From the results (subsection 4.1.1; chapter 4) one can conclude that a period of 6 months, with all criteria used in the calibration process, ensures similar performance to calibration over 10 years of data using one criterion (e.g. the NSE). Calibration over 3 months worth of data also ensured similar performance after the implementation of extra knowledge (subsection 4.1.2). For a period-length of 3 months, sections 4.1 and 4.2 showed that the remaining parameter-sets, could ensure both performance (figures 4.5 & 4.6; table 4.3) and consistency (figures 4.9 & 4.11; table 4.4).

The additional research question read:

> *"Is the addition of multiple evaluation criteria enough to see the model perform similarly, while calibrated over a smaller period, to the same model that has been calibrated over a period of 10 years?"*

When 6 months worth of data are used then it can be concluded that this in fact is the case, after looking at the results (subsection 4.1.1). As mentioned in the answer to the main research question, calibration over 3 month of data will require some additional information. The difference between figures 4.2 and 4.5 showed that knowledge concerning the calibration starting month is vital in this regard. Thereby comes the fact that during this thesis no use has been made of parameter- and process constraints, with the idea of viewing the influence of additional evaluation criteria. However, as mentioned in sections 5.3 and 5.4 and seen in figure 5.3, this could give rise to problem of equifinality.

Equifinality could not be prevented by evaluation criteria alone when the calibration period length is as small as 3 months. Equifinality appears to be already less of a concern when 6 months of data were used in the calibration process (figure D.4; appendix D).

6.2. Recommendations

Weighted criteria
For further research, a first recommendation would be to use weighted evaluation criteria. Mentioned in section 4.1, the euclidean distance, with equal weight for all the criteria, is not an adequate tool for the determination of the best performing parameter-set. As the evaluation criteria are measurements for testing different hydrological signatures and hydrological signatures differ from one another between catchments, it would therefore seem logical to weigh the evaluation criteria. A catchment where the main contributory to the main stream-flow is rapid overland-flow would see a higher weight for the Nash-Sutliffe efficiency criterion, as rapid overland-flow produces high and sharp peaks. This in turn coincides well with the NSE (section 3.1).

Parameter identification
The development of a method for parameter-set identification is something to be developed in the future. Within the space of this thesis, there did not appear to be a clear-cut method of isolating the parameter-set Par_x, which visually seemed to outperform those parameter-sets that could be identified, like e.g. the sole remaining parameter-set and Par_z.

Expert knowledge
Although the weighing of criteria is a form of expert knowledge, this paragraph concerns itself with parameter and process constraints. Sections 5.3 and 5.4 mentioned the apparent problem of equifinality, which signifies multiple 'behavioural' parameter-sets. Section 6.1 stated that evaluation criteria alone cannot stop equifinality from occurring, when the calibration period becomes small. Kelleher et al. (2017) showed that equifinality could be reduced by applying parameter- and process constraints to the model and the calibration process. This could e.g. be the determination of a smaller interval for I_{max} from literature, where different magnitudes are mentioned for different respective vegetation and ground coverage.

Calibration method
As the Monte-Carlo method is brute-force, it tends to take its toll on the calculation time. This especially is a concern for older devices. Therefore a more sophisticated calibration method is recommended.

Sampling
Much like Wang et al. (2017), the results of this thesis suggest that there is a informative period to collect stream-flow data for calibration-purposes or at least a period in which data should not be collected. Section 4.1 and figure 4.4 showed that periods of meteorological change should be avoided when collecting data over a small period for calibration.

References

C. Aguilar and M. J. Polo. Generating reference evapotranspiration surfaces from the hargreaves equation at watershed scale. *Hydrology and Earth System Sciences*, 15(8):2495–2508, 2011. doi: 10.5194/hess-15-2495-2011. URL https://www.hydrol-earth-syst-sci.net/15/2495/2011/.

W. R. Berghuijs, M. Sivapalan, R. A. Woods, and H. H. G. Savenije. Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales. *Water Resources Research*, 50(7):5638–5661, 2014. doi: 10.1002/2014WR015692. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014WR015692.

S. Bergström. *Development and Application of a Conceptual Runoff Model for Scandinavian Catchments*, volume 134 pp. 01 1976.

P. Blair and W. Buytaert. Socio-hydrological modelling: a review asking "why, what and how?". *Hydrology and Earth System Sciences*, 20(1):443–478, 2016. doi: 10.5194/hess-20-443-2016. URL https://www.hydrol-earth-syst-sci.net/20/443/2016/.

T Chapman. A comparison of algorithms for stream flow recession and baseflow separation. *Hydrological Processes*, 13(5):701–714, 4 1999. doi: 10.1002/(SICI)1099-1085(19990415)13:5<701::AID-HYP774>3.0.CO;2-2. URL https://doi.org/10.1002/(SICI)1099-1085(19990415)13:5<701::AID-HYP774>3.0.CO;2-2.

I. F. Creed, A. T. Spargo, J. A. Jones, J. M. Buttle, M. B. Adams, F. D. Beall, E. G. Booth, J. L. Campbell, D. Clow, K. Elder, M. B. Green, N. B. Grimm, C. Miniat, P. Ramlal, A. Saha, S. Sebestyen, D. Spittlehouse, S. Sterling, M. W. Williams, R. Winkler, and H. Yao. Changing forest water yields in response to climate warming: results from long-term experimental watershed sites across north america. *Global Change Biology*, 20(10):3191–3208, 2014. doi: 10.1111/gcb.12615. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.12615.

T. Euser, H. C. Winsemius, M. Hrachowitz, F. Fenicia, S. Uhlenbrook, and H. H. G. Savenije. A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences*, 17(5):1893–1912, 2013. doi: 10.5194/hess-17-1893-2013. URL https://www.hydrol-earth-syst-sci.net/17/1893/2013/.

F. Fenicia, H. H. G. Savenije, P. Matgen, and L. Pfister. Is the groundwater reservoir linear? learning from data in hydrological modelling. *Hydrology and Earth System Sciences*, 10(1):139–150, 2006. doi: 10.5194/hess-10-139-2006. URL https://www.hydrol-earth-syst-sci.net/10/139/2006/.

F. Fenicia, H. H. G. Savenije, P. Matgen, and L. Pfister. Understanding catchment behavior through stepwise model concept improvement. *Water Resources Research*, 44(1), 2008. doi: 10.1029/2006WR005563. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006WR005563.

F. Fenicia, D. Kavetski, and H. H. G. Savenije. Elements of a flexible approach for conceptual hydrological modeling: 1. motivation and theoretical development. *Water Resources Research*, 47(11), 11 2011. doi: 10.1029/2010WR010174. URL https://doi.org/10.1029/2010WR010174.

J. E. Freer, H. McMillan, J. J. McDonnell, and K. J. Beven. Constraining dynamic topmodel responses for imprecise water table information using fuzzy rule based performance measures. *Journal of Hydrology*, 291(3):254 – 277, 2004. doi: https://doi.org/10.1016/j.jhydrol.2003.12.037. URL http://www.sciencedirect.com/science/article/pii/S0022169404000356.

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

M. Guimberteau, P. Ciais, A. Ducharne, J. P. Boisier, A. P. Dutra Aguiar, H. Biemans, H. De Deurwaerder, D. Galbraith, B. Kruijt, F. Langerwisch, G. Poveda, A. Rammig, D. A. Rodriguez, G. Tejada, K. Thonicke, C. Von Randow, R. C. S. Von Randow, K. Zhang, and H. Verbeeck. Impacts of future deforestation and climate change on the hydrology of the amazon basin: a multi-model analysis with a new set of land-cover change scenarios. *Hydrology and Earth System Sciences*, 21(3):1455–1475, 2017. doi: 10.5194/hess-21-1455-2017. URL https://www.hydrol-earth-syst-sci.net/21/1455/2017/.

G. H. Hargreaves and G. A. Richard. History and evaluation of hargreaves evapotranspiration equation. *Journal of Irrigation and Drainage Engineering*, 129(1):53–63, 2003. doi: 10.1061/(ASCE)0733-9437(2003)129:1(53). URL https://ascelibrary.org/doi/abs/10.1061/.

Z. H. He, F. Q. Tian, H. V. Gupta, H. C. Hu, and H. P. Hu. Diagnostic calibration of a hydrological model in a mountain area by hydrograph partitioning. *Hydrology and Earth System Sciences*, 19(4):1807–1826, 2015. doi: 10.5194/hess-19-1807-2015. URL https://www.hydrol-earth-syst-sci.net/19/1807/2015/.

M. Hrachowitz and M. P. Clark. Hess opinions: The complementary merits of competing modelling philosophies in hydrology. *Hydrology and Earth System Sciences*, 21(8):3953–3973, 2017. doi: 10.5194/hess-21-3953-2017. URL https://www.hydrol-earth-syst-sci.net/21/3953/2017/.

M. Hrachowitz, O. Fovet, L. Ruiz, T. Euser, S. Gharari, R. Nijzink, J. Freer, H. H. G. Savenije, and C. Gascuel-Odoux. Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. *Water Resources Research*, 50(9):7445–7469, 2014. doi: 10.1002/2014WR015484. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014WR015484.

S. K. Jain and K. P. Sudheer. Fitting of hydrologic models: A close look at the nash–sutcliffe index. *Journal of Hydrologic Engineering*, 13(10):981–986, 2008. doi: 10.1061/(ASCE)1084-0699(2008)13:10(981). URL https://ascelibrary.org/doi/abs/10.1061/.

C. Jothityangkoon, M. Sivapalan, and D. L. Farmer. Process controls of water balance variability in a large semi-arid catchment: downward approach to hydrological model development. *Journal of Hydrology*, 254(1):174 – 198, 2001. doi: https://doi.org/10.1016/S0022-1694(01)00496-6. URL http://www.sciencedirect.com/science/article/pii/S0022169401004966.

J. Juston, J. Seibert, and P. Johansson. Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment. *Hydrological Processes*, 23(21):3093–3109, 2009. doi: 10.1002/hyp.7421. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.7421.

C. Kelleher, B. McGlynn, and T. Wagener. Characterizing and reducing equifinality by constraining a distributed catchment model with regional signatures, local observations, and process understanding. *Hydrology and Earth System Sciences*, 21(7):3325–3352, 2017. doi: 10.5194/hess-21-3325-2017. URL https://www.hydrol-earth-syst-sci.net/21/3325/2017/.

J. W. Kirchner. Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. *Water Resources Research*, 45(2), 2009. doi: 10.1029/2008WR006912. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008WR006912.

F. M. Kleissen, M. B. Beck, and H. S. Wheater. The identifiability of conceptual hydrochemical models. *Water Resources Research*, 26(12):2979–2992, 1990. doi: 10.1029/WR026i012p02979. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR026i012p02979.

M.I. L'vovich and G.F. White. Use and transformation of terrestrial water systems. *The earth as transformed by human action: global and regional changes in the biosphere over the past 300 years*, pages 235–252, 01 1990.

W. H. Maes, P. Gentine, N. E. C. Verhoest, and D. G. Miralles. Potential evaporation at eddy-covariance sites across the globe. *Hydrology and Earth System Sciences*, 23(2):925–948, 2019. doi: 10.5194/hess-23-925-2019. URL https://www.hydrol-earth-syst-sci.net/23/925/2019/.

A. Montanari and E. Toth. Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins? *Water Resources Research*, 43(5), 5 2007. doi: 10.1029/2006WR005184. URL https://doi.org/10.1029/2006WR005184.

R. D. Moore. Storage-outflow modelling of streamflow recessions, with application to a shallow-soil forested catchment. *Journal of Hydrology*, 198(1):260–270, 1997. doi: https://doi.org/10.1016/S0022-1694(96)03287-8. URL http://www.sciencedirect.com/science/article/pii/S0022169496032878.

E. Morin, K. P. Georgakakos, U. Shamir, R. Garti, and Y. Enzel. Objective, observations-based, automatic estimation of the catchment response timescale. *Water Resources Research*, 38(10):30–1–30–16, 2002. doi: 10.1029/2001WR000808. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001WR000808.

R. Moussa and N. Chahinian. Comparison of different multi-objective calibration criteria using a conceptual rainfall-runoff model of flood events. *Hydrology and Earth System Sciences*, 13(4):519–535, 2009. doi: 10.5194/hess-13-519-2009. URL https://www.hydrol-earth-syst-sci.net/13/519/2009/.

M. K. Muleta. Model performance sensitivity to objective function during automated calibrations. *Journal of Hydrologic Engineering*, 17(6), 2011. doi: 10.1061/(ASCE)HE.1943-5584.0000497. URL https://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0000497.

J. E. Nash and J.V. Sutcliffe. River flow forecasting through conceptual models part i — a discussion of principles. *Journal of Hydrology*, 10(3):282 – 290, 1970. doi: https://doi.org/10.1016/0022-1694(70)90255-6. URL http://www.sciencedirect.com/science/article/pii/0022169470902556.

R. Nijzink, C. Hutton, I. Pechlivanidis, R. Capell, B. Arheimer, J. Freer, D. Han, T. Wagener, K. McGuire, H. Savenije, and M. Hrachowitz. The evolution of root-zone moisture capacities after deforestation: a step towards hydrological predictions under change? *Hydrology and Earth System Sciences*, 20(12):4775–4799, 2016. doi: 10.5194/hess-20-4775-2016. URL https://www.hydrol-earth-syst-sci.net/20/4775/2016/.

Y. A. Pachepsky, G. Martinez, F. Pan, T. Wagener, and T. Nicholson. Evaluating hydrological model performance using information theory-based metrics. *Hydrology and Earth System Sciences Discussions*, 2016:1–24, 2016. doi: 10.5194/hess-2016-46. URL https://www.hydrol-earth-syst-sci-discuss.net/hess-2016-46/.

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

S. Pool, M. J. P. Vis, R. R. Knight, and J. Seibert. Streamflow characteristics from modeled runoff time series – importance of calibration criteria selection. *Hydrology and Earth System Sciences*, 21(11):5443–5457, 2017. doi: 10.5194/hess-21-5443-2017. URL https://www.hydrol-earth-syst-sci.net/21/5443/2017/.

S. Razavi and B. A. Tolson. An efficient framework for hydrologic model calibration on long data periods. *Water Resources Research*, 49(12):8418–8431, 2013. doi: 10.

1002/2012WR013442. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2012WR013442.

L. Santos, G. Thirel, and C. Perrin. Technical note: Pitfalls in using log-transformed flows within the kge criterion. *Hydrology and Earth System Sciences*, 22(8):4583–4591, 2018. doi: 10.5194/hess-22-4583-2018. URL https://www.hydrol-earth-syst-sci.net/22/4583/2018/.

M. Sattari, A. Rezazadeh-Joudi, and A. Kusiak. Assessment of different methods for estimation of missing data in precipitation studies. *Hydrology Research*, 48, 09 2016. doi: 10.2166/nh.2016.364.

H. H. G. Savenije, A. Y. Hoekstra, and P. van der Zaag. Evolving water science in the anthropocene. *Hydrology and Earth System Sciences*, 18(1):319–332, 2014. doi: 10.5194/hess-18-319-2014. URL https://www.hydrol-earth-syst-sci.net/18/319/2014/.

K. Sawicz, T. Wagener, M. Sivapalan, P. A. Troch, and G. Carrillo. Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern usa. *Hydrology and Earth System Sciences*, 15(9):2895–2911, 2011. doi: 10.5194/hess-15-2895-2011. URL https://www.hydrol-earth-syst-sci.net/15/2895/2011/.

S. J. Schymanski and D. Or. Leaf-scale experiments reveal an important omission in the penman–monteith equation. *Hydrology and Earth System Sciences*, 21(2):685–706, 2017. doi: 10.5194/hess-21-685-2017. URL https://www.hydrol-earth-syst-sci.net/21/685/2017/.

J. Seibert and K. J. Beven. Gauging the ungauged basin: how many discharge measurements are needed? *Hydrology and Earth System Sciences*, 13(6):883–892, 2009. doi: 10.5194/hess-13-883-2009. URL https://www.hydrol-earth-syst-sci.net/13/883/2009/.

J. Seibert and J. J. McDonnell. On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research*, 38(11):23–1–23–14, 2002. doi: 10.1029/2001WR000978. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001WR000978.

E. Shamir, B. Imam, E. Morin, H. V. Gupta, and S. Sorooshian. The role of hydrograph indices in parameter estimation of rainfall–runoff models. *Hydrological Processes*, 19 (11):2187–2207, 2005. doi: 10.1002/hyp.5676. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.5676.

R. C. Spear and G. M. Hornberger. Eutrophication in peel inlet—ii. identification of critical uncertainties via generalized sensitivity analysis. *Water Research*, 14(1):43 – 49, 1980. ISSN 0043-1354. doi: https://doi.org/10.1016/0043-1354(80)90040-8. URL http://www.sciencedirect.com/science/article/pii/0043135480900408.

L. M. Tallaksen. A review of baseflow recession analysis. *Journal of Hydrology*, 165(1): 349–370, 1995. doi: https://doi.org/10.1016/0022-1694(94)02540-R. URL http://www.sciencedirect.com/science/article/pii/002216949402540R.

E. Tarawneh, J. Bridge, and N. Macdonald. A pre-calibration approach to select optimum inputs for hydrological models in data-scarce regions. *Hydrology and Earth System Sciences*, 20(10):4391–4407, 2016. doi: 10.5194/hess-20-4391-2016. URL https://www.hydrol-earth-syst-sci.net/20/4391/2016/.

R. J. Thayyen and A. P. Dimri. Factors controlling slope environmental lapse rate (selr) of temperature in the monsoon and cold-arid glacio-hydrological regimes of the himalaya. *The Cryosphere Discussions*, 8:5645–5686, 2014. doi: 10.5194/tcd-8-5645-2014. URL https://www.the-cryosphere-discuss.net/8/5645/2014/.

F. Tian, Y. Sun, H. Hu, and H. Li. Searching for an optimized single-objective function matching multiple objectives with automatic calibration of hydrological models. *Hydrology and Earth System Sciences Discussions*, 2016:1–33, 2016. doi: 10.5194/hess-2016-88. URL https://www.hydrol-earth-syst-sci-discuss.net/hess-2016-88/.

USFS. Hj andrews experimental forest: long-term ecological research, July 2018. URL https://andrewsforest.oregonstate.edu.

T. Wagener, M. Sivapalan, P. A. Troch, B. L. McGlynn, C. J. Harman, H. V. Gupta, P. Kumar, P. S. C. Rao, N. B. Basu, and J. S. Wilson. The future of hydrology: An evolving science for a changing world. *Water Resources Research*, 46(5), 2010. doi: 10.1029/2009WR008906. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009WR008906.

L. Wang, H. J. van Meerveld, and J. Seibert. When should stream water be sampled to be most informative for event-based, multi-criteria model calibration? *Hydrology Research*, 48(6):1566–1584, 2017. doi: 10.2166/nh.2017.197. URL http://dx.doi.org/10.2166/nh.2017.197.

I. K. Westerberg and H. K. McMillan. Uncertainty in hydrological signatures. *Hydrology and Earth System Sciences*, 19(9):3951–3968, 2015. doi: 10.5194/hess-19-3951-2015. URL https://www.hydrol-earth-syst-sci.net/19/3951/2015/.

H. C. Winsemius, B. Schaefli, A. Montanari, and H. H. G. Savenije. On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. *Water Resources Research*, 45(12), 2009. doi: 10.1029/2009WR007706. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009WR007706.

M. Yadav, T. Wagener, and H. Gupta. Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, 30(8):1756 – 1774, 2007. ISSN 0309-1708. doi: https://doi.org/10.1016/j.advwatres.2007.01.005. URL http://www.sciencedirect.com/science/article/pii/S0309170807000140.

D. Yu, J. Yang, L. Shi, Q. Zhang, K. Huang, Y. Fang, and Y. Zha. On the uncertainty of initial condition and initialization approaches in variably saturated flow modeling. *Hydrology and Earth System Sciences Discussions*, 2018:1–42, 2018. doi: 10.5194/hess-2018-557. URL https://www.hydrol-earth-syst-sci-discuss.net/hess-2018-557/.

A

Flex_{nd} model

A.1. Constitutive Equations

Reservoir	**Water Balance Equation**	**Constitutive Equations**
Snow	$\frac{dS_n}{dt} = P - P_{eff}$ (A.1)	$P_{eff} = \begin{cases} C_o \times (T_a - T_{thresh}) & \text{if } T_a > T_{thresh} \\ 0 & \text{if } T_a \leq T_{thresh} \end{cases}$ (A.2)
Interception	$\frac{dS_i}{dt} = P_{eff} - E_i - P_s$ (A.3)	$P_s = max(0, S_i + P_{eff} - I_{max})$ (A.4) $E_i = min(E_{pot}, S_i - P_s)$ (A.5)
Unsaturated or Soil Moisture	$\frac{dS_u}{dt} = P_s - E_t - R_u - R_i$ (A.6)	$S_{u,m} = (1+\beta) S_{u,max} \left(1 - \left(1 - \frac{S_u}{S_{u,max}}\right)^{1/(1+\beta)}\right)$ (A.7) $R_u = P_s - S_{u,max} + S_u + S_{u,max} \left(1 - \frac{P_s + S_{u,m}}{(1+\beta) S_{u,max}}\right)^{1+\beta}$ (A.8) $E_t = \begin{cases} E_{pot} \frac{S_u}{0.5 \times S_{u,max}} & \text{if } S_u \leq 0.5 \times S_{u,max} \\ min(E_{pot}, S_u) & \text{if } S_u > 0.5 \times S_{u,max} \end{cases}$ (A.9) $R_i = R_{i,max} \frac{S_u}{S_{u,max}}$ (A.10)

Fast	$\frac{dS_f}{dt} = (1-D) \times R_u - Q_f$	(A.11)	$Q_f = S_f - S_f \times e^{K_f \times \Delta t}$ (A.12)
Slow	$\frac{dS_s}{dt} = D \times R_u + R_i - Q_s$	(A.13)	$Q_s = S_s - S_s \times e^{K_s \times \Delta t}$ (A.14)

Table A.1: Constitutive equation of the $FLEX_{nd}$-model.

A.2. Lag-function

Listing A.1: Weights of the lag function

```
function [ Weigths ] = Weigfun( Tlag )
nmax=ceil(Tlag);
if nmax==1
    Weigths=1;
else
    Weigths=zeros(1,nmax);
    th=Tlag/2;
    nh=floor(th);
    for i=1:nh
        Weigths(i)=(i-.5)/th;
    end
    i=nh+1;
    Weigths(i)=(1+(i-1)/th)*(th-floor(th))/2+(1+(Tlag-i)/th)
       *(floor(th)+1-th)/2;
    for i=nh+2:floor(Tlag)
        Weigths(i)=(Tlag-i+.5)/th;
    end
    if Tlag>floor(Tlag)
        Weigths(floor(Tlag)+1)=(Tlag-floor(Tlag)).^2/(2*th);
    end
end
Weigths=Weigths/sum(Weigths);
```

A.3. Matlab code of the model

Listing A.2: Code behind the FLEX$_{nd}$-model

```
function [Qm] = UsaMod2( Par, ExtraPar )
%Thesis model
Imax=Par(1);
Ce=Par(2);
Sumax=Par(3);
beta=Par(4);
Pmax=Par(5);
Tlag=Par(6);
Kf=Par(7);
Ks=Par(8);
Melt=Par(9);
Tth=Par(10);
D=Par(11);

Prec=ExtraPar.Forcing(:,1);
Ta=ExtraPar.Forcing(:,2);
Etp=ExtraPar.Forcing(:,3);
Sect=ExtraPar.Sect(1,:);
Tdis=ExtraPar.Sect(2,:);

tmax=length(Prec);
Sn=zeros(tmax,length(Sect(1,:)));
Si=zeros(tmax,1);
Su=zeros(tmax,1);
Sf=zeros(tmax,1);
Ss=zeros(tmax,1);
Eidt=zeros(tmax,1);
Eadt=zeros(tmax,1);
Pet=zeros(tmax,length(Sect(1,:)));
Qtotdt=zeros(tmax,1);

Sn(1,:)=ExtraPar.Sin(1);
Si(1)=ExtraPar.Sin(2);
Su(1)=ExtraPar.Sin(3);
Sf(1)=ExtraPar.Sin(4);
Ss(1)=ExtraPar.Sin(5);

dt=1;

%%
% Flex Model_nd
for j=1:tmax
    Pdt=Prec(j)*dt;
```

```
44      Epdt=Etp(j)*dt;
45      % Snow Reservoir
46      Tin=Tdis+Ta(j);
47      for z=1:length(Sect(1,:))
48          if Tin(z) > Tth
49              Pes=min(Sn(j,z),Melt*(Tin(z)-Tth));
50              Sn(j,z)=Sn(j,z)-Pes;
51              Pet(j,z)=(Pes+Pdt)*Sect(z);
52          else
53              Sn(j,z)=Sn(j,z)+Pdt;
54              Pet(j,z)=0;
55          end
56      end
57      Pes=sum(Pet(j,:));
58      if j<tmax
59          Sn(j+1,:)=Sn(j,:);
60      end
61      % Interception Reservoir
62      if Pes>0
63          Si(j)=Si(j)+Pes;
64          Pedt=max(0,Si(j)-Imax);
65          Si(j)=Si(j)-Pedt;
66          Eidt(j)=0;
67      else
68          % Evaporation only when there is no rainfall
69          Pedt=0;
70          Eidt(j)=min(Epdt,Si(j));
71          Si(j)=Si(j)-Eidt(j);
72      end
73      if j<tmax
74          Si(j+1)=Si(j);
75      end
76      %unsaturated reservoir
77      if Su(j)>Sumax
78          Sudt=Su(j)-Sumax;
79          Su(j)=Sumax;
80          Qufdt=Pedt+Sudt;
81      else
82          Sum=(1+beta)*Sumax*(1-(1-(Su(j)/Sumax))^(1/(1+beta))
                );
83          if Pedt+Sum>(1+beta)*Sumax
84              Qufdt=Pedt-Sumax+Su(j);
85              Su(j)=Su(j)+Pedt-Qufdt;
86          else
87              Qufdt=Pedt-Sumax+Su(j)+Sumax*(1-(Pedt+Sum)/((1+
                    beta)*Sumax))^(1+beta);
```

```
88              Su(j)=Su(j)+Pedt-Qufdt;
89          end
90      end
91      % Transpiration
92      Epdt=max(0,Epdt-Eidt(j));
93      if Su(j)>0.5*Sumax
94          Eadt(j)=min(Su(j),Epdt);
95          Su(j)=Su(j)-Eadt(j);
96      else
97          Eadt(j)=min(Su(j),Epdt*(Su(j)/(Sumax*Ce)));
98          Su(j)=Su(j)-Eadt(j);
99      end
100     % Percolation
101     Qusdt=min(Su(j),(Su(j)/Sumax)*Pmax*dt);
102     Su(j)=Su(j)-Qusdt;
103     if j<tmax
104         Su(j+1)=Su(j);
105     end
106     % Fast Reservoir
107     Sf(j)=Sf(j)+(1-D)*Qufdt;
108     Qfdt=Sf(j)-Sf(j)*exp(-dt*Kf);
109     Sf(j)=Sf(j)-Qfdt;
110     if j<tmax
111         Sf(j+1)=Sf(j);
112     end
113     % Slow Reservoir
114     Ss(j)=Ss(j)+Qusdt+D*Qufdt;
115     Qsdt=Ss(j)-Ss(j)*exp(-dt*Ks);
116     Ss(j)=Ss(j)-Qsdt;
117     if j<tmax
118         Ss(j+1)=Ss(j);
119     end
120     Qtotdt(j)=Qsdt+Qfdt;
121 end
122
123 % Check Water Balance
124 Sfn=sum(Sect.*Sn(tmax,:))+Si(tmax)+Ss(tmax)+Sf(tmax)+Su(tmax
        );
125 Sin=sum(ExtraPar.Sin);
126 WB=sum(Prec)-sum(Eidt)-sum(Eadt)-sum(Qtotdt)-Sfn+Sin;
127 %disp(WB)
128
129 Weigths=Weigfun(Tlag);
130 Qm=conv(Qtotdt,Weigths);
131 Qm=Qm(1:tmax);
```

B

Evalutation criteria & calibration

B.1. Criteria

B.1.1. Matlab code

Listing B.1: Creating a quantitative evaluation criterion out the rising limp density.

```
function [E] = RLD(Qmodel,Qdata,xst)

Thres=(nanmedian(Qdata)/nanmean(Qdata));

[dym,dfm] = RLDm(Qmodel,xst,Thres);
[dyd,dfd] = RLDm(Qdata,xst,Thres);

Lm=sum(dym-dfm)/length(dym);
Ld=sum(dyd-dfd)/length(dyd);

E=1-abs(1-Lm/Ld);
```

Listing B.2: The algorithm behind determining the rising limp density for a given period.

```
function [dy,df,x,Qf] = RLDm(Qdata,xst,Thres)
%% Data gets loaded here in preperation for the
   determination of the necessary minima and maxima.
[x,y] = findpeaks(Qdata);
TF = islocalmin(Qdata);
days=linspace(1,length(Qdata),length(Qdata))+xst-1;
df=days(TF);
dy=days(y);
Qf=Qdata(TF);
count=0;
```

```
10 if length(dy) > length(df)
11     dy(1)=[];
12     x(1)=[];
13 end
14 if length(df) > length(dy)
15     df(end)=[];
16     Qf(end)=[];
17 end
18 if isempty(df)==1 || isempty(dy)==1
19     dy=0;
20     df=0;
21 end
22 if df(1) > dy(1)
23     dy(1)=[];
24     x(1)=[];
25     df(end)=[];
26     Qf(end)=[];
27 end
28 if isempty(df)==1 || isempty(dy)==1
29     dy=0;
30     df=0;
31 end
32
33 %% determine maxima and minima
34
35 for j=1:length(x)
36     j=j-count;
37     if j>length(x)
38         break
39     end
40     if j>1
41         if 2*(x(j-1)-Qf(j)) < (x(j-1)-Qf(j-1)) && x(j)>x(j
               -1)
42             x(j-1)=[];
43             dy(j-1)=[];
44             df(j)=[];
45             Qf(j)=[];
46             count=count+1;
47         end
48     end
49 end
50 for j=1:length(x)
51     if j>length(x)
52         break
53     end
54     while x(j)-Qf(j)<Thres
```

```
55          x(j)=[];
56          dy(j)=[];
57          if  j+1>length(Qf)
58              Qf(end)=[];
59              df(end)=[];
60              break
61          end
62          if  Qf(j) > Qf(j+1) || Qf(j)/Qf(j+1)<1
63              Qf(j)=[];
64              df(j)=[];
65          else
66              Qf(j+1)=[];
67              df(j+1)=[];
68          end
69      end
70  end
```

Listing B.3: How creating a series of autocorrelations is formulated in code.

```
1   function [E] = AutoCor(Qmodel,Qdata)
2
3   n=45;
4   z1=nan*ones(length(Qmodel),n);
5   z2=nan*ones(length(Qmodel),n);
6
7   for  j=1:45
8       z1z=circshift(Qmodel,-(j-1));
9       z2z=circshift(Qdata,-(j-1));
10      z1(1:end-(j-1),j)=z1z(1:end-(j-1));
11      z2(1:end-(j-1),j)=z2z(1:end-(j-1));
12
13
14  end
15
16  ac1=corr(Qmodel,z1,'rows','pairwise').';
17  ac2=corr(Qdata,z2,'rows','pairwise').';
18
19  E=Nash(ac1,ac2);
```

B.1.2. Figure(s)

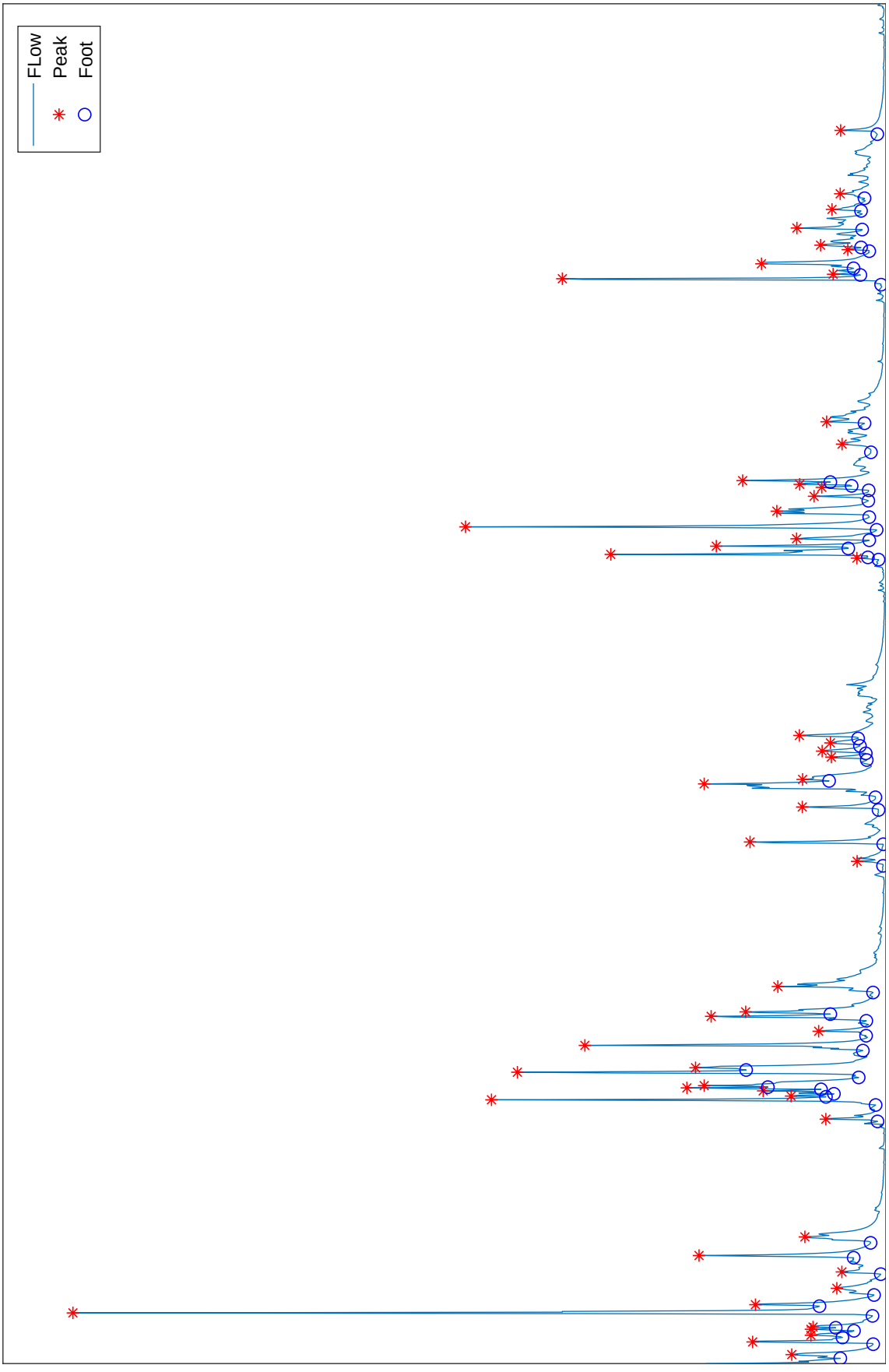


Figure B.1: An example of the RLD algorithm determining what the peaks and their corresponding foots are.

B.2. Calibration

B.2.1. Matlab Code

Listing B.4: Very first calibration step

```
clear all

%% Load Data
Dir=pwd;
loadmap='BasicFiles';
Precip=csvread('Precipnew.csv',1,1);
EvaT=csvread('EvaTnew2.csv');
Qdata=csvread('Discharge2.csv',1,1);
Temp=csvread('Temp.csv',1,1);
Sect=load(sprintf('%s\\%s\\%s',Dir,loadmap,'catchsize2.txt'));
perc95=readtable(sprintf('%s\\%s\\%s',Dir,loadmap,'95 percentile.txt'));
Temp=Temp(:,1);
MaskQ=Reviewdata(Qdata);

data(:,1)=Precip;
data(:,2)=Temp;
data(:,3)=EvaT;


%% Define Parameter ranges

%                Imax   Ce   Sumax   beta   Pmax     Tlag      Kf
          Ks          Melt    Tthresh     D
ParRange.minn = [0     .5    1       .1      1       .1       .1
   .05659*0.75        1       -1        0   ];
ParRange.maxn = [5     .5    1000     4      10      3        3
   .05659*1.25        5         1         1.0];
Sin = [0 0 100 0 5];

% Make up input data
ndata=8*365+3*366;
MaskQ=MaskQ(1:ndata);
ExtraPar.Forcing=data(1:ndata,:);
ExtraPar.Sin=Sin;
ExtraPar.Sect=Sect;

%% Calibration Section
scal=365+1;
n=100000;
```

```
A=[];
lijst=[];
h = waitbar(0,'Please wait...');
for i=1:n
    Random=rand(1,11);
    Par=(ParRange.maxn-ParRange.minn).*Random+ParRange.minn
       ;
    Qm=UsaMod2(Par,ExtraPar);
    eps=Nash(Qm(scal:end),MaskQ(scal:end));
    A=[A;[Par eps]];
    if eps>perc95.Var2(1)
        lijst=[lijst;[Par eps]];
    end
    waitbar(i/n)
end
close(h)

%% Further editing data
savemap = 'BasicFiles';
saveDir = sprintf('%s\\%s',Dir,savemap);
MakeDir(saveDir);
savename = 'Benchmark100k.txt';
savefile = sprintf('%s\\%s',saveDir,savename);
savename2 = 'OriginalPar100k.txt';
savefile2 = sprintf('%s\\%s',saveDir,savename2);

%save(savefile,'lijst','-ascii');
%save(savefile2,'A','-ascii');

%% making plot of the best set
Best = find(max(A(:,12))==A(:,12));
BestPar = A(Best,1:11);
Days=scal:1100;
Qm = UsaMod2(BestPar,ExtraPar);
fig = figure('Position', get(0, 'Screensize'));
plot(Days,Qm(scal:1100),'r',Days,Qdata(scal:1100),'b');
legend('Model outcome','Collected data','Location','north')
xlim([scal,1100]);
savename = 'BestFitNSE100k';

%saveas(fig,fullfile(saveDir,savename), 'png');
```

Listing B.5: Moving window calibration

```
clear all

%% Load Data
Dir = pwd;
loadmap='BasicFiles';
Precip = csvread('Precipnew.csv',1,1);
EvaT = csvread('EvaTnew2.csv');
Qdata = csvread('Discharge2.csv',1,1);
Temp = csvread('Temp.csv',1,1);
Calib=load(sprintf('%s\\%s\\%s',Dir,loadmap,'OriginalPar100k.txt'));
Sect=load(sprintf('%s\\%s\\%s',Dir,loadmap,'catchsize.txt'));
perc95=readtable(sprintf('%s\\%s\\%s',Dir,loadmap,'95percentile.txt'));
Temp = Temp(:,1);
MaskQ = Reviewdata(Qdata);

data (:,1) = Precip;
data (:,2) = Temp;
data (:,3) = EvaT;

%% Make up input data
Sin = [0 0 100 0 5];
ndata = 8*365 + 3 * 366;
MaskQ = MaskQ(1:ndata);
Qdata = Qdata(1:ndata);
ExtraPar.Forcing=data(1:ndata,:);
ExtraPar.Sin=Sin;
ExtraPar.Sect=Sect;

%% define Lengths
deeix = {'5years', '2years', '1year', '6months', '3months'};
dx    = [  1826       730       365       182          91   ];
mnd   = [   31   28   31   30   31   30   31   31   30   31   30   31   ];
nt = size(Calib,1);
scal = 365+1;
mndu=[];
for j=1:10
    mndu=[mndu,mnd];
end

%% Make Calibrations map
```

```
savemap ='TNSE100k';
saveDir =sprintf('%s\\%s',Dir,savemap);
MakeDir(saveDir);
mapname='CalibrationLogNSE';
Path=sprintf('%s\\%s',saveDir,mapname);
MakeDir(Path)

%% Receiving corresponding coefficients
h = waitbar(0,'Please wait...');
bestlijst=[];
getin=[];
for j=1:length(dx)
    lijst=[];
    num = round((ndata-scal+1)/mean(mnd));
    num = num - round(dx(j)/mean(mnd)) + 1 ;
    pareto=zeros(nt,num);
    waitmes = sprintf('An calibration period of %s is
       processed' ,deeix{j});
    g = waitbar(0,waitmes);
    pos_w1=get(h,'position');
    pos_w2=[pos_w1(1) pos_w1(2)+pos_w1(4) pos_w1(3) pos_w1
       (4)];
    set(g,'position',pos_w2,'doublebuffer','on')
    for k=1:nt
        Par = Calib(k,1:11);
        Qm = UsaMod2(Par,ExtraPar);
        for z=1:num
            eps=zeros(1,7);
            inm=Qm((scal+sum(mndu(1:(z-1)))):(scal+sum(mndu
               (1:(z-1)))+dx(j)-1));
            ind=MaskQ((scal+sum(mndu(1:(z-1)))):(scal+sum(
               mndu(1:(z-1)))+dx(j)-1));
            inq=Qdata((scal+sum(mndu(1:(z-1)))):(scal+sum(
               mndu(1:(z-1)))+dx(j)-1));
            Rain=Precip((scal+sum(mndu(1:(z-1)))):(scal+sum(
               mndu(1:(z-1)))+dx(j)-1));
            eps(1)=Nash(inm,ind);
            eps(2)=logNash(inm,ind);
            if eps(1)>perc95.Var2(1) && eps(2)>perc95.Var2
               (2)
               lijst=[lijst;[k z Calib(k,12)]];
            end
        end
        waitbar(k/nt)
    end
    close(g)
```

```
80      savename = deeix{j};
81      savefile = sprintf('%s\\%s.txt',Path,savename);
82      %save(savefile,'lijst','-ascii');
83      waitbar(j/length(dx));
84  end
85  close(h)
```

B.2.2. Figures

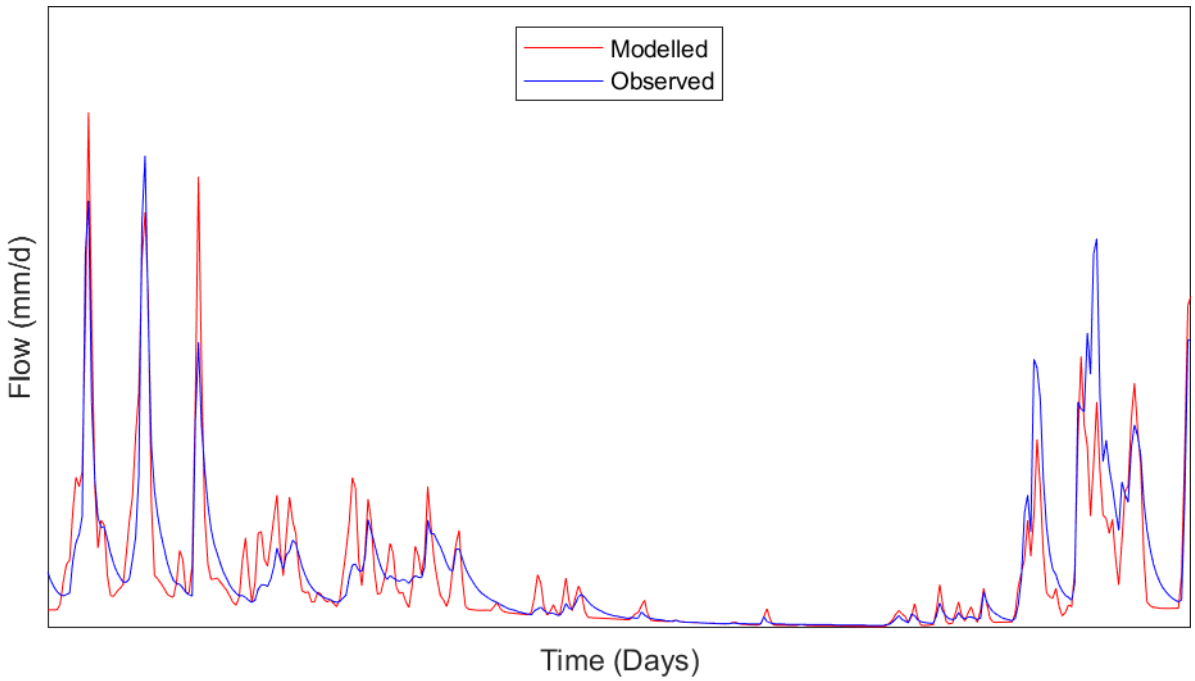


Figure B.2: Depiction of the modelled flow versus the observed flow. The modelled flow used the parameter-set which had the smallest euclidean distance, where each of the 7 evaluation criteria had equal weight.

C

Results supplement

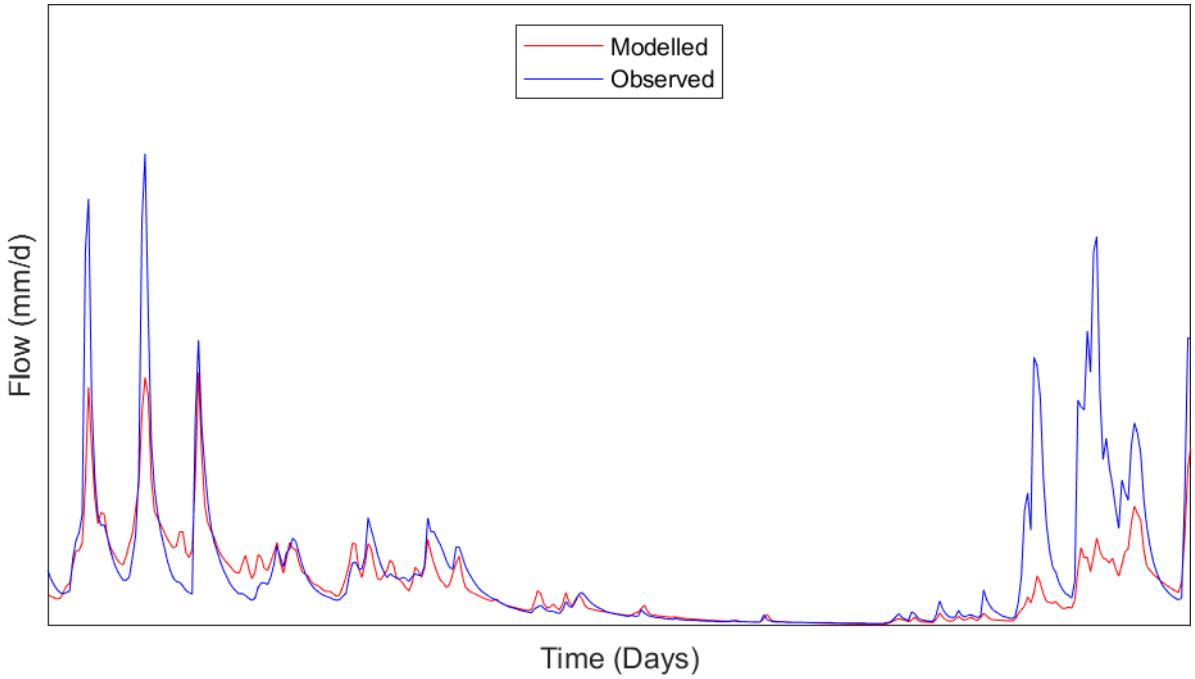


Figure C.1: A depiction of the modelled flow versus the observed. The model used the parameter-set that had the heighest NSE-value while only being evaluated against the NSE over a calibration period of 3 months. This NSE values is not the equivalent performance in terms of NSE, but the performance of the 3 months period.

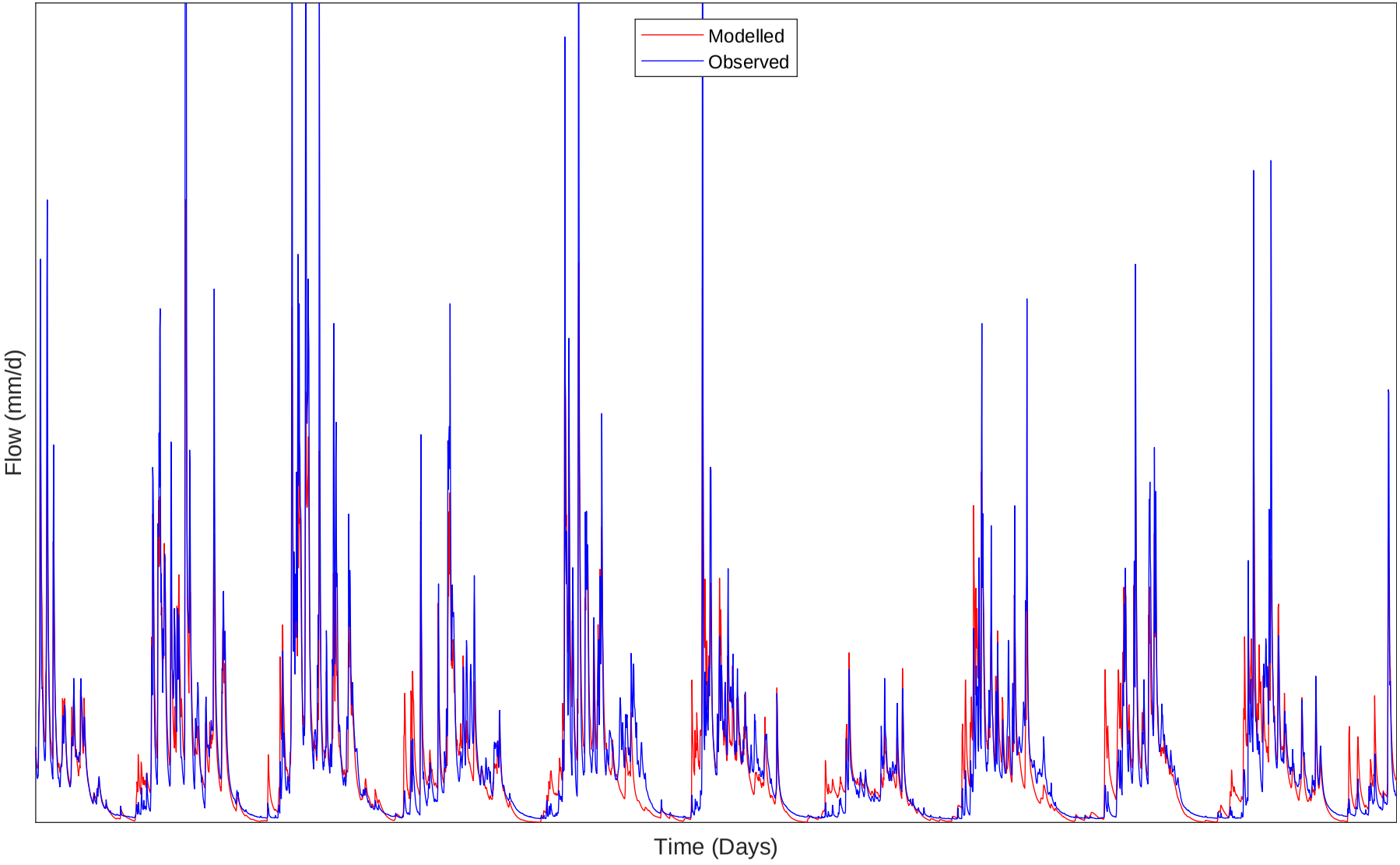


Figure C.2: A depiction of the modelled flow versus the observed. The model used the last remaining parameter-set after the addition of the last evaluation criterion. This graph shows the comparison over the entire benchmark period of 10 years.

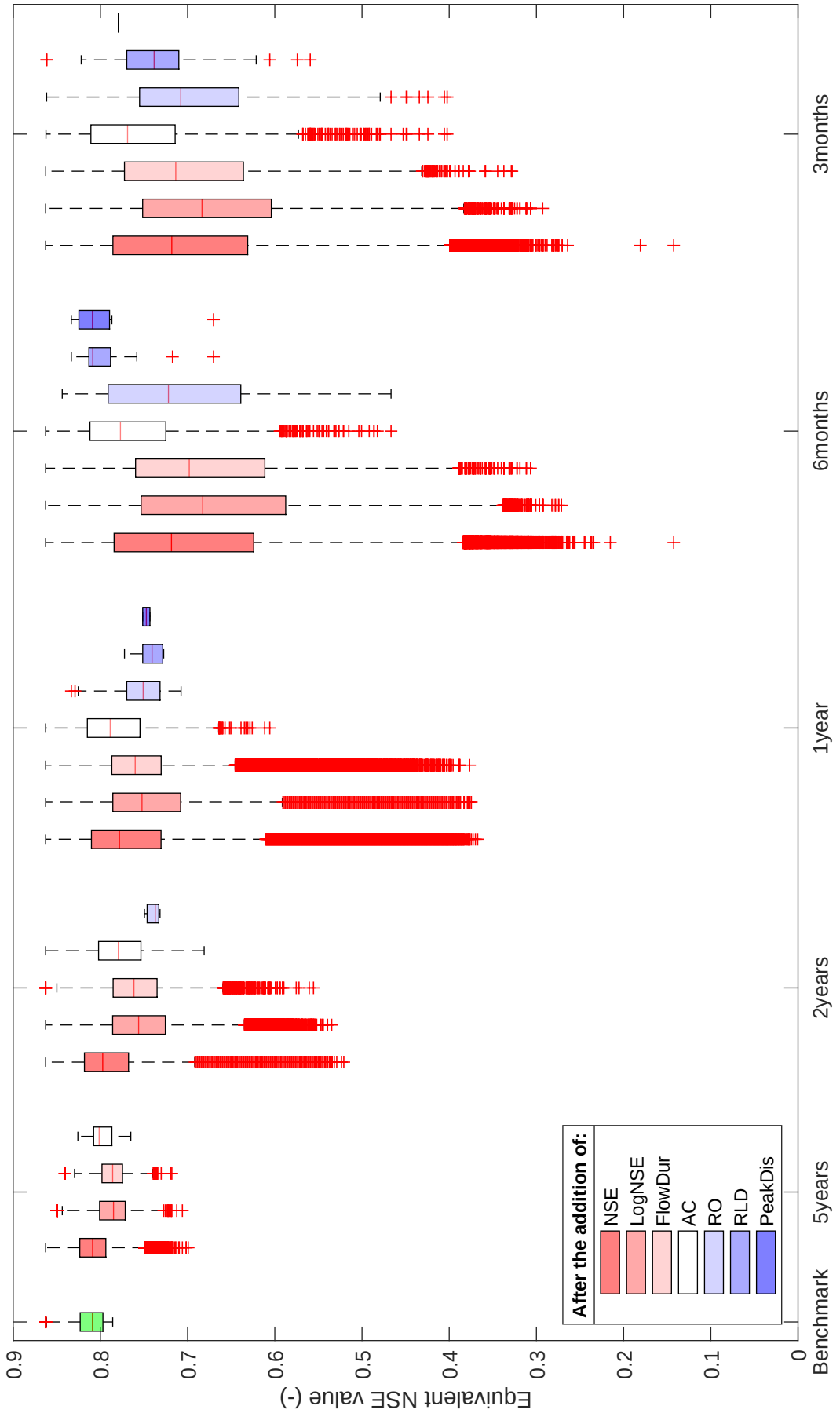


Figure C.3: The equivalent performance, in terms of Nash-Sutcliffe efficiency over the entire benchmark calibration period, of the model using the parameter-sets that made it through the calibration process.

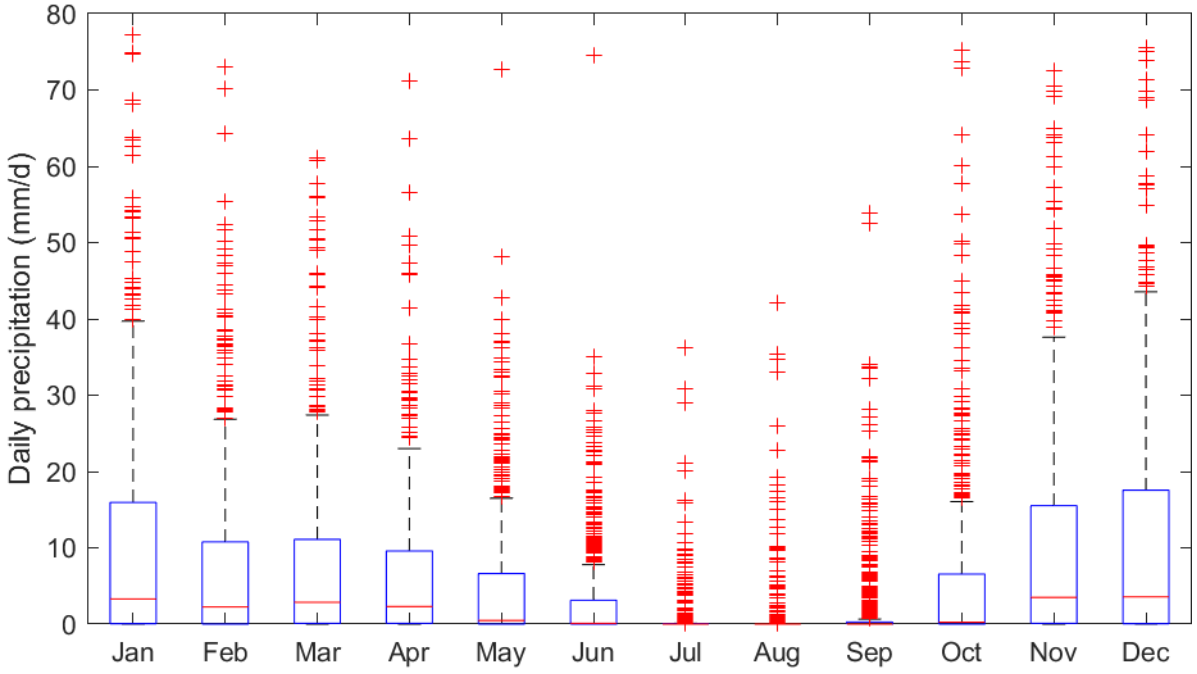


Figure C.4: Daily percipiation catergorised per month.

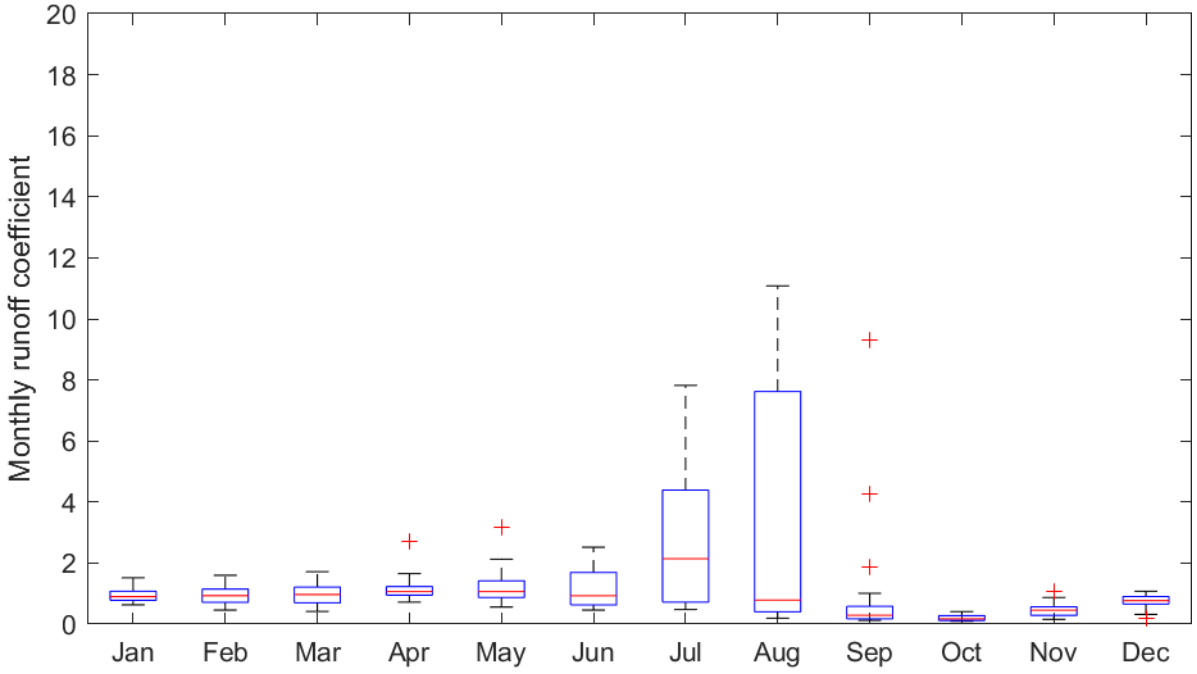


Figure C.5: Monthly mean runoff coefficient.

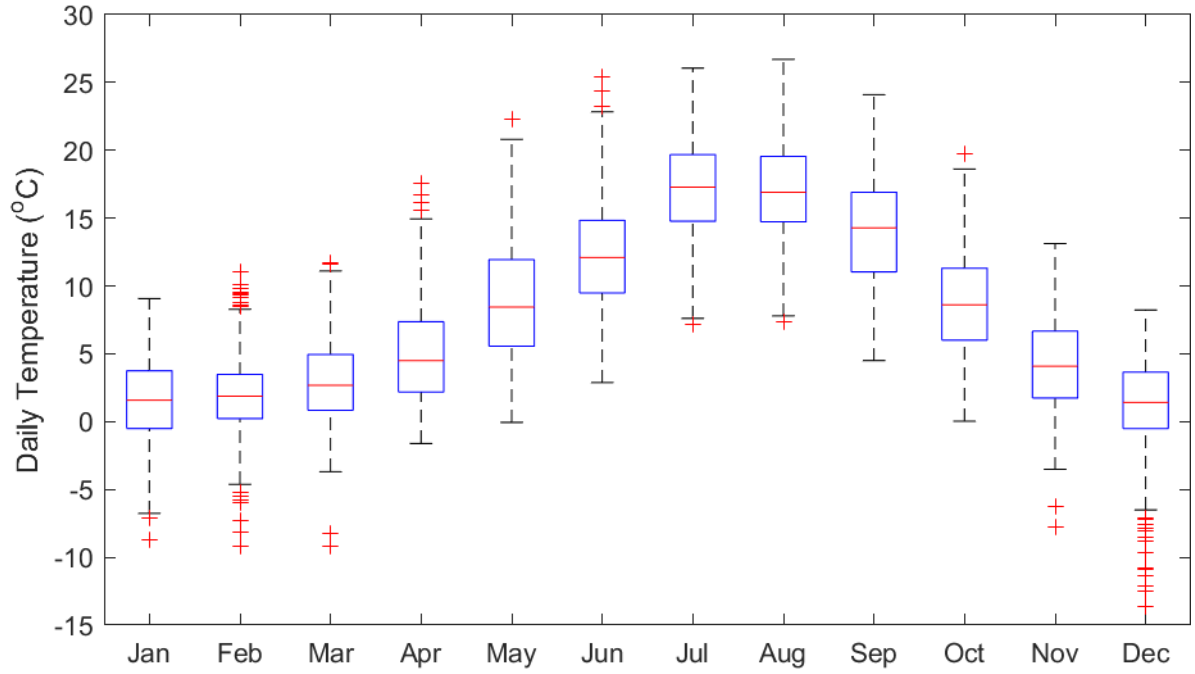


Figure C.6: Daily temperature catergorised per month.

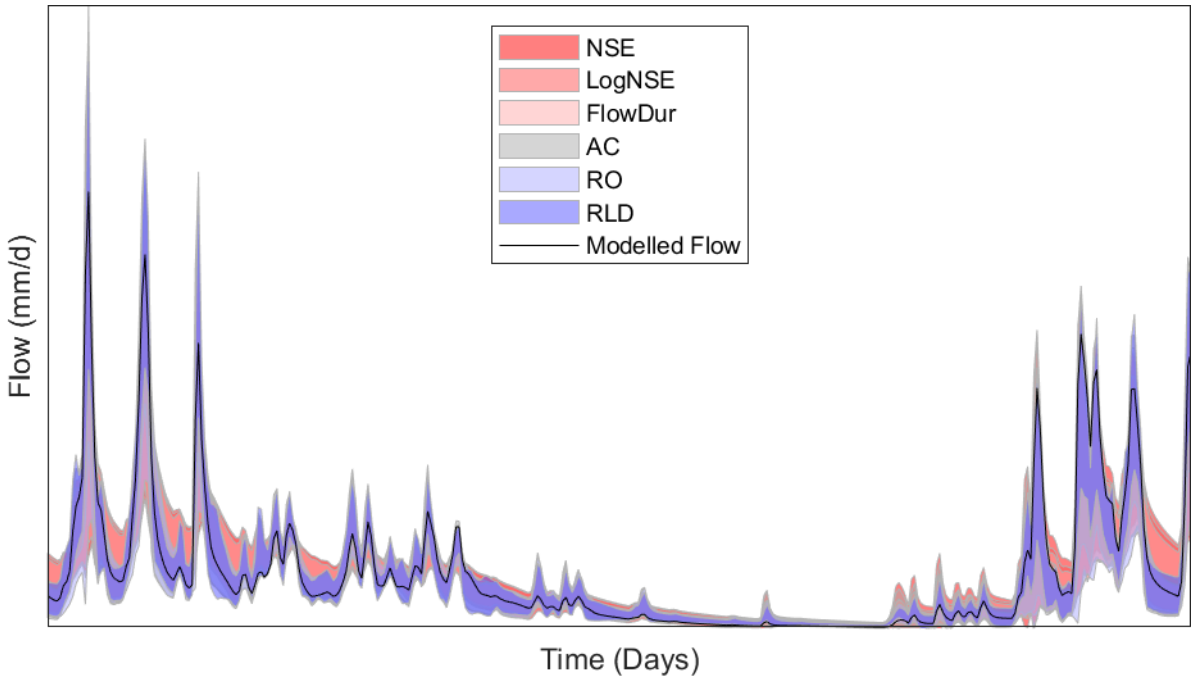


Figure C.7: The prediction intervals after each criterion addition. The parameter-sets were evaluated over a 6 month period, i.e. the calibration period-length.

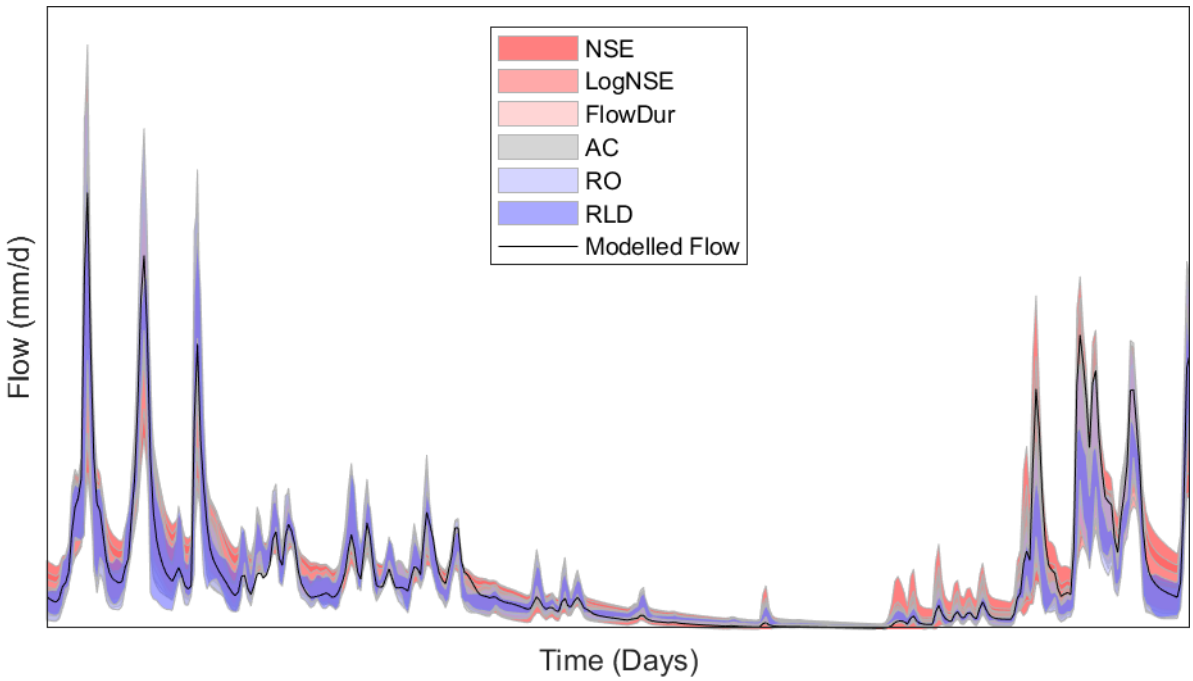


Figure C.8: The prediction intervals after each criterion addition. The parameter-sets were evaluated over a 1 year period, i.e. the calibration period-length.

D

Discussion supplement

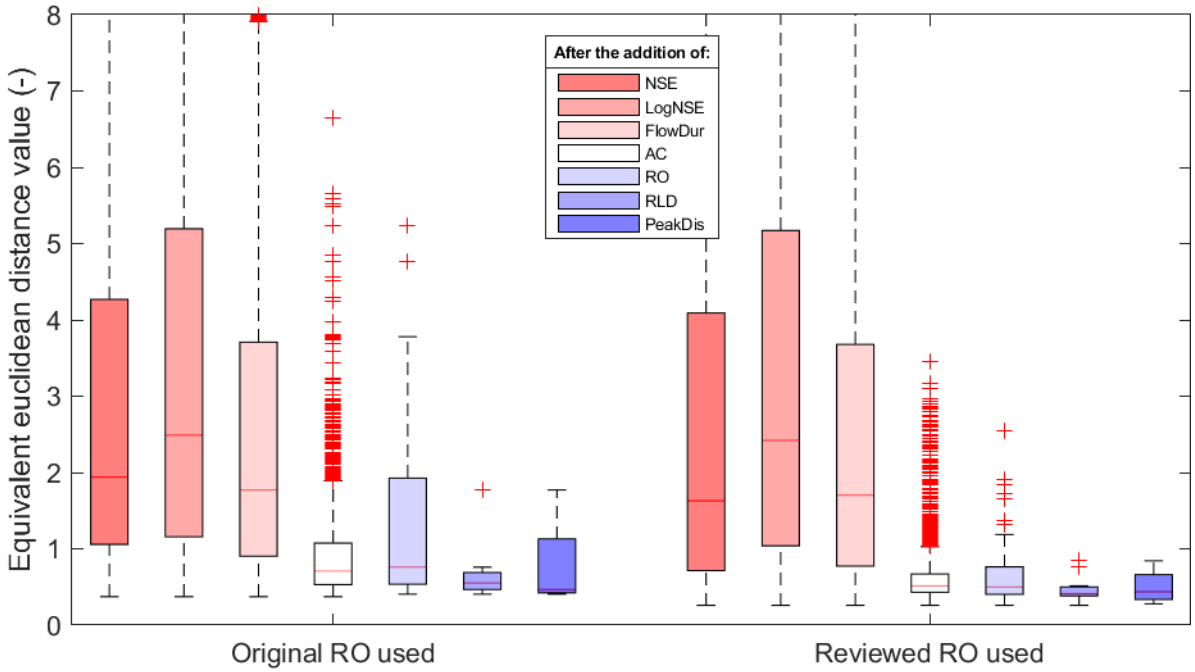


Figure D.1: Equivalent perfomance of the parameter-sets that passed each addition of evaluation criteria over a 6 month period. The cluster of boxplots on the left used the RO criterion according to equations (3.6) & (3.7) and the cluster to the right according to equations (5.1) & (5.2).

Parameter	**Par$_x$**	**Par$_z$**
C_o	1.3550	4.7190
$S_{u,max}$	207.01	628.70
β	0.3385	2.4324

Table D.1: Parameter comparion between Par$_x$ and Par$_z$.

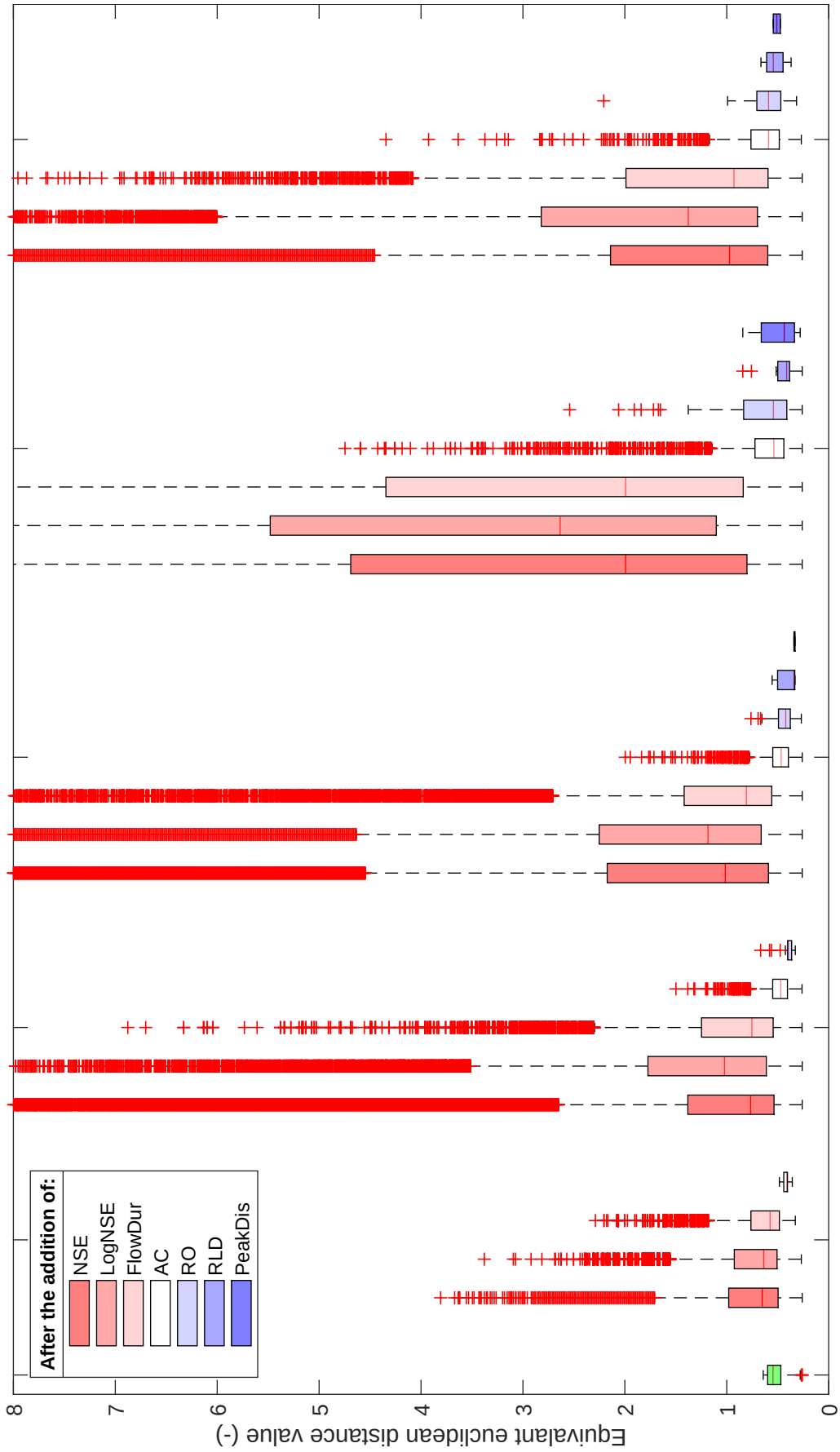


Figure D.2: The equivalent performance, in terms of euclidean distance over the entire benchmark calibration period, of the model using the parameter-sets that made it through the calibration process. For this figure use had been made of the reviewed runoff coefficient criterion.

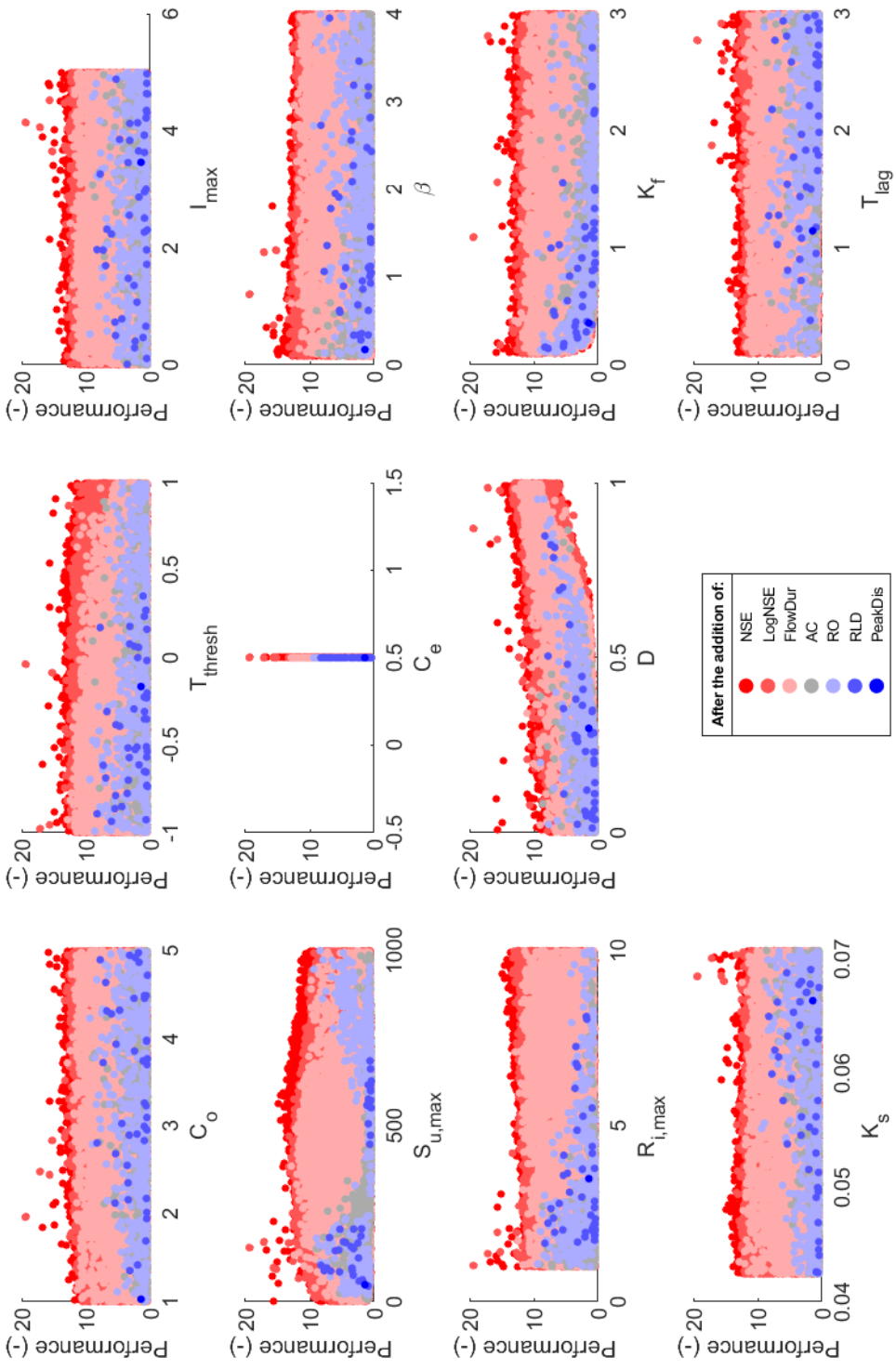


Figure D.3: Sensitivity of each parameter is displayed after each criterion addition. The performance shown in this figure is the equivalent euclidean distance. The parameters used in this figure are from the remaining parameter-sets of the 3 months period-length calibration.

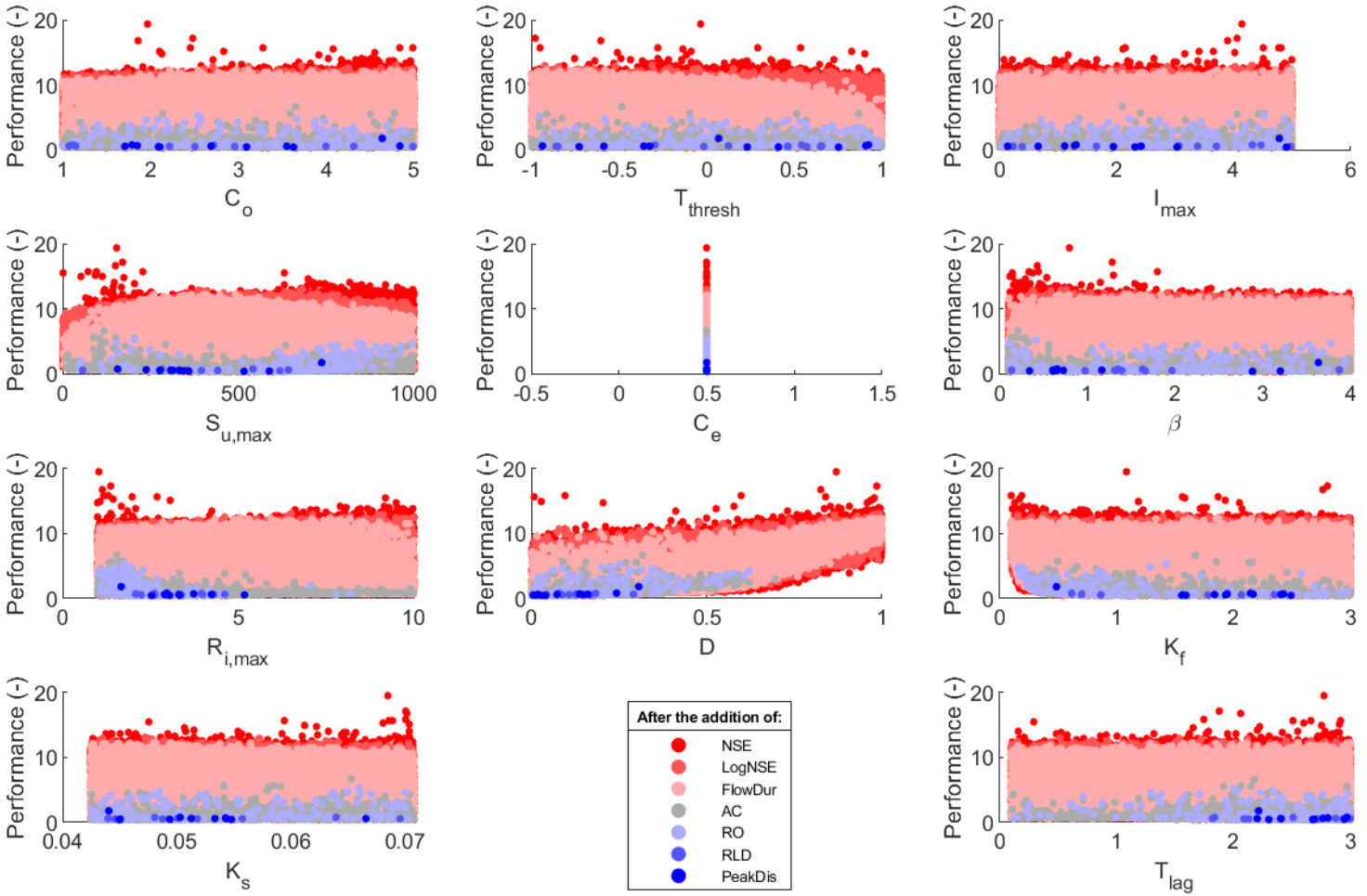
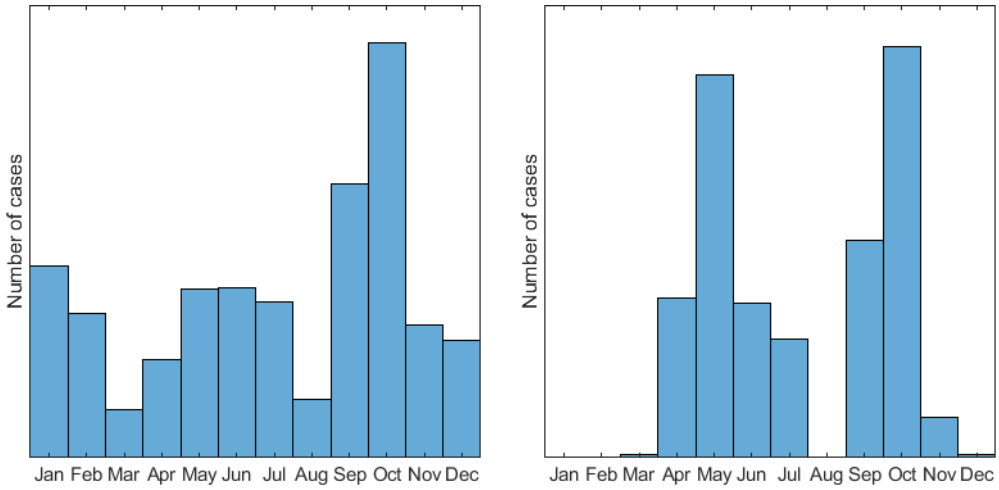
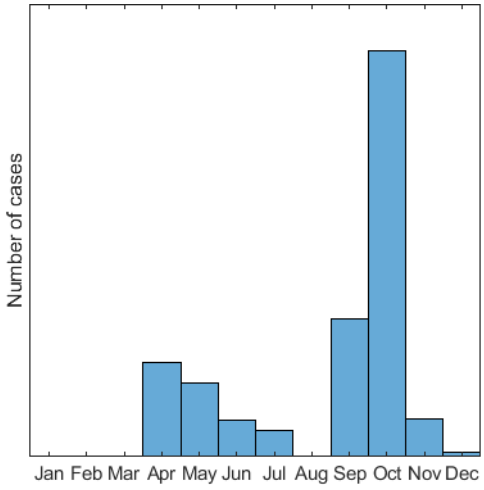


Figure D.4: Sensitivity of each parameter is displayed after each criterion addition. The performance shown in this figure is the equivalent euclidean distance. The parameters used in this figure are from the remaining parameter-sets of the 6 months period-length calibration.

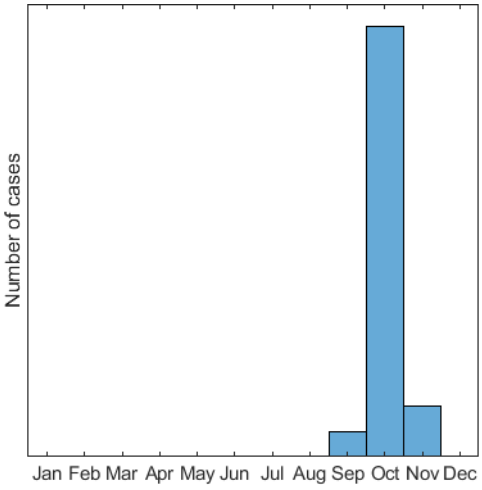


(a) NSE

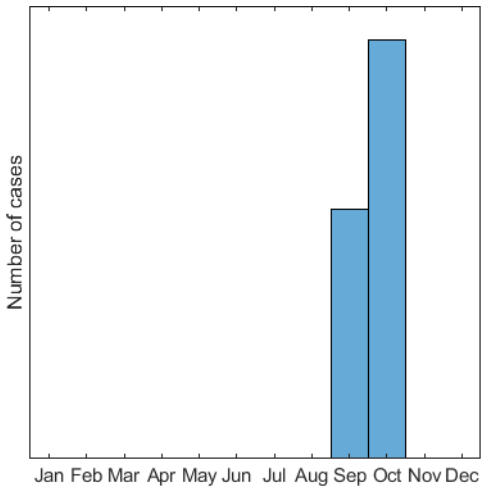
(b) LogNSE



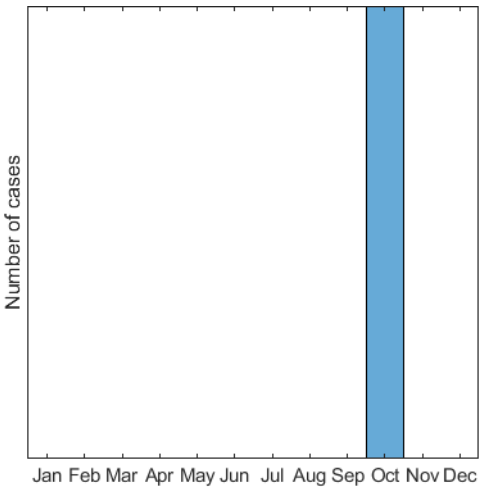
(c) FlowDur



(d) AC

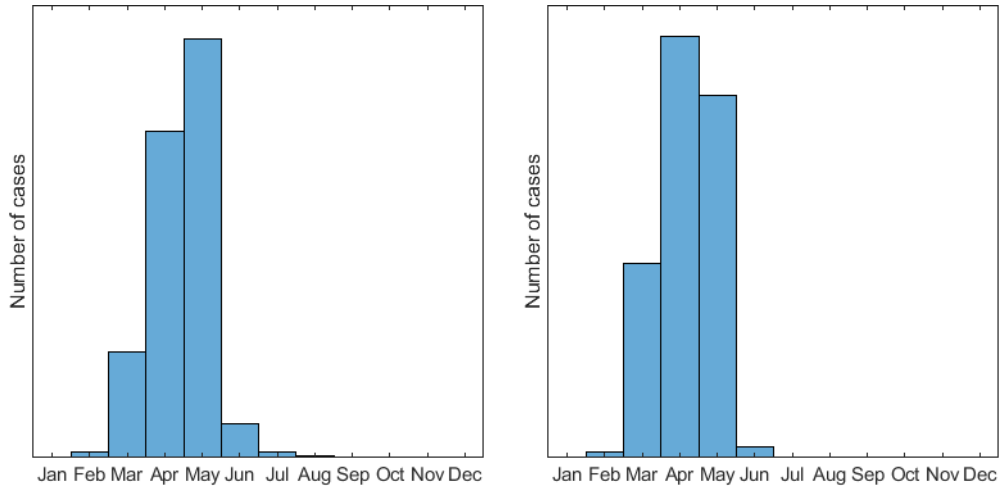


(e) RO



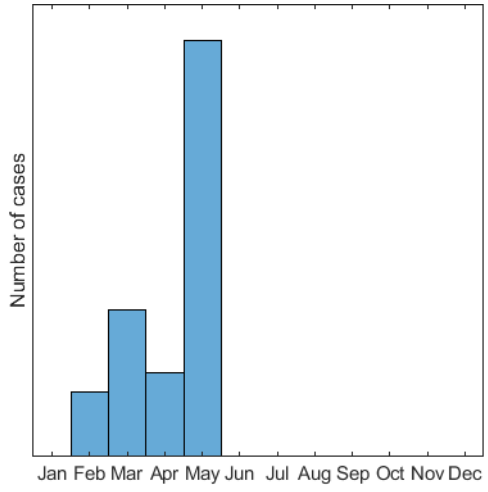
(f) RLD

Figure D.5: Best 10 percent of the equivalent euclidean distance values after each evaluation criteria addition for the 3 month period-length.

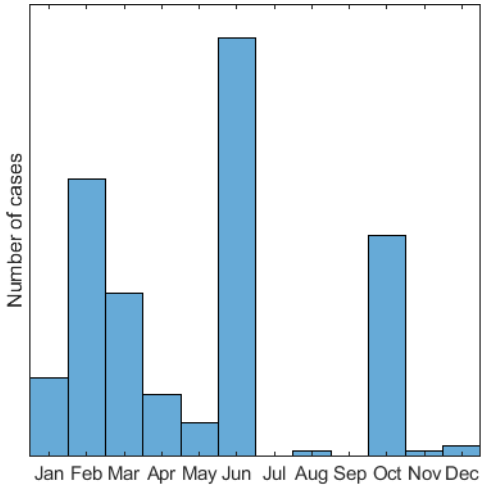


(a) NSE

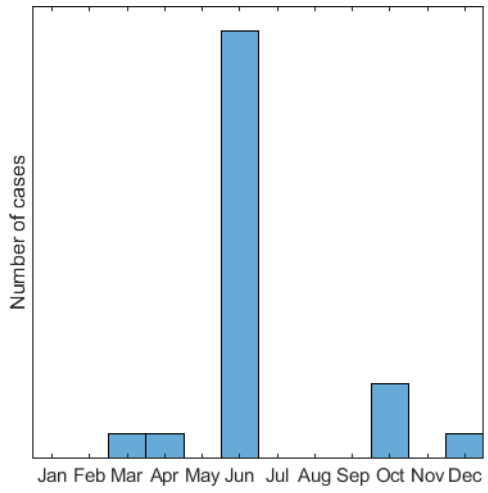
(b) LogNSE



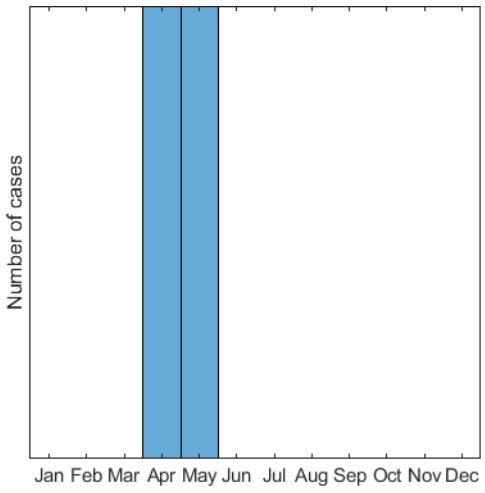
(c) FlowDur



(d) AC

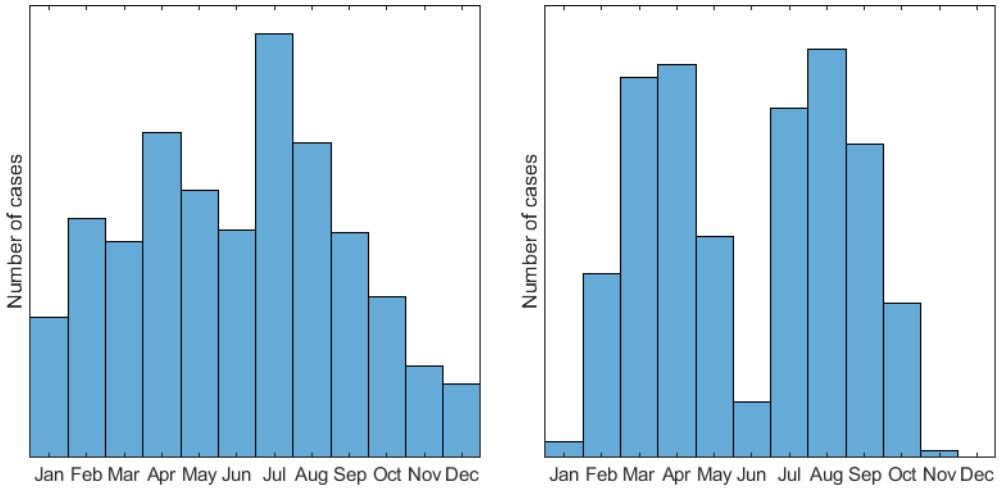


(e) RO



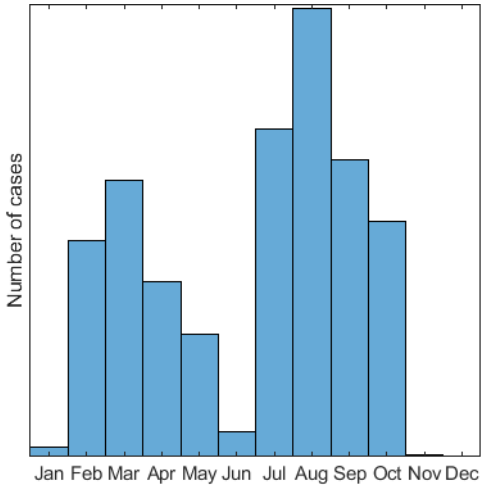
(f) RLD

Figure D.6: Worst 10 percent of the equivalent euclidean distance values after each evaluation criteria addition for the 6 month period-length.

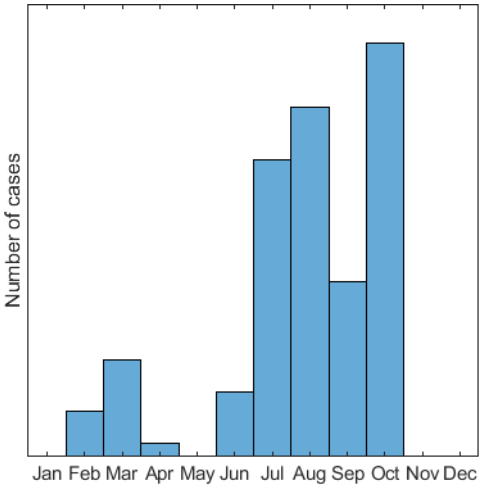


(a) NSE

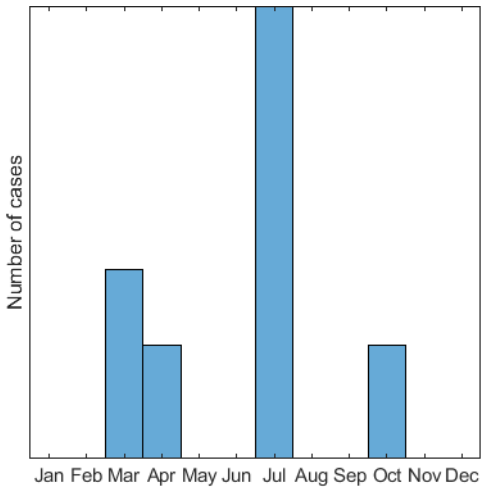
(b) LogNSE



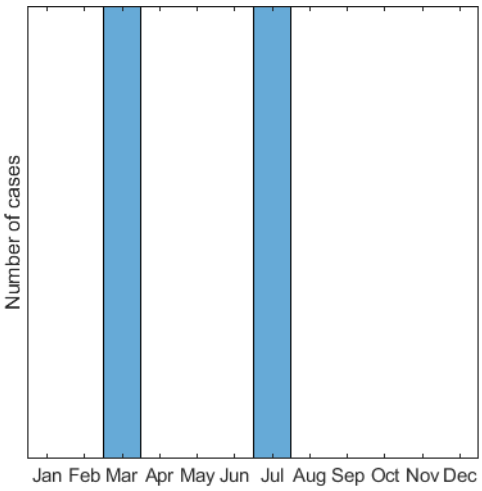
(c) FlowDur



(d) AC

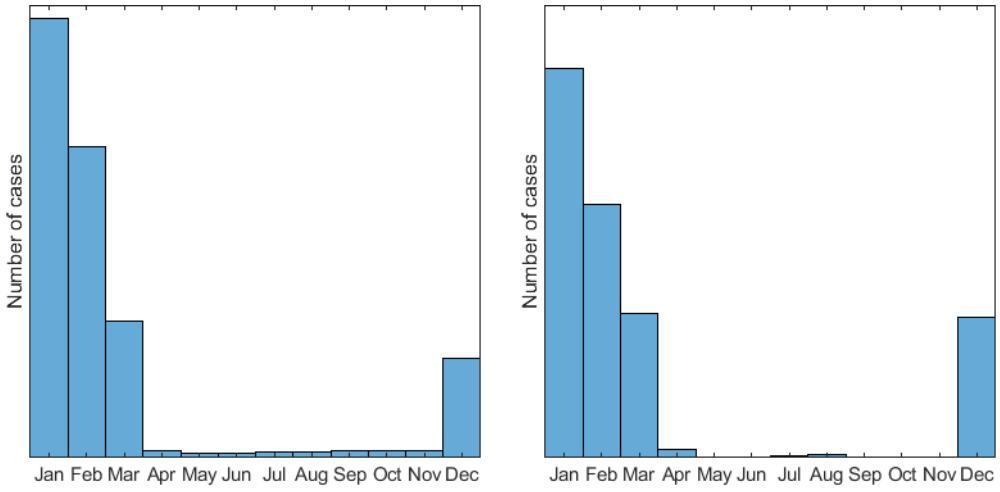


(e) RO



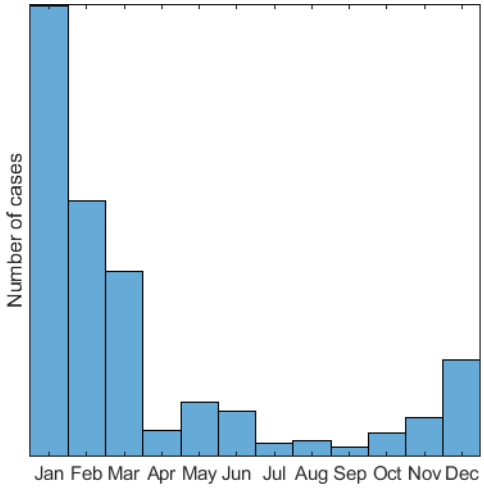
(f) RLD

Figure D.7: Best 10 percent of the equivalent euclidean distance values after each evaluation criteria addition for the 6 month period-length.

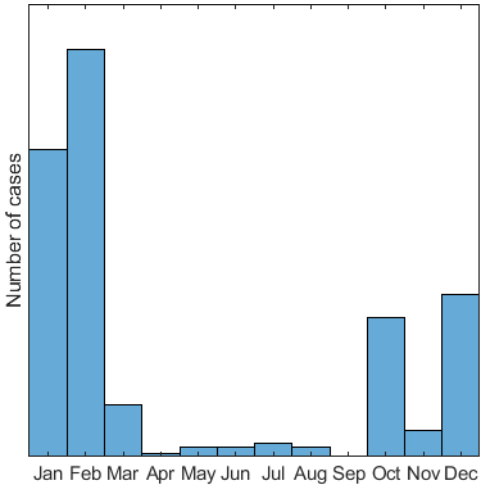


(a) NSE

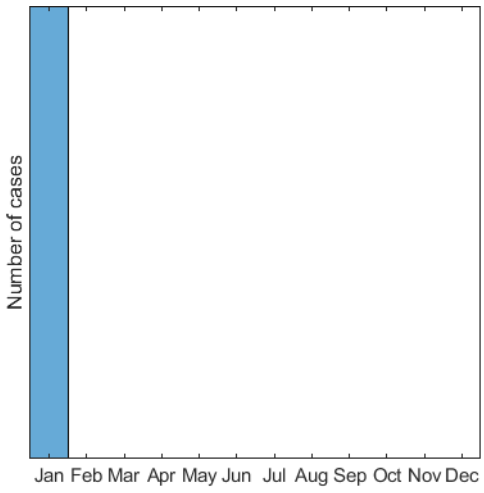
(b) LogNSE



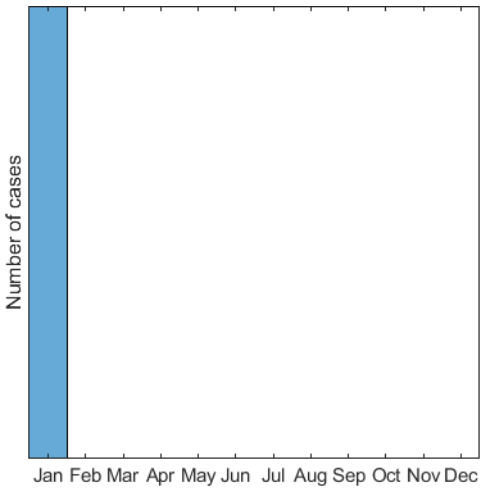
(c) FlowDur



(d) AC

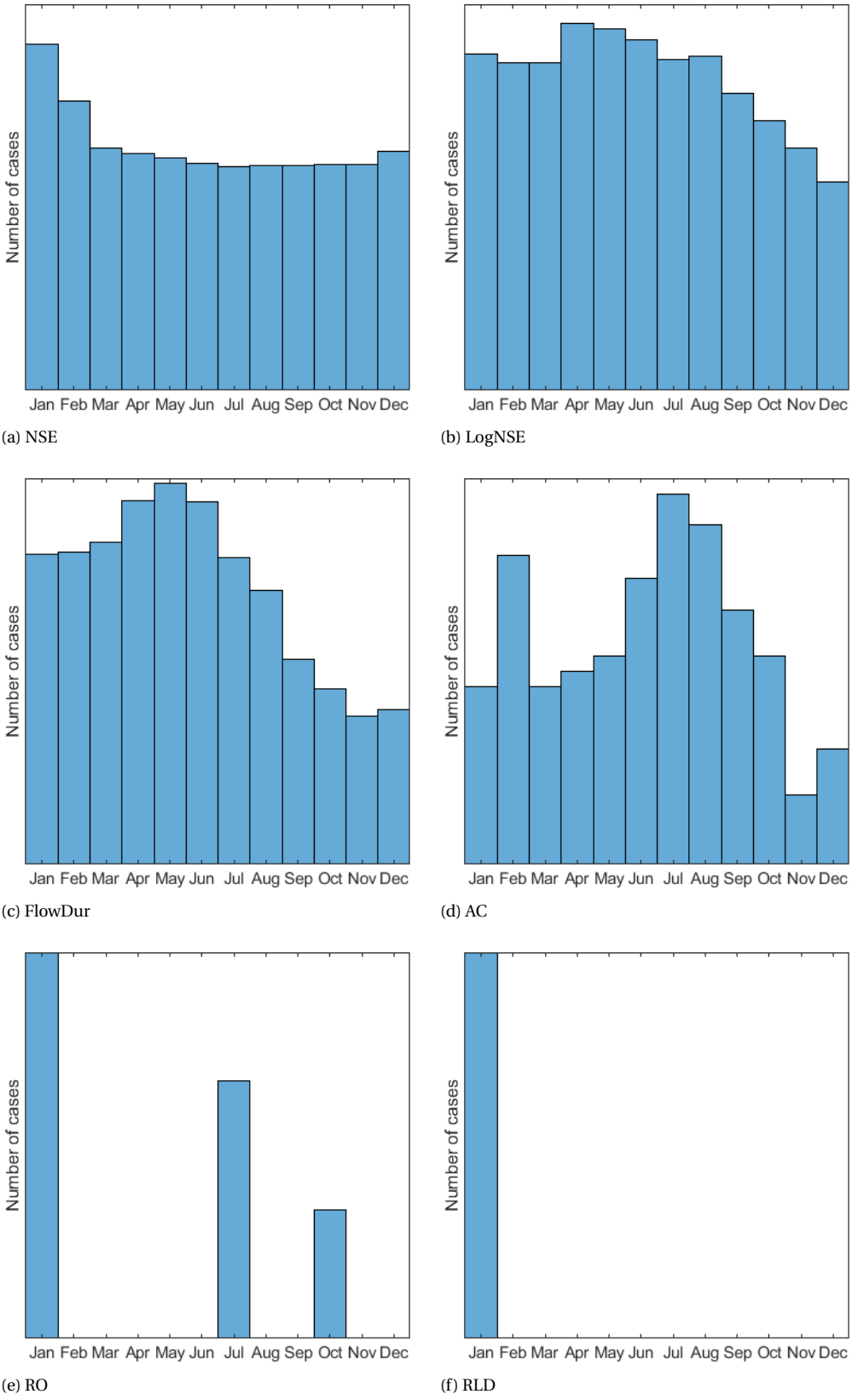


(e) RO



(f) RLD

Figure D.8: Worst 10 percent of the equivalent euclidean distance values after each evaluation criteria addition for the 1 year period-length.



(a) NSE

(b) LogNSE

(c) FlowDur

(d) AC

(e) RO

(f) RLD

Figure D.9: Best 10 percent of the equivalent euclidean distance values after each evaluation criteria addition for the 1 year period-length.