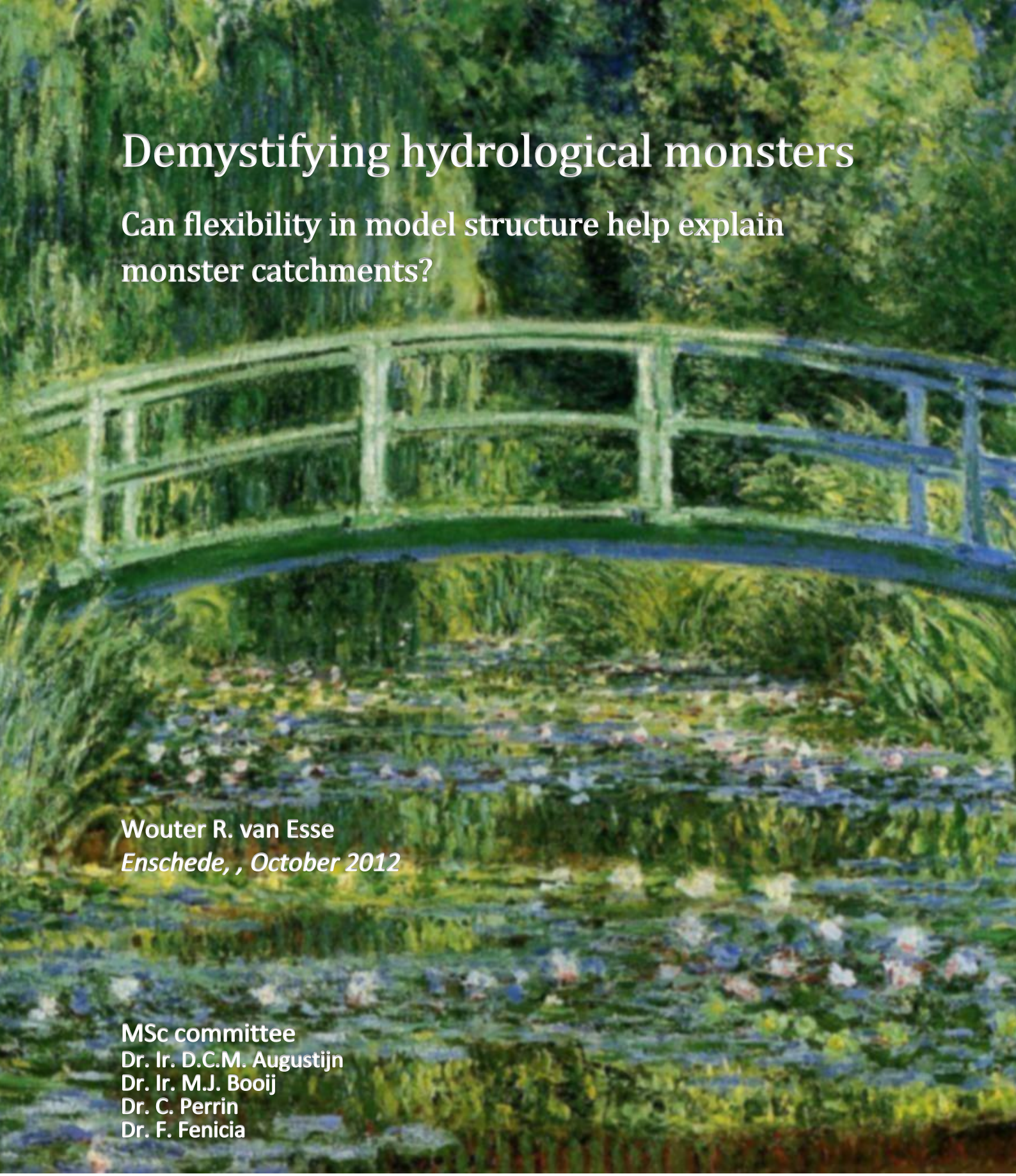

Demystifying hydrological monsters

Can flexibility in model structure help explain monster catchments?

Wouter R. van Esse

Enschede, , October 2012

MSc committee
Dr. Ir. D.C.M. Augustijn
Dr. Ir. M.J. Booij
Dr. C. Perrin
Dr. F. Fenicia

Centre de Recherche Public Gabriel Lippmann

irstea

UNIVERSITY OF TWENTE.

Demystifying hydrological monsters

Can flexibility in model structure help explain monster catchments?

Wouter R. van Esse
Enschede, October 2012

MSc committee:
Dr. Ir. D.C.M. Augustijn
Dr. Ir. M.J. Booij
Dr. C. Perrin
Dr. F. Fenicia

Cover photo: The Japanese Bridge (The Water-Lily Pond) – Claude Monet – 1899.

Centre de Recherche Public
Gabriel Lippmann

UNIVERSITY OF TWENTE.

Colophon

Author
Wouter R. van Esse

Student Civil Engineering and Management | w.r.vanesse@alumnus.utwente.nl

Members of the MSc Thesis committee:
Dr. Ir. D.C.M. Augustijn [1]

Associate Professor | d.c.m.augustijn@ctw.utwente.nl

Dr. Ir. M.J. Booij [1]

Assistant Professor | m.j.booij@ctw.utwente.nl

Dr. C. Perrin [2]

Hydrologist |charles.perrin@irstea.fr

Dr. F. Fenicia [3]

Researcher | fenicia@lippmann.lu

[1] **University of Twente**
Faculty of Engineering Technology
Department of Water Engineering and Management
P.O. Box 217
7500AE Enschede
The Netherlands

[2] **Irstea**
Hydrosystems and Bioprocesses Research Unit
1, rue Pierre-Gilles de Gennes
CS 10030, 92761 Antony Cedex
France

[3] **Centre de Recherche Public – Gabriel Lippmann**
Department of Environment and Agro-Biotechnologies
41, rue du Brill
L-4422 Belvaux
Luxembourg

Institutions	: University of Twente
	Irstea
	CRP Gabriel Lippmann
Project	: MSc thesis
Extent document	: 110 pages
Author	: Wouter R. van Esse
Date	: 11 October 2012

Summary

Rainfall-runoff hydrological models are commonly used to investigate and simulate catchment behaviour and predict discharges. The simulation of the discharge is never perfect and in some cases a hydrological monster is created, a combination of a model and a catchment that together result in a poor simulation. This study compares two reservoir based modelling approaches to investigate the role of model structure on model performance and ultimately explain the hydrological monsters.

237 French catchments are modelled using the fixed GR4H and flexible SUPERFLEX approach. GR4H is a single model structure that is calibrated using four parameters and generally shows good average performance on a wide range of catchments. In the SUPERFLEX approach, model components and functions can be combined in any way to a create specific model for each catchment. Twelve SUPERFLEX structures with varying complexity are used to analyse the influence of model structure. All models are calibrated using a split sample test, ten year time series where split in two periods. Calibration took place on the first and second period and validation on the second and first respectively. Inconsistency between parameter sets or model structures (SUPERFLEX) is considered as a failure of the approach.

This study found that relatively simple model structures with some key components can lead to a good simulation of the discharge. The analysis of the thirteen individual model structures (GR4H + 12 SUPERFLEX structures) showed that:

- The use of a power function to describe reservoir outflow significantly increases mean model performance on the catchment set,
- Independently calibrated parallel reservoirs increase model performance in permeable catchments with dominant base flow,
- A lag-function between reservoirs in SUPERFLEX structures leads to no significant increase in mean model performance on the catchment set, and
- Increasing model complexity beyond a certain point does not lead to higher average performance. However, complex structures do show a smaller range in performance, meaning there are less catchments for which they perform very poor and that fewer monsters are created.

On the whole catchment set, the flexible modelling approach does not provide better results than the fixed modelling approach on average. However, it manages to provide consistent results between test periods on a larger number of catchments. On 69 catchments, both modelling approaches performed poorly or inconsistently. These catchments were selected as monsters in this study and were classified into three groups:

- Catchments where severe climatic differences between calibration and validation periods are too large for the models to correctly simulate discharge across these periods,
- Catchments where models are unable to simulate the extreme flashy behaviour, and
- Catchments where small scale disturbances in flow or measurement errors hinder good simulation.

Generally, selecting an individual model structure for each catchment helps to rehabilitate some hydrological monsters but adding complexity is no guarantee for better results.

Preface

This report on hydrological modelling is my final work of my master Civil Engineering and Management at the University of Twente. During my work on this report I spend time at two research institutes in Europe: two months in France at Irstea and two months in Luxembourg at the CRP Gabriel Lippmann.

At both institutes I had some great experiences, learned a lot and made great friends. I would like to thank my colleagues at both institutes for welcoming me into their country and their support of my work. I give special thanks to Florent Obligeois for helping me with my research at Irstea and the great times we had playing Frisbee or Volleyball with all the colleagues. I also want to give special thanks to my colleagues at the CRP Lippmann for their hospitality in allowing me to be part of their group and showing me around in Luxembourg.

I acknowledge Météo France and SCHAPI and the Ministère de l'Ecologie et du Développement Durable for providing precipitation and discharge data and the European Environment Agency for various data on the catchments I used in this research. I want to thank Dmitri Kavetski for his support with the BATEA programme and my supervisors Charles Perrin (Irstea), Fabrizio Fenicia (CRP Lippmann), Martijn Booij and Denie Augustijn (University of Twente), from whom I learned a tremendous amount.

Finally, I thank my parents for their love and support in giving me this opportunity, my girlfriend for her love and strength during my time abroad and my friends for the great times we had.

Enschede, October 2012

Wouter van Esse

Table of Contents

List of Figures

List of Tables

1 Introduction

Rainfall-runoff hydrological models are commonly used to investigate and simulate catchment behaviour and predict discharges. Rainfall-runoff models can generally be categorised as empirical or physical, lumped or distributed and deterministic or stochastic (Beven, 2001; Diermanse, 2001). This study uses lumped deterministic models that are empirical or conceptual. Conceptual models are based on a modeller's concept of the physical or hydrological processes in a catchment. Given the limitations of data availability, hydrological models are often used to learn more about hydrological processes like runoff generation (Beven, 2001).

Besides research, rainfall-runoff models are widely used as management tools (Beven, 2001; Diermanse, 2001; Jakeman et al., 2006). Decision makers base flood prevention measures on the results of rainfall-runoff models after they are adapted by modellers to a specific catchment and purpose, such as the impact assessment of land use or climate change. The reliability, accuracy, correctness etc. of these results is therefore very important (Jakeman et al., 2006). Selecting the correct model for a purpose and available resources, is a key step in the modelling process. Once a model is selected, the modelling process generally consists of calibrating and validating the model on available data and using the model to answer any questions at hand (e.g. make future predictions).

1.1 The modelling process

The calibration of a rainfall-runoff model is the procedure of adjusting model parameters to reproduce the observed runoff from a catchment using rainfall and evaporation as the main inputs (Refsgaard & Henriksen, 2004). The parameter values can be calibrated one by one visually or all at the same time by optimising the fit between model output and measured discharge data through the use of one or more objective functions. Additionally, information about the parameters obtained by direct measurement or derivation can be used to establish the value of parameters (Jakeman et al., 2006).

Model validation should be performed on an unused part of the time series (split-sample test) or on time series of another catchment (proxy-basin test; Klemes, 1986). Validation substantiates *that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model* (Refsgaard & Henriksen, 2004). The validation is used as a measure of the trust that can be put in the model results.

The importance of rainfall-runoff models drives research to improve their performance. Researchers report on the use of models in different parts of the world and attempt to improve models with new data and information. In the special issue of Hydrological Sciences Journal entitled *The Court of Miracles of Hydrology* and the preceding workshop, it was shown that in the current literature on hydrological modelling success stories are given great prominence while there is little written about negative results (Andréassian et al., 2010). The authors state that we are reluctant or unable to diagnose our negative results and therefore simply reject them, and so negative results are not seen as results. In other cases, publications on failures get rejected for publication and thus never reach a large public (Browman, 1999). This may have led to a limited amount of qualitative improvement in hydrological modelling compared to other fields (Schertzer et al., 2010) and therefore Andréassian et

al. (2010) advocate that *giving greater prominence to the analysis of failures would more fruitfully advance the hydrological sciences*.

1.2 Modelling failures

Within the special issue mentioned above, several failure stories were presented. The analysis of the modelling or model failure shows the possible value of publishing failure stories. Failure in rainfall-runoff modelling can be the consequence of, for example, the incorrect use of the model, errors in data or errors in the model itself.

What model to use is still largely specific for each modelling exercise as no globally united view exists on what the rainfall-runoff model should look like (Wagener et al., 2007). Despite the importance of selecting the correct model for the purpose, often the wrong model is selected or used in ways it was not intended for (e.g. without proper concern of the assumptions on which the model is based; Jakeman et al., 2006). Limited resources and knowledge of the modelled catchment can also lead to failure, like in the case described by Refsgaard and Hansen (2010) aimed at identifying cost-effective measures to reduce nitrate loads in Denmark. This exercise failed because the selected distributed model was unable to cope with the geological heterogeneity later found in the study area. This example shows how the use of data with a resolution or scale different from the one for which the model is designed, can lead to the failure of any model (Booij, 2003).

The importance of the quality of data for the performance of a model is considered very large; research by Valéry et al. (2010) showed that correcting rainfall data in high altitude, snow covered areas using the water balance, increases model performance. Boughton (2006) even states that model performance depends more on the quality of the data than on the model because models always perform well with good quality data and bad with poor quality data. But there are cases where the model should be blamed for the modelling failure.

One example of a real error in a model is that of the Soil Moisture Accounting and Routing (SMAR) model discovered when used on the Fergus River in Ireland (Goswami & O'Connor, 2010). In this case the model actually performed very well compared to other models while as a conservative model, it should not have been able to deal with the inter-catchment groundwater flows in this specific catchment. The model was later found to have a structural error in the surface routing component that caused a water balance error which was consequently resolved by correcting the model structure.

In most cases, a model structure is not able to deal with every possible process taking place in a catchment. If one such a process becomes important (like inter-catchment groundwater flow), the model should be blamed for not representing these processes (Le Moine et al., 2007a). Selecting the correct model is therefore very important but still often the wrong model is selected, like in the examples of Bredehoeft (2005). A special case is that of the groundwater flow model in Los Angeles described by this author. Modellers were so convinced about the (very complex) model they created themselves that they failed to see its flawed fundaments.

Many modelling failures are caused by the incomplete understanding of the hydrological processes and the errors or subjectivity of the conceptualisation (Refsgaard & Hansen, 2010; Troldborg et al., 2007). In fact, any model based on the Unit Hydrograph principle could be considered flawed because no such thing exist in reality (Szöllösi-Nagy, 2009). Taking this analysis to the level of verification (or realism) would require very detailed knowledge of the real hydrological processes and

probably not aid to advancing hydrological modelling. One thing literature agrees on is that models have a potential *for revealing the implications of assumptions, estimating the impact of interactions, changes and uncertainties on outcomes, and enhancing communication between researchers from different backgrounds and between researchers and the broader community*, even if a poor model is used (Jakeman et al., 2006). We should thus continue to learn from unexpected model behaviour and confirm our knowledge of hydrology through hydrological models (Beven, 2001), like what led to the improvement of the SMAR model and the discovery about the study area and its model in the example from Denmark, both described above.

1.3 Hydrological monsters

The examples above show that many things can cause the modelling process to fail; a hydrological monster is a specific case of model failure. The term 'monster' catchment as used during the workshop and in the special issue, is a catchment for which a model gives a poor performance (Andréassian et al., 2010). As different models represent hydrological processes differently and all of them are imperfect (Duan et al., 2006), any model can have any number of monster catchments while for other models other catchments can be monsters. Therefore, a hydrological monster is defined as the combination of a model and a catchment that together give a poor result.

Modellers create models by translating their perceptual models into mathematical descriptions of the hydrological processes. In order to create a viable model, modellers simplify and make assumptions about the perceptual model. This is necessary as some processes may be too complex to solve mathematically or too little is known about them as many of the hydrological processes in a catchment are not or cannot be measured (Wagener et al., 2007). Variety in model objective and the modeller's perception have led to many different models (Jakeman et al., 2006) and modellers continue to adapt models based on new objectives or specific catchments while it remains difficult to find the right model for the intended use. The special issue also discusses the sources of hydrological monsters and reducing structural uncertainty is mentioned as one of the key steps in increasing the confidence in model results (Andréassian et al., 2010).

To find out to what extent modelling "monstrosity" can be related to structural error, this study compares a fixed modelling approach to a flexible one. Here, only lumped deterministic rainfall-runoff models with model structures representing different hydrological processes and various levels of complexity are considered. We will focus on the hydrological monsters of the fixed GR4 rainfall-runoff model (GR4J for daily and GR4H for hourly time step; Perrin et al., 2003) and the flexible modelling approach SUPERFLEX (Fenicia et al., 2011; Kavetski & Fenicia, 2011).

The GR4 model was developed empirically and found to perform well on a large set of catchments (Le Moine, 2008; Le Moine et al., 2007b; Perrin et al., 2003). Because of its high average performance on a large range of catchments this model is used as a reference model structure. Despite its good reputation, like all models GR4 does have hydrological monsters (Andréassian et al., 2010). Le Moine et al. (2008) report that the hydrological processes taking place in a karstic basin may be too complex to be modelled with the GR4J model. Wu et al. (2010) show that the model over- or underestimates streamflow and that in some cases, this can be reduced by using soil moisture observations if available. Kavetski and Fenicia. (2011) show that the hourly version of the model (GR4H) was unable to capture seasonal dynamics and the switch in hydrological response from wet to dry conditions in the Wollefsbach catchment (Luxembourg).

Although there may be many causes of the failure of this model, structural inadequacy or the lack of flexibility in model structure is indicated to be and generally seen as one of the main reasons for its failure (Perrin et al., 2011; Perrin et al., 2003; Wagener, 2003). Identifying the role of GR4's structure in the failure of the model is however very difficult because of its fixed empirical nature. The SUPERFLEX approach on the other hand, is designed to allow for flexibility in the model structure.

The SUPERFLEX approach uses generic model components that can be combined in any fashion or order and is therefore suitable for testing different hypotheses of model structure. The limited amount of examples of the approach and the uncertainty about the amount of resources required to set up a working model are reasons for modellers to prefer a readymade conceptual model like GR4. However, the large flexibility and the generic components of the SUPERFLEX approach allow for the evaluation of different conceptualisations and link model components to hydrological processes. These insights may help reduce errors or subjectivity in model conceptualisation and perhaps pinpoint them in other models.

1.4 Problem definition

The previous discussions showed that hydrologists are often unable to accurately describe hydrological processes taking place in a catchment because of the lack of knowledge or ability to measure the needed parameters. The results of modelling exercises are generally good, but assumptions and simplifications made by hydrological modellers sometimes lead to hydrological monsters, even for the well performing and often used GR4 model. Despite the opportunities for learning from model failure, they are given little attention in scientific publications. This means little is known about when and why a catchment becomes a hydrological monster.

In this research the emphasis is placed on the hydrological monsters of the fixed GR4 model and the flexible SUPERFLEX approach and the role of model structure in model results. The findings of Kavetski and Fenicia (2011) are based on a limited amount of catchments which means that little is known about SUPERFLEX results on a large range of catchments compared to the fixed modelling approach GR4 or if flexibility can explain why certain catchments become hydrological monsters.

1.5 Objective and research questions

Given the problem definition in the previous paragraph the objective of this research is:

To find and clarify the hydrological monsters of a fixed and a flexible modelling approach by investigating which model structures perform well on average and why catchments become monsters for some structures.

As described in the previous paragraph, GR4H represents the fixed modelling approach and SUPERFLEX the flexible approach in this study. The objective is translated in the following research question and sub-questions:

When and why do a catchment and a model become a hydrological monster?

- *What are the effects of increasing model complexity on model performance on different types of catchments?*
- *What are successful model structures and what are hydrological monsters when the fixed and the flexible modelling approaches are strictly followed?*
- *What do the monster catchments look like why do model structures perform poorly on these catchments?*

1.6 Research outline

The next chapter (2) describes the catchment set of 250 French catchments available for this study. On these catchments a fixed and a flexible modelling approach are applied to compare the performance of different model structures. The model structure(s) used in both approaches are described in chapter 3. Chapter 4 describes the calibration and validation; each model structure is calibrated and validated in the same way to ensure the results are comparable. Chapter 5 discusses the performance of the model structures on the used catchment set and examines the influence of several model components. In chapter 6, the fixed and the flexible modelling approaches are compared in a more strict way. These results are used to select the best model structures and the hydrological monsters. In chapter 7 the results of chapter 5 and 6 are used to demystify the hydrological monsters. Chapter 8 and 9 contain the discussion and the conclusions and recommendations drawn from this study.

2 Data

This chapter describes the observed rainfall, potential evapotranspiration and discharge data (2.1) of the available catchment set. Additionally, several analyses where done to determine the quality of the data (2.2) and find possible monster catchments without any modelling (2.3). To aid analysis of the model results in later stages, the catchments are classified using catchment characteristics (2.4).

2.1 Catchment data

Figure 2—1 shows the location of the 250 catchments that are available for this research. The catchments are spread throughout France and range in size from 16 km^2 to 6836 km^2, with an average of 567 km^2. To be better able to simulate flood events in small catchments, hourly time series are used. For each catchment, hourly time series of rainfall, potential evapotranspiration (Penman-Monteith, PET) and discharge are available between 1997 and 2007. Additional daily rainfall and evaporation data (disaggregated at the hourly time step using a uniform distribution) for the 1994-1997 period were used for model warm-up. Meteorological data were provided by Météo-France (n.d.) and hydrological data originates from the national flow archive (banquet HYDRO managed by SCHAPI, MEDD, 2007). Physical catchment characteristics such as elevation and land cover maps are also available (Bourgin et al., 2011; European Environment Agency, 2006).

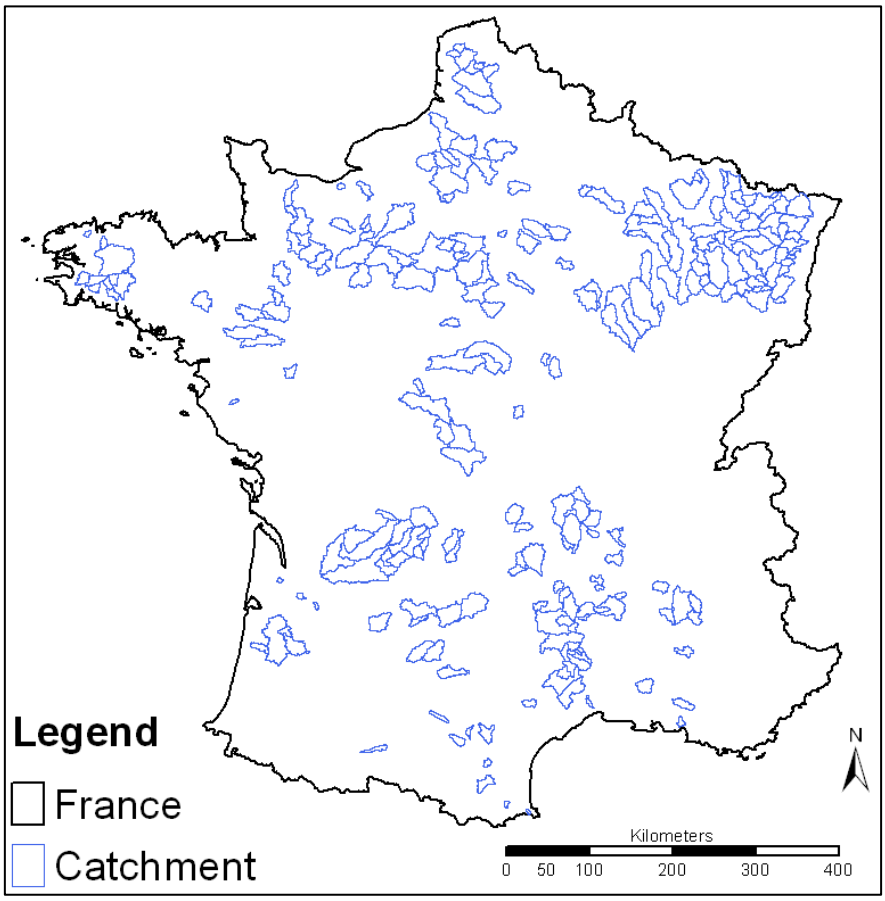

Figure 2—1. Overview of 250 catchments used in this research.

The catchments show a wide range of meteorological and physiographic properties including:

- mean annual precipitation ranging from 61 to 1961 mm/y (mean: 988 mm/y),
- mean annual ranging discharge from 84 to 1329 mm/y (383 mm/y),
- runoff coefficients (mean discharge over mean rainfall) from 0.11 to 0.83 (0.37),
- the Wetness Index (or Aridity Index, mean rainfall over mean PET) ranging from 0.41 to 3.03 (1.37) shows that both wet and dry catchments are represented,
- annual mean temperatures range from 6.8 to 14.5°C (10.5°C) with mean temperature during three winter months between -0.4 and 7.4°C (3.5°C), and
- mean catchment altitude ranges from 41 to 1276 m above sea level (383 m a+sl),
- slopes ranging from 0.01% to 0.42% (0.10%),
- river length varies from 5 to 138 km (28 km).

2.2 Data quality

The quality of the data is reviewed and documented to exclude it as a source of poor model performance. The available rainfall and PET data are considered of high quality since they have been pre-processed by Météo-France. The discharge series contain missing data and parts of the data have been interpolated. Missing data are easily pinpointed by negative values for discharge and skipped in model calibration and validation. However, interpolations are more difficult to detect.

Interpolations may have been created during the recording of the discharge data and can either be the consequence of malfunctioning of the measuring devices (floater accuracy) or errors during data processing (i.e. rating curve: transforming floater measurements of water level to hourly discharge measurements). In this research an attempt was made to detect and remove interpolations from the time series so model calibration and validation are performed on real data only.

Especially during low flow when the river shows little variability, interpolations are difficult to detect and difficult to distinguish from constant flow situations. To reduce the loss of valuable data, the minimum length of a interpolation was set at 48 hours. Any interpolated parts of the time series where omitted from the series and considered as missing data.

Measurement errors are even more difficult to detect, especially without visual inspection of the time series. Measurement errors can exist both in the meteorological and discharge data and can be of different scale. Small errors may occur as the result of malfunctioning measurement equipment and river regulation or distortions. The data were visually checked for small scaled spiky behaviour in the discharge and this was documented for each catchment. This behaviour is not considered to be of major influence in model performance but may explain poor model performance for low flow.

Larger scale errors may be the results of systematic errors in rainfall or discharge measurements and will affect the water balance of the catchments. The data are checked for these errors through the use of a non-dimensional plot of the water balance and visual comparison of neighbour catchments through cumulative QQ-plots. The non-dimensional plot of the water balance (runoff coefficient plotted against Wetness index) allows to easily pinpointing catchments with high or low discharge compared to rainfall and PET. The cumulative QQ-plots compare the behaviour of one catchment to its neighbours enabling the detection of systematic errors or any sudden changes in a catchment which could point to a change in the river at hand or to the measuring device (e.g. river regulation or repositioning of measuring device).

From the 250 catchments available, twelve catchments are rejected due to over 15% missing data and one catchment due to obvious incorrect data. A total of 237 catchments remain for further analysis. Appendix A shows examples of the data quality analyses and the full results.

2.3 Suspected monster catchments

The results of the analyses described in the previous section may point to catchments for which it is unlikely that any hydrological model will show good performance. Hydrological models can generally not cope with large changes to the catchment without changes to the model structure or parameters. To investigate the influence of the anomalies found in the data, these anomalies are reported so they can be used in a later stage. Table 2—1 shows some example catchments selected as possible monsters due to data errors.

Table 2—1. Suspected monster catchments based on observed data

Catchment description			Data Quality			Possible monster	
Code	Name of measurement station	Surface (km^2)	Missing data (%)	Inter-polated >48h (%)	Spikes	Non-Dimension. plot	Neigh-bour conflicts
A9021010	La Sarre à Sarrebourg	307	0.0%	3.4%		x	x
B1092010	Le Mouzon à Circourt-sur-Mouzon [Villars]	401	0.8%	0.8%		x	x
E3518510	La Laquette à Witternesse	81	2.5%	2.4%		x	x
H8012010	L'Epte à Gournay-en-Bray	247	1.0%	0.1%			x
H8043310	L'Aubette de Magny à Ambleville	101	2.7%	3.2%			x
I2213610	L'Ancre à Cricqueville-en-Auge	60	0.1%	5.1%			x
P7261510	L'Isle à Abzac	3758	0.0%	0.4%	x		x
V6035010	Le Toulourenc à Malaucène [Veaux]	157	0.0%	2.1%			x
Y1345010	Le Lampy à Raissac-sur-Lampy	58	2.2%	6.6%			x

2.4 Catchment classification

To help characterize the catchments in this study and enable the selection of similar catchments, four characteristics are selected by which the catchments are classified: catchment area, Wetness Index, permeability and the ratio between summer and winter runoff coefficient ($RC_{S/W}$). Table 2—2 shows the correlation between the selected characteristics. The low correlation between the characteristics ensures that they are independent and thus truly show a different aspect of each catchment.

Table 2—2. Pearson correlation coefficients between the classification characteristics.

	Area	Wetness Index	Permeability	$RC_{S/W}$
Catchment area	1			
Wetness Index	-0.01	1		
Permeability	-0.02	0.32	1	
$RC_{S/W}$	-0.01	-0.01	-0.28	1

Catchment area is the first characteristics by which the catchments are classed. The 237 catchments are divided into three groups with approximately equal number of catchments in the Small, Medium and Large class.

The *Wetness Index* is used as an indication of the meteorological conditions in a catchment. The index is calculated as the ratio between rainfall and potential evapotranspiration; it is therefore independent of the physical properties of a catchment. The index is calculated using only the ten years of real rainfall and PET data from 1997 up to 2007. The catchments are classified Wet, Moist or Dry.

Permeability is used to classify the catchments according to physical characteristics. This characteristic is based on the type of bedrock underlying the catchments and is classified in Permeable, Semi-permeable or Impermeable by the European Environment Agency (2006). Permeability is linked to catchment response (Hellebrand et al., 2008) and depends on both catchment slope and drainage density (Le Moine, 2008). It is therefore considered to be a good representative of the physiographic properties in a catchment.

$RC_{S/W}$ is considered as a more integral property of climate and catchment characteristics, but still remains largely uncorrelated with the other characteristics. The runoff coefficient is defined as the ratio between rainfall and discharge and in this case is calculated separately for three summer (July to September) and three winter (January to March) months (averaged over ten years of data):

$$RC_{S/W} = \frac{RC_S}{RC_W} = \frac{P_S/Q_S}{P_W/Q_W} \qquad \text{Eq. 2—1}$$

The ratio of these two coefficients describes the amount of rainfall compared to flow during summer normalised by that of during winter and is best explained by two extreme cases:

1. High $RC_{S/W}$: when runoff compared to rainfall is high in summer, it is likely that the river is fed by additional groundwater that was stored in winter. This means that some rainfall in winter did not go to runoff and thus lowering the runoff coefficient in winter. This case is classified as groundwater dominated runoff (GW runoff).
2. Low $RC_{S/W}$: when runoff compared to rainfall in summer is low, water might be lost to PET and little water from storage will flow in the river. Rainfall in winter is then more likely to flow in the river more directly as there is little storage. This case is classified as Direct runoff.

A middle class is added named Mixed and again all three classes are given approximately the same number of catchments. Note that since this fourth characteristic uses observed streamflow, it could not be used as such in an ungauged basin perspective, if one intended to link models and catchment types.

The qualitative scales chosen here may not match the scales often found in the literature for these four characteristics. Here the intention was simply to distinguish between catchments with below-median, median or above-median characteristics on the catchment set.

Table 2—3 shows an overview of the classes that are explained above.

Table 2—3. Classification ranges and number of catchments in each range.

Property	Classification	Range	# Catchments
Catchment Area ***[km^2]***		16 – 6836	(237)
	Small	< 200	85
	Medium	200 – 600	79
	Large	> 600	73
Wetness Index ***[-]***		0.41 – 2.03	(237)
	Dry	*< 1.2*	78
	Moist	*1.2 – 1.5*	83
	Wet	*> 1.5*	76
Permeability		Classified by European Environment Agency (2006)	(237)
	Impermeable		89
	Semi-permeable		85
	Permeable		63
$RC_{S/W}$ ***[-]***		0.03 – 1.00	(237)
	Direct runoff	< 0.15	87
	Mixed	0.15 – 0.24	74
	GW runoff	>0.24	76

3 Models and modelling process

The GR4H model and the structures used in the SUPERFLEX study are described briefly in this chapter while appendices B and C contain a more detailed overview of both models. This chapter focuses on the differences between both modelling approaches that may be important for the performance of the models.

3.1 GR4H model

The GR4H model has a single fixed structure that lumps hydrological processes while only four parameters are calibrated. Figure 3—1 shows the structure of the GR4 model that uses rainfall and PET as input and generates discharge as output.

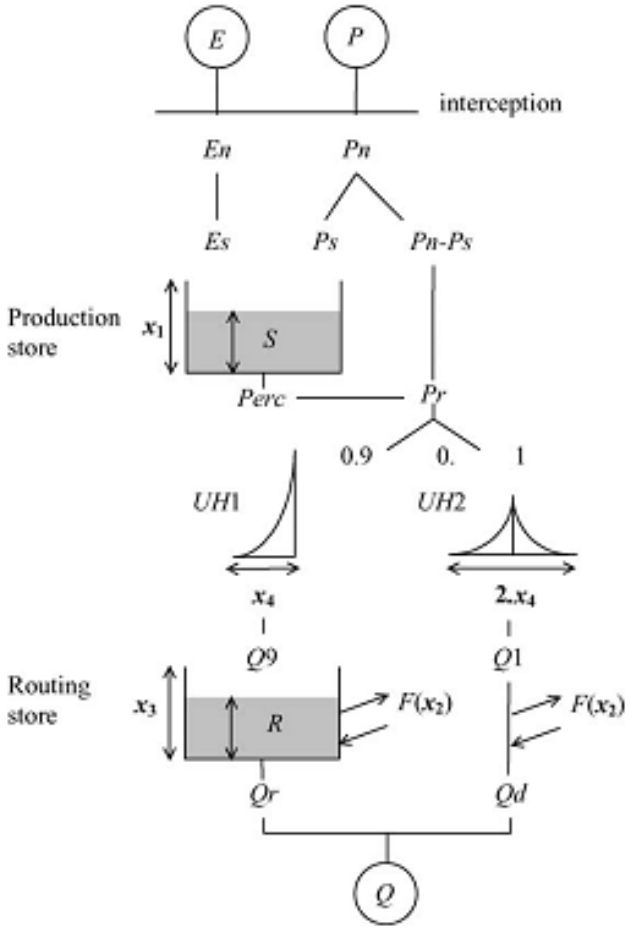

Figure 3—1. Diagram of the GR4 model (Perrin et al., 2003)

Rainfall and PET are subtracted to find the net precipitation P_n or net evaporation E_n. P_n is partitioned between storage into a soil moisture accounting (SMA) reservoir named the Production store *S*, and effective rainfall (P_n- P_s). The Production store is depleted by a percolation function *Perc* that is added to effective rainfall. Effective rainfall is then routed to the outlet via a two-branch routing module. The first branch (10% of effective rainfall) is routed by a single unit hydrograph. The other 90% are routed by a unit hydrograph and a non-linear reservoir named Routing store *R*. A water exchange function *F* is applied to the two flow components, to simulate import or export of groundwater with the underlying aquifer or neighbouring catchments.

3.2 SUPERFLEX structures

The SUPERFLEX framework allows for the construction of different model structures by combining different components. These structures can be hypothesized based on the knowledge of catchment behaviour or calibrated to provide the best fit between simulated and observed data. This study uses twelve structures hypothesised by Fenicia et al. (2012) to differ in a controlled way and cover a broad range of model complexities. Figure 3—2 shows the twelve structures used that differ stepwise allowing for the analysis of the influence of individual components.

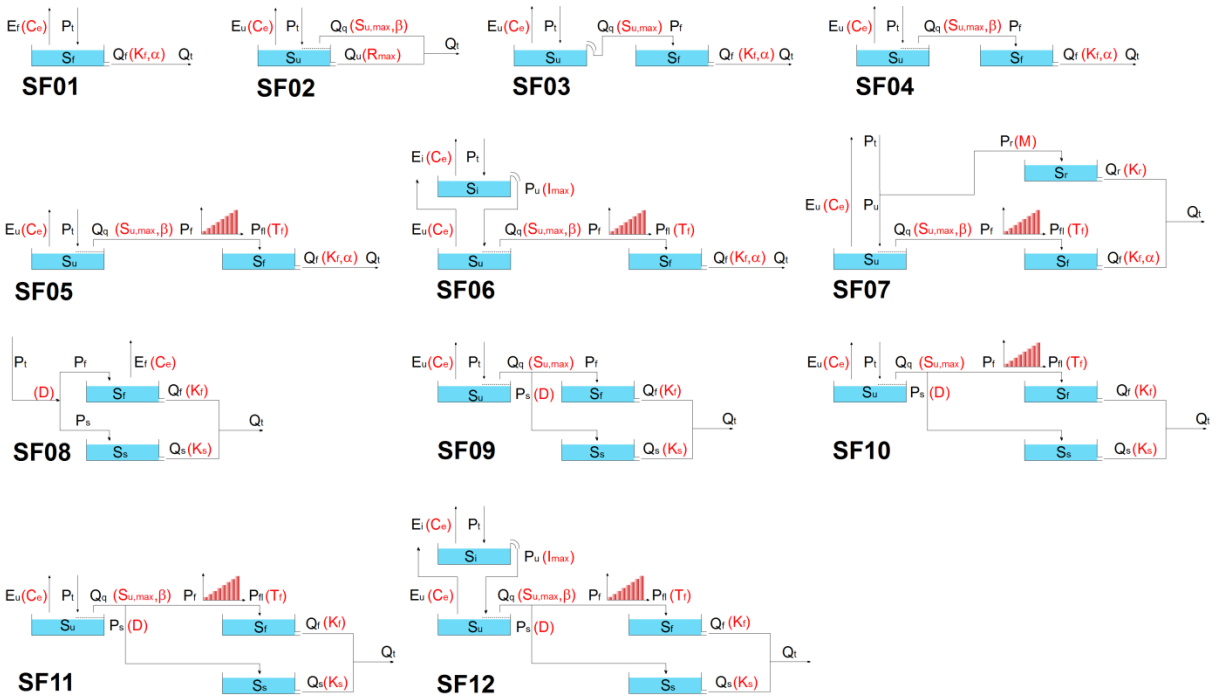

Figure 3—2. Twelve model hypotheses generated using the SUPERFLEX model-development framework (Fenicia et al., 2012).

The structures differ in the number and type of reservoirs, lag-functions and junction elements so that the influence of individual components can be assessed. The structures differ in complexity starting with SF01 and SF02, the two most simple single-reservoir models. SF01 is a single reservoir model with a non-linear discharge relation. The fast reservoir (FR) is filled by rainfall *Pt* end emptied by PET and a flow *Qf*. The potential evapotranspiration used as input is corrected by a fixed ratio *Ce* that is calibrated and is shown in red. The flow *Qf* depends on two parameters, *Kf* and α. *Kf* is the residence time in the FR and *alpha* a power that controls the shape of the produced flow. SF02 uses a single reservoir as well, but differs in the way flow is generated depending on the water level in the reservoir. Again a flux of water is denoted by a letter in black and calibrated parameters are between brackets in red.

In structures SF03 to SF05 an unsaturated reservoir (UR) is added and connected in series to the FR. These three structures vary in the number of calibrated parameters and use functions to describe flow to and from the reservoirs. In SF05, a lag-function is introduced for the first time which distributes flow over multiple time steps. Structure SF06 to SF11 all use three reservoirs, but of different types and with the introduction of more complex connections and functions. SF07 is the first structure with reservoirs connected in parallel allowing for more independent flow components along two flow paths. SF08 is a simple structure with parallel reservoirs: a slow reservoir SR is introduced with a residence time *Ks* independent to that of the FR. From this structure onward

complexity is again increased up to the introduction of a fourth reservoir in structure SF12. Describing all structures in detail would require too much space, but appendix C describes SF12, the most complex structure in detail and gives an overview of all functions used in all the structures.

3.3 GR4H and SUPERFLEX

Since SUPERFLEX allows for the combination of various reservoirs and functions, the GR4H model can be recreated in SUPERFLEX. Figure 3—3 shows the GR4H model in SUPERFLEX components. Like Figure 3—1 this is a simplified image since the complexity of the used functions is not depicted. This image is only used to clarify the differences between GR4H and the other SUPERFLEX structures.

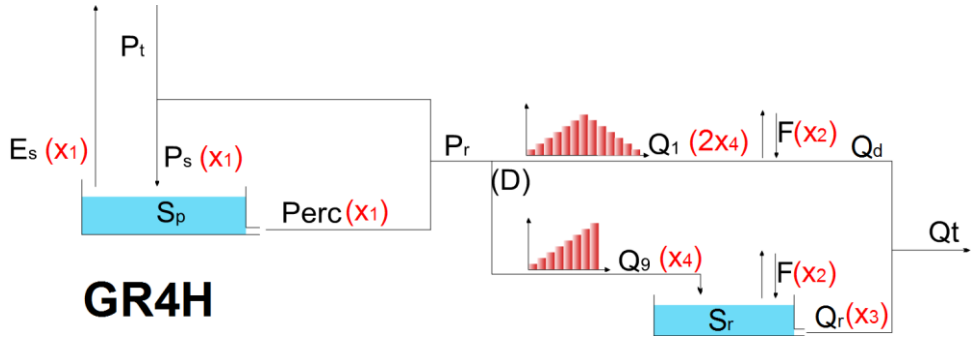

Figure 3—3. The GR4H model in SUPERFLEX components.

The GR4H model is most similar to the SF05 model but there are still some differences. The main differences between the GR4H model and the SUPERFLEX structures are in the used connections and functions:

- The first difference is the correction applied to the PET entering the first reservoir. In GR4H this is dependent on the relative amount of water in the first reservoir (saturation coefficient) while SUPERFLEX uses a constant that is calibrated for the whole time series.
- The addition of a direct connection from rainfall to the second part of the GR4H model can be compared with the amount of water going to a riparian reservoir in SF07. GR4H uses a more complex function to derive the ratio between water entering the production store or the direct route which is dependent on the saturation coefficient of the first reservoir.
- Percolation from the Production store in GR4H is generated by a power function similar to the ones used in SUPERFLEX, but with the addition of fixed empirical parameters.
- The combined direct and indirect flow in GR4H is divided by a fixed ratio while in SUPERFLEX, parameter *D* is calibrated and thus adds to model flexibility.
- GR4H contains two parallel lag-functions of which one leads directly to runoff while the other fills a slow reservoir. This is different from the parallel structures used in SUPERFLEX where only parallel reservoirs are implemented.
- In GR4H, the function that determines the flow from the routing store does use a power function, but the power value was empirically fixed and is not calibrated.
- Finally, GR4H introduces a flow component that can correct for intercatchment groundwater flow, which is not applied in any of the SUPERFLEX structures.

The SF05 model is the closest match to the GR4H model. The amount of flow going through the upper connection (parallel direct runoff) in GR4H is only 10%, which makes the routing store more important. This part is comparable with the combination of UR and FR in series like in SF05.

3.4 Structural complexity

To compare the GR4H model and the SUPERFLEX structures it is important to keep track of the complexity of each structure. Some model structures use threshold or lag-functions while others only use linear functions. These distinctions can be useful in the analysis of differences in performance and are shown in Table 3—1. To quantify complexity of a model the number of calibrated parameters (N_θ) is added to the number of states (N_s) , where states are the number of reservoirs and lag-functions that represent the hydrological processes in a catchment. In case two structures have the same complexity score, the type of functions will determine which is more complex (e.g. a power function is more complex than a linear relation).

Table 3—1. Distinctions between the twelve SUPERFLEX structures and GR4H with N_{res}, N_θ and N_s for number of reservoirs, calibrated parameters, and states respectively and complexity as the sum of N_s and N_θ.

Structure	N_{res}	Connection	Type of function(s)	N_θ	N_s	Complexity
SF01	1	-	Power	3	1	4
SF02	1	-	Power + Linear	4	1	5
SF03	2	Series only	Threshold + Power	4	2	6
SF04	2	Series only	Power	5	2	7
SF05	2	Series only	Lag + Power	6	3	9
SF06	3	Series only	Threshold + Lag + Power	7	4	11
SF07	3	Series and parallel	Lag + Power + Linear	8	4	12
SF08	2	Parallel only	Linear	4	2	6
SF09	3	Series and parallel	Linear	5	3	8
SF10	3	Series and parallel	Lag + Linear	6	4	10
SF11	3	Series and parallel	Lag + Power + Linear	7	4	11
SF12	4	Series and parallel	Threshold + Lag + Power + Linear	8	5	13
GR4H	2	Series and parallel	Lag + Power + Linear	4	4	8

The GR4H model also uses some hidden parameters that were based on empirical research. These parameters have shown good performance in a large group of catchments during the development of the model (Perrin et al., 2003) and are now kept fixed. These parameters are not incorporated in the complexity measure. Another difference with the SUPERFLEX structures is the numerical implementation of the GR4H approach. All equations used in a SUPERFLEX structure are solved implicitly for each time step, so all at once. The equations in GR4H are solved sequentially, so one after the other, which can lead to numerical instability in some cases (Kavetski and Clark, 2010).

4 Methods

This chapter describes the methods used to calibrate the models (4.1), how the model results are compared (4.2) and how the monster catchments will be selected (4.2.3).

4.1 BATEA and calibration

BATEA (An Integrated Bayesian Inference and Prediction Environment) is a programme developed and used for Bayesian inference and prediction in hydrological modelling (Kavetski & Evin, 2011; Kavetski et al., 2006). This research uses the programme for calibration as it holds both GR4H and SUPERFLEX models, so all models are calibrated in exactly the same way which provides an objective comparison.

The BATEA programme allows for the use of different objective functions and optimization algorithms commonly used in hydrological modelling. In this case, GR4H and SUPERFLEX are calibrated on a Weighted Least Square (WLS) objective function using a Quasi-Newton optimization method with twenty multi-starts. This method was found to be a fast and reliable way to calibrate a hydrological model (Kavetski et al., 2007). As all models use the same data and are calibrated in the same way, these aspects can be excluded as a source of model failure (Troldborg et al., 2007) and leave the model structure as the main variable under investigation.

4.1.1 Objective function

The WLS differs from the Standard Least Square in that it does not assume that the error in flow is identical over the entire time series but can vary with time, flow or any other known variable. In the case of rainfall-runoff models it is often observed that prediction errors are larger during high flow and this knowledge can be implemented in the calibration of the model. BATEA uses the Bayes method to include this information in the optimisation approach which maximises the log Likelihood (the objective function).

The principle of the Bayesian analysis is to derive knowledge about model parameters from prior knowledge (some assumptions or expectations) and any available data (Kavetski et al., 2006). In the Bayesian analysis, knowledge is defined as some probability density for the parameters given the available data $p(\theta|Q_{obs})$, where θ represents the parameter set and Q_{obs} the observed discharge data. The Bayes equation is given by:

$$p(\theta|Q_{obs}) = \frac{p(Q_{obs}|\theta)p(\theta)}{p(Q_{obs})} \propto p(Q_{obs}|\theta)p(\theta) \propto p(Q_{obs}|\theta) = L \qquad \text{Eq. 4—1}$$

where $p(Q_{obs}|\theta)$ is the probability of the data given the parameter set, $p(\theta)$ holds any prior knowledge of the distribution of the parameters and $p(Q_{obs})$ the probability density of the observed data. The latter can be left out for simplicity since it is independent of the chosen parameter set and thus a constant through the calibration (a maximum of the terms $p(Q_{obs}|\theta)p(\theta)$ will also yield a maximum of $p(\theta|Q_{obs})$).

In this study no prior knowledge of the distribution of the parameters is known, except the minimum and maximum values. The minimum and maximum values are set as bounds for the calibration meaning that the prior $p(\theta)$ is a uniform distribution. This then only leaves the probability of the data given the parameter set $p(Q_{obs}|\theta)$ as shown in Eq. 4—1, which is also called Likelihood L.

To compute the Likelihood, some assumptions must be made about the nature of errors. This study assumes that the errors between observed and simulated discharge are independent from time step to time step and normally distributed with zero mean, the likelihood then becomes:

$$L(Q_{obs}|\theta) = \prod_{i=1}^{N} N\left(Q_{obs,i} - Q_{sim,i}|0, \sigma_i\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2}\frac{\left(Q_{obs,i} - Q_{sim,i}\right)^2}{\sigma_i^2}\right) \quad \text{Eq. 4—2}$$

$$L(Q_{obs}|\theta) = \left(\frac{1}{\sqrt{2\pi}\bar{\sigma}}\right)^N \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\frac{\left(Q_{obs,i} - Q_{sim,i}\right)^2}{\sigma_i^2}\right) \quad \text{Eq. 4—3}$$

where L is the Likelihood, i the time step, N the total number of time steps, σ_i the standard deviation at time step i and $Q_{obs,i}$ and $Q_{sim,i}$ are the observed and simulated stream flow at time step i.

Instead of the absolute likelihood, BATEA uses the natural logarithm of the likelihood because it is easier to use and avoids problems with very large or very small values. Eq. 4—3 then becomes:

$$\text{Ln L} = -N \,\text{Ln}\sqrt{2\pi} - \sum_{i=1}^{N} \text{Ln}\,\sigma_i - \frac{1}{2}\sum_{i=1}^{N}\frac{\left(Q_{obs,i} - Q_{sim,i}\right)^2}{\sigma_i^2} \quad \text{Eq. 4—4}$$

in which the first term is always constant, the second term varies through the variable standard deviation or error in each time step and the third term by the sum of squared errors and the variable standard deviation. The variable standard deviation corresponds to the main assumption of the WLS method that error is not constant over the whole time series, but can vary per time step. In this case σ is assumed to vary with discharge so more emphasis can be placed on low flow.

Eq. 4—5 gives the equation for the log likelihood according to the WLS method, omitting the constant term from Eq. 4—4 for simplicity. In the second part of Eq. 4—5, the standard deviation in the denominator acts as a weight to the error in prediction. Since we assumed error in prediction is higher in high flow, the standard deviation will be higher in a time step with high flow. This will reduce the weight of the error in prediction during high flow compared to that during low flow. An error in low flow now becomes more important compared to that in high flow.

$$\text{Ln}\,L \propto -N \,\text{Ln}\,\sigma_i - \frac{1}{2}\sum_{i=1}^{N}\frac{\left(Q_{obs,i} - Q_{sim,i}\right)^2}{\sigma_i^2} \quad \text{Eq. 4—5}$$

$$\sigma_i = a + bQ_{sim,i} \quad \text{Eq. 4—6}$$

Eq. 4—6 shows the linear relation between σ and simulated flow, simulated flow is used to increase the speed with which the Likelihood converges to a maximum.

4.1.2 Optimization method

Newton optimization algorithms are stepwise local optimization algorithms that use the gradient of the objective function to choose the direction of the next step in parameter optimization (Sorooshian & Gupta, 1995). Multi-starts are used to prevent finding only a local optimum. The local optimizer is applied multiple times using different initial values across the parameter space saving only the best performance. The Quasi-Newton method simplifies the way the gradient of the objective function is calculated to reduce computation time (Schoenberg, 2001) and was shown to be effective and efficient in finding the optimal parameter values of conceptual hydrological models using multi-starts on a smoothed parameter space (Kavetski & Kuczera, 2007). The objective function is smooth as a

result of the model implementation in BATEA, the constitutive functions ensure continuous behaviour instead of sudden changes in model behaviour (i.e. when a reservoir dries up from one time step to the next, see appendix C). In preliminary testing using twenty randomly selected catchments, it was found that with the smoothed parameter space, twenty multi-starts are sufficient to find the global optimum (i.e. this optimum was found multiple times in twenty starts for all test catchments).

4.1.3 Calibration periods

Calibration is performed following the full split-sample test in which the data record is split into two parts and used twice (Klemes, 1986). To reduce initialisation problems, an initial value was given to the level in the reservoirs (as a fraction of reservoir capacity) and three years of data were used as warm-up period for the model. These settings were selected based on expert knowledge of the GR4H model and previous experiences with these catchments (Le Moine, 2008).

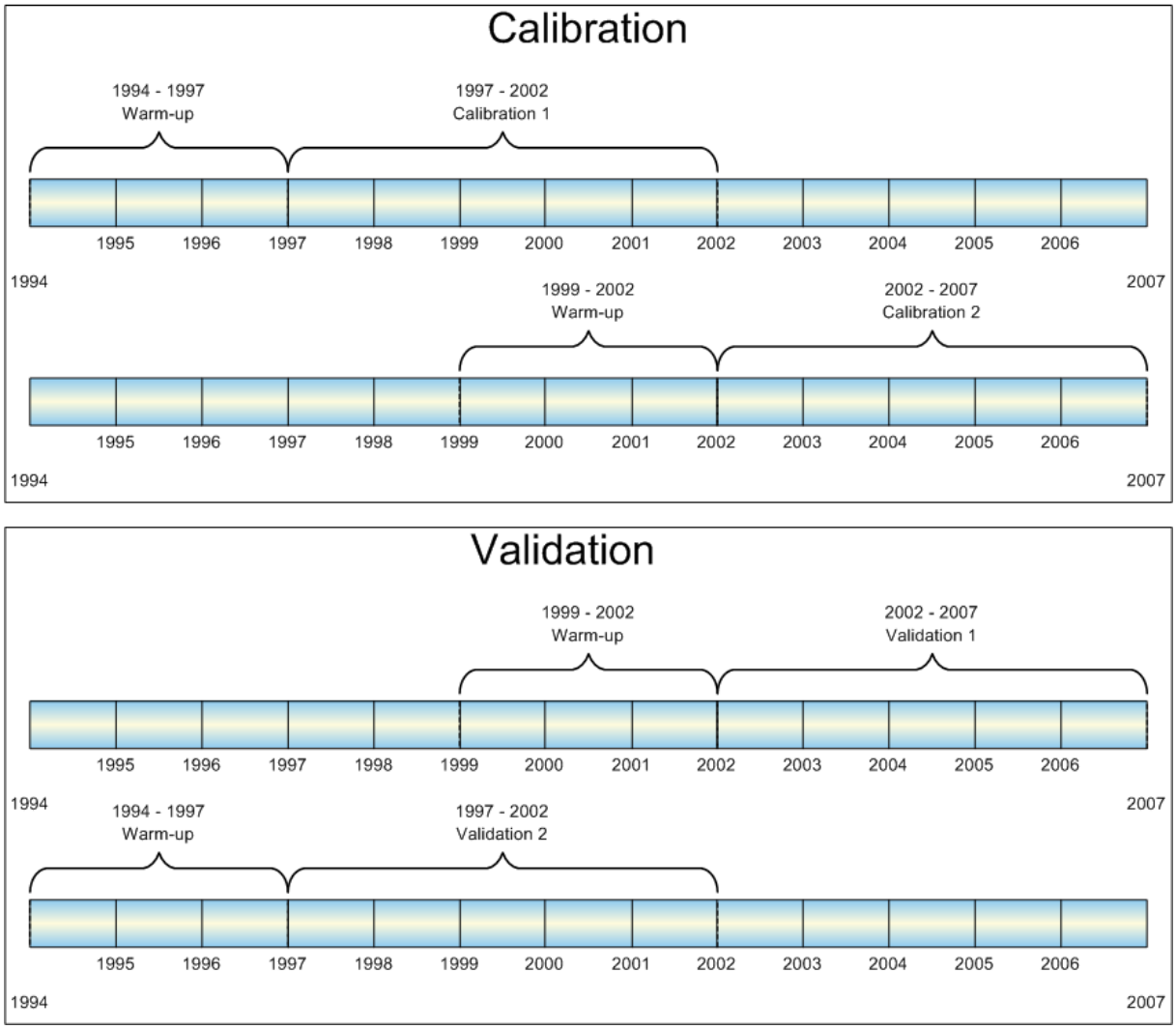

Figure 4—1. Application of full split-sample test on available data.

The ten years of real data are split in two equal periods of five years, calibration is performed on the first five years of data while the second five years are used for validation. The model is then calibrated again for the second five years of data and then validated on the first five years (Figure 4—1). This means that the model is calibrated twice on each catchment and the results of each

calibration can be compared as an additional quality check, especially important in the case of the SUPERFLEX approach when both structure and parameters may be different.

The results of the full split-sample test are two data series with ten years of simulated discharge, from now on referred to as the two splits. The first split has been calibrated on the first five years of observed data and the second calibrated on the second five years; in each case, the remaining five-year period is used for validation.

More on the calibration of both models (parameter bounds, functions and structures) can be found in appendices B and C.

4.2 Comparison protocol

The comparison protocol is designed to compare the fixed with the flexible modelling approach and derive as much information as possible about the behaviour of the models from their performance. The comparison comprises of two part. In the first part all SUPERFLEX structures are treated as separate models to give more insight in the influence of structural components on performance. In the second part the fixed approach with GR4H is compared to the flexible approach SUPERFLEX to show whether or not the fixed and the flexible method are comparable and both yield viable results on all the catchments (i.e. are consistent in their representation of a catchment).

4.2.1 Comparing 13 model structures

In the first step of the comparison the twelve SUPERFLEX structures are treated as separate models, so including GR4H thirteen model structures will be compared. The comparison will be based on two measures: performance and robustness. Additionally, the catchment classification (section 2.4) will be used to investigate the performance of different model structures on different types of catchments. The stepwise structural differences in the SUPERFLEX structures and any differences among catchment classes defined in section 2.4 will help to identify which model components are important for which type of catchment.

Performance

A set of evaluation criteria was selected to evaluate the *performance* of the model structures. Performance is a measure of goodness of fit between simulated and observed discharge. The criteria will be calculated in the same way for all models and are used to give better insight in the model performance, making different catchments comparable and be able to focus on different aspects of model performance (i.e. high or low flow).

The performance of each model structure is evaluated using the score of four evaluation criteria calculated in the validation for each split-sample test. Model evaluation in validation is more relevant than in calibration since models are used in validation mode in practice, when no observed data are available. Each of the evaluation criteria is selected to focus on a different quality of the simulated discharge (high flow, low flow, volume error and variability of prediction). Eq. 4—7 to Eq. 4—10 give the evaluation criteria.

$$CR1 = 1 - \frac{\sum_{i=1}^{N}\left(Q_{obs,i} - Q_{sim,i}\right)^2}{\sum_{i=1}^{N}\left(Q_{obs,i} - \overline{Q_{obs}}\right)^2} \quad \text{Eq. 4—7}$$

$$CR2 = 1 - \frac{\sum_{i=1}^{N}\left(\frac{1}{Q_{obs,i}+\epsilon} - \frac{1}{Q_{sim,i}+\epsilon}\right)^2}{\sum_{i=1}^{N}\left(\frac{1}{Q_{obs,i}+\epsilon} - \overline{\frac{1}{Q_{obs}+\epsilon}}\right)^2} \quad \text{Eq. 4—8}$$

$$CR3 = 1 - \left|\sqrt{\frac{\sum_{i=1}^{N} Q_{sim,i}}{\sum_{i=1}^{N} Q_{obs,i}}} - \sqrt{\frac{\sum_{i=1}^{n} Q_{obs,i}}{\sum_{i=1}^{n} Q_{sim,i}}}\right| \quad \text{Eq. 4—9}$$

$$CR4 = \begin{cases} -1 + 2\frac{\sigma_{sim}}{\sigma_{obs}}, & \sigma_{obs} > \sigma_{sim} \\ -1 + 2\frac{\sigma_{obs}}{\sigma_{sim}}, & \sigma_{obs} < \sigma_{sim} \end{cases} \quad \text{Eq. 4—10}$$

where $Q_{obs,i}$ and $Q_{sim,i}$ are the observed and simulated discharge at time step i, $\overline{Q_{obs}}$ and $\overline{\frac{1}{Q_{obs}+\epsilon}}$ are the mean values over the selected period, N the number of time steps, ϵ a small constant (one-hundredth of mean flow, see Pushpalatha et al., 2012, p. 178 for more information) and σ is the standard deviation of observed or simulated discharge over a selected period.

Criterion 1 (CR1) is the well-known and often used Nash-Sutcliffe efficiency (Nash & Sutcliffe, 1970) which is most sensitive to peaks in discharge (Perrin et al., 2003). Criterion 2 (CR2) is the Nash-Sutcliffe efficiency based on the inversed discharge emphasizing low flow error (Pushpalatha et al., 2012). Criterion 3 (CR3) is based on the Relative Volume error and thus emphasizes any error in the water balance between observed and simulated discharge (Perrin et al., 2003).

CR1 to CR3 can have values between 1 (perfect fit) and -∞. The values are transformed to a value between 1 and -1 to avoid the influence of very low negative values on the calculation of mean performance (Mathevet et al., 2006; Pushpalatha et al., 2012):

$$CR^* = \frac{CR}{2 - CR} \quad \text{Eq. 4—11}$$

where CR^* is the new value of the criterion now with range [1 to -1] and CR is the original value of the same criterion.

The fourth criterion (CR4) is the ratio between standard deviations of observed and simulated discharges with a maximum value of 1 meaning the simulated discharge was able to reproduce the variability in the observed discharge and a minimum value of -1 when the difference in standard deviation becomes very large (Gupta et al., 2009).

The measure for performance is the average of CR1-CR4 over the validation periods which will be used as the primary criterion for the comparison of the models. Additionally, all four criteria are used to find catchments where models perform poorly for a specific quality of discharge.

Robustness

Apart from performance, robustness is a valuable tool for evaluation (Klemes, 1986) and especially important when comparing models on a large number of catchments. Model *robustness* measures the difference between errors in calibration and validation and is a measure of the correspondence between the catchment and the used model and parameter set.

To avoid any bias caused by differences in flow between calibration and validation periods, the ratio between the squared errors of calibration and validation are used instead of any of the Nash-Sutcliffe criteria as they use the mean discharge of a period (Coron et al., 2012). Eq. 4—12 shows the formula for the calculation of the Robustness, the larger error is always placed in the denominator so the values for R stay between one and zero for better comparison. Note that the second case in Eq. 4—12 corresponding to better performance in validation than in calibration should seldom happen since it indicates that the calibration algorithm was stuck on a secondary optimum. Figure 4—2 shows how the simulated time series (calibration and validation) are used to calculate robustness R.

$$R = \begin{cases} \dfrac{\sum_{i=1}^{N}\left(Q_{obs,i} - Q_{sim,i}\right)^2_{cal}}{\sum_{i=1}^{N}\left(Q_{obs,i} - Q_{sim,i}\right)^2_{val}}, & cal > val \\ \dfrac{\sum_{i=1}^{N}\left(Q_{obs,i} - Q_{sim,i}\right)^2_{val}}{\sum_{i=1}^{N}\left(Q_{obs,i} - Q_{sim,i}\right)^2_{cal}}, & val > cal \end{cases} \qquad \text{Eq. 4—12}$$

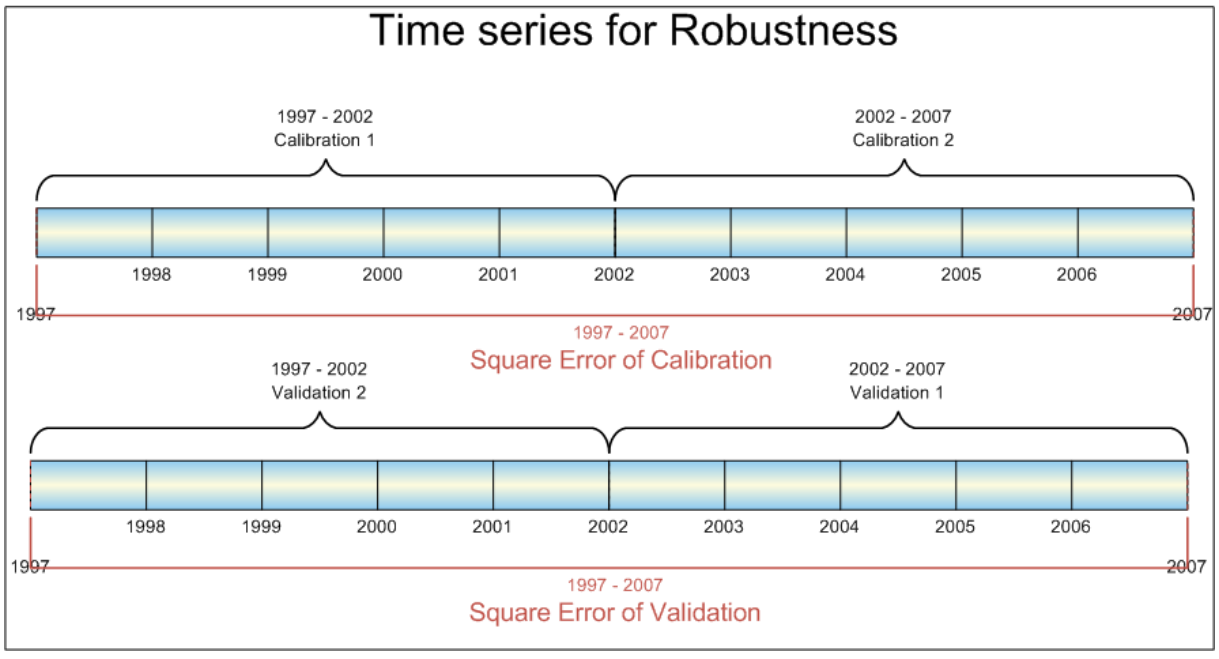

Figure 4—2. Time series showing how the Squared Errors for Robustness are computed.

A large difference between the error in calibration and validation shows that the calibrated parameter set is not suited to predict discharges in another period. This indicates that the model structure is not able to capture the actual hydrological processes taking place in the catchment.

4.2.2 Comparing 2 approaches

Comparing the fixed with the flexible modelling approach means strictly following the modelling process as intended by each approach. For the fixed approach this is very straightforward: apply the GR4H model structure on all catchments and find the parameter set that best represents the catchment behaviour (as described in section 4.1, this means maximising the likelihood in two splits of the time series). The fixed modelling approach is subject to parameter inconsistency when the optimum parameter sets in both splits is very different (see below).

The SUPERFLEX approach advocates adapting not only the parameter set, but also the model structure. Since in this study, the model structure is not left completely free but twelve alternatives are available, the best of these twelve structures is selected based on the objective function. Some structures only differ slightly and may yield similar results because they show a stepwise increase in complexity. To favour parsimony, the selection is made including the measure of complexity described in section 3.4. The simpler structure is selected in case two structures perform very similar. Two models are considered to perform similarly when their respective objective functions differ by less than 5%. The flexible modelling approach can be subject to parameter inconsistency and structural inconsistency.

Parameter inconsistency

In case a model yields two very different parameter sets, this will be considered a modelling failure, since the model is not able to describe the catchment in a consistent way. Whenever one of the model parameters varies by more than 50% between the two splits, the parameters are labelled inconsistent:

$$\frac{|\theta_1 - \theta_2|}{\frac{1}{2}(\theta_1 + \theta_2)} < 0.5 \qquad \text{Eq. 4—13}$$

where θ_1 and θ_2 are parameter values in the first and second split for a given catchment.

Structural inconsistency

In case the two calibrations yield two different best performing model structures, the flexible approach fails to robustly represent the catchment. Therefore any model structure that gives very different scores for the objective function between the splits is considered modelling failure. The likelihood of a structure is allowed to differ by 25%:

$$\frac{|L_1 - L_2|}{\frac{1}{2}(L_1 + L_2)} < 0.25 \qquad \text{Eq. 4—14}$$

where L_1 and L_2 are the likelihoods of a given structure in split 1 and 2 for a given catchment.

The remaining, consistent SUPERFLEX structures and GR4H models will be compared using the performance and robustness measures as defined in the previous paragraph.

4.2.3 Selecting monsters

A catchment and a model structure will be selected as a hydrological monster based on the results of the comparison between the fixed and the flexible approach. A low average performance over the two validation periods or modelling failure because of inconsistency identifies a monster catchment. The comparison will show how many catchments become monsters and whether they can be grouped according to the above mentioned classification (section 2.4).

5 Results I: Comparing 13 model structures

In the first part of the results, the twelve SUPERFLEX structures and GR4H are treated as thirteen separate model structures. The performance in validation of each structure is compared over the 237 catchments, no model structures are rejected based on inconsistency of the parameter set across the two validation periods. Section 5.1 discusses average performance and section 5.2 the structural differences between the model structures. Section 5.3 discusses the performance of each model structure on CR1-CR4 and section 5.4 discusses the average performance on groups of catchments classified using the four characteristics. The above comparisons lead to four hypotheses about four model components that are further investigated in section 5.5. In section 5.6 the robustness of the thirteen model structures is compared and in section 5.7 the main findings of this chapter are summarized.

5.1 Average performance across models

Figure 5—1 shows the distribution of performance for all model structures on the 237 selected catchments. Performance in this figure is the average of CR1-CR4 in both splits. The figure shows that six of the model structures (SF01, SF02, SF03, SF08, SF09 and SF10) perform poorly compared to the best ones and do not seem to be good candidate structures in the perspective of a fixed model structure approach. It can also be noticed from Figure 5—1 that the other seven model structures (GR4H, SF04-SF07, SF11 and SF12) show very similar average performance despite their differences in complexity.

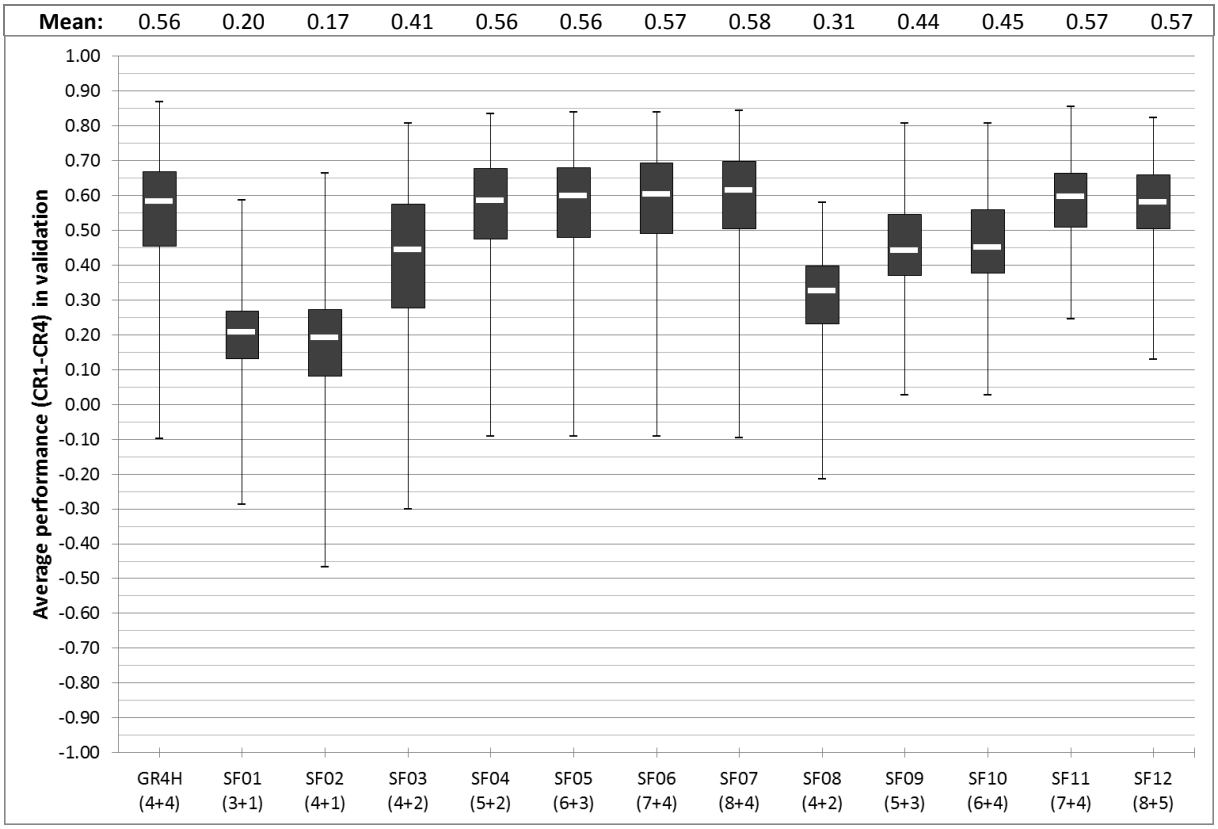

Figure 5—1. Boxplots (maximum, 75th percentile, median, 25th percentile and minimum) of CR1-CR4 values obtained by all model structures in validation on the 237 catchments. The x-axis shows the twelve SUPERFLEX structures plus GR4H, the value between parenthesis denotes the complexity measure (nr. of calibrated parameters + nr. of states, Table 3—1). At the top of the figure the mean values for model performance are given.

Figure 5—1 confirms that GR4H has a relatively good performance on average with an equally high performance range as the most complex SUPERFLEX structures. SF04 and SF05 are models similar to GR4H in terms of complexity and reach the same level of performance making them some of the best performing models for SUPERFLEX. The fixed power functions describing reservoir outflow in GR4H are expected to increase performance just as in the SUPERFLEX structures. Only SF11 has a smaller range in performance but no better mean, performing relatively well for a wider range of catchments.

The increased complexity of model structures SF05-SF07 compared to SF04, SF10 compared to SF09 and SF12 compared to SF11 shows no significant increase in mean performance. Generally a complex model structure provides a closer fit to observed data in calibration. On average, this is what happens for these models (appendix D). However, this does not mean that a more complex model performs better in validation. The high performance in calibration does not mean that the model structure is a better representation of the hydrological processes in a catchment.

5.2 Structural differences between models

Figure 5—2 repeats Figure 5—1 but includes some notes on structural differences between the SUPERFLEX models. The stepwise increase in complexity in models allows for the analysis of the influence of individual components, like reservoirs or functions. Figure 5—3 to Figure 5—5 are used to remind the reader about the differences between the structures.

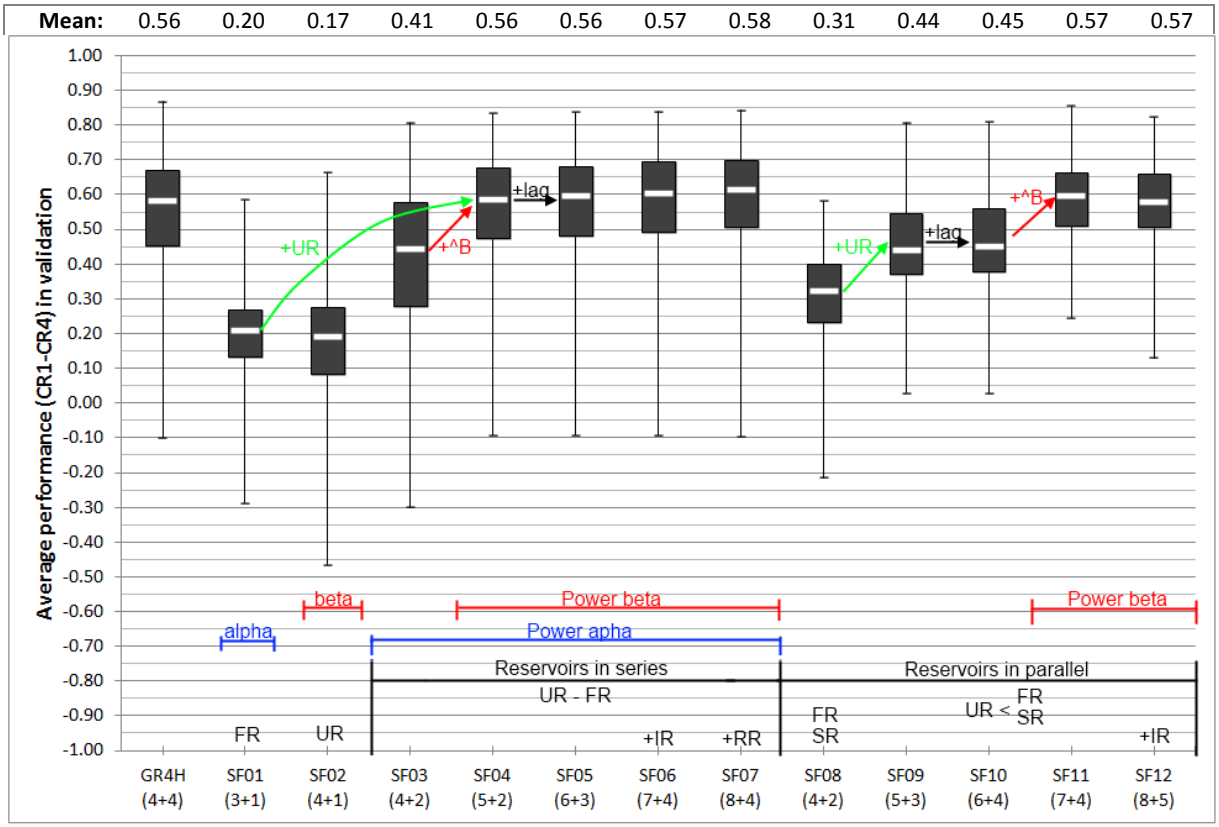

Figure 5—2. Boxplots of CR1-CR4 including notes on differences in SUPERFLEX structures. Where '+^B' means adding a power function β, '+lag' means adding a lag-function, UR = Unsaturated zone Reservoir, FR = Fast Reservoir, SR = Slow Reservoir, IR = Interception Reservoir and RR = Riparian zone reservoir (see also section 3.2).

The two simplest model structures SF01 and SF02 use only a single reservoir and give the poorest performance. The use of only a single fast reservoir (FR) in SF01 or an unsaturated reservoir (UR) in SF02 appears insufficient to model most catchments. The use of the more complex runoff from a single reservoir in SF02 decreases mean performance compared to the single reservoir and power function used in SF01. Introducing a second reservoir in parallel (like in SF08) does not increase performance as much as introducing it in series (SF03-05). However, the complexity of the used functions in SF03-SF05 can also be the cause of this better performance.

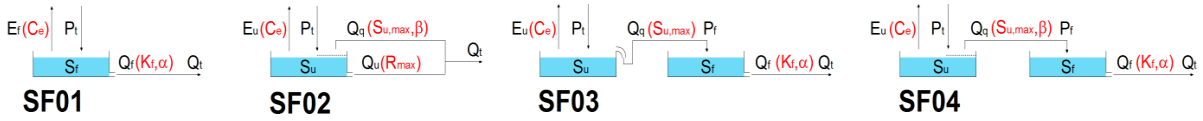

Figure 5—3. SUPERFLEX structures SF01 to SF04.

The threshold reservoir in SF03 used in combination with FR like in SF01 shows the first real increase in performance across the model structures. The relatively wide range of the SF03 performance is an indication of particular types of catchments working well with a threshold while others do not at all. A simple UR and FR in series like in SF04 lead to a large jump in performance and to one of the best performing model structures (from SF01 to SF04).

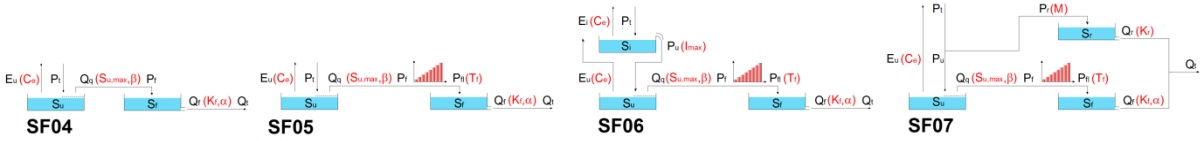

Figure 5—4. SUPERFLEX structures SF04 to SF07.

Model structures SF04 to SF07 show very similar performance despite the increasing complexity. In SF05 a lag-function is introduced between the UR and FR, but this does not lead to improvement on average. In addition to the lag-function, SF06 and SF07 introduce an interception reservoir and a riparian zone reservoir, respectively. The use of the extra reservoirs only marginally increases average performance and varies across the catchments between better and worse, showing again that increasing complexity is no guarantee for better performance.

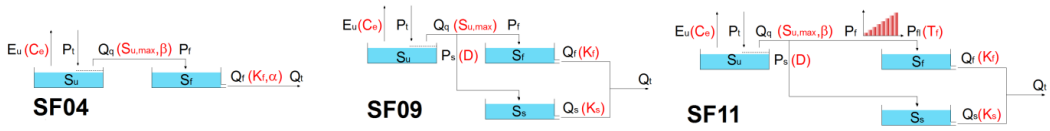

Figure 5—5. SUPERFLEX structures SF04, SF09 & SF11.

SF09 and SF10 introduce an unsaturated reservoir (UR) to the parallel fast and slow reservoir (FR + SR) of SF08 and show a clear jump in performance. The lag-function introduced in SF10 shows little to no increase in performance compared to SF09, much like in SF05. Interesting is the difference between SF04 and SF09: although SF09 is the more complex structure and uses an extra reservoir it performs worse. An apparently important difference between these two structures is the use of a power function (powers α & β) to calculate the runoff generated by the reservoirs. SF04 uses two different power functions for both reservoirs while SF09 uses none. Only when the power function is re-introduced in SF11, between the UR and the FR and SR, performance reaches a similar level as SF04. The effect of yet another IR reservoir in SF12 shows no increase in mean performance.

The differences found in the average performance of the SUPERFLEX models indicate the importance of the use of multiple reservoirs in combination with power functions to link them. The average performance and performance range of the GR4H model is close to that of models SF04-SF07. The more complex models SF11 and SF12 decrease the range in performance, by limiting the number of strong model failures, which is very valuable. However they are not able to outperform the simpler models (including GR4H) on average.

5.3 Performance in CR1-CR4 across models

Figure 5—6 shows the performance of the model structures for CR1-CR4 separately. For most criteria, the same pattern of good and poor performing models can more or less be observed, especially for the seven high performing model structures. The poor performing model structures (SF01, SF02, SF03, SF08, SF09, SF10) have lower scores on at least one of the criteria. For SF01 for example, CR2 is especially low while CR1 scores are relatively high. The single fast reservoir (FR) in SF01 is thus fairly good at simulating high flow, but poor for low flow. With an unsaturated reservoir (UR) (models SF03-SF07) this difference is no longer observable, indicating that combining reservoirs is important for simulating different types of flow.

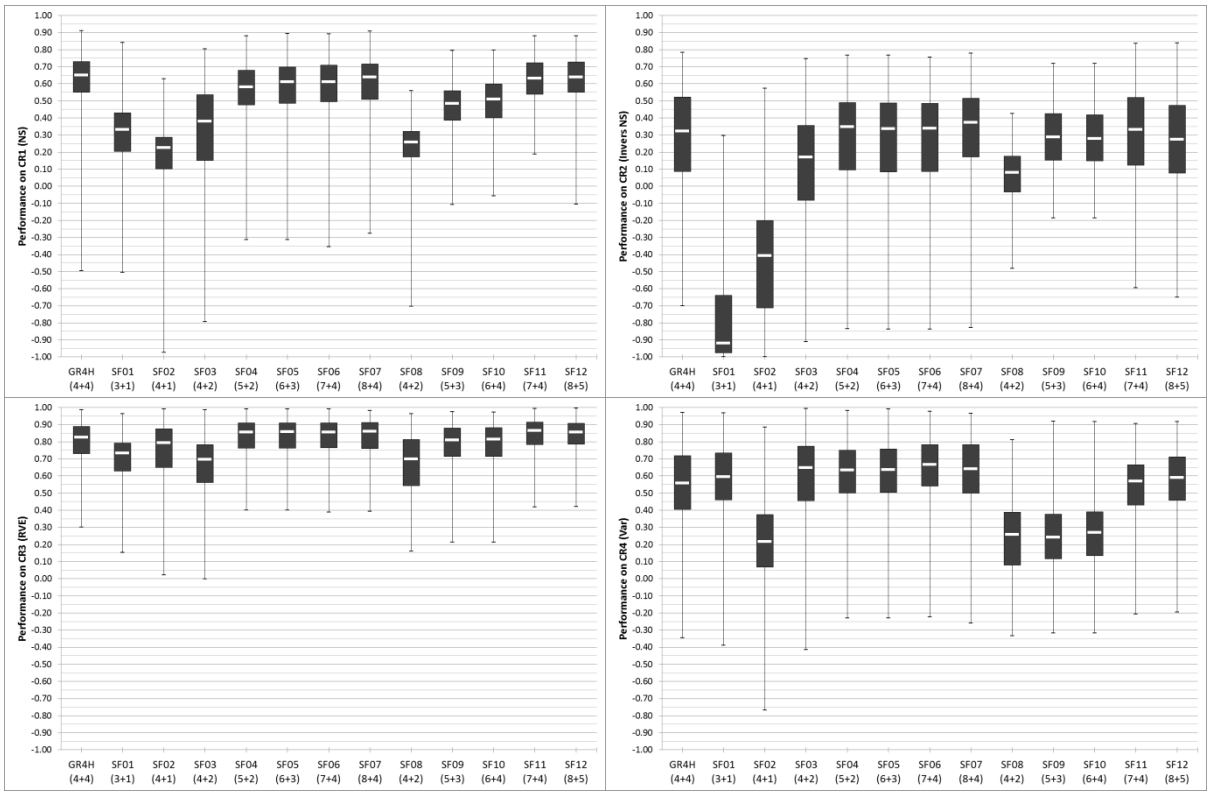

Figure 5—6. Distribution in model performance on the individual criteria CR1-CR4 (upper left to lower right) for all models. CR1 (upper left) is the Nash-Sutcliffe criterion (NS) sensitive to high/peak flow, CR2 (upper right) uses the inverse NS sensitive to low flow. CR3 (lower left) is the water balance criterion and CR4 (lower right) is based on the difference in variability between observed and simulated flow.

Low performance for the low flow criterion (CR2) across all model structures indicates that low flows are more not as well simulated as high flows. The range in this criterion is also wider than that of the other criteria, indicating that low flow is simulated very differently across the catchments. This can be partly due to the nature of the objective function (sum of squares still takes important role emphasizing high instead of low flow) but also a result of the models' structure.

Model structures SF08, SF09 and SF10 show small ranges on CR2. These models all have separate fast and slow reservoirs that are linearly filled. SF11 and SF12 also have separate fast and slow reservoirs but use a power function between the upper unsaturated reservoir and these lower reservoirs. The presence of the power function shows an increase in average performance but at the expense of the low flow simulation in some catchments.

Relatively high performance on the water balance criterion CR3 and limited differences between the models can be expected as simulating the water balance is a relatively easy part of simulation. Also, all models have a parameter that is aimed at correcting the water balance error (groundwater exchange coefficient x_2 in GR4H and multiplication of PET by *Ce* in SUPERFLEX).

For CR4 the pattern looks slightly different, SF02 and SF08-SF10 perform relatively poorer than the other models. This indicates that these models are unable to mimic the variability in the observed flow, the simulated flow will be either too smooth or too jagged. Closer examination of the results shows that the difference in variance varies over the 237 catchments for all four models, so no model is always too smooth or too jagged.

The performance of GR4H is very similar to that of model SF04-FS07 for all criteria, as observed for the average performance. On CR1, SF11 has a smaller range in performance than GR4H, so SF11 describes high flow better. On CR2, SF09 and SF10 have smaller ranges than GR4H, so these model structures perform better in the sense of performing relatively well on a wider range of catchments. These two differences are further investigated in section 5.5 where we look at the parameter distributions.

5.4 Average performance in catchment classes

To investigate the effect of catchment characteristics on model performance, average performance is analysed in four catchment classes defined in section 2.4: catchment area, Wetness Index, permeability and the ratio of runoff coefficients in summer and winter.

5.4.1 Catchment area

Figure 5—7 shows the average performance of each model structures divided into three area classes; small, medium and large. The standard deviation for each class is shown as well to give some indication of the confidence of the average performance. The average performance of all models is better in the larger catchments than the small catchments. The standard deviation is lower in large catchments than in small catchments for all models, so generally there is less chance of a poor performance in a large catchment. The performance of lumped conceptual models is thus better on large catchments than on smaller ones. Apparently, for catchments where hydrological processes are mixed and that have smoother response, lumped model are better able to reproduce the rainfall-runoff relationship. This corroborates previous findings by Merz et al. (2009) on a large set of Austrian catchments. These authors showed that model failures were less likely on large catchments.

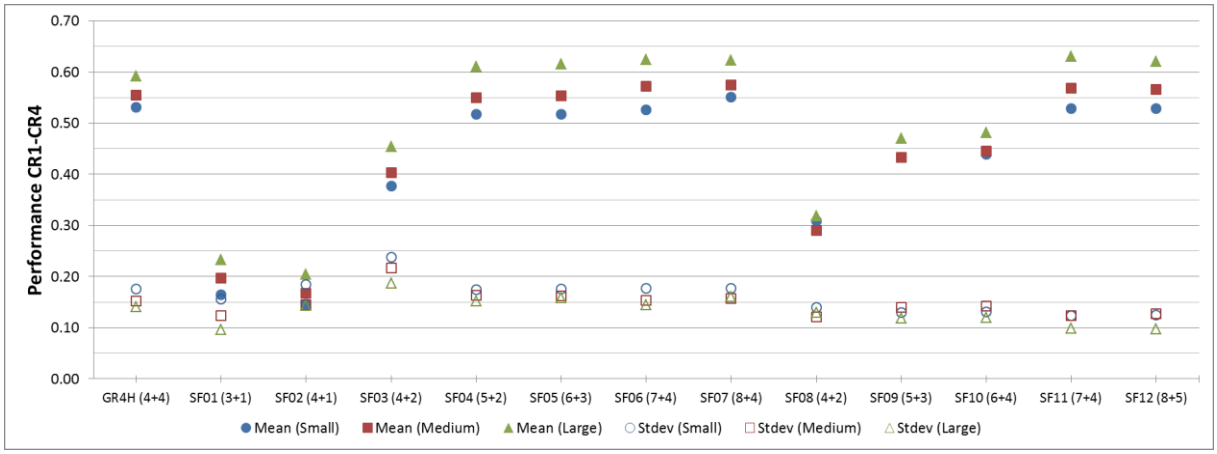

Figure 5—7. Mean and standard deviation of overall performance (CR1-CR4) over the three catchment area classes (small, medium and large catchment areas).

The similar behaviour across all model structures means little can be derived from the structural differences between the models. Only in structures SF08-SF10, the difference between medium and small catchments is very small. These model structures are the only ones without any sort of power relationship, but it remains difficult to derive a relation with catchment response.

5.4.2 Wetness Index

Figure 5—8 shows the performance over three Wetness Index classes, where all models including GR4H perform better in wet catchments than in dry catchments. This is in agreement with general results of the literature showing that dry catchments are generally more difficult to model due to higher non linearity in processes. Only SF11 and SF12 show little difference in performance over the three classes (however their performance is still lower for dry catchments). The parallel fast and slow reservoir plus the unsaturated zone reservoir with a power function relation seems better able to simulate both wet and dry catchments. As observed in the average performance (section 5.1), SF11 performs most equally on different types of catchments.

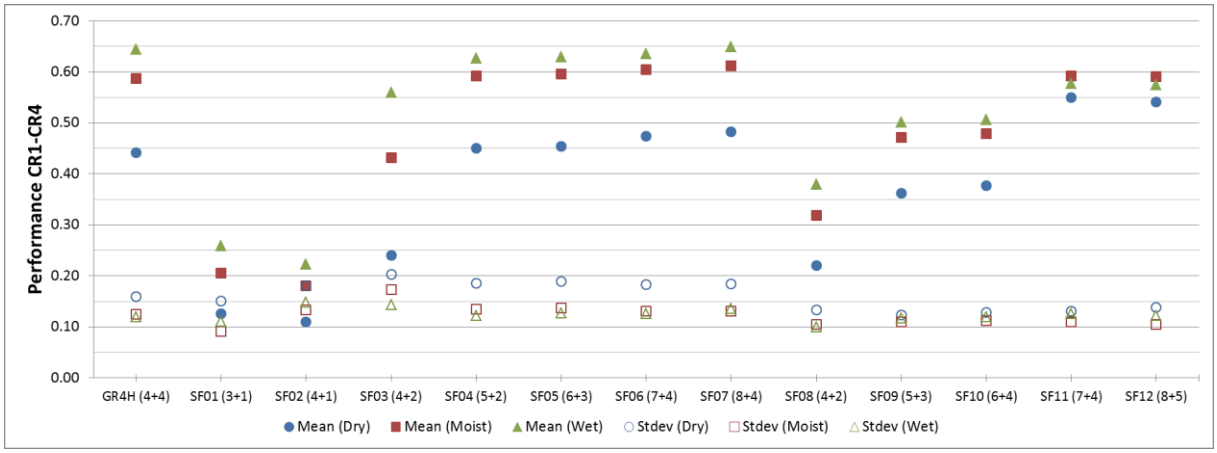

Figure 5—8. Mean and standard deviation of overall performance (CR1-CR4) over the three Wetness Index classes (dry, moist and wet catchments).

Figure 5—8 also shows that the standard deviation in dry catchments decreases in the more complex models which confirms that these models are better able to simulate a wider range of catchments. However, as the average performance does not increase, the added complexity has a negative impact on some catchments as well; especially the wet catchments where the simpler models perform marginally better. The use of SF11 or SF12 can be justified when catchments are dry or when the Wetness Index is unknown, as these models perform relatively well in all cases.

5.4.3 Permeability

Figure 5—9 shows the average performance of classes according to the permeability based on geological data. Models with two reservoirs in series (SF03-SF07) show a much better performance in impermeable catchments than in the other classes. The improved performance is clearly visible in the standard deviation of SF04 to SF07 as well, indicating that these models work well for most impermeable catchments. Adding a slow reservoir in models SF08 to SF12 decreases performance for the impermeable catchments without really increasing the performance in the other classes. Only for SF11 and SF12 the performance of the most permeable catchments slightly increases, they benefit from a power function that was introduced in these models.

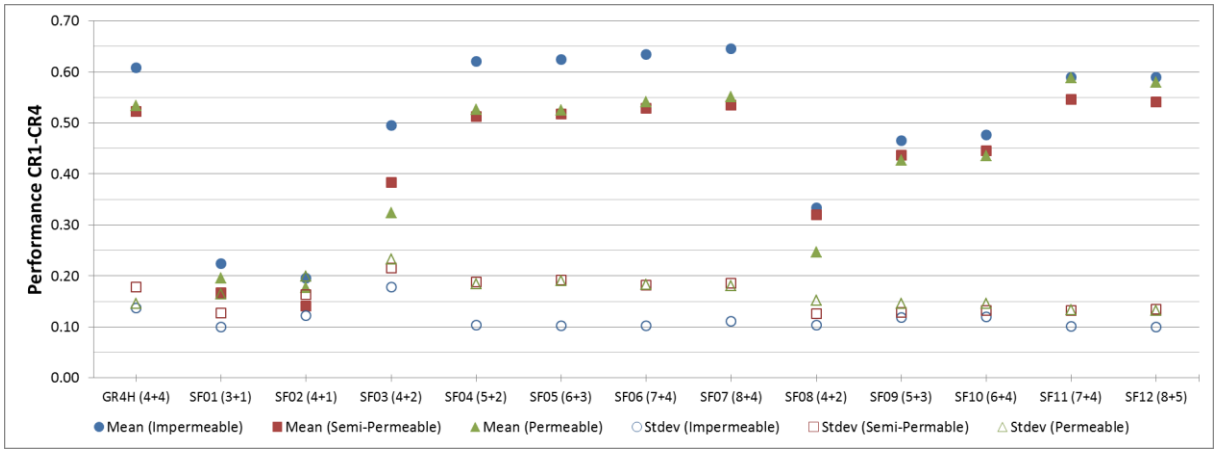

Figure 5—9. Mean and standard deviation of overall performance (CR1-CR4) over the three permeability classes (impermeable, semi-permeable and permeable).

The behaviour of the GR4H model is again comparable to structurally similar models SF04 to SF07. GR4H does contain parallel flow paths like SF11, but permeable and impermeable catchments still perform different. The independent calibration of the flow paths in SF11 could be an important difference for permeable catchments. SF08's poor performance in permeable catchments shows that the unsaturated reservoir is an important model component for these catchments.

5.4.4 Runoff coefficient

Figure 5—10 shows the average performance of all models for three runoff coefficient classes. $RC_{S/W}$ is the ratio between the summer runoff coefficient and the winter runoff coefficient and classified into groundwater dominated runoff, mixed and direct runoff as described in section 2.4.

This classification will separate the catchments based on the following principle: a catchment with much storage is likely to be groundwater dominated. A groundwater dominated catchment has relatively high flows in summer as a large portion of groundwater will feed the stream. In winter, runoff is relatively low as a large portion of rainfall is used to refill the groundwater storage. In a catchment with less storage (and thus more direct runoff) rain will flow to the stream quicker, especially in winter when saturation is high. During summer, flow will be lower because there is less storage that can feed the stream.

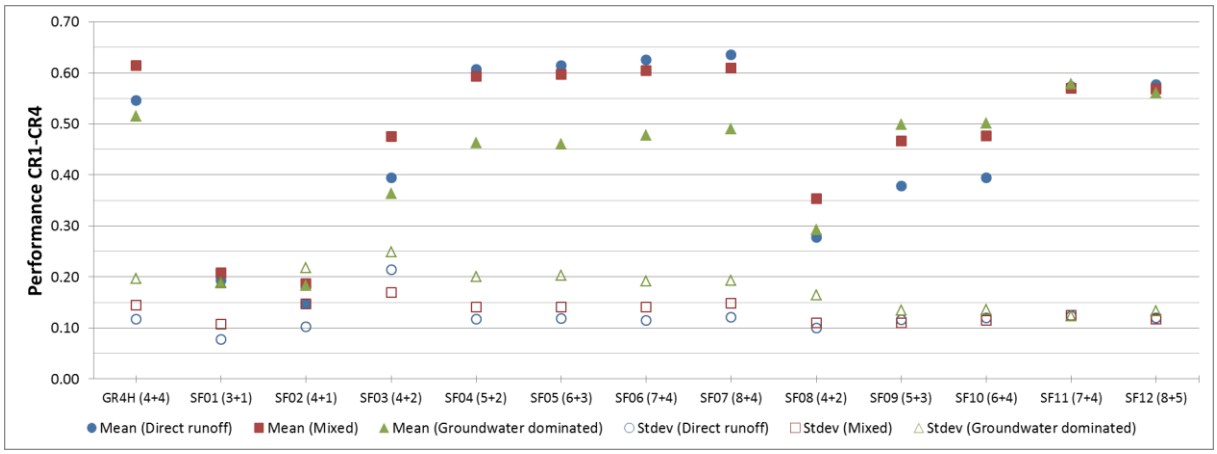

Figure 5—10. Mean and standard deviation of overall performance (CR1-CR4) over the three $RC_{S/W}$ classes (direct runoff, mixed and groundwater dominated).

The $RC_{S/W}$ classification shows some interesting differences between the models. In models SF04 to SF07, the groundwater driven catchments perform much worse than the other catchments while in models SF09 and SF10, these catchments are simulated best of the three classes. This reversed order of performance can be explained by the parallel slow reservoir component in SF09 and SF10. This component allows independent fast and slow flow, so enabling an independent groundwater component while maintaining the ability to produce high flow in case of a storm event.

However, the performances of SF09 and SF10 are still lower than that of the models without a slow reservoir (SF11 and SF12). This may be the effect of the missing power function in SF09 and SF10. Re-introducing the power function (β) in SF11 (the only difference compared to SF10, Figure 5—11) makes all three $RC_{S/W}$ classes perform the same, again with SF11 as the most constantly performing model.

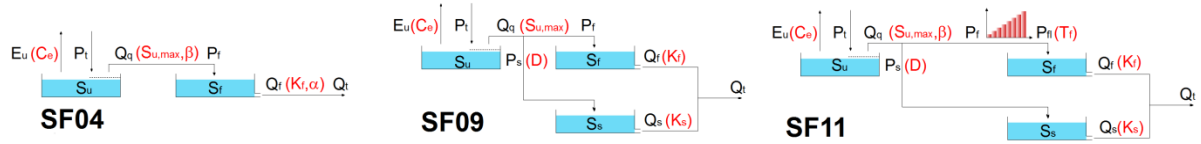

Figure 5—11. SUPERFLEX structures SF04, SF09 & SF11.

GR4H and SF03 (Figure 5—12) perform differently from the other SUPERFLEX models, compared to SF04-SF07, SF11 and SF12 it gives the highest performance on mixed catchments. The relatively poor performance of GR4H on direct runoff catchments may be explained by the relative large ratio (90%) of water that is directed through the Routing store. However, the model performs poorly on groundwater dominated catchments as well, so the model may additionally be limited by the missing reservoir in the fast flow path or the fixed nature of the division between the parallel flows (both reduce calibration freedom). Low performance on the direct runoff in SF03 could be the result of the upper threshold reservoir, this means there can be no flow when the storage in this reservoir is low. The relatively high performance in groundwater dominated catchments may controversially benefit from the lack of response to individual rainfall events in summer periods or simply mean that the level in the reservoir is constantly above the threshold.

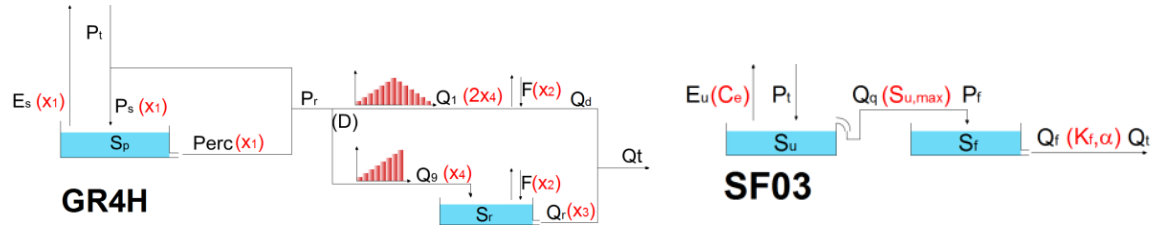

Figure 5—12. GR4H and SF03 model structure.

5.5 Four statements about model structure

The previous sections several components of the model structures have been discussed. From several components the effects on model performance seem very large while other seem to have little influence. To further increase the understanding of four of these components, the following four statements are investigated in the sections below:

- The analysis of the average performance has indicated that the presence of power functions in the SUPERFLEX structures likely increases model performance across all catchments. What cannot be derived from the previous analysis is, if the value of the power parameter varies across catchments or if the mere presence of a (more or less equal) power increases performance alone, if the power parameters are sensitive to the model and at what location in the structure the power function is most effective.
- The analysis of average performance also showed little to no performance increase from the lag-function in the SUPERFLEX structures, suggesting that the role of lag-functions is redundant when combined with the existing functions in the model.
- The flexible ratio D between the fast and slow flow components in SUPERFLEX models helps increase average performance. The flexible ratio between fast and slow flow paths make the model more adaptable to catchment specific processes than the use of a fixed division in GR4H.
- Average performance in the groundwater dominated catchments was notably higher in models with parallel fast and slow flow components, which suggest that these components are especially important for this type of catchments.

5.5.1 The importance of a power function

The comparison of the SUPERFLEX structures showed that the presence of power functions in the SUPERFLEX models increases performance (SF03 to SF04 and SF10 to SF11). What could not be derived from this comparison is whether the mere presence of a power, the flexibility of the power function or the location of the power in the model is important. Figure 5—13 shows the model structure of SF04 as an example. SF04 has two power functions, β for describing flow from the unsaturated reservoir and α for describing flow from the lower (fast) reservoir. These powers are used in different combinations in the other SUPERFLEX structures.

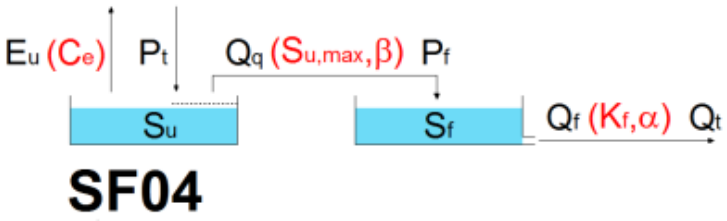

Figure 5—13. SUPERFLEX structure SF04.

Figure 5—14 shows the calibrated values of power β of all catchments in both splits plotted against each other. The frequency of the values for both splits is also given by bars. Note that the dots in these plots correspond to calibrated values of the power parameter in different catchments. When dots are on the one to one line it means the value of the parameter is the same in both splits. Any variation from this line means different values for the same catchment were found for the two periods. This can be the result of climatic differences between the calibration periods, but also because the parameter is poorly identifiable or insensitive.

In case of the figure with power β, the scatters and bars show that for models SF02 and SF04 to SF07 values cluster around $\beta = 1$, indicating that the power function is not really used. Correlation between the values of the different splits is also very low which could indicate that the values are not sensitive to catchment characteristics. It seems also that in many cases, there are outliers that make the correlation coefficient drop. Models SF04 to SF07 also include power α which is discussed below. In SF11 and SF12 power β is the only power function in the model. For these models, the range of the calibrated values is wider. This may mean that the value is more sensitive to differences between the catchments and more useful to calibrate, but it may also mean that the model results are not sensitive to this parameter, with little influence on model performance. Correlation of about 0.5 does show better parameter identifiability, and the wider range indicate some sensitivity to catchment characteristics. Given the strong reaction in average performance seen in section 5.2 the insensitivity of model results to this parameter is less likely and power β therefore appears an important parameter for these models.

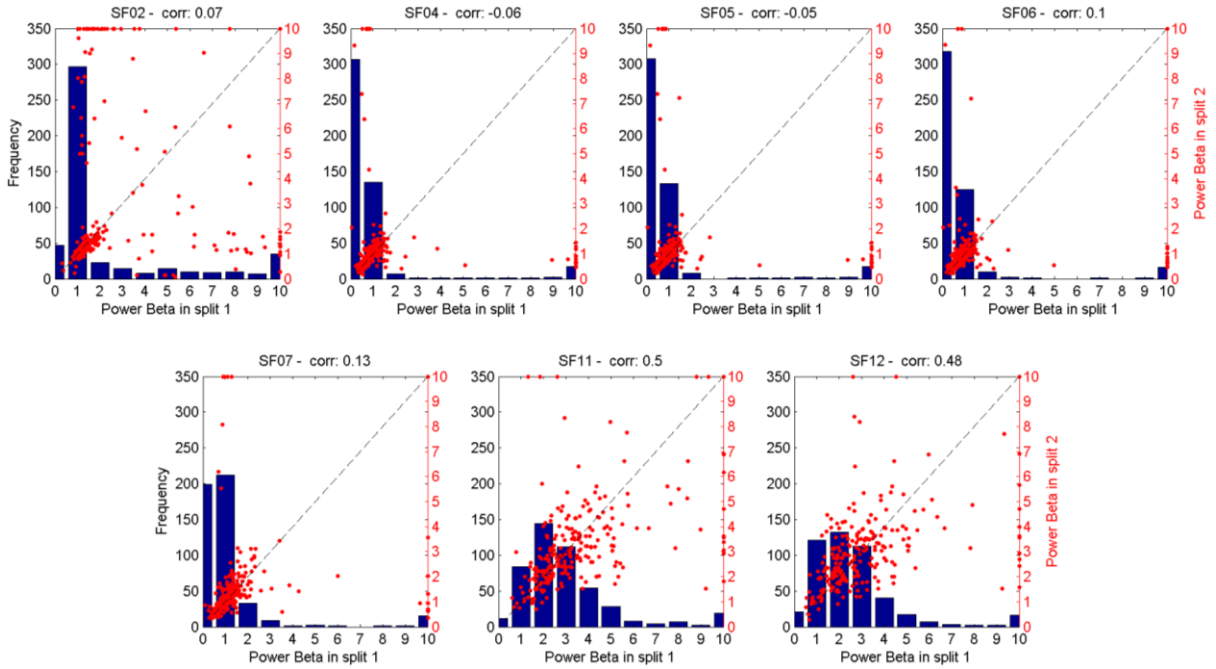

Figure 5—14. Calibrated values of power beta in seven of twelve SUPERFLEX structures. The bars show the frequency of values of both splits within the bounds. The dots are actual calibrated values: for each catchment the value from split 1 was plot against that of split 2. In the title of each plot the model structure is indicated along with the Pearson correlation between values of both splits.

Figure 5—15 shows the values of the power α of both splits against each other, again with frequency bars and correlation given in the figure. The calibrated values of α for SF01 are highly correlated and concentrated around the value of 2. This indicates that the parameter could also be fixed for this structure without considerable performance loss (this was done in GR4H with the power values in the model). For the other models in Figure 5—15 the power α varies more between catchments, indicating that it is a more sensitive parameter, even for those models that already contain power β (SF04-SF07).

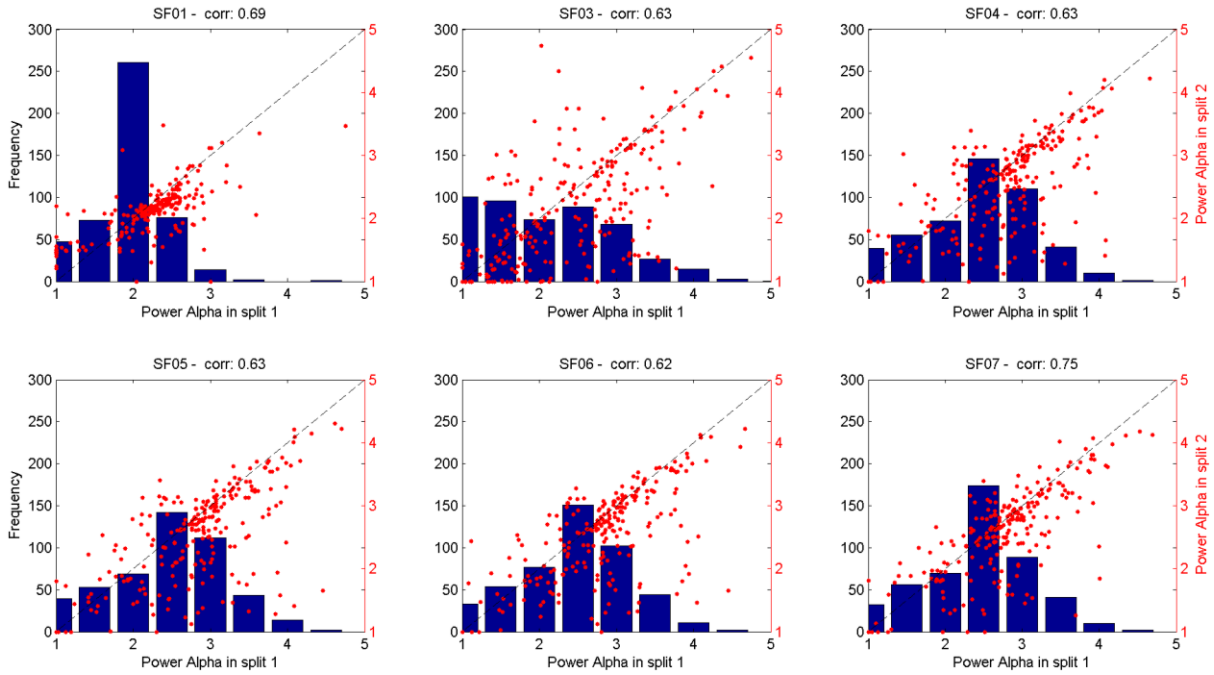

Figure 5—15. Calibrated values of power alpha in six of twelve SUPERFLEX structures. The bars show the frequency of values within the bounds. The dots are actual calibrated values: for each catchment the value from split 1 was plot against that of split 2. In the title of each plot the model structure is indicated along with the Pearson correlation between values of both splits.

From the above figures, it can be derived that in the high performing models like SF04 at least one power function is an efficient calibration parameter for adapting the model to a specific catchment. The variation in calibrated parameters plus the high average performance shows that using a flexible power enables the model to adapt to specific catchments. Actually, the added value of having a flexible power compared to a fixed one could be further tested by comparing a given structure with flexible power with the same structure in which the power is fixed to the median power value found over the catchment set. The clustering of values around a single value like in SF01 shows that in some cases, a power function does not vary much over different catchments. These results also show that when using two flexible power functions, one can become redundant and that a power function connected to a lower reservoir (one that directly leads to flow) is more effective than one in intermediary reservoirs.

5.5.2 The redundant lag-function

Section 5.2 has shown that including a lag-function in the SUPERFLEX model structure has little to no influence on the average performance. Figure 5—16 shows the performance of two model structures without a lag-function against the performance of two models with a lag-function. The compared models, SF04 against SF05 and SF09 against SF10, only differ by the lag-function. Figure 5—16 shows that there is very little difference in the performance in validation of the model structures with or without a lag-function on the individual catchments, which is also observed in calibration (not shown).

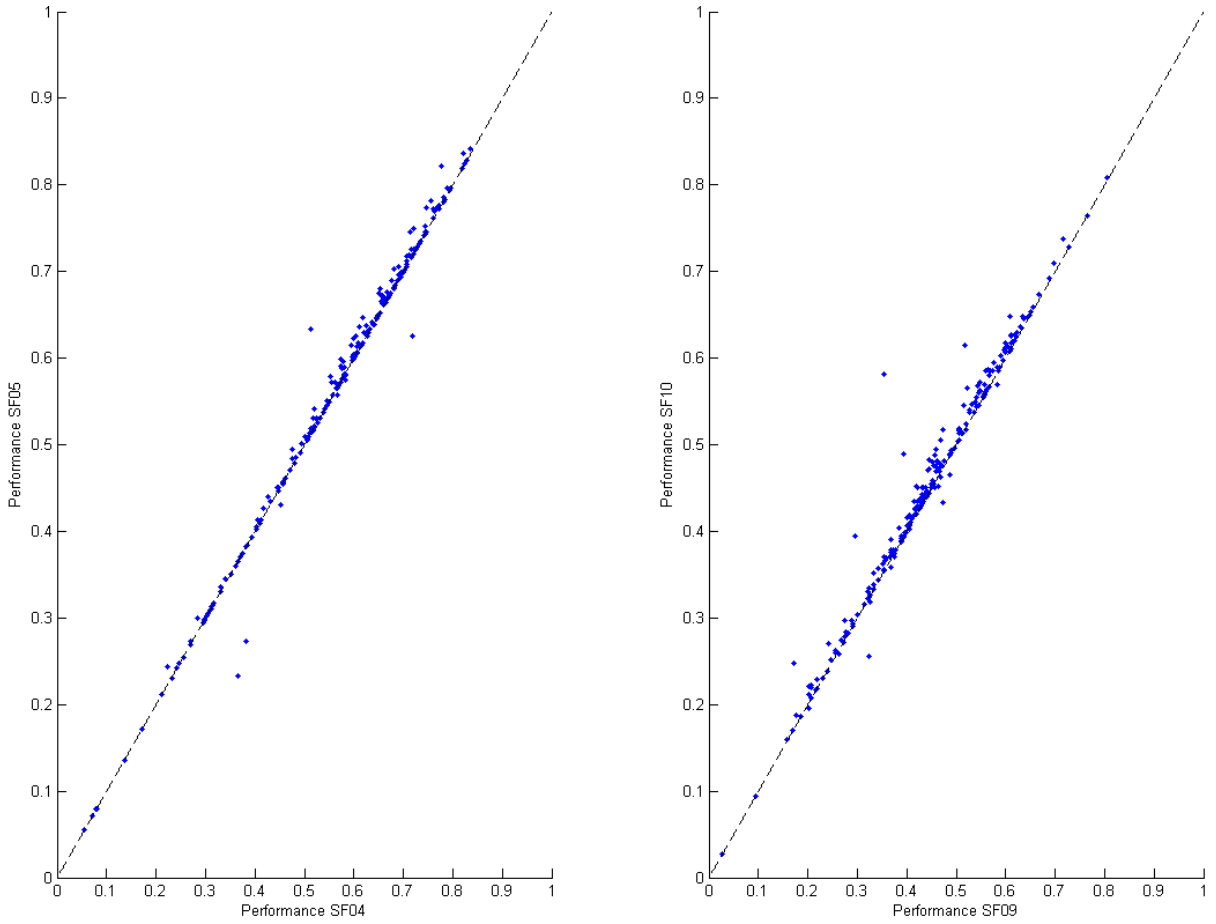

Figure 5—16. Performance of model structures without lag-functions (SF04 and SF09 on the x-axis) compared to model structures with a lag-function (SF05 and SF10) on individual catchments. SF04 and SF05 have the same average performance of 0.56, SF09 and SF10 have average performance of 0.44 and 0.45 respectively. The dashed line gives the one to one performance.

Figure 5—17 shows the calibrated time base of the lag-function in seven model structures. It shows that the time base is somewhat sensitive to differences in catchments (range of calibrated values). Apparently this only leads to very small differences in performance making the lag-function largely redundant with the delaying effect of reservoirs already existing in the structures.

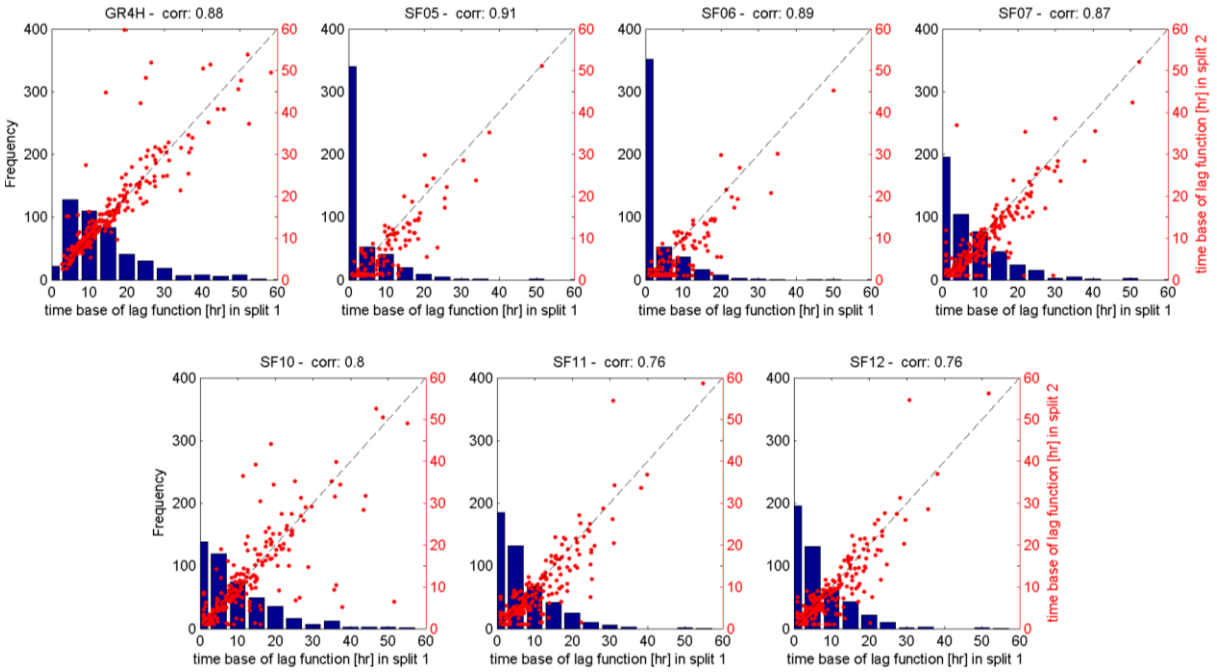

Figure 5—17. Calibrated values of lag time base in GR4H and six of twelve SUPERFLEX structures. The bars show the frequency of values within the bounds. The dots are actual calibrated values: for each catchment the value from split 1 was plot against that of split 2. Most catchments have a time base of less than 60 hours, those with longer time bases are left out to increase visibility in these plots. In the title of each plot the model structure is indicated along with the Pearson correlation between values of both splits.

Figure 5—17 also shows that the time base of the lag-functions in the GR4H model (x_4) is sensitive to differences between the catchments. In this study, GR4H was not compared to a version without the lag-functions so its influence cannot directly be investigated. Also, it must be noted that the implementation of the time lag in GR4H and SUPERFLEX is different (see appendix B and C). But previous tests made at the Irstea research institute on the hourly version of the model indicate that this function actually adds significant performance gains. This is the reason why it was kept in the model structure. Also, it must be noted that the implementation of the time lag in GR4H and SUPERFLEX is different (see appendix B and C).

5.5.3 The flexible ratio *D*

The ratio of water split between the fast and the slow flow paths of the SUPERFLEX structures is determined by the parameter *D*. In SUPERFLEX, this parameter is calibrated while in GR4H the ratio is kept fixed at 0.9 to the Routing store and 0.1 to the direct flow path with lag-function. Figure 5—18 shows the calibrated values of ratio *D* in the SUPERFLEX structures with both a fast and a slow reservoir. The variation in the calibration values shows the sensitivity of the ratio to the differences in the 237 catchments. The high correlation in the more complex structures shows that this parameter is quite well identifiable during calibration and quite stable between. Individual values can lie quite far apart but over the whole range, the correlation is clearly visible.

It must be noted that the residence time of the fast and slow reservoirs is also calibrated and can vary between the splits, making it more difficult to find the exact same value for parameter *D*. These extra parameters can be an advantage to the flexible approach but can also lead to equifinality; different parameter sets giving the same results. However, given the high correlation of the calibrated values between splits and the high average performance on most catchments of SF11 (small and high performance range, see section 5.1) it is likely that a flexible ratio D is an important model component.

The fast and slow reservoir residence time in SUPERFLEX are also bounded to ensure that the fast flow path is in fact faster than the slow path. In GR4H some precaution is taken to prevent this switching of flow paths by linking the fast and slow path through the lag-function, but still cases are known (including in this research) where the role of the paths switches (see also section 7.1.1).

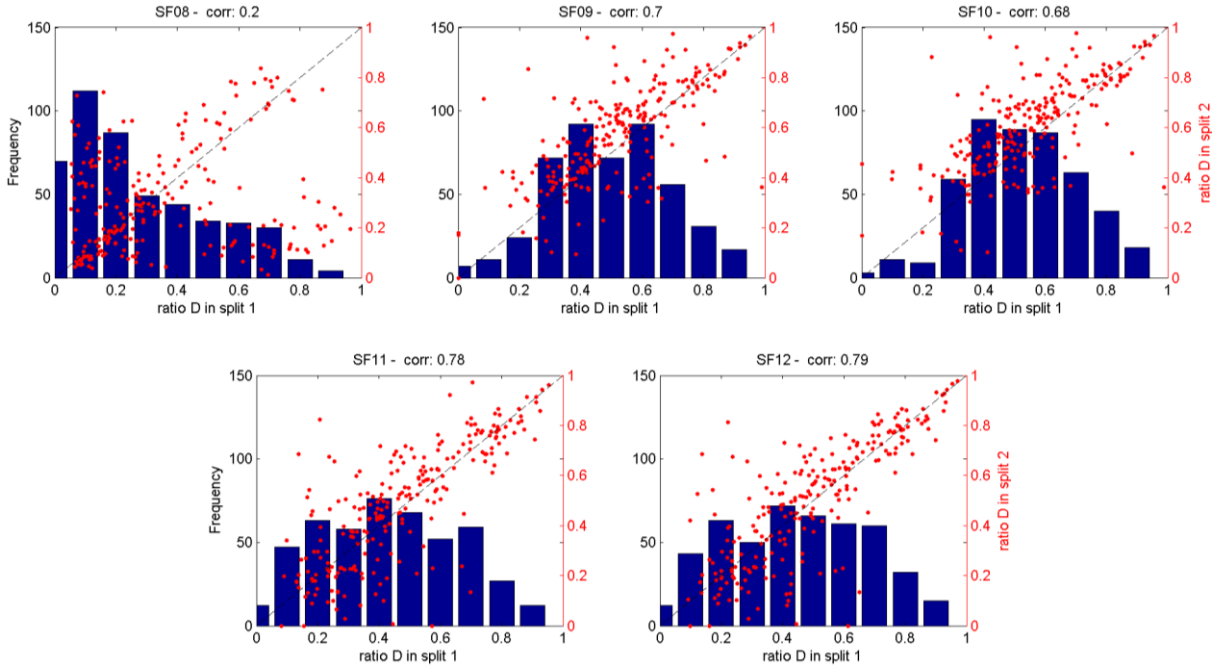

Figure 5—18. Calibrated values of ratio *D* in five of twelve SUPERFLEX models. The bars show the frequency of values within the bounds. The dots are actual calibrated values: for each catchment the value from split 1 was plot against that of split 2. In the title of each plot the model is indicated along with the Pearson correlation between values of both splits.

5.5.4 Slow reservoirs in groundwater driven catchments

The comparison of average performance in catchment classes showed that SUPERFLEX models with a slow reservoir performed better on groundwater dominated catchments. Figure 5—19 shows the average performance of SF05 against that of SF11, two identical models apart from the use of a slow reservoir in SF11. The figure confirms that the more complex SF11 does not perform better than SF05 overall (mean performance of 0.57 and 0.56 respectively), but that it significantly improves performance in the groundwater dominated catchments. When SF05 performs very poor (<0.5), SF11 performs notably better indicating that the slow reservoir does indeed increase performance for these type of catchments. Incidentally, SF11 performs poorer than SF05 on a few groundwater-dominated catchments. The reasons for this behaviour should be analysed in more detail.

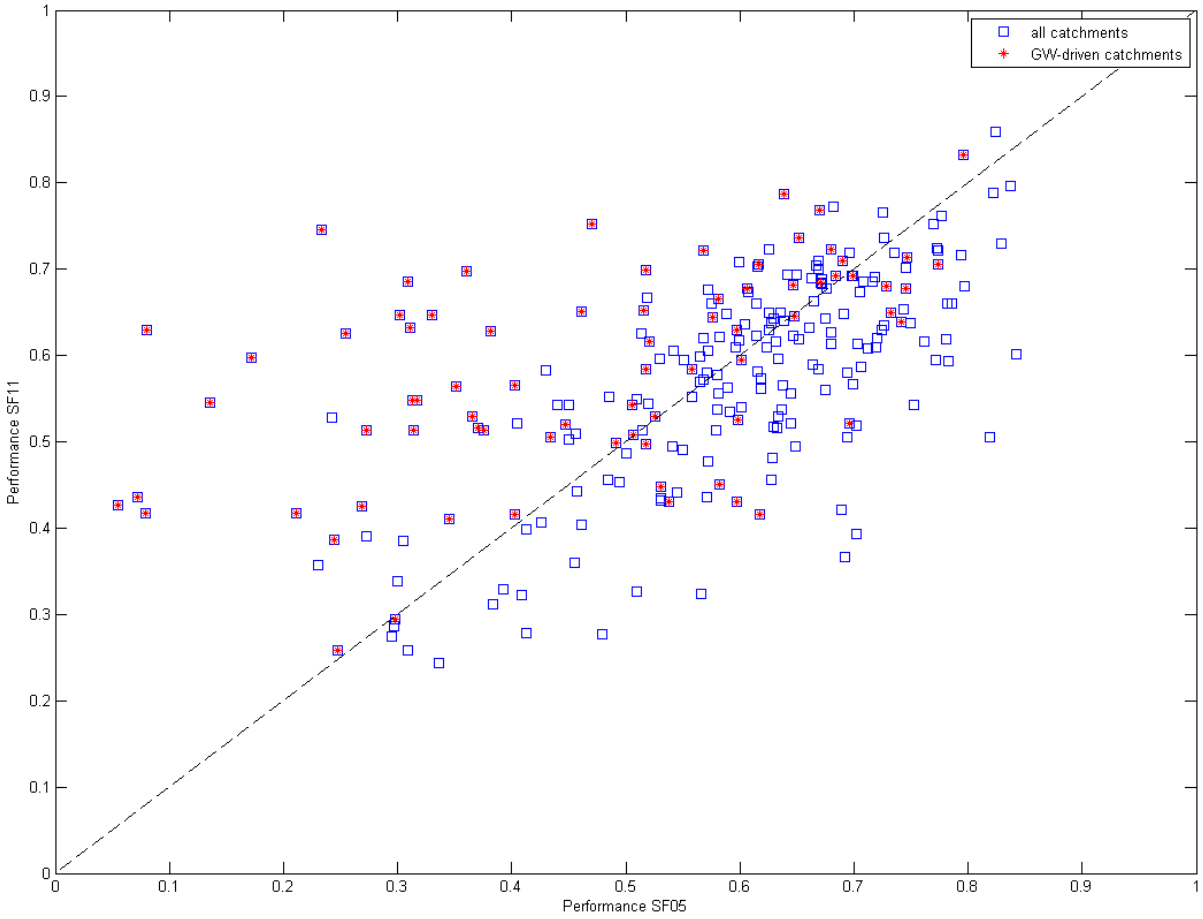

Figure 5—19. Average performance of model structure SF05 against SF11. Groundwater dominated catchments are marked to show the specific performance of these catchments compared to the others.

5.6 Robustness across models

Figure 5—20 shows the robustness of all the models as measured by the ratio in total error between calibration and validation (section 4.2.1). The figure shows that no model is really much more robust than the others because the averages are close together. The range varies across the models but without a clear link to model complexity.

Mean robustness is quite high, even models with low performance give high robustness. This may be explained by the lower complexity in the low performing models, since less freedom for the model means less chance of large differences between periods.

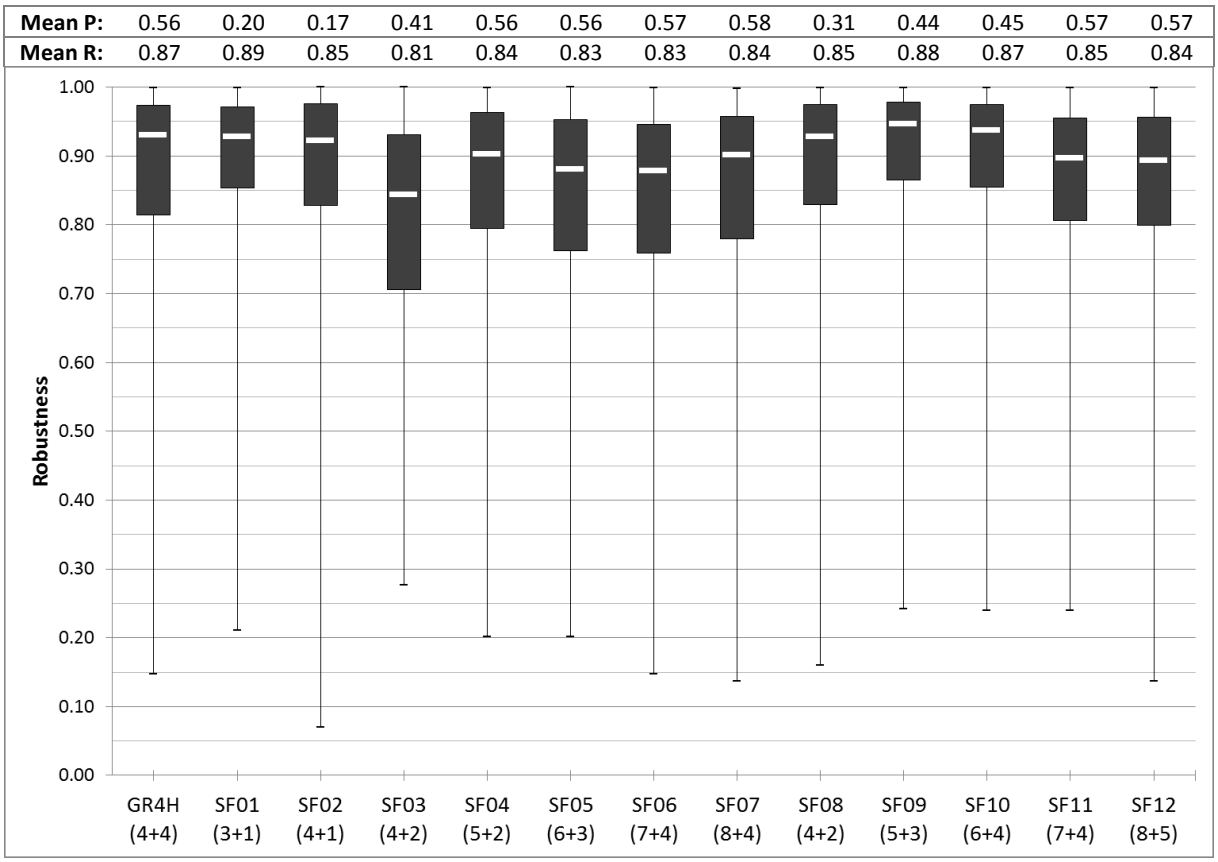

Figure 5—20. Average robustness of all model structures validated on 237 catchments including bounds. The x-axis shows the twelve SUPERFLEX structures plus GR4H, the value between brackets denotes the complexity measure (nr. of states + nr. of calibrated parameters). On the top, the first line gives mean performance and second line gives mean robustness.

Neither the average robustness nor the robustness in catchment classes (as discussed for performance) give much insight or show differences in any of the models and is therefore not further discussed.

5.7 Concluding remarks

The main conclusion that can be drawn from the average performance of the thirteen different model structures, is that there are seven structures (GR4H, SF04-SF07, SF11 and SF12) that perform significantly better compared to the six remaining ones (SF01-SF03, SF08-SF10). The seven high performing structures perform very similarly despite their structural differences. Increasing complexity does not always lead to higher average performance in these structures: the added complexity over models SF04 to SF07 or SF11 to SF12 does not lead to better model performance. In other cases adding a component does lead to a clear performance increase: adding the upper unsaturated reservoir (SF01 to SF04 and SF08 to SF09) or implementing a power function (SF03 to SF04 and SF10 to SF11) increase average model performance.

Analysis of CR1-CR4 shows that the six poor performing structures score low on at least one aspect of flow, which means that these model structures perform poor on at least one aspect of flow. The seven high performing structures perform similarly on the different criteria. The analysis of performance in different catchment types shows some differences between the high performing structures:

- Model performance in small catchments is generally lower than in large catchments,
- Model performance in wet catchments is generally higher than in dry catchments
- Models with independent parallel flow paths perform better in permeable or groundwater dominated catchments, and
- Models with only reservoirs in series perform better on impermeable catchments where runoff is more direct.

Detailed analysis of four of the model components has shown that:

- Calibrating the value of at least one power function in a model structure increases performance. The best location for the power function in the investigated model structures is at the outflow of a lower reservoir. Adding a second power function (between reservoirs) is less effective,
- The lag-function used in the SUPERFLEX structures proved redundant,
- The calibrated ratio *D* that determines division of water between the fast and slow reservoir in the SUPERFLEX structures is an important flexible parameter, unlike the fixed division used in the GR4h model, and that
- Model structures with independently calibrated parallel flow paths do indeed perform better on groundwater dominated catchments.

Average robustness of all model structures is high and seems unrelated to average performance. Robustness could not be linked to differences in model structure or catchment types.

Overall, there seems to be a limit to the complexity that can be added to lumped conceptual models that is still useful. Some relatively simple models already perform very well, while adding complexity does not yield higher performance in validation. This chapter also shows that the GR4H model structure is one of the best performing structures and that on average, several SUPERFLEX structures can match this model.

New systematic tests would be needed to further investigate the sensitivity of model results to the various modelling options identified as relevant here.

Interestingly, in the case of the GR4 model, recent investigations independent of the present study showed that:

- the addition of a free parameter in the formulation of the water exchange function significantly improves the simulation of low flows (see Le Moine, 2008), and
- the addition of a second routing store (with an additional parameters) in parallel to the existing ones also significantly improves results (see Pushpalatha et al., 2011).

These results corroborates to some extent the above conclusions.

6 Results II: Comparing 2 approaches

In the second part of the results, the intended modelling processes of the fixed and the flexible modelling approaches are strictly followed. For the fixed approach this means applying the fixed GR4H model structure on all catchments and calibrating the four parameters x_1-x_4. In the flexible approach the model structure can also be adapted. This essentially means that not only the parameters but also the model structure must be calibrated. To simplify the calibration, this study uses only twelve alternative SUPERFLEX structures from which the best (in calibration) is selected. Additionally, both approaches must result in a single best structure and a similar parameter set for both calibration periods, otherwise the approach is considered inconsistent. Chapter 4 described the comparison and inconsistency in more detail.

Section 6.1 investigates whether the both approaches are equivalent and section 6.2 tries to find the best model structures. Section 6.3 selects the hydrological monsters based on both modelling approaches.

6.1 Consistent model results

In calibration, the fixed approach (GR4H) found a consistent parameter set in 170 of 237 catchments. The SUPERFLEX approach was able to select at least one structure with consistent parameter set for 215 catchments. For 67 and 22 catchments respectively the approaches are unable to select a consistent parameter set or single structure (Table 6—2). The greater number of consistent models in the flexible approach shows that with a flexible structure, a wider range of catchments can be modelled.

Table 6—1. Number of consistent and inconsistent catchments per approach.

Approach	Fixed	Flexible
Consistent	170	215
Inconsistent	67	22
Total	237	237

Figure 6—1 shows the performance and robustness of the consistent catchments compared to the results without consistency constraints. Note that in this figure, each box represents a different number of catchments. The distribution in performance that include inconsistent results show that there is almost no difference in performance between the approaches. In the consistent results the fixed approach performs better than the flexible one, but mainly because lower performing catchments are removed from the distribution. For the flexible approach, the consistent results are slightly lower than the results including inconsistent structures. Some high performing structures are apparently inconsistent and replaced by consistent structures with lower performance. The distributions in different criteria show similar results (appendix D). The distribution in robustness shows that both approaches become more robust when the inconsistent structures are removed from the distribution.

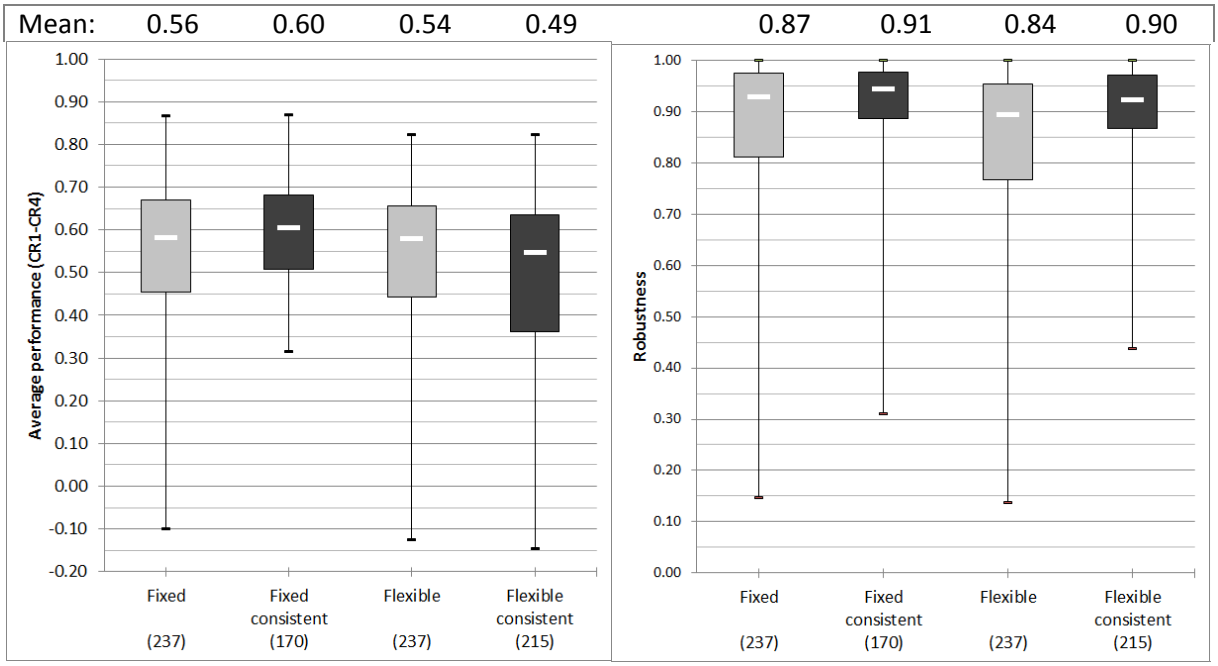

Figure 6—1. Distribution in average performance in validation (left) and robustness for the fixed GR4H approach and the flexible SUPERFLEX approach. These boxplots represent the minimum and maximum (whiskers), the area between the 25th and 75th percentile (box) and the median (bar)The grey boxes represent the distributions including inconsistent results, the black boxes represent only consistent results.. Number of catchments in each group noted between brackets.

These results show that the approaches differ little on the used catchment set, apart from the amount of catchments that are simulated consistently. This is a clear advantage of the flexible approach, in which a matching structure can be selected for each catchment. However this does not yield better performance on average on the whole catchment set.

6.2 The best structures

To find out more about which of the flexible model structures perform well, we look at the number of times a structure is selected as best. A structure is selected as best when it has the highest score for the objective function in calibration, or when it is within 5% of the highest score but simpler than the highest scoring structure. Each structure was given a complexity measure equal to the number of calibrated parameters plus the number of states (section 3.4). This measure is used to determine which structure is simpler. Table 6—2 shows how many times each structure is selected for the flexible approach alone and when GR4H is included as one of thirteen structures.

Table 6—2. Number of times each structure is selected for the flexible approach alone and when GR4H is considered as one of thirteen structures. The complexity of each structure is shown by N_θ+N_S, the number of calibrated parameters plus the number of states used in each model.

Model (N_θ+N_S)	Flexible	Combined
GR4H (4+4)	-	111
SF01 (3+1)	39	10
SF02 (4+1)	7	2
SF03 (4+2)	27	14
SF04 (5+2)	52	45
SF05 (6+3)	2	2
SF06 (7+4)	0	0
SF07 (8+4)	6	4
SF08 (4+2)	5	4
SF09 (5+3)	51	16
SF10 (6+4)	3	1
SF11 (7+4)	22	14
SF12 (8+5)	1	0
Total SUPERFLEX	215	112
Inconsistent	22	14
Total	237	237

Table 6—2 shows that, for the flexible approach, structures SF01, SF03, SF04, SF09 and SF11 are clearly selected more often than the other structures. This is somewhat surprising since SF01, SF03 and SF09 are structures that on average perform poorly compared to SF04 and SF11 (section 5.1). Apparently, these structures are able to perform close to the more complex structures (within 5%) or the complex structures were inconsistent, which makes selecting the best structure difficult. Other structures with high average performance are SF05-SF07 and SF12, but are selected in very few cases. This can be explained by the stepwise increase in complexity in structures SF04 to SF07 and from SF11 to SF12. In section 5.1 we have already seen that the increased complexity led to very little increase in average performance and thus are the simpler structures SF04 and SF11 favoured. Incidentally, one can notice that all SUPERFLEX structures (except SF06) are selected as best on at least one catchment.

Table 6—2 also shows that if the GR4H model would be one of thirteen structures, it is better than all other structures together for almost half of the catchments. This confirms that the empirical GR4H model performs very well on many different catchments. All the SUPERFLEX structures are selected less often. SF01 and SF03 are simpler structures than GR4H and apparently GR4H can outperform them by more than 5% on the respective catchments. SF09 and SF11 are much more complex

models, but again GR4H is able to match or outperform them on most catchments. Only SF04 is still selected quite often, it is slightly simpler than GR4H and performs relatively well compared to the other SUPERFLEX structures.

Table 6—2 also shows that even when combining both approaches, there are still fourteen catchments for which no structure was consistent. These catchments are likely monsters, but also other catchments with low performance can be considered as monsters.

6.3 The hydrological monsters

The hydrological monsters are selected based on the results of both approaches. Figure 6—2 shows the average performance for each catchment of the fixed approach against that of the flexible approach. No catchment performs below zero and catchments that give inconsistent results are given the value -0.3 for performance to make them appear on the edges of Figure 6—2. The catchments on the left side of the figure are those for which GR4H is inconsistent, strikingly the SUPERFLEX structures are able to give high performance on many of these. On the bottom of the figure the catchments with inconsistent SUPERFLEX catchments are shown, GR4H also performs high for some of these. The point (-0.3;-0.3) represents fourteen inconsistent catchments.

Catchments for which both modelling approaches are consistent form the cloud of points in the upper right part of the figure. The dark dashed line represents the line where both approaches would perform equally. Catchments far from this line indicate different levels of performance for the two modelling approaches. This group of catchments and the group for which one of the two approaches is inconsistent are interesting for analysing differences between the two approaches. However, in this study only focused on the hydrological monsters. Therefore, the catchments for which both approaches perform poorly or inconsistent are selected as monster catchments and the non-monster catchments are not further investigated.

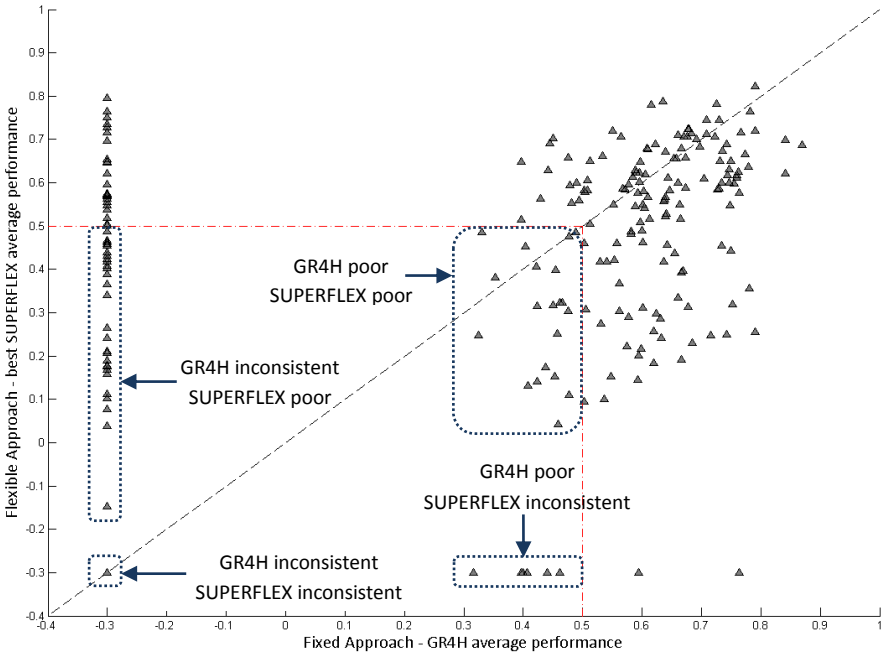

Figure 6—2. Average performance of the fixed approach vs. the average performance of the flexible approach for individual catchments. Inconsistent model results are given the value of -0.3 and are shown on the edges of the figure. Catchments below the red dashed/dotted lines are selected as hydrological monsters.

Poor performance is defined as performance below 0.5. In Figure 6—2 two red dashed/dotted lines indicate the area where both approaches perform below 0.5. Based on this criterion, 69 from 237 catchments are selected as hydrological monsters. Among the 69 selected hydrological monsters:

- for 14 catchments, both modelling approaches are inconsistent,
- for 29 catchments, the GR4H model is inconsistent while SUPERFLEX performs poor,
- for 6 catchments, SUPERFLEX is inconsistent while GR4H performs poor, and
- for 20 catchments, both approaches give a poor result.

These hydrological monsters are investigated further in the next chapter.

6.4 Concluding remarks

This chapter has shown that the fixed and flexible modelling approach perform very similarly on average. Only when the consistency rules are used, the advantage of the flexible approach becomes clear. The fixed approach is inconsistent on more catchments than the flexible approach which means that the flexibility in model structure helps to better describe some catchments. The fixed approach does perform high on those catchments for which it is consistent, showing the strength of the GR4H model. The flexible approach performs lower but is consistent for more catchments.

Closer examination of the selected SUPERFLEX structures shows that five of the twelve structures are selected for most of the catchments. Among these, there are some very simple structures that apparently perform well compared to more complex structures or are consistent instead of inconsistent. SF04, SF09 and SF11 are selected often, these structures are simple compared to SF05-SF07, SF10 and SF12 respectively. The stepwise increase in complexity in these structures only sometimes leads to better results, like shown for average performance in chapter 5. When GR4H is compared with the SUPERFLEX structures, it is selected most often. The GR4H model is a high performing model despite its simplicity. SF04 is a simple and high performing SUPERFLEX structure that is selected often as well.

When the performance of both approaches on individual catchments are compared, it becomes clear there are some catchments for which the approaches perform very differently. These cases could be very interesting for further investigation into the differences between the two approaches. In this study however, the focus is on the hydrological monsters of both approaches. Therefore, those catchments for which both approaches are inconsistent or perform poor are selected as monster catchments.

7 Results III: Demystifying hydrological monsters

In this chapter the hydrological monster from section 6.3 are demystified using characteristics of the observed discharge in section 7.1. Section 7.2 discusses the possibility of predicting which catchments will become monsters and section 7.3 summarizes the main finding of this chapter.

7.1 Three groups of monster catchments

In this section the selected hydrological monsters are further investigated. There can be different reasons why a catchment becomes a monster or why a model gives a poor result. The monster catchments form a heterogeneous group of characteristics, location and size that makes separating them based on the catchment classes difficult. By looking at the observed hydrographs of the monster catchments, the 69 catchments can be characterised in three groups: (1) Severe climatic differences between calibration and validation periods, (2) flashy flow and (3) disturbed flow measurements. More details about the hydrological monsters are given in appendix E.

7.1.1 Severe climatic differences between calibration and validation periods

In 37 of the 69 monster catchments, a pattern of wet years can be recognised in the observed hydrographs. This pattern is the result of three consecutive years with high rainfall followed by five years of relative dry or average years. The wet years are observed in catchments in the northern half of France in the years 1999-2001 (Figure 7—1), with rainfall between 20 and 40% higher than average (see also appendix E).

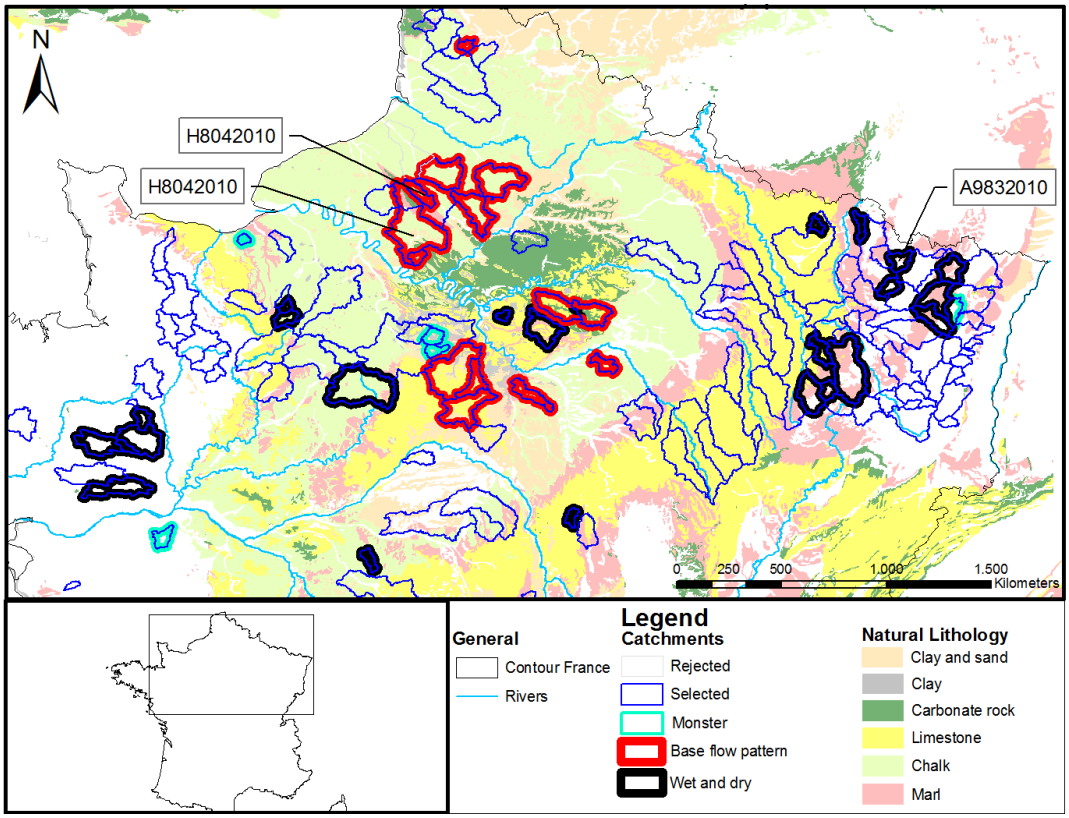

Figure 7—1. Location and lithology of monster catchments in the north of France.

Catchment H8042010 (Epte at Fourges, 1386 km², a tributary of the Seine basin) is a good example of a catchment where the wet years lead to an inter-annual pattern in base flow (BFP), which is clearly observable in 13 of the 37 northern monster catchments. The classification shows that all but one of these catchments are groundwater dominated and are permeable to semi-permeable. Figure 7—2 shows the observed hydrograph of catchment H8042010. Through the relatively wet years 1999-2001, the base flow rises while it decreases in the following years.

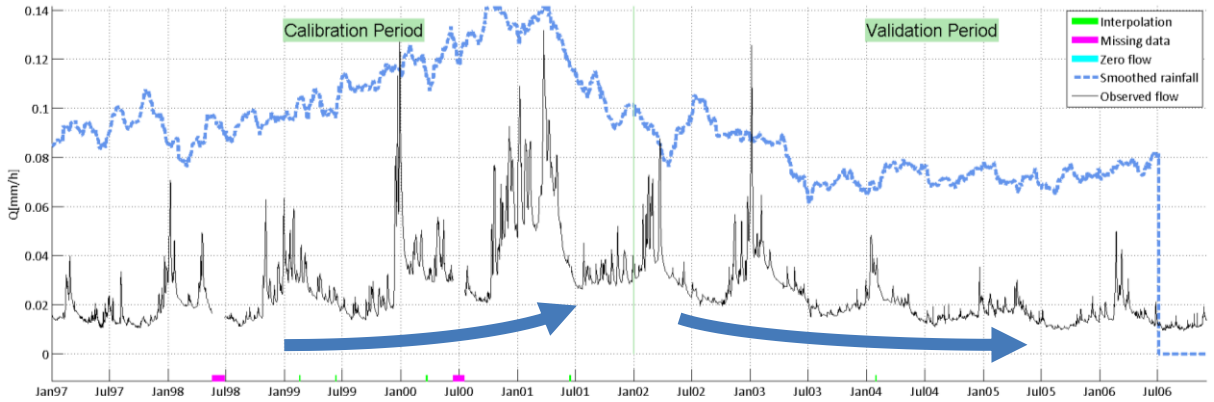

Figure 7—2. Observed hydrograph and smoothed rainfall (moving average of 360 days) of catchment H8042010 – Epte at Fourges – 1386 km².

In the remaining 24 monster catchments in the north of France, there is no BFP like the one above, but the difference between the wet and the dry years can still be observed. Catchments in this group are rarely classed as groundwater dominated and have varying permeability, size and wetness. Figure 7—3 shows catchment A9832010 (Nied Allemande at Faulquemont, 203 km²) as an example. Even though the difference between the wet and dry years may be less clear in this catchment, the mean flow on the first half of the hydrograph is clearly higher (0.047 mm/h) than on the second half (0.022 mm/h) and shows fewer peaks.

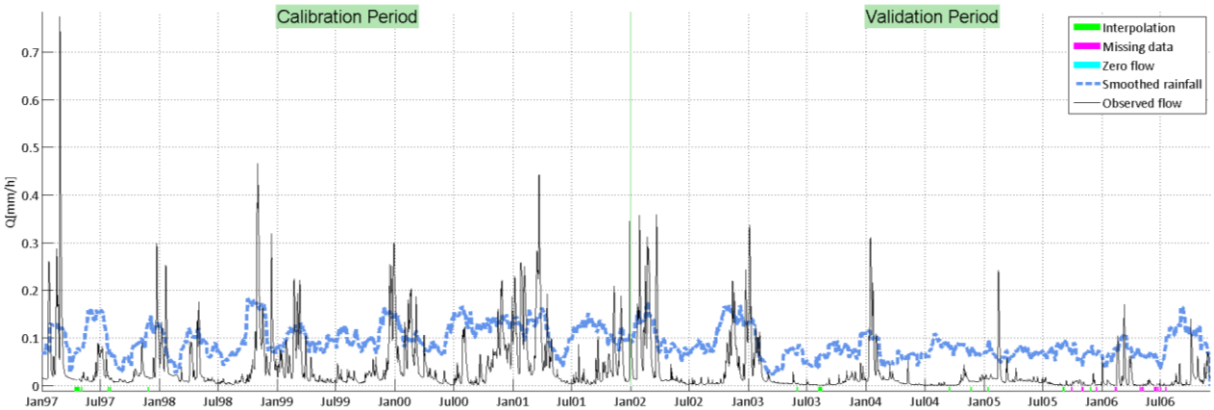

Figure 7—3. Observed hydrograph and smoothed rainfall (moving average of 180 days) of catchment A9832010 – Nied Allemande at Faulquemont – 203 km².

Modelling these monster catchments

Both modelling approaches have difficulties in finding a parameter set that can simulate the severe climatic differences. The wet period is located near the end of the first period used for calibration or validation (the first five years of the time series, S1). The drier years are located in the second calibration or validation period (the second five years of the time series, S2). The selection of the calibration and validation period seems very demanding in this respect: models are either calibrated on the rising or the decreasing pattern, not on both. Another selection of periods may have left the models' sensitivity to the calibration conditions unnoticed.

Climatic differences between periods over which a model is calibrated and validated are known to reduce model performance (Coron et al., 2012). This sensitivity of model performance to calibration conditions reduces the transferability of parameter sets across different periods. In a dry period not all or different hydrological processes may be active than in a wet period reducing the identifiability of some parameters. Since most of these catchments with BFP are permeable and groundwater plays an important role, groundwater storage and flow generation from this storage is a likely process that is simulated poorly.

Other possible reasons for model failure are errors or trends in the water balance, which are difficult to solve for models. Errors in the water balance may be the consequence of errors in data, especially when these errors are not constant in time. Inter-catchment groundwater flow is a likely cause in catchments with a strong trend in base flow. These type of catchments are likely in the chalky areas in the north-west of France (Figure 7—1) where subsurface transport of water across the topographic borders of catchments is suspected (Le Moine, 2008, chapter 3). However, these catchments were no outliers in the non-dimensional plot that was designed to detect catchments that are leaking or gaining water (see appendix A).

In catchments with BFP, the calibration has led to very different parameter sets in the two periods: for 9 of the 13 monster catchments, none of the models is consistent. Figure 7—4 shows catchment H8042010 (Epte at Fourges) with the simulated hydrograph of GR4H. In this figure, the underestimation of flow in the validation period is very clear. Figure 7—5 shows the simulated hydrograph of GR4H calibrated on the second period. In this case, the model response to short rainfall events is very poor and base flow is over-estimated in the validation period.

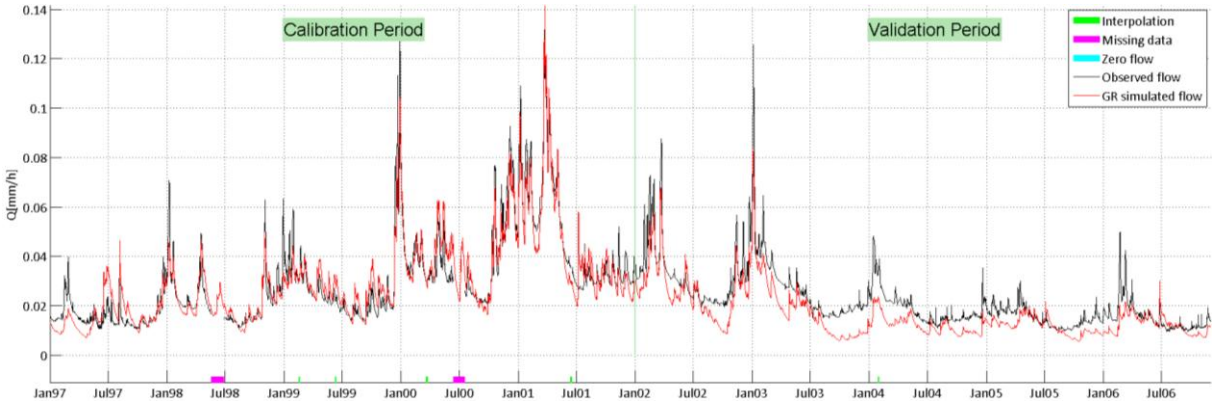

Figure 7—4. Observed and by GR4H simulated hydrograph calibrated on period S1 of catchment H8042010 – Epte at Fourges – 1386 km².

Figure 7—5. Observed and by GR4H simulated hydrograph calibrated on period S2 of catchment H8042010 – Epte at Fourges – 1386 km².

In the remaining 24 monster catchments where the pattern is less prominent (like A9832010), performance is generally low for both approaches. In these catchments, flow is generally overestimated in the dry period and under-estimated in the wet period.

Differences between the fixed and flexible approach

In catchments with BFP, the GR4H model is consistent in only two of 13 cases, in which the model performance and robustness are low. The flexible approach finds a very simple model on two other catchments, both with very low performance but reasonable robustness. Robustness increases with simplicity, but simulated hydrographs look very poor (i.e. poor response to individual rainfall events). In monsters with climatic differences (without BFP), the fixed approach fails in 12 out of 24 cases while the flexible fails in only 2. The fixed GR4H outperforms the SUPERFLEX structures when it is consistent, but works on fewer catchments. Simpler SUPERFLEX structures are selected giving relatively low performance. See appendix E for full results.

In the case of example catchment H8042010, a BFP monster, neither approach resulted in a consistent model. This can be observed in the large difference between the two simulated hydrographs in Figure 7—4 and Figure 7—5. The analysis of the internal model processes shows that the reservoir stores in the GR4H model differs greatly between the two calibration periods. The model calibrated on the wet period has a very large production store (1770 mm) compared to that in the dry period (53 mm). The model calibrated on the dry years has a very large routing store (7200 mm) that mimics the slowly decreasing base flow (271 mm in wet years, see appendix E for the time series of the production and routing store of H8042010). These differences in parameter values can be the consequence of different dominant hydrological processes in the two periods or a problem of the calibration algorithm that cannot find a suitable parameter set for the catchment. The latter can be the case, when the model structure cannot be adapted to this specific catchment.

There are eight catchments near these monster catchments that are not selected as monsters. In these catchments, some similar patterns in base flow are visible and in four cases the GR4H model failed based on consistency. The SUPERFLEX approach is able to produce consistent results for these catchments with performance up to 0.7. Catchment H7742010 (Thérain at Bauvais, 754 km^2) is one of the neighbour catchments that shows BFP. SUPERFLEX structure SF09 has an average performance of 0.73 and robustness of 0.53 while GR4H fails to give a consistent parameter set. Figure 7—6 and Figure 7—7 show a hydrograph of a dry and a wet year in validation. Comparing these two figures explains the low robustness: the structure calibrated on the dry years (Figure 7—7) underestimates peaks and overestimates low flow, while it does less so when calibrated on the wet years. Despite this, the model performs relatively well.

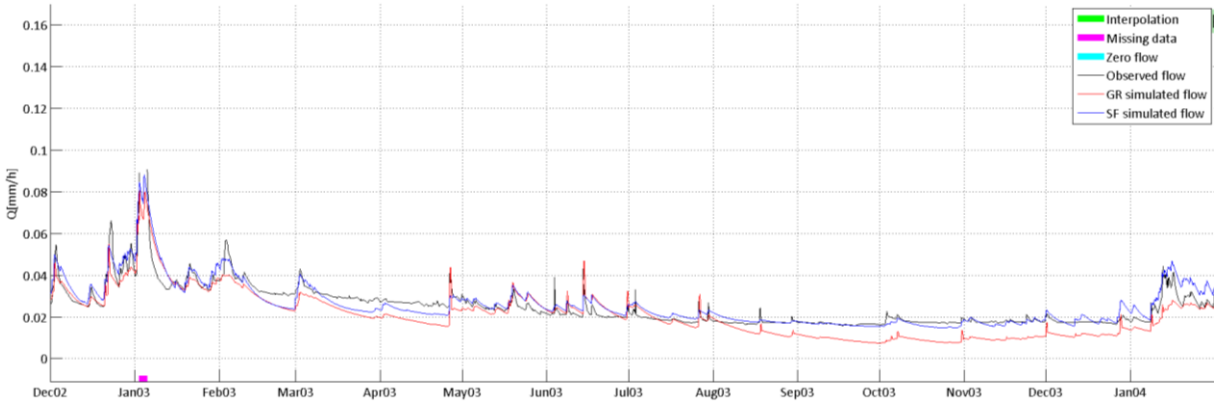

Figure 7—6. Hydrographs of catchment H7742010 –Thérain at Beauvais – 754 km^2 , a zoom of a dry year 2003 with model results (SF09 and GR4H) calibrated on the wet period.

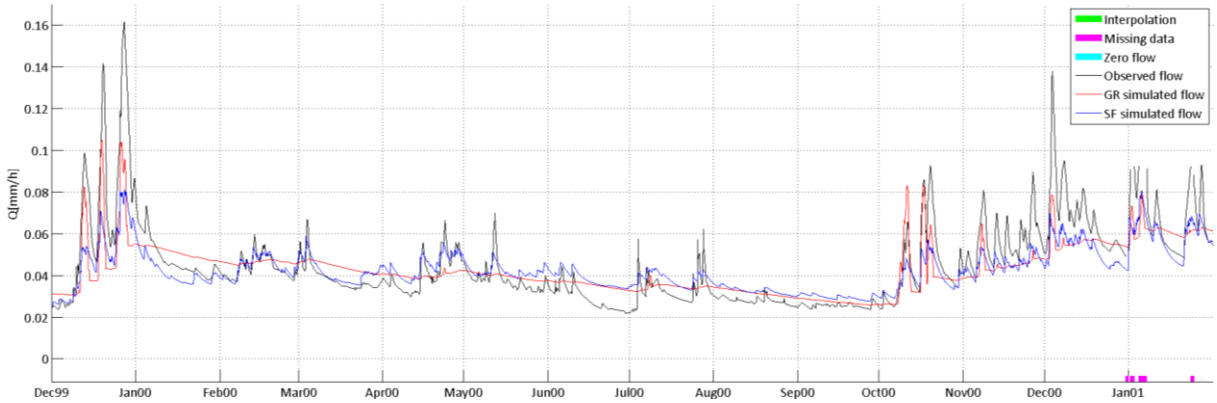

Figure 7—7. Hydrographs of catchment H7742010 –Thérain at Beauvais – 754 km^2, a zoom of a wet year 2000 with model results (SF09 and GR4H) calibrated on the dry period.

In the SF09 structure, water from the upper unsaturated reservoir is divided into a fast and a slow reservoir. In case of catchment H7742010 and other neighbours with BFP, the slow reservoir plays an important role in flow simulation. A large portion of flow (67%-88%) is directed through the slow reservoir. In this reservoir, BFP can be clearly seen (Figure 7—8 and Figure 7—9). The fast reservoir can mimic the catchment's response to individual rainfall events and is unaffected by the processes in the slow reservoir. The GR4H model appears to have too little flexibility to adjust for this type of flow, since the division of flow is fixed.

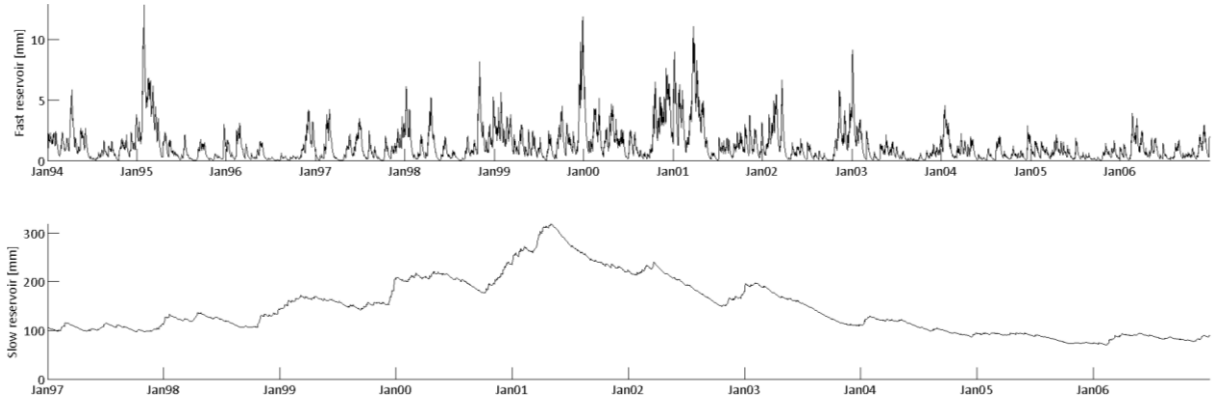

Figure 7—8. Water level in fast and slow reservoir of structure SF09 calibrated on S1 of catchment H7742010.

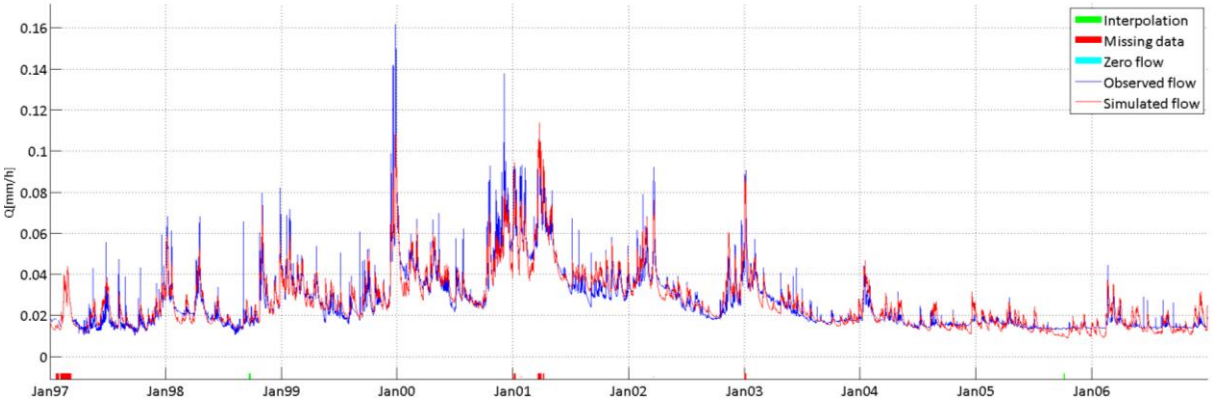

Figure 7—9. Observed and simulated hydrograph by SF09 calibrated on period S1 of catchment H7742010.

The introduction of a second routing reservoir in the GR4H structure should reduce these problems as, discussed by Le Moine, 2008; Pushpalatha et al. (2012). The size of this existing routing reservoir is calibrated, much like the residence time of the slow reservoir in SF09 is calibrated. In monster H8042010 (Figure 7—5) and in neighbour catchment H7742010 (Figure 7—7), GR4H fails to respond to individual events while SF09 does much better for the neighbours and seems better able at identifying the important hydrological processes in both calibration periods. The independent calibration of the fast flow path in SF09 allows for a good response to the individual events and is a key advantage in this case.

The number of cases where any of the SUPERFLEX structures with a slow and a fast reservoir work well on catchments like the one above, is limited. This means there is little support for any general statement about the advantages of these structures, although they support the findings of Kavetski and Fenicia (2011). In the monster catchments with BFP, these structures are rejected based on consistency. Perhaps the differences in calibration conditions are still too large for these structures to find a robust parameter set.

7.1.2 Flashy flow

In 18 of the 69 monster catchments, the reason for the monsters is most likely linked to the flashy nature of flow. Many of these catchments are situated in the south of France near the Mediterranean sea (Figure 7—10) where climate conditions (like large rainfall intensities) are responsible for flashy behaviour. Catchments further from the Mediterranean sea are relatively small or lay in impermeable areas. These conditions also increase the chance of flashy flow since a single rainfall event can have a large and fast effect on total discharge.

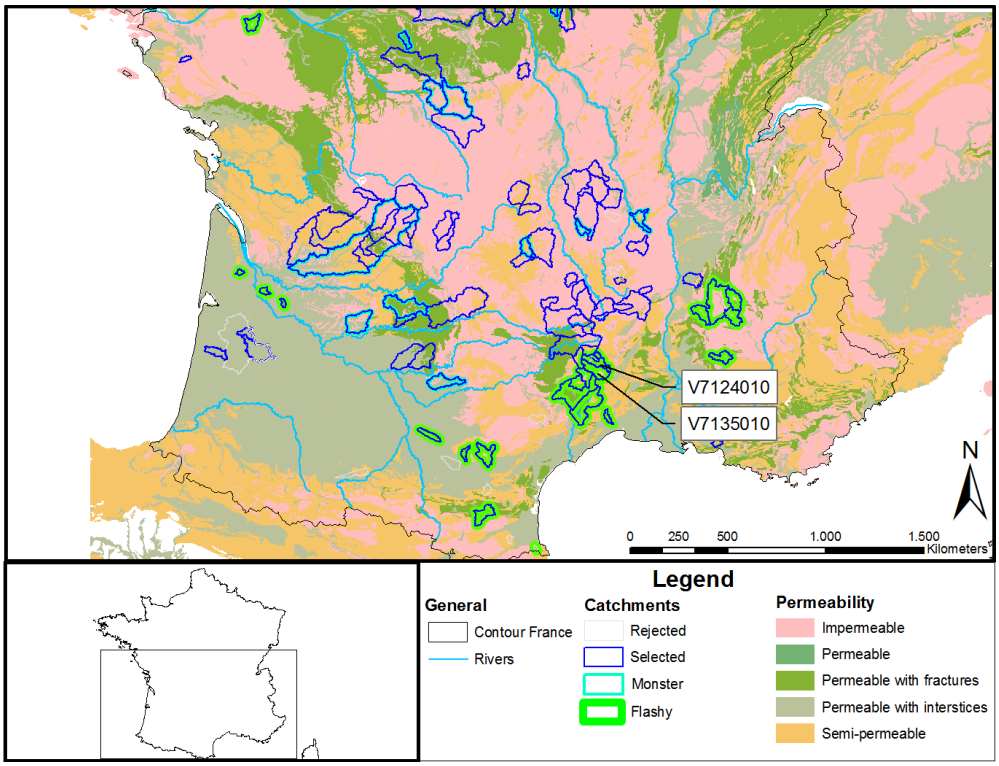

Figure 7—10. Location and permeability of monster catchments in the south of France.

The flashy flow monsters have relative high and sharp peaks while on average flow is low and without much response to smaller rainfall events. Table 7—1 attempts to illustrate this by giving the average 'flashiness' of three groups of catchments, which is much higher in the monster catchments than in their neighbours. The table also shows that flashiness in monster catchments is higher than over all catchments. The monsters are thus especially flashy compared to the region they are in. These finding are complemented by the 24-hour autocorrelation of flow, which characterises the catchment dynamics in response to rainfall events. The flashy catchments respond more sharply than the other catchments, again emphasizing the flashy behaviour of these catchments.

Table 7—1. 'Flashiness' of flashy flow monsters and their neighbours compared to the whole catchment set. Here flashiness is defined as the average of maximum discharge in each of the ten years in the time series over the average discharge. Also the average 24-hour autocorrelation is given to show the sharp response of the flashy catchments.

	Average discharge [mm/hr $\cdot 10^{-2}$]	Peak measurement [mm/hr]	'Flashiness' [-]	24h auto-correlation [-]
Flashy Flow monsters	4.20	1.91	42.78	0.60
Flashy Flow neighbours	4.24	1.49	28.53	0.67
All catchments	4.07	0.86	16.10	0.80

Modelling these monster catchments

The performance of both modelling approaches is poor on the 18 flashy catchments. Again the flexible approach is able to find a consistent parameter set more often, but with simpler low performing structures. Two neighbouring catchments in the Mediterranean area were used as examples of the flashy flow catchments. They are both selected as monsters: V7124010 (Gardon de Mialet at Générargues [Roucan], 240 km^2) and V7135010 (Gardon de Saint-Jean at Corbès [Roc Courbe], 262 km^2). The catchments are located in the Cévennes mountains, which is prone to extreme rainfall events at the end of the summer. They have a very similar shape, lie next to each other and flow in the same direction. Both have schist bedrock and are semi-permeable. Figure 7—11 and Figure 7—12 show hydrographs of these two catchments in the spring and summer of 1999. The major flashy flood events occur typically in fall, winter and spring time while summer is characterised by very low flow.

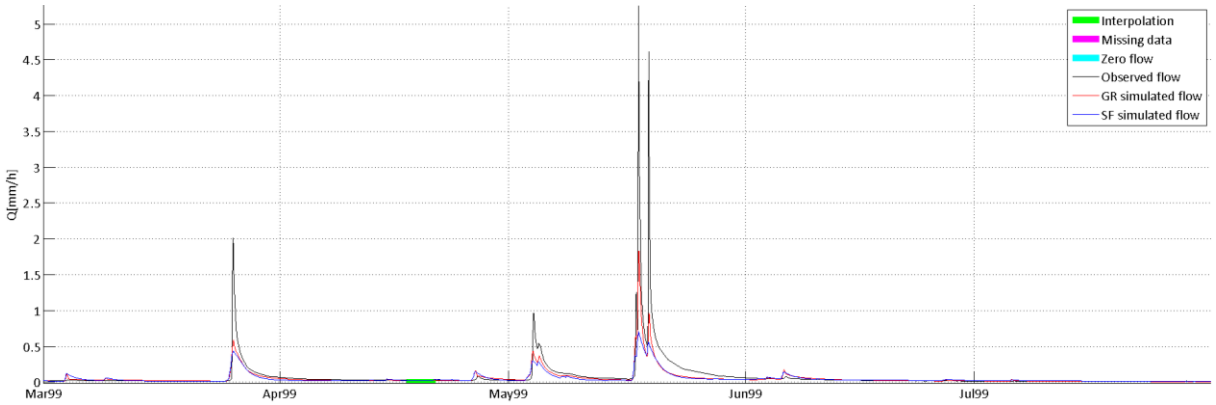

Figure 7—11. Hydrographs of catchment V7124010 – Gardon de Mialet at Générargues [Roucan] – 240 km^2, a zoom of spring and summer 1999 with model results (SF09 and GR4H) in validation.

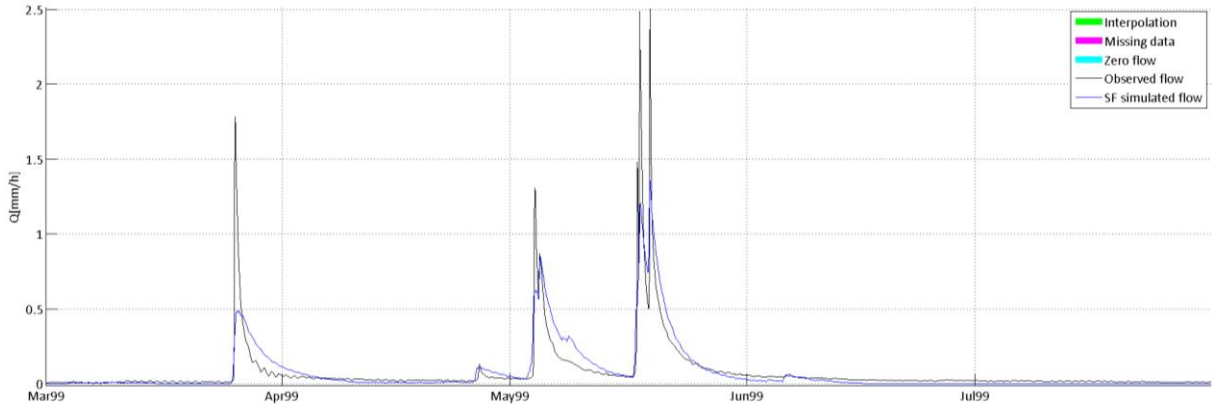

Figure 7—12. Hydrographs of catchment V7135010 – Gardon de Saint-Jean at Corbès [Roc Courbe] – 262 km^2, a zoom of spring and summer 1999 with model results (SF01, GR4H is inconsistent) in validation.

Figure 7—11 and Figure 7—12 show simulated hydrographs that are unable to mimic the height of the peaks in the observed flow. In the hydrograph of catchment V7124010 (Figure 7—11), SF09 and GR4H both give consistent parameter sets but with relative low performance, 0.32 and 0.46 respectively. CR4 (for variability) values are especially low for both models, meaning that the models are not able to mimic the full range in the observed flow, in this case the high peaks. SUPERFLEX structure SF01 is the only consistent structure in catchment V7135010 (Figure 7—12). SF01 is a single reservoir model with a residence time and a power function. This simple model is able to simulate flow reasonably well: apart from CR2 (low flow) the model scores relatively high on all criteria.

A relatively large number of catchments in this area is rejected due to data issues. Low flow measurements are sensitive to measurement errors and disturbances near the measurement station. Measurements of peak discharge are sensitive to measurement errors in the stream's rating curve. The relation between water height and discharge is rarely measured during peak discharge and thus interpolated on few points and may change over time (MEDD, 2007).

Additionally, there are problems linked to convective rainfall events near the Mediterranean sea that produce large peak flows after summer. The lumped approach may have some difficulties with the spatial variability of these rainfall events or the infiltration in soils may respond differently after a long dry period than during repeated rainfall, so called wetting up of the catchment (Piñol et al., 1997). Figure 7—13 shows an example of a peak discharge after summer which is highly underestimated by both models. Two rainfall events later, the GR4H model is able to simulate the peak discharge much better.

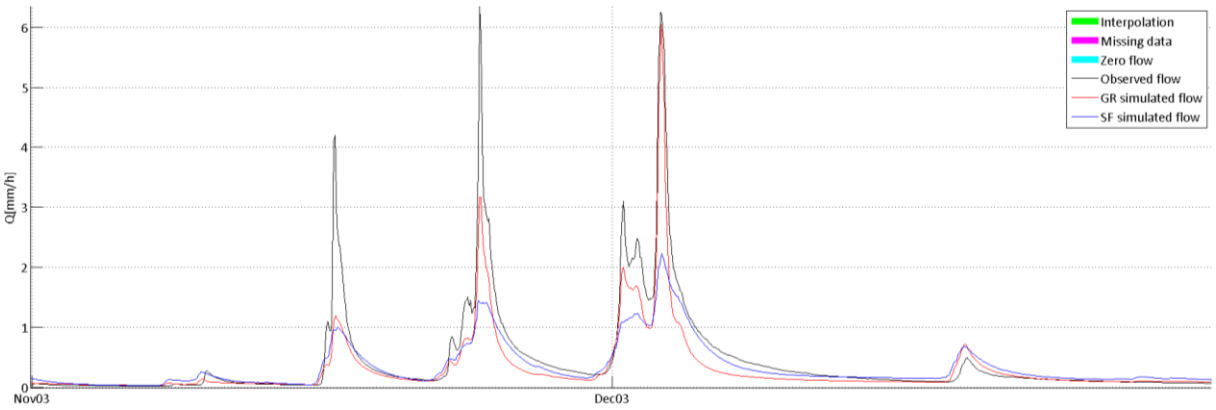

Figure 7—13. Hydrographs of catchment V7124010 – Gardon de Mialet at Générargues [Roucan] – 240 km^2, a zoom of November and December 2003 with model results (SF09 and GR4H) in validation.

Differences between the fixed and flexible approach

For 13 out of 18 flashy flow catchments SUPERFLEX finds a consistent structure In most cases, structures that are simpler than GR4H are selected with low performance and high robustness. In three cases, structure SF09 is selected, a slightly more complex structure than GR4H (but performance remains low). These results are not in accordance with those of Kavetski and Fenicia (2011), who argue that threshold type model components (like in SF03) increase performance in catchments with flashy flow.

In 5 out of 18 catchments, a consistent parameter set for GR4H is found. Average performance is close to the 0.5 threshold under which models are marked as monster catchments. Both approaches generally perform better on flashy flow catchments than on those with the BFP. The fixed GR4H still fails on more catchments but in those cases, the sizes of the stores do not vary much between periods. The limited differences between calibration and validation make pinpointing plausible causes of model failure more difficult.

The relatively small number of events in a calibration period combined with problems of spatial variability and wetting up of the catchments could explain the model's inability to correctly identify the dominant hydrological processes. Simple models like SF01 are able to describe some part of the flow more consistently, which is plausible as they will need less information to find an optimal parameter set.

7.1.3 Disturbed flow measurements

There are 14 monster catchments left that are difficult to allocate to either of the previous groups. In these catchments the most likely reason for a model's low performance or failure is disturbance in flow measurements. Figure 7—14 shows a hydrograph of catchment H4223110 (Remarde at Saint-Cyr-sous-Dourdan, 151 km^2). The observed recession (April 2003) looks unnatural in this catchment. Instead of slowly decreasing, flow remains constant and shows small variations possibly caused by downstream influences. Although this station is not reported to have unreliable data (MEDD, 2007), the extremely low discharge may be very sensitive to (downstream) disturbances. With flow of 0.02 mm/h and area of 151 km^2, discharge is only 0.84 m^3/s. Small disturbances like deposition of branches or vegetation development (reported in MEDD, 2007) can have large effects on the hydrograph, which makes modelling more challenging than with high discharges.

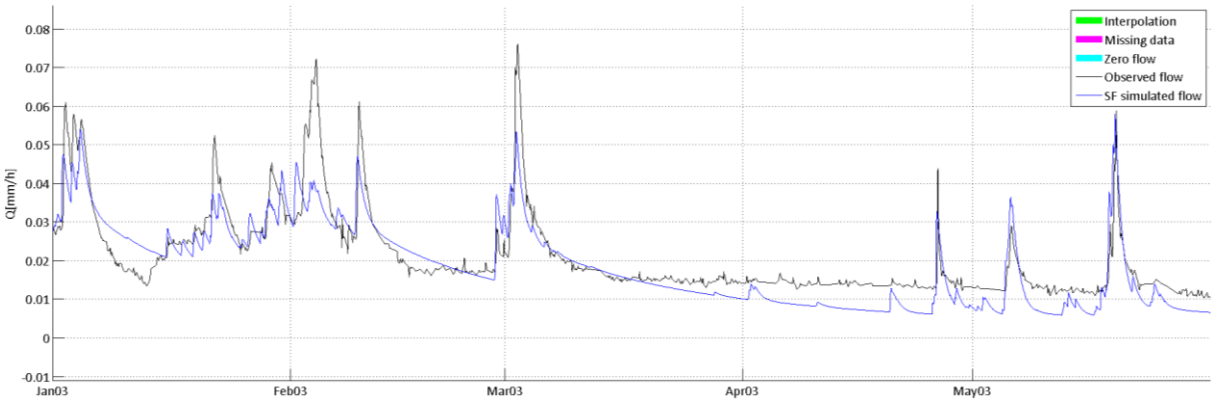

Figure 7—14. Hydrographs of catchment H4223110 – Remarde at Saint-Cyr-sous-Dourdan – 151 km^2, a zoom of January through June 2003 with model results (SF07) in validation. An example of unnatural recession.

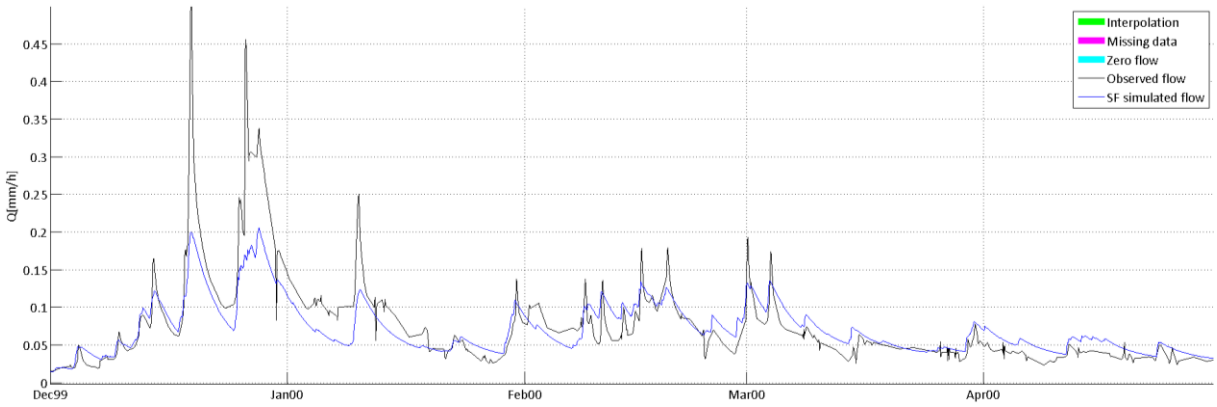

Figure 7—15. Hydrographs of catchment A3422010 – Zorn at Saverne [Schinderthal] – 183 km^2, a zoom of December 1999 through May 2000 with model results (SF09) in validation with examples of sudden downward spikes.

Figure 7—15 shows the hydrographs of A3422010 (Zorn at Saverne [Schinderthal], 183 km^2) where observed flow shows sudden downward spikes. In this case these disturbances appear not only on low flow and have only a short effect. The influence on the calibration result is difficult to evaluate without further tests, but apart from these spikes, the observed flow in catchment A3422010 appears natural. Poor performance in the remaining catchments can be the result of one of the above reasons but pinpointing the main one remains difficult.

7.2 Predicting monster catchments

In section 2.3, the possibility of predicting which catchments will be monsters was discussed. Observed data were analysed for water balance errors and cumulative discharge curves were compared with the curves from neighbouring catchments. These techniques were aimed at finding out which catchments showed anomalies and could be expected to be monster catchments. Figure 7—16 shows the average performance of the fixed approach against that of the flexible approach four times. In each plot a different indicator is used to predict monster catchments, catchments that are predicted to be monster are circled.

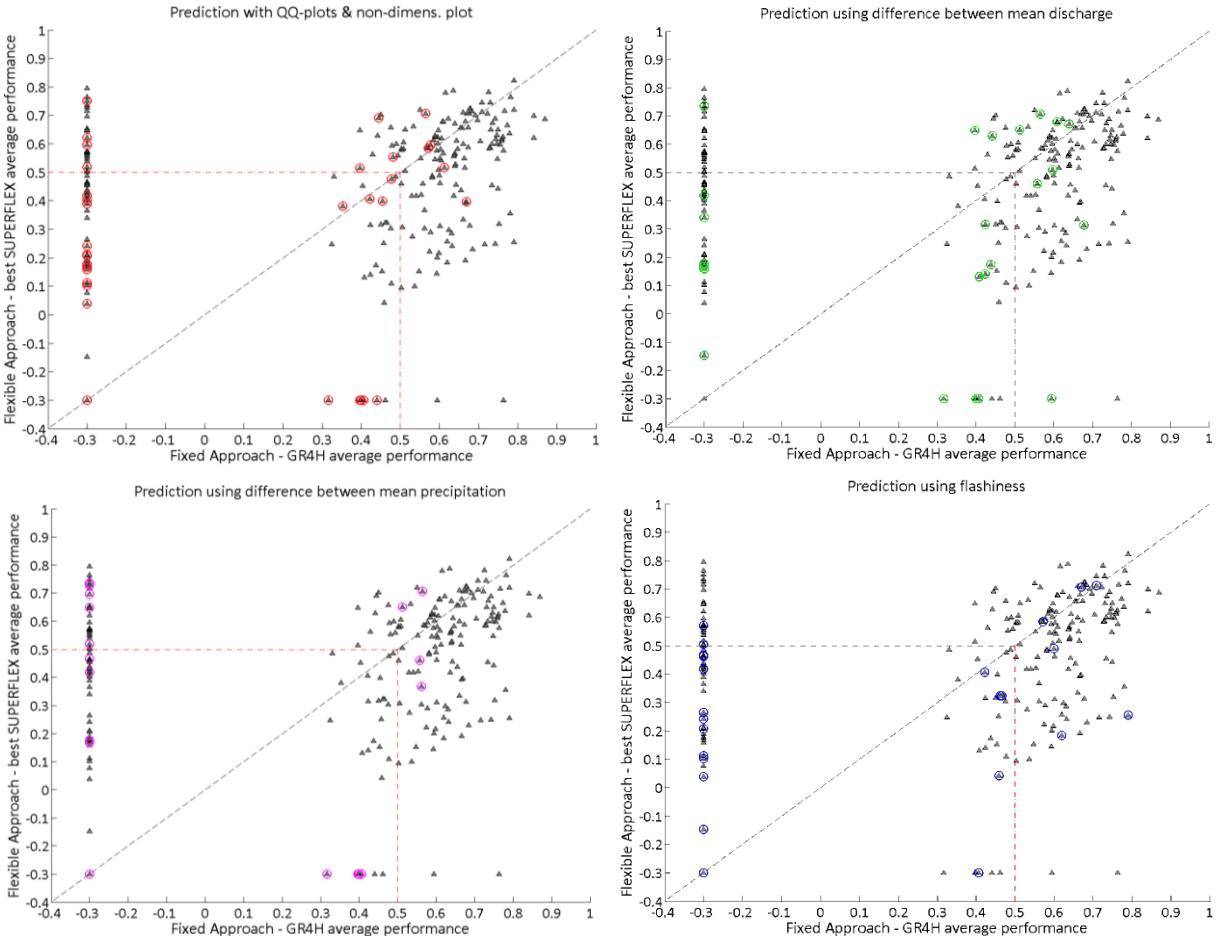

Figure 7—16. Average performance of the fixed approach vs. the average performance of the flexible approach for individual catchments. Inconsistent model results are given the value of -0.3 and are shown on the edges of the figure. Catchments below the red dashed/dotted lines are selected as hydrological monsters. In each plot a different indicator for predicting monster catchments was used. Top left: QQ-plots and non-dimensional plot (section 2.3, 44 selected). Top right: difference in mean discharge between the two periods (24 selected). Bottom left: difference in mean precipitation between the two periods (24 selected). Bottom right: flashiness (24 selected). Predicted catchments are circled.

The suspected monster catchments based on QQ-plots and the non-dimensional plot (section 2.3) do not accurately match actual monster catchments. Eight of the 44 suspected monster catchments are in fact high performing catchments for both modelling approaches. Also many monster catchments were not predicted and these techniques were therefore mostly unsuccessful.

Given the groups in which the monsters are classified in section 7.1, the wet and dry periods in precipitation or discharge and flashiness could be good indicators to predict monster catchments. For each indicator, 24 catchments were selected to see if they could be used for predicting monster

catchments. The remaining plots in Figure 7—16 show that for all indicators prediction is poor: for all three indicators several catchments that perform high for both approaches are selected. It seems it is impossible to predict monster catchments with a single indicator. Combining these indicators used here is unlikely to yield better results, because the indicators are probably unrelated (e.g. the flashy behaviour was not observed in the monsters with severe climatic variability). Other (combinations of) indicators should be investigated and results preferably validated on an independent catchment set. This study has shown that predicting hydrological monsters is difficult using some apparently obvious indicators.

7.3 Concluding remarks

In this chapter the hydrological monsters were further investigated to find the causes of the poor or inconsistent modelling results. It was found that the catchments could be grouped into three groups based on the likely cause of poor model performance:

1. Catchments where wet and dry periods have led to severe differences in observed flow between calibration and validation periods. These catchments proved to be difficult to model, especially when the effect of one wet or dry year lasts over multiple years. Many structures give inconsistent results on these catchments, but those with independent parallel flow paths have a higher success rate.
2. Other monster catchments showed very flashy flow and are generally small, located near the Mediterranean sea and are small or located in impermeable areas. Structures are not able to correctly simulate the flashy nature of the flow from these catchments. Many structures were inconsistent or gave poor performance due to poor simulation of the high sharp peaks. On these catchments, simple structures perform relatively well and are often selected as best.
3. The last group contains catchments where disturbances in measured flow are the most likely reason for poor performance. Several disturbances can be observed in observed flow which are difficult to explain. Because no other likely cause for these monster catchments could be found, these disturbances are expected to cause hydrological monsters by influencing the evaluation criteria or the calibration process.

Finally, the possibility to predict monster catchments was tested. Unfortunately, these results show that it is difficult to predict which catchments will be monsters only based on the water balance or when observed flow is compared against neighbouring catchments. The use of some flow characteristics that correspond to groups into which the monsters were divided, also proved unsuccessful.

8 Discussion

Calibration was performed in the BATEA programme to ensure that all thirteen model structures were treated in the same way. The calibration optimised the objective function based on the weighted least square (WLS) method. The WLS method was used to put less emphasis on the calibration of peak discharges by assuming a flow dependent error model. The error model assumes the errors to be normally distributed and independent from time step to time step. Emphasis was put on low flows by a linear relation to discharge. However, this study found that the WLS method is still more sensitive to high flows as all model structures scored poorly on the low flow criterion.

The validation was carried out using four evaluation criteria, aiming to help identify which type of flow causes poor model performance. The disadvantage of these criteria is that they differ from the objective function for calibration, which can be confusing. Calibration on the four criteria was not possible due to time and modelling constraints. Double calibration and validation of each catchment and the consistency rules have made sure that only models that give a consistent representation of a catchment are considered. Despite being somewhat arbitrary, the consistency rules have made the validation test more severe.

This study is the first to apply twelve SUPERFLEX structures on a large catchment set. Both the structures and parameter bounds are based on expert knowledge of hydrology and experience with the SUPERFLEX approach. Although the range of structures remains limited compared to the possibilities of numerical models, the twelve structures cover a wide range of common conceptual models. They were designed to differ stepwise in complexity which has made the analysis of individual model components possible. Calibration, especially of the most complex structures, was time consuming. This should be considered when selecting a modelling approach. Also, the complex models are more prone to parameter equifinality and have a higher chance of resulting in an inconsistent parameter set.

The GR4H model has not been created out of SUPERFLEX building blocks and could therefore not be evaluated in exactly the same way. The BATEA programme uses an implicit scheme to evaluate all equations of a single time step at once while in GR4H the equations are evaluated sequentially. The effects on the results of these differences cannot be determined at this point, but they are expected to be small as sequential evaluation generally remains stable. Implementation in SYPERFLEX would prevent occasional instability and make both approaches better comparable.

The large catchment set used in this study contains a wide range of catchments that should make the results quite general and extent the work of Kavetski and Fenicia (2011). A relatively large number of monster catchments was analysed in which three groups could be distinguished. The different trends in the hydrographs of the catchments in each group could be clearly observed. The analysis of the monster catchments showed some leads towards the workings of different model components on these catchments, but results were mixed.

9 Conclusions & Recommendations

This chapter summarizes the main findings of this study in the conclusions (section 9.1) and recommendations (section 9.2).

9.1 Conclusions

This study aimed to *find and clarify the hydrological monsters of a fixed and a flexible modelling approach by investigating which model structures perform well on average and why catchments become monsters for some structure. As a conclusion we provide* short answer to each research question raised at the beginning of this report:

What are the effects of increasing model complexity on the performance of different catchments?

The average performance on the 237 catchments has shown that seven of the thirteen used model structures perform better than the remaining six. The seven structures perform very similarly despite their varying complexity (in terms of parameters or functions), showing that increasing complexity does not automatically mean higher performance. This corroborates previous findings in the literature (e.g. Jakeman et al., 1993, Perrin et al., 2001, among others). Comparison of the used structures also showed that a power function increases average performance on all catchments and that the lag function used in SUPERFLEX is largely redundant with the other delaying functions existing in the structure (especially reservoirs).

Model performance in catchment classes showed that for all structures, small catchments are more difficult to model, which is in agreement with the results by Merz et al. (2009). Performance in wet catchments is generally higher. It also showed that structures with independent parallel flow paths have an advantage in permeable and groundwater dominated catchments that can be linked to hydrological processes in these catchments. Some model components are always important, but selecting those that fit the type of catchment can considerably increase model performance.

What are successful model structures and what are hydrological monsters when the fixed and flexible modelling approaches are strictly followed?

The fixed and flexible approaches perform very similarly on the used catchment set when all catchments are considered. The SUPERFLEX structures SF04, SF09 and SF11 perform well on average and are selected as best model on many catchments. SF04 uses two reservoirs with power functions to describe reservoir outflow, SF09 uses three reservoirs with two independent flow paths and SF11 combines both. The use of these key components results in good average performance of these model structures on the catchment set. The stepwise increase in complexity among the SUPERFLEX structures led to very small performance differences between those structures that expanded on the best three. This shows the difficulty of selecting a single best structure, especially when a model structure must be chosen without simulation.

Several of the SUPERFLEX structures taken individually perform very similarly to GR4H, including SF04 which is considered slightly less complex than the GR4H model. The GR4H model is still more often selected as best, proving its high average performance despite the simplicity of the model structure. Only when the consistency rules are used, the advantage of the flexible approach becomes clear: The flexible approach is consistent for more catchments than the fixed approach. In these cases, the flexible structure can be better adjusted to fit the catchment characteristics.

When looking at the performance on individual catchments, it becomes clear that both approaches perform very differently on some catchments. In this study however, the focus is on those catchments for which both modelling approaches are inconsistent or perform low. This has led to the selection of 69 hydrological monsters out of a total of 237 catchments.

What do the monster catchments look like why do model structures perform poorly on these catchments?

Three groups can be distinguished in the hydrological monsters based on the observed hydrographs of the monster catchments: catchments with severe climatic differences between calibration and validation periods, catchments with flashy flows and catchments with small scale disturbances.

Catchments with severe climatic differences between calibration and validation period

Wet years in the first period and dry years in the second period lead to differences in flow which are too large to be simulated by most models. Especially in permeable, groundwater dominated catchments, flow differences between the periods are large and even leads to a pattern of increasing base flow in wet years and decreasing base flow in dry years. Models with independent parallel flow components are in some cases able to simulate these inter-annual patterns, but calibration conditions often remain too different from validation for correct simulation. This result can be linked to the recent studies on model robustness under contrasted conditions (see e.g. Merz et al. 2011, Coron et al., 2012).

Catchments with flashy flow

Long periods of low flow interrupted by very steep and high peak flows are poorly simulated by all model structures. These flashy flow catchments are mainly impermeable, small and situated near the Mediterranean sea. Poor model simulation is linked to the influence of catchment saturation on the response to individual rainfall events and poor gauging of convective rainfall events. Some very simple single-reservoir models are able to give reasonable results.

Catchments with small scale disturbances

Disturbances in observed flow, either caused by measurement errors or actual (downstream) influences on the stream water level, hinder good simulation. In some catchments observed flow is very small leading to relatively large influences of downstream disturbances, such as vegetation or fallen logs. Downstream locks or dams can influence larger streams, especially during recession or low flow. The used models are not equipped to mimic these disturbances while general behaviour can be quite good. The reasons for poor performance can be linked to a response of the calibration or over-sensitivity of the evaluation criteria (especially for low flow) to the disturbances.

The main research question, '*When and why do a catchment and a model become a hydrological monster?'*, can now be answered. This study has shown that most hydrological monsters can be explained by looking at the observed hydrograph and are mostly linked to the catchment type, not the model. The similar average performance of many model structures shows that different conceptualisations can give similar results. This also shows that the role of the model structure may be limited in the hydrological monsters. However, some of the model components in the SUPERFLEX structures did cause a significant increase in performance for specific types of catchments, showing the importance of selecting the right structure for each catchment.

9.2 Recommendations

This study showed that several different model structures perform very similarly on average. For the SUPERFLEX approach in particular, some model components (power functions and parallel flow paths) significantly increase average model performance while other components (lag-function and interception store) proved largely redundant with other model functions. More research towards the sensitivity or implementation of some parameters of the SUPERFLEX structures is needed to confirm their value for the model.

This study also showed that flexibility in model structure can help a model to better adapt to specific catchments, such as independent parallel flow path for groundwater dominated catchments. Some recent work towards improving the GR4H model (modification of the groundwater exchange function and introduction of a second parallel routing store, see Le Moine, 2008 and Pushpalatha et al., 2011) showed that the model gained in versatility, with a significant improvement of low-flow simulation. The results presented here suggest that the introduction of a flexible power in the existing non-linear routing store may yield valuable improvements for these specific catchments. Another way the GR4H model might be improved is by ensuring that the role of the production and routing store cannot be switched during calibration. Full implementation of the GR4H model into the SUPERFLEX approach, including implicit evaluation of model equations, will allow for easy investigation of changes to the GR4H and is therefore recommended.

This study analysed the monster catchments of both modelling approaches. In this analysis the role of some model components were singled out, but not examined closely. This study provides some leads to what type of model structure is necessary for what type of catchments, but further research is needed given the existing level of equifinality between various model structures. Research towards those catchments where different model structures give a very different result or when a structure works very well on a specific (type of) catchment can provide a better understanding of the hydrological processes taking place in a catchment.

Predicting a priori which catchments will become monster catchments was unsuccessful in this study. The classification of the monster catchments do provides some notion towards how monster catchments could be predicted. For instance, flashy flow catchments could be selected and modelled with a specific model. However, the a priori selection of monsters requires more research. The best chance of preventing hydrological monsters is with a more complex model structure (like SF11) or by selecting the best structure for each catchment.

Finally, calibration conditions are of great influence on model simulation and care should be taken into selection of these periods (e.g. to contain several types of flow).

References

Andréassian, V., Perrin, C., Parent, E. & Bárdossy, A. (2010). The Court of Miracles of Hydrology: can failure stories contribute to hydrological science? *Hydrological Sciences Journal, 55*(6), 849-856.

Beven, K. J. (2001). *Rainfall-Runoff Modelling - The Primer*. Chichester, England: John Wiley & Sons Ltd.

Booij, M. J. (2003). Determination and integration of appropriate spatial scales for river basin modelling. *Hydrological Processes, 17*(13), 2581-2598.

Boughton, W. (2006). Calibrations of a daily rainfall-runoff model with poor quality data. *Environmental Modelling & Software, 21*(8), 1114-1128.

Bourgin, P. Y., Lobligeois, F., Peschard, J., Andréassian, V., Le Moine, N., Coron, L., Perrin, C., Ramos, H. & Khalifa, A. (2011). *Description des caractéristiques morphologiques, climatiques et hydrologiques de 4436 bassins versants français – Guide d'utilisation de la base de données hydro-climatique.* Antony: Internal document, Unité HBAN, Equipe Hydrologie, Irstea, Antony.

Bredehoeft, J. (2005). The conceptualization model problem—surprise. *Hydrogeology Journal, 13*(1), 37-46.

Browman, H. I. (1999). Negative results. *MARINE ECOLOGY PROGRESS SERIES, 191*, 301-309.

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., et al. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 australian catchments. *Water Resources Research, 48*(5), W05552.

Diermanse, F. L. M. (2001). *Physically based modelling of rainfall runoff processes.* Universiteit Delft, Delft, The Netherlands.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. & Wood, E. F. (2006). Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *Journal of Hydrology, 320*(1–2), 3-17.

Edijatno, Nascimento, N. D., Yang, X. L., Makhlouf, Z. & Michel, C. (1999). GR3J: a daily watershed model with three free parameters. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques, 44*(2), 263-277.

European Environment Agency. (2006). Corine Land Cover 2006 raster data from EEA: http://www.eea.europa.eu//themes/landuse/dc

Fenicia, F., Kavetski, D. & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research, 47*(11), W11510.

Fenicia, F., Kavetski, D., Savenije, H. H. G., Schoups, G., Pfister, L., Clark, M. P. & Freer, J. (2012). Catchment properties and conceptual model structure: is there a correspondance? Part 2. Modelling and hypothesis testing. *Hydrological Processes*, under review.

Goswami, M. & O'Connor, K. M. (2010). A "monster" that made the SMAR conceptual model "right for the wrong reasons". *Hydrological Sciences Journal, 55*(6), 913-927.

Gupta, H. V., Kling, H., Yilmaz, K. K. & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology, 377*(1-2), 80-91.

Hellebrand, H., van den Bos, R., Hoffmann, L., Juilleret, J. & Pfister, L. (2008). The potential of winter stormflow coefficients for hydrological regionalization purposes in poorly gauged basins of the middle Rhine region. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques, 53*(4), 773-788.

Jakeman, A. J. & Hornberger, G. M. (1993). How Much Complexity Is Warranted in a Rainfall-Runoff Model. *Water Resources Research, 29*(8), 2637-2649.

Jakeman, A. J., Letcher, R. A. & Norton, J. P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software, 21*(5), 602-614.

Kavetski, D. & Clark, M. P. (2010). Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction. Water Resources Research, 46(10).

Kavetski, D. & Evin, G. (2011). *BATEA The User Guide (v 7.020.005)*: University of Newcastle (Australia).

Kavetski, D. & Fenicia, F. (2011). Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. *Water Resources Research, 47*(11), W11511.

Kavetski, D. & Kuczera, G. (2007). Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration. *Water Resources Research, 43*(3), -.

Kavetski, D., Kuczera, G. & Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research, 42*(3).

Kavetski, D., Kuczera, G., Thyer, M. & Renard, B. (2007). *Multistart Newton-type optimisation methods for the calibration of conceptual hydrological models.* Paper presented at the MODSIM07 - Land, Water and Environmental Management: Integrated Systems for Sustainability, Proceedings.

Klemes, V. (1986). Operational Testing of Hydrological Simulation-Models. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques, 31*(1), 13-24.

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

Le Moine, N., Andréassian, V. & Mathevet, T. (2008). Confronting surface- and groundwater balances on the La Rochefoucauld-Touvre karstic system (Charente, France). *Water Resources Research, 44*(3), W03403.

Le Moine, N., Andreassian, V., Perrin, C. & Michel, C. (2007). "Outlier" catchments: What can we learn from them in terms of prediction uncertainty in rainfall-runoff modelling? *IAHS-AISH Publication, 313*, 195-203.

Le Moine, N., Andréassian, V., Perrin, C. & Michel, C. (2007b). How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments. *Water Resources Research, 43*(6), W06428.

MEDD. (2007). Banque Hydro. Retrieved 2012, from Le Ministère de l'Ecologie et du Développement Durable (MEDD), Direction de l'Eau et le SCHAPI: www.hydro.eaufrance.fr/

Mathevet, T., Michel, C., Andréassian, V. & Perrin, C. (2006). A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins. *Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment–MOPEX, IAHS Publ. 307*, 211-219.

Merz, R., Parajka, J. & Bloschl, G. (2009). Scale effects in conceptual hydrological modeling. *Water Resources Research*, 45, W09405.

Merz, R., Parajka, J. & Blöschl, G. (2011). Time stability of catchment model parameters: Implications for climate impact analyses. *Water Resources Research, 47*(2), W02531.

Météo France. (n.d.). Prévisions météo de Météo-France. https://public.meteofrance.com/

Nash, J. E. & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology, 10*(3), 282-290.

Perrin, C., Fenicia, F., Kavetski, D., Andréassian, V. & Booij, M. J. (2011). Can flexible hydrological models demystify the hydrological "monsters" of a fixed model structure? Investigation on a large set of catchments. Enschede: University of Twente.

Perrin, C., Michel, C. & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology, 279*(1-4), 275-289.

Piñol, J., Beven, K. & Freer, J. (1997). Modelling the hydrological response of mediterranean catchments, Prades, Catalonia. The use of distributed models as aids to hypothesis formulation. *Hydrological Processes, 11*(9), 1287-1306.

Pushpalatha, R., Perrin, C., Le Moine, N. & Andréassian, V. (2012). A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology, 420-421*(2012), 171-182.

Refsgaard, J. C. & Hansen, J. R. (2010). A good-looking catchment can turn into a modeller's nightmare. *Hydrological Sciences Journal*, 55(6), 899-912.

Refsgaard, J. C. & Henriksen, H. J. (2004). Modelling guidelines - terminology and guiding principles. *Advances in Water Resources*, 27(1), 71-82.

Schertzer, D., Tchiguirinskaia, I., Lovejoy, S. & Hubert, P. (2010). No monsters, no miracles: In nonlinear sciences hydrology is not an outlier! *Hydrological Sciences Journal, 55*(6), 965-979.

Schoenberg, R. (2001). Optimization with the Quasi-Newton Method. Aptech Systems, Inc. Maple Valley, WA.

Sorooshian, S. & Gupta, V. K. (1995). Model calibration. In V. P. Singh (Ed.), *Computer Models of Watershed Hydrology* (pp. 23-68). Colorado: Water Resources Publications.

Szöllösi-Nagy, A. (2009). *LEARN FROM YOUR ERROR - IF YOU CAN! Reflections on the value of hydrological forecasting models*. Delft, The Netherlands: UNESCO-IHE.

Troldborg, L., Refsgaard, J., Jensen, K. & Engesgaard, P. (2007). The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system. *Hydrogeology Journal*, 15(5), 843-860.

Valéry, A., Andréassian, V. & Perrin, C. (2010). Regionalization of preciptation and air temperature over high altitude catchments - learning from outliers. *Hydrological Sciences Journal*, 55(6), 928-940.

Wagener, T. (2003). Evaluation of catchment models. *Hydrological Processes, 17*(16), 3375-3378.

Wagener, T., Lees, M. J. & Wheater, H. S. (2001). A toolkit for the development and application of parsimonious hydrological models. In V. P. Singh, R. Frevert & D. Meyers (Eds.), *Mathematical models of small watershed hydrology* (Vol. 2). LLC, USA: Water Resources Publications.

Wagener, T., Sivapalan, M., Troch, P. & Woods, R. (2007). Catchment Classification and Hydrologic Similarity. *Geography Compass, 1*(4), 901-931.

Wu, W., Clark, J. S. & Vose, J. M. (2010). Assimilating multi-source uncertainties of a parsimonious conceptual hydrological model using hierarchical Bayesian modeling. *Journal of Hydrology, 394*(3-4), 436-446.

Appendices

A. Data analysis

This appendix gives the list of the catchments used in this study and presents examples of how data were analysed prior to model simulation. This appendix discusses missing and interpolated data, measurement errors and contains a full list of all catchments used in this study.

A.I. Missing data and interpolations

Figure A—1 shows part of the hydrograph of one of the 250 catchments in which part of the discharge data was interpolated. Any series of interpolated data longer than 48 hours is detected and will not be used for calibration or validation. Missing data are treated the same way, but they are much easier to detect since they are already replaced by negative values. A specific column with quality code was added to the data file that is read by the BATEA program so it knows which data to use and which to skip. The percentages for missing data and detected interpolations are given in Table A—1 below.

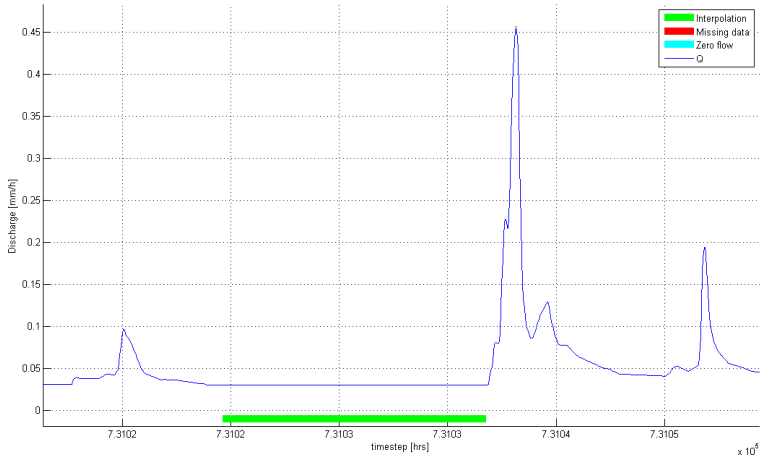

Figure A—1. Example of an interpolation in the hydrograph of catchment P6222510 – Auvézère at Lubersac – 115 km^2.

A.II. Measurement errors

Figure A—2 shows an example of a hydrograph where small sudden spikes appear during low flow. The downward spikes and low flow make the accuracy of these spikes unlikely. There are a number of catchments showing these spikes with varying frequency. Possible reasons are measurement errors, river regulations, locks or water extractions for industrial or agricultural use. Catchments showing a high frequency of spikes are documented for use in a later stage of the analysis.

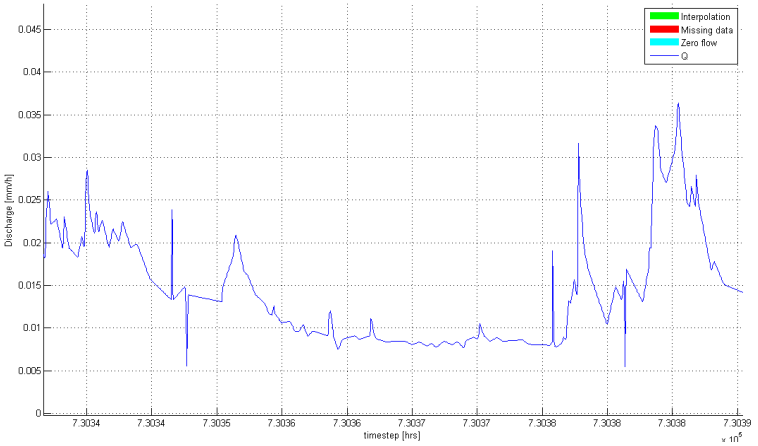

Figure A—2. Example of small sudden spikes during low flow in catchment R1132510 – Tardoire at Maisonnais-sur-Tardoire – 136 km^2.

Larger errors in measurement will have an impact on the water balance and most likely on model performance. These catchments may be monsters simply because of these errors. A way to find catchments with an unrealistic water balance is by using a non-dimensional plot of the ratio between discharge and rainfall (Q/P) against the ratio between rainfall and potential evapotranspiration (P/PET, Le Moine et al., 2007). Catchments outside the bands Q/P=1 and Q/P=1-1/(P/PE) correspond to specific water balance characteristics (catchments gaining water – e.g. karstic catchments – or losing water – e.g. leaky catchments) or to problems in measurements (over/underestimation of rainfall, potential evapotranspiration or streamflow). Figure A—3 shows that from the 250 catchments 26 fall outside the bands. These are also reported in Table A—1, which may be the indication of specific hydrological behaviour or problems in data.

Note that there may be catchments within the limits that also exhibit problems in data.

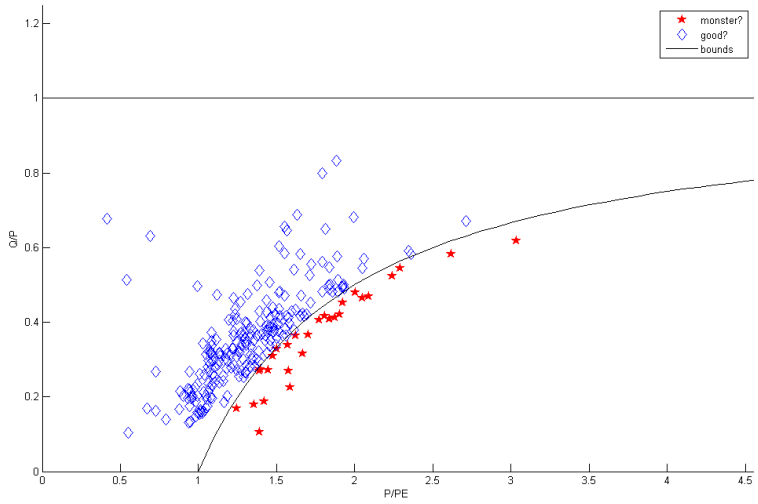

Figure A—3. Non-dimensional plot with 250 catchments for checking water balance.

Additionally, all catchments are checked visually by comparing the cumulative discharge of each catchment to that of its three nearest neighbours. This procedure was implemented at Irstea by Laurent Coron (personal communication). It shows whether one catchment behaves significantly different from its neighbours, which may point to errors in measurements. In Figure A—4 catchment A8322010 is compared with three of its neighbours which all show different behaviour indicating that this catchment may be more difficult to model, especially because no specific features appear in the rainfall plots. The upper right plot in this figure shows a sudden jump, while neither of its neighbours nor the rainfall shows any sign of a large flood. When this specific catchment was compared to its three nearest neighbours, none of them showed this sudden change indicating a possible measurement error in this catchment.

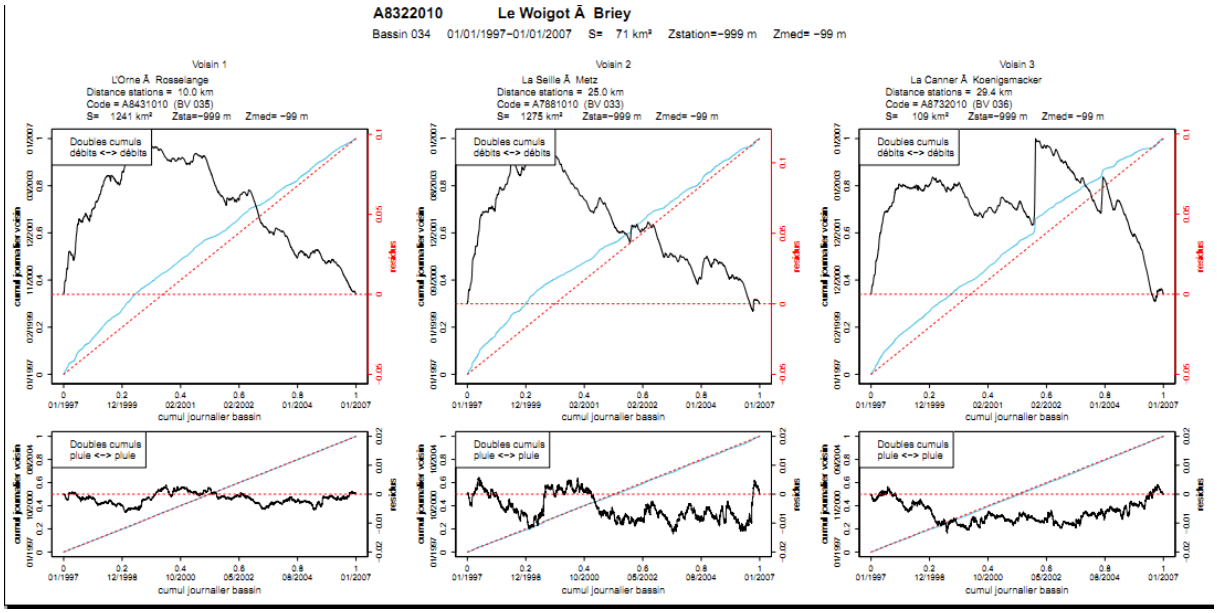

Figure A—4. Example of how the cumulative discharges of neighbour catchments are compared, in this case an example of unexpected behaviour. The three upper plots show the cumulative QQ-plots (diagonal blue line) and the residuals between the blue line and a linear between the first and the last point. The three lower plots show the same but for rainfall.

Table A—1 discussed in the next section also indicates the catchments that show very different behaviour from their neighbours. In total 48 catchments are flagged. They show little correspondence to the catchments selected by the non-dimensional plot. Only four catchments are both outside the non-dimensional plot and show strange behaviour compared to their neighbours. The value of either of the analyses therefore remains largely unknown; they will therefore be used with caution in the analysis of possible monsters.

A.III. List of available catchments

The 250 catchments are coded with letter and number. The letter code corresponds to the region in which the catchments is situated. Figure A—5 shows the number code per region. The remaining code consists of six digits for individual catchments. Table A—1 shows a full list of the catchments along with the catchment name, area and results from the data analysis conducted in this study.

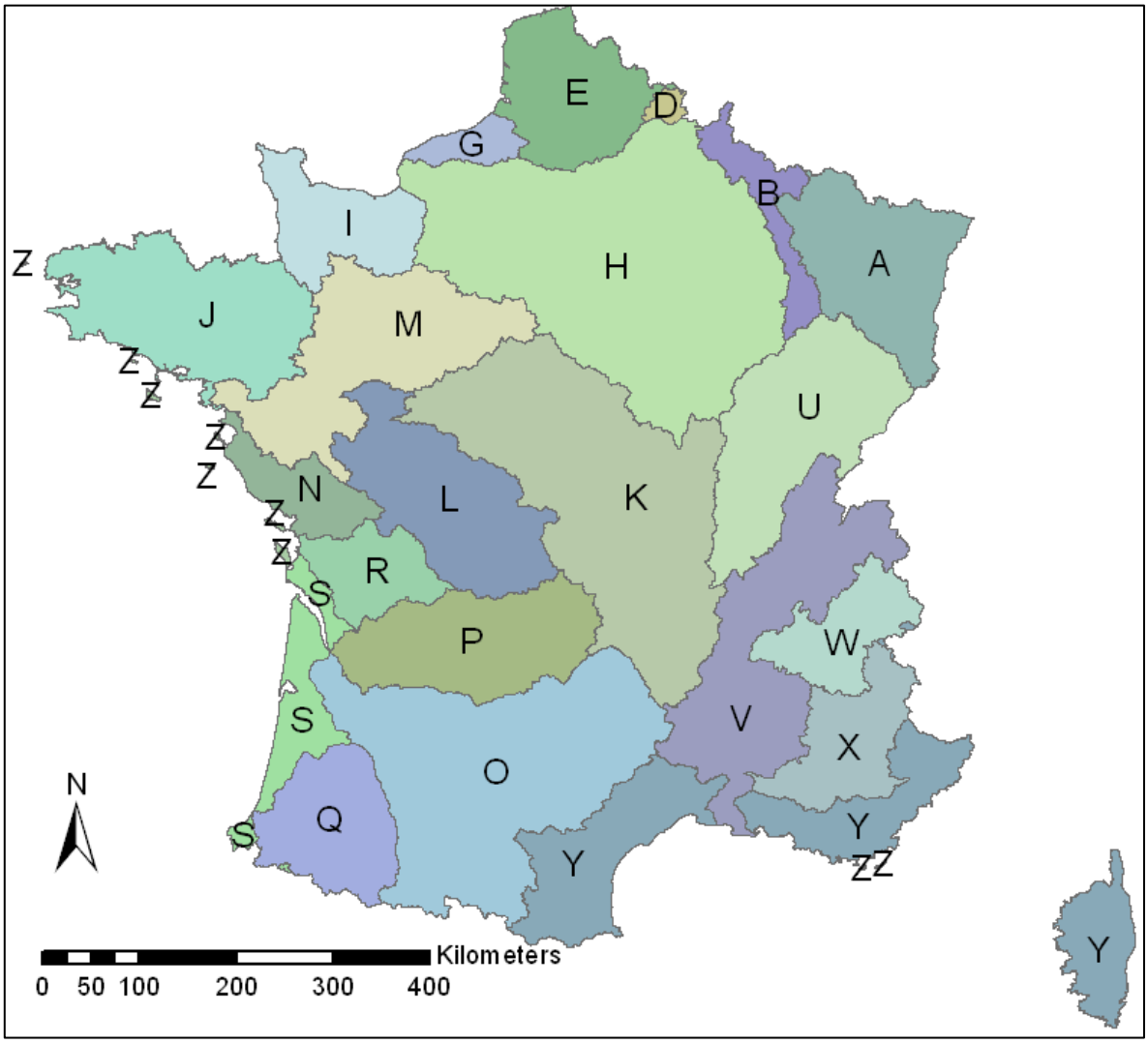

Figure A—5. Positions of catchment letter codes in France (MEDD, 2007).

Table A—1. Overview of the 250 catchments used in this study with the results of the data quality and possible monster analysis

Catchment description			Data Quality			Possible monster	
Code	Name of measurement station	Surface (km2)	Missing data (%)	Interpolated >48h (%)	Spikes	Non-Dimens. Plot	Conflicting neighbours
A1522020	La Lauch à Guebwiller	80	2.3%	1.1%		x	
A2052020	La Fecht à Ostheim	460	0.0%	0.7%		x	
A2312020	Le Giessen à Thanvillé	115	0.0%	1.7%		x	
A2332110	La Lièpvrette à Lièpvre	107	0.0%	1.1%		x	
A2512010	L'Andlau à Andlau	41	0.0%	2.1%			
A2612010	L'Ehn à Niedernai	57	0.0%	0.4%		x	
A2732010	La Bruche à Russ [Wisches]	223	0.0%	1.1%		x	
A2842010	La Mossig à Soultz-les-Bains	170	0.0%	2.1%	x		x
A2860110	La Bruche à Holtzheim [2]	684	0.0%	1.1%		x	
A3301010	La Moder à Schweighouse-sur-Moder [aval]	622	0.0%	0.5%			
A3422010	La Zorn à Saverne [Schinderthal]	183	0.0%	0.2%			
A3472010	La Zorn à Waltenheim-sur-Zorn	683	0.0%	1.1%			
A4173010	La Cleurie à Cleurie	65	1.6%	1.1%			
A4200630	La Moselle à Saint-Nabord [Noirgueux]	627	2.1%	0.0%		x	
A4250640	La Moselle à Épinal	1218	0.0%	0.0%		x	
A4333010	Le Neuné à Laveline-devant-Bruyères	94	0.0%	1.0%		x	
A4362030	La Vologne à Cheniménil [2]	355	0.1%	0.2%			
A5431010	Le Madon à Pulligny	950	0.0%	0.1%			
A5500610	La Moselle à Pont-Saint-Vincent	3079	0.7%	0.1%			
A5730610	La Moselle à Toul	3345	0.0%	0.0%			
A6051020	La Meurthe à Saint-Dié	369	0.0%	1.0%		x	
A6221010	La Meurthe à Azerailles	962	0.5%	0.0%			
A6571110	La Vezouze à Lunéville	566	0.0%	0.8%			
A6731220	La Mortagne à Gerbéviller	494	0.0%	2.7%			
A6761010	La Meurthe à Damelevières	2316	0.6%	0.5%			
A6872010	Le Sanon à Dombasle-sur-Meurthe	286	0.6%	0.4%			
A6921010	La Meurthe à Laneuveville-devant-Nancy	2788	1.5%	1.2%			
A6941020	La Meurthe à Malzéville [2]	2925	4.3%	0.0%			
A7010610	La Moselle à Custines	6836	0.0%	0.0%			
A7581020	La Seille à Moyenvic	349	1.6%	1.5%			
A7642010	La Petite Seille à Château-Salins	153	0.2%	5.5%			

Catchment description			Data Quality			Possible monster	
Code	Name of measurement station	Surface (km2)	Missing data (%)	Interpolated >48h (%)	Spikes	Non-Dimens. Plot	Conflicting neighbours
A7821010	La Seille à Nomeny	926	1.5%	0.6%			
A7881010	La Seille à Metz	1275	0.0%	1.0%			
A8322010	Le Woigot à Briey	71	0.0%	6.3%			x
A8431010	L'Orne à Rosselange	1241	1.0%	1.0%			
A8732010	La Canner à Koenigsmacker	109	0.0%	2.7%			
A9021010	La Sarre à Sarrebourg	307	0.0%	3.4%		x	x
A9091050	La Sarre à Keskastel	887	0.1%	0.5%			
A9091060	La Sarre à Diedendorf	737	0.6%	1.1%		x	
A9221010	La Sarre à Sarreinsming	1760	0.0%	0.1%			
A9301010	La Sarre à Wittring	1717	0.0%	0.0%			
A9352050	L'Eichel à Oermingen	280	1.1%	2.0%			x
A9832010	La Nied Allemande à Faulquemont	203	0.7%	1.2%			x
A9862010	La Nied Allemande à Varize	366	0.4%	0.4%			
A9942010	La Nied à Bouzonville	1150	9.7%	0.6%			
B1092010	Le Mouzon à Circourt-sur-Mouzon [Villars]	401	0.8%	0.8%		x	x
B1282010	Le Vair à Soulosse-sous-Saint-Élophe	443	0.9%	0.7%			x
B1322010	Le Vair à Belmont-sur-Vair	139	0.0%	0.1%			
B2042010	L'Aroffe à Vannes-le-Châtel	197	1.0%	10.9%		x	x
B2220010	La Meuse à Saint-Mihiel	2550	0.0%	0.5%			
E3511220	La Lys à Delettes	162	0.8%	1.2%			
E3518510	La Laquette à Witternesse	81	2.5%	2.4%		x	x
E3646210	La Clarence à Robecq	224	1.8%	2.5%		x	
E4035710	L'Aa à Wizernes	393	1.6%	0.7%	x		
E4306010	La Hem à Tournehem-sur-la-Hem [Guémy]	106	1.9%	0.5%		x	
E5300210	La Liane à Wirwignes	104	0.4%	4.9%			
E5400310	La Canche à Brimeux	917	1.3%	1.4%	x		
E5505720	L'Authie à Dompierre-sur-Authie	788	5.0%	0.9%	x		
E6406010	L'Avre à Moreuil	621	3.0%	0.9%			x
E6426010	La Selle à Plachy-Buyon	546	0.0%	0.7%	x		x
G0402020*	La Bresle à Ponts-et-Marais [Ponts-et-Marais]	686	29.1%	0.7%	x		
H0203030	La Laignes aux Riceys	648	0.0%	0.3%		x	
H0321030	L'Ource à Autricourt	582	0.0%	0.3%			

Catchment description			Data Quality			Possible monster	
Code	Name of measurement station	Surface (km2)	Missing data (%)	Interpolated >48h (%)	Spikes	Non-Dimens. Plot	Conflicting neighbours
H0400010	La Seine à Bar-sur-Seine	2353	0.3%	0.4%	x		
H0503010	L'Hozain à Buchères [Courgerennes]	240	0.0%	0.3%			
H1201010	L'Aube à Bar-sur-Aube	1299	0.0%	0.5%			
H1713010	L'Ardusson à Saint-Aubin	158	0.0%	1.0%			x
H2062010	Le Beuvron à Ouagne [Champmoreau]	265	0.0%	0.1%			
H2073110	Le Sauzay à Corvol-l'Orgueilleux	89	0.0%	0.2%			
H2473010	L'Armance à Chessy-les-Prés	480	1.5%	0.1%			
H3322010	La Bezonde à Pannes	343	0.1%	0.1%			x
H3613020	Le Lunain à Épisy	247	0.0%	0.6%			
H4022020	L'Essonne à Guigneville-sur-Essonne [La Mothe]	850	0.0%	0.1%	x		
H4022030	L'Essonne à Boulancourt	585	0.0%	0.5%	x		x
H4042010	L'Essonne à Ballancourt-sur-Essonne	1858	0.0%	0.4%	x		
H4202020	L'Orge à Saint-Chéron [Saint-Évroult]	111	0.0%	5.0%			
H4223110	La Remarde à Saint-Cyr-sous-Dourdan	151	0.0%	0.8%			
H4252010	L'Orge à Morsang-sur-Orge	942	0.1%	0.1%			x
H4322030	L'Yerres à Courtomer [Paradis]	426	0.0%	1.9%			x
H4333410	Le Réveillon à Férolles-Attilly [La Jonchère]	56	2.2%	0.4%			x
H5033310	La Suize à Villiers-sur-Suize	83	0.0%	0.1%			
H5062010	Le Rognon à Doulaincourt-Saucourt	619	0.2%	0.0%			
H5083050	La Blaise à Louvemont [Pont-Varin]	469	2.3%	0.4%			
H5122340	L'Ornain à Tronville-en-Barrois	674	0.0%	0.5%		x	
H5142620	La Chée à Bettancourt-la-Longue	234	6.6%	0.1%	x		
H5172010	La Saulx à Vitry-en-Perthois	2117	0.0%	0.0%			
H5173110	Le Bruxenelle à Brusson	129	0.0%	1.7%			
H5702010	Le Grand Morin à Meilleray	350	0.9%	0.7%			
H5732010	Le Grand Morin à Pommeuse	762	0.9%	0.1%			
H6021020	L'Aisne à Verrières	384	8.9%	0.0%			
H6122010	L'Aire à Varennes-en-Argonne	633	1.6%	0.1%			
H6402030	La Vesle à Puisieulx	610	8.3%	0.5%			
H6412010	La Vesle à Saint-Brice-Courcelles	739	8.3%	0.1%	x		
H7513010	L'Automne à Saintines	286	2.1%	0.4%	x		x
H7602010	La Brêche à Nogent-sur-Oise	463	0.1%	0.6%	x		x

Catchment description			Data Quality			Possible monster	
Code	Name of measurement station	Surface (km2)	Missing data (%)	Interpolated >48h (%)	Spikes	Non-Dimens. Plot	Conflicting neighbours
H7702010	Le Thérain à Bonnières	207	0.0%	1.3%	x		
H7742010	Le Thérain à Beauvais	754	1.7%	0.2%	x		
H8012010	L'Epte à Gournay-en-Bray	247	1.0%	0.1%			x
H8042010	L'Epte à Fourges	1386	2.1%	0.4%			x
H8043310	L'Aubette de Magny à Ambleville	101	2.7%	3.2%			x
H8212010	L'Andelle à Vascoeuil	377	0.0%	1.0%	x		
H9021010	L'Eure à Saint-Luperce	318	0.7%	0.0%			x
H9113001	La Drouette à Saint-Martin-de-Nigelles	229	0.7%	0.0%	x		
H9121010	L'Eure à Charpont	2021	0.0%	0.0%	x		
H9402030	L'Iton à Normanville	1030	1.6%	0.6%	x		x
I0011010	La Risle à Rai	143	2.0%	0.7%			
I0102010	La Charentonne à Bocquencé	67	1.6%	2.6%		x	
I0113010	Le Guiel à Montreuil-l'Argillé	86	1.4%	0.9%			
I1203010	La Calonne aux Authieux-sur-Calonne	171	3.5%	0.8%			
I2021010	La Dives à Beaumais	279	0.5%	0.4%			
I2213610	L'Ancre à Cricqueville-en-Auge	60	0.1%	5.1%			x
I3462010	Le Noireau à Cahan [Les Planches - CD 911]	525	6.0%	0.2%			
I5053010	La Souleuvre à Carville	116	2.6%	0.5%			
I5352010	La Drôme à Sully	240	3.5%	0.6%			
J2614020	Le Queffleuth à Plourin-lès-Morlaix [Les Trois Chênes]	94	2.3%	1.0%			
J3024010	Le Guillec à Trézilidé	42	1.3%	1.4%			
J3811810	L'Aulne à Châteauneuf-du-Faou [Pont Pol ty Glass]	1223	1.7%	0.3%			
J4211910	L'Odet à Ergué-Gabéric [Tréodet]	206	0.0%	2.4%			
J4224010	Le Jet à Ergué-Gabéric	109	0.9%	0.2%			
J4313010	Le Steir à Guengat [Ty Planche]	182	1.1%	1.7%			
J4734010	L'Inam au Faouët [Pont Priant]	116	0.0%	1.6%		x	
J4742010	L'Éllé à Arzano [Pont Ty Nadan]	574	2.1%	1.2%		x	
J4803010	L'Isole à Scaër [Stang Boudilin]	97	0.0%	2.1%			
J4813010	L'Isole à Quimperlé [Place des Anciennes Fonderies]	226	4.4%	0.8%			
J7024010	La Valière à Erbrée [Pont D 110]	30	0.0%	4.2%			
J7633010	Le Semnon à Bain-de-Bretagne [Rochereuil]	415	0.0%	3.7%			
J7963010	Le Don à Guémené-Penfao [Juzet]	605	0.4%	6.8%			

Catchment description			Data Quality			Possible monster	
Code	Name of measurement station	Surface (km2)	Missing data (%)	Interpolated >48h (%)	Spikes	Non-Dimens. Plot	Conflicting neighbours
J8632410	L'Aff à Quelneuc [La rivière]	347	0.0%	2.9%			x
K0114020	La Gazeille à la Besseyre-Saint-Mary	52	2.8%	4.9%			
K0214010*	La Gagne à Saint-Germain-Laprade [Les Pandreaux]	111	15.1%	0.2%			
K0454010	La Dunières à Sainte-Sigolène [Vaubarlet]	218	2.9%	10.2%			
K0614010	Le Furan à Andrézieux-Bouthéon	178	0.9%	0.9%			
K0643110*	La Mare à Saint-Marcellin-en-Forez [Vérines]	95	4.4%	21.4%			x
K0753210	Le Lignon du Forez à Boën	374	0.2%	1.0%			
K0773220	Le Lignon de Chalmazel à Poncins [2]	662	0.9%	1.7%			
K0813020	L'Aix à Saint-Germain-Laval	196	0.9%	1.9%			
K2010820	L'Allier à Laveyrune [Rogleton 2]	48	1.0%	1.0%			
K2070810	L'Allier à Langogne	326	1.0%	0.7%			
K2173020	Le Chapeauroux à Saint-Bonnet-de-Montauroux	385	1.3%	0.1%			
K2223030	L'Ance du Sud à Saint-Préjet-d'Allier	159	3.5%	10.3%			
K2254010	La Seuge à Saugues	115	2.0%	0.6%			
K2514010	L'Allanche à Joursac [Pont du Vernet]	156	0.8%	0.6%		x	
K2523010	L'Alagnon à Joursac [Le Vialard]	323	0.6%	0.1%			
K2593010	L'Alagnon à Lempdes	995	0.5%	1.3%			
K2674010	La Couze Chambon à Montaigut-le-Blanc [Champeix]	159	0.6%	8.2%	x		
K2821910	La Dore à Dore-l'Église	106	3.2%	2.5%			
K2871910	La Dore à Saint-Gervais-sous-Meymont [Maison du Parc / Giroux-Dore]	789	3.1%	0.1%	x		
K2884010	La Faye à Olliergues [Giroux-Faye]	76	0.7%	4.2%	x		
K2981910	La Dore à Dorat	1518	2.3%	0.2%	x		
K3222010	La Sioule à Pontgibaud	355	0.4%	1.3%			
K4013010	L'Aubois à Grossouvre [Trézy]	135	1.5%	0.0%			
K4443010	L'Ardoux à Lailly-en-Val	164	0.0%	0.1%	x		
K6373020	La Petite Sauldre à Ménétréol-sur-Sauldre	346	0.0%	0.1%			
K6402510	La Sauldre à Salbris	1239	8.5%	0.0%			
K6492510	La Sauldre à Selles-sur-Cher	2297	2.9%	0.1%	x		
K7312610	L'Indre à Saint-Cyran-du-Jambot	1701	0.1%	1.3%			
K7414010	La Tourmente à Villeloin-Coulangé [Coulangé]	106	1.1%	1.0%			x
K7777777	La Loire à Ardentes [station test bidon]	687	9.7%	0.9%			
L0563010	La Briance à Condat-sur-Vienne [Chambon Veyrinas]	607	0.0%	0.0%			

Catchment description			Data Quality			Possible monster	
Code	Name of measurement station	Surface (km2)	Missing data (%)	Interpolated >48h (%)	Spikes	Non-Dimens. Plot	Conflicting neighbours
L4411710	La Petite Creuse à Fresselines [Puy Rageaud]	853	0.0%	0.5%			
L4653010	La Bouzanne à Velles [Forges]	438	1.0%	0.1%			
M0050620	La Sarthe à Saint-Céneri-le-Gérei [Moulin du Désert]	907	0.0%	2.1%	x		
M0114910	Le Merdereau à Saint-Paul-le-Gaultier [Chiantin]	118	0.0%	0.2%	x		
M0243010	L'Orne Saosnoise à Montbizot [Moulin Neuf Cidrerie]	501	3.6%	0.2%			
M0361510	L'Huisne à Nogent-le-Rotrou [Pont de bois]	842	0.4%	0.1%	x		x
M1041610	Le Loir à Saint-Maur-sur-le-Loir	1085	0.0%	2.7%			
M3020910	La Mayenne à Madré	329	0.3%	0.6%			
M3060910	La Mayenne à Ambrières-les-Vallées [Cigné]	832	0.6%	0.3%			
M3103010	La Varenne à Domfront	201	4.6%	0.0%			
M3133010	La Varenne à Saint-Fraimbault [Moulin Crinais]	513	0.2%	0.6%			
M3323010	L'Ernée à Andouillé [Les Vaugeois]	378	0.0%	0.0%			
M3711810	L'Oudon à Cossé-le-Vivien	135	0.0%	3.3%			
M3771810	L'Oudon à Châtelais [Marcillé]	732	0.0%	0.7%			
M3774010	Le Chéran à la Boissière	78	0.0%	0.6%			
M5214020	L'Hyrome à Saint-Lambert-du-Lattay [Chauveau]	153	0.0%	1.6%			x
M6333020	L'Erdre à Nort-sur-Erdre [Moulin de Vault]	457	0.0%	1.9%	x		
M8144020	La Logne à Legé [Le Paradis]	44	0.1%	0.4%			
O0295310*	La Noue à Laffite-Toupière	121	0.9%	38.0%			x
O2215010	La Saune à Quint-Fonsegrives	109	0.0%	11.7%	x		x
O3064010	Le Tarnon à Florac	133	0.0%	5.8%			
O3084320	La Mimente à Florac	127	0.0%	5.7%			
O3141010	Le Tarn à Mostuéjouls [La Muse]	945	1.0%	1.0%			
O5754020	La Vère à Bruniquel [La Gauterie]	314	0.0%	1.7%			
O5854010	La Lère à Réalville	388	0.0%	1.0%			
O5964020	Le Lemboulas à Lafrançaise [Lunel]	401	0.0%	7.7%			
O7001510	Le Lot à Bagnols-les-Bains	92	0.0%	11.5%	x		
O7021530	Le Lot à Mende [aval]	288	0.0%	1.8%			
O7101510	Le Lot à Banassac [La Mothe]	1161	0.0%	1.9%			
O7234010	La Rimeize à Rimeize	117	0.0%	3.1%			
O8113510*	Le Célé à Figeac [Merlançon]	672	20.0%	1.1%	x		
O8133520	Le Célé à Orniac [Les Amis du Célé]	1254	0.0%	0.1%			

Catchment description			Data Quality			Possible monster	
Code	Name of measurement station	Surface (km2)	Missing data (%)	Interpolated >48h (%)	Spikes	Non-Dimens. Plot	Conflicting neighbours
O8255010	Le Vert à Labastide-du-Vert [Les Campagnes]	116	0.0%	2.5%			
O8584010	La Lède à Casseneuil	413	2.0%	0.0%			
O9685310	La Pimpine à Cénac	52	0.0%	2.6%			
O9785310	La Jalle de Ludon au Pian-Médoc	32	0.6%	10.9%			x
P2484010	Le Céou à Saint-Cybranet	573	0.0%	2.5%			
P3234010	La Loyre à Voutezac [Pont de l'Aumonerie]	103	0.0%	1.0%			
P3274010	La Loyre à Saint-Viance [Pont de Burg]	250	0.0%	1.0%			
P3322510	La Corrèze à Saint-Yrieix-le-Déjalat [Pont de Lanour]	58	0.0%	0.8%			
P3502510	La Corrèze à Tulle [Pont des soldats]	356	10.0%	0.1%	x		
P3674010	La Montane à Laguenne [Pont de la Pierre]	213	0.0%	0.2%			
P5715010	L'Engranne à Baigneaux	28	0.0%	10.2%			
P6081510	L'Isle à Corgnac-sur-l'Isle	447	0.0%	2.6%			
P6161510	L'Isle à Mayac	804	3.4%	0.5%			
P6222510	L'Auvézère à Lubersac	115	0.0%	2.0%		x	
P6342510	L'Auvézère à Cherveix-Cubas	587	3.1%	10.0%			
P6382510	L'Auvézère au Change [Aubarède]	883	0.0%	0.7%			
P7001510	L'Isle à Bassilac [Charrieras]	1863	0.0%	0.6%			
P7041510	L'Isle à Périgueux	2112	0.0%	0.7%			
P7261510	L'Isle à Abzac	3758	0.0%	0.4%	x		x
P8284010	La Lizonne à Saint-Séverin [Le Marchais]	622	1.0%	0.2%	x		
P8312520	La Dronne à Bonnes	1915	0.0%	1.0%	x		
P8462510	La Dronne à Coutras	2790	0.0%	9.5%			
R1132510	La Tardoire à Maisonnais-sur-Tardoire	136	2.5%	2.0%			
S2134010	La Petite Leyre à Belhade	420	0.0%	0.5%			
S2224610	Le Grand Arriou à Moustey [Biganon]	115	0.0%	4.1%			
S2242510*	L'Eyre à Salles	1676	15.7%	0.4%			
S3214010	Le Canteloup à Saint-Paul-en-Born [Talucat]	153	0.5%	0.5%			
U0020010	La Saône à Monthureux-sur-Saône	232	0.0%	4.6%			
U0124010	Le Coney à Fontenoy-le-Château	316	0.0%	3.5%	x	x	
V4034020	La Véore à Beaumont-lès-Valence [Laye]	196	0.0%	0.2%			x
V4145210	La Glueyre à Gluiras [Tisoneche]	72	0.5%	0.2%			
V4214010	La Drôme à Luc-en-Diois	191	0.0%	0.3%	x		

Catchment description			Data Quality			Possible monster	
Code	Name of measurement station	Surface (km2)	Missing data (%)	Interpolated >48h (%)	Spikes	Non-Dimens. Plot	Conflicting neighbours
V4225010	Le Bez à Châtillon-en-Diois	225	0.7%	0.1%			
V4264010	La Drôme à Saillans	1131	4.9%	0.0%			
V4414010	Le Roubion à Soyans	186	1.2%	0.5%			
V5004030	L'Ardèche à Meyras [Pont Barutel]	97	1.1%	1.9%			
V5014010	L'Ardèche à Vogüé	617	0.0%	0.4%			
V6035010	Le Toulourenc à Malaucène [Veaux]	157	0.0%	2.1%			x
V7104010	Le Gardon de Saint-Martin à Saint-Étienne-Vallée-Française [Roq.]	29	0.0%	11.7%			
V7105210	Le Gardon de Saint-Germain à Saint-Germain-de-Calberte [Bastide]	32	0.0%	11.2%			
V7115010*	Le Gardon de Sainte-Croix à Gabriac [Pont Ravagers]	52	0.8%	18.2%			
V7124010	Le Gardon de Mialet à Générargues [Roucan]	240	0.0%	1.3%			
V7135010	Le Gardon de Saint-Jean à Corbès [Roc Courbe]	262	0.0%	1.1%			
V7216510	Le Vigueirat à Tarascon [Saint-Gabriel]	247	0.0%	0.1%			x
Y0115410*	La Massane à Argelès-sur-Mer [Mas d'en Tourens]	16	3.6%	18.5%			x
Y0325010*	La Canterrane à Terrats [Moulin d'en Canterrane]	35	0.0%	33.9%			x
Y0624020	L'Agly à Saint-Paul-de-Fenouillet [Clue de la Fou]	225	0.0%	3.0%			x
Y1225010*	Le Lauquet à Greffeil	67	0.0%	19.0%			x
Y1325010*	Le Treboul à Villepinte	137	0.1%	16.5%			x
Y1345010	Le Lampy à Raissac-sur-Lampy	58	2.2%	6.6%			x
Y1415020	L'Orbiel à Bouilhonnac [Villedubert]	239	1.7%	1.2%			x
Y2015010	L'Arre au Vigan [La Terrisse]	157	0.1%	6.1%			
Y2035010	La Vis à Saint-Laurent-le-Minier	309	3.1%	1.1%			
Y2102010	L'Hérault à Laroque	918	3.1%	0.4%			
Y2142010	L'Hérault à Gignac	1432	0.0%	0.1%			
Y2525010+	La Mare au Pradal	115	1.3%	2.0%			x
Y3315080	Le Salaison à Mauguio	55	0.0%	0.8%			x
Y4214010*	La Touloubre à la Barben [La Savonnière]	208	1.1%	15.0%			x
Y4225610	La Cadière à Marignane [stade Saint-Pierre]	76	0.4%	1.0%			x

* Catchment rejected based on missing data >15%

+ Catchment rejected based on obvious incorrect data

B. GR4H Model

The GR4J (modèle du Génie Rural à 4 paramètres Journalier) model uses a daily time step and can be considered as a soil moisture accounting model. It is based on the GR3J model (Edijatno et al., 1999) and was developed empirically, which explains the structure and functions used (see below). The model is considered parsimonious (as advocated by Wagener et al., 2001) since only four parameters are calibrated. This research uses the hourly version of the GR4J model, GR4H. This model is very similar to the daily version and was tested to yield similar performances with only minor changes (Le Moine, 2008). In this appendix, the model is set out and the changes from daily to hourly are highlighted. The description is derived from the one given by Perrin et al. (2003).

B.I. Model description

Figure 3—1 shows a diagram of the GR4 model which uses only the precipitation *P* [mm] and potential evapotranspiration *E* [mm] as input. First E is subtracted from *P* to find the net precipitation P_n [mm] or net evapotranspiration E_n [mm].

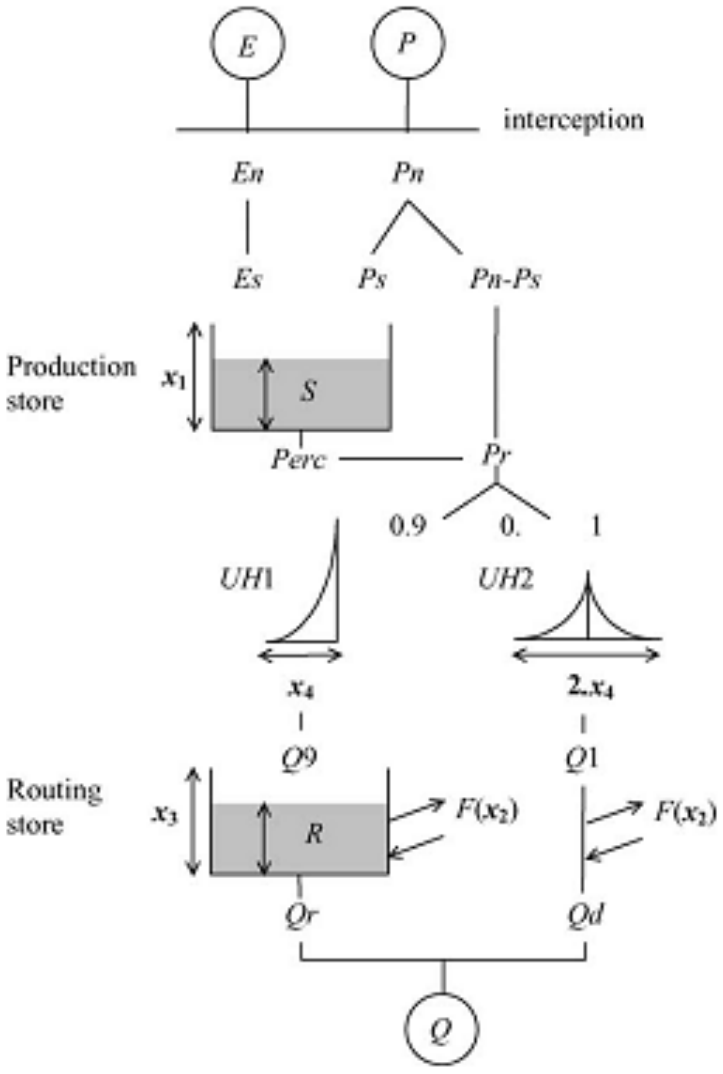

Figure B—1. Diagram of the GR4 model (Perrin et al., 2003).

When there is net precipitation, it partly fills the production store P_s [mm]. The fraction is determined by the level of the production store *S* [mm] and the first parameter to be calibrated; the maximum capacity of the production store x_1 [mm]:

$$P_s = \frac{x_1\left(1-\left(\frac{S}{x_1}\right)^2\right)\tanh\left(\frac{P_n}{x_1}\right)}{1+\left(1-\frac{S}{x_1}\right)\tanh\left(\frac{P_n}{x_1}\right)} \qquad \text{Eq. B—1}$$

In case of non-zero net evapotranspiration, the actual evapotranspiration is determined as a function of the capacity and level in the production store:

$$E_s = \frac{S\left(2-\frac{S}{x_1}\right)\tanh\left(\frac{E_n}{x_1}\right)}{1+\left(1-\frac{S}{x_1}\right)\tanh\left(\frac{E_n}{x_1}\right)} \qquad \text{Eq. B—2}$$

The level in the production store is then updated with:

$$S = S - E_s + P_s \qquad 0 < S < x_1 \qquad \text{Eq. B—3}$$

Then the percolation leakage from the production store *Perc* [mm] can be calculated:

$$Perc = S\left\{1-\left[1+\left(\frac{1}{5}\frac{S}{x_1}\right)^4\right]^{-\frac{1}{4}}\right\} \qquad \text{Eq. B—4}$$

This equation was slightly changed for the hourly version of the model; originally the daily GR4J model uses $\frac{4}{9}$ instead of $\frac{1}{5}$ as coefficient. It was changed because the parameter of the percolation function depends on the time step. Perrin et al. (2003) report that given the power law in Eq. B—4, the percolation does not contribute much to streamflow and is mainly of interest in low flow situations. With the percolation the production store is updated again:

$$S = S - Perc \qquad 0 < S < x_1 \qquad \text{Eq. B—5}$$

The total amount of water P_r [mm] reaching the next stage, is given by:

$$P_r = Perc + (P_n - P_s) \qquad \text{Eq. B—6}$$

There are two unit hydrographs UH1 and UH2 that take 90% and 10% of P_r respectively. UH1 leads to a non-linear routing store. The time base of both hydrographs x_4 [hours] is calibrated. The time base of UH1 equals x_4 and of UH2 it equals twice x_4. The ordinates from both hydrographs correspond to S-curves SH1 and SH2, defined along daily time step *t*:

$$For\ t \leq 0,\ \ SH1(t) = 0 \qquad \text{Eq. B—7}$$

$$For\ 0 < t < x_4,\ \ SH1(t) = \left(\frac{t}{x_4}\right)^{\frac{5}{4}} \qquad \text{Eq. B—8}$$

$$For\ t \geq x_4,\ \ SH1(t) = 1 \qquad \text{Eq. B—9}$$

$$For\ t \leq 0,\ \ SH2(t) = 0 \qquad \text{Eq. B—10}$$

$$For\ 0 < t < x_4,\ \ SH2(t) = \frac{1}{2}\left(\frac{t}{x_4}\right)^{\frac{5}{4}} \qquad \text{Eq. B—11}$$

$$For\ 0 < t < 2x_4,\ \ SH2(t) = 1 - \frac{1}{2}\left(2 - \frac{t}{x_4}\right)^{\frac{5}{4}} \qquad \text{Eq. B—12}$$

$$For\ t \geq 2x_4,\ \ SH2(t) = 1 \qquad \text{Eq. B—13}$$

In Eq. B—8, Eq. B—11 and Eq. B—12 the coefficient $\frac{5}{4}$ was found specifically for the hourly version instead of $\frac{5}{2}$ in the daily version, which corresponds to a more smoothed unit hydrograph. The ordinates of UH1 and UH2 are then calculated (using integer *j*):

$$UH1(j) = SH1(j) - SH1(j-1) \qquad \text{Eq. B—14}$$

$$UH2(j) = SH2(j) - SH2(j-1) \qquad \text{Eq. B—15}$$

A groundwater exchange term *F* [mm] that acts on both flow components is then introduced. It uses the level in the routing store *R* [mm], the routing store reference capacity x_3 [mm] and the water exchange coefficient x_2 [mm]:

$$F = x_2\left(\frac{R}{x_3}\right)^{\frac{7}{2}} \qquad \text{Eq. B—16}$$

The water exchange coefficient can be either negative or positive corresponding to water export or import. As *R* cannot exceed x_3, x_2 equals the maximum value of *F*. The level in the non-linear routing store is updated as followed:

$$R = \max(0; R + Q9 + F) \qquad \text{Eq. B—17}$$

where *Q9* [mm] is the total water output of UH1 at a given day. The outflow of the routing store Q_r [mm/day] is then calculated as:

$$Q_r = R\left\{1 - \left[1 + \left(\frac{R}{x_3}\right)^4\right]^{-\frac{1}{4}}\right\} \qquad \text{Eq. B—18}$$

The level in the routing store is updated as:

$$R = R - Q_r \qquad \text{Eq. B—19}$$

The flow component Q_d [mm/day] is calculated using the total water output of UH2 at a given time step *Q1* [mm] and the water exchange term *F*:

$$Q_d = \max(0; Q1 + F) \qquad \text{Eq. B—20}$$

Finally, the total stream flow *Q* [mm]is calculated as:

$$Q = Q_r + Q_d \qquad \text{Eq. B—21}$$

Table B–1 shows an overview of the symbols used in the model, due to the daily time step and the total discharge in mm/day, the unit off all the internal parameters is mm, except the time base of the unit hydrographs. The four parameters x_1 to x_4 are calibrated while a number of other parameters are left fixed. The values of these fixed parameters are based on the results of a large number of catchments during the development of the model. The used functions are also derived empirically from these data.

Table B—1. Overview symbols in GR4H model.

Model variables	
P	Rainfall depth [mm]
E	Potential evapotranspiration (PE) [mm]
P_n	Net rainfall [mm]
E_n	Net potential evapotranspiration [mm]
P_s	Part of P_n to fill production store [mm]
S	Level in production store [mm]
E_s	Actual evapotranspiration [mm]
Perc	Percolation leakage from production store [mm]
P_r	Total quantity of water reaching the routing function [mm]
UH1	Unit Hydrograph 1 [fraction/hour]
UH2	Unit Hydrograph 2 [fraction/hour]
Q9	Output of UH1 [mm]
Q1	Output of UH2 [mm]
F	Catchment water exchange [mm]
R	Level routing store [mm]
Q_r	Outflow of routing store [mm]
Q_d	Outflow Q1 and F (≥ 0) [mm]
Q	Total streamflow

Calibration parameters	
x_1	Maximum capacity of the production store [mm]
x_2	Groundwater exchange coefficient [mm]
x_3	One day ahead maximum capacity of the routing store [mm]
x_4	Time base of unit hydrograph [hour]

B.II. Calibration

The GR4H model is calibrated with parameters using the techniques described in section 4.1. Realistic values for the GR4J model parameters (so the daily version) are based on the work of Perrin et al. (2003) on a large set of catchments (Table B—2).

Table B—2. Values of calibration parameters for the GR4J model (Edijatno et al., 1999; Perrin et al., 2003).

GR4J	Median Value	80% Confidence Interval
x_1 [mm]	350	100 – 1200
x_2 [mm]	0	-5 – 3
x_3 [mm]	90	20 – 300
x_4 [days]	1.7	1.1 – 2.9

For the hourly version, Le Moine (2008) investigated the transformation of the parameter values. He found clear correlation for the maximum capacity of the production store and the groundwater exchange coefficient, values for both parameters fell in the same range. Correlation between both models for the capacity of the routing store was found when the hourly parameter was multiplied by a factor 2.21. For the time base of the unit hydrograph, no clear correlation could be found. Table B—3 lists the transformed parameter values and the somewhat arbitrary bounds for calibration.

Table B—3. Values of calibration parameters for the GR4H model after Le Moine (2008).

GR4H	Median Value	80% Confidence Interval	Calibration bounds
x_1 [mm]	350	100 – 1200	1 – 20000
x_2 [mm]	0	-5 – 3	-100 – 100
x_3 [mm]	199	44 – 663	1 – 20000
x_4 [hours]	-	-	0.5 – 96

C. SUPERFLEX structures

In the SUPERFLEX approach the structure of the model can be built up using any combination of generic model components. In this research the number of combinations has been limited to twelve structures described in section 3.2. This appendix gives details of all the structures used (section C.I) and the most complex model is explained in full to increase the understanding of how the models work (section C.II).

C.I. Details of all SUPERFLEX structures

The generic modelling components or elements can approximate three functions known in hydrological modelling; a reservoir, a lag-function or a junction element. One or more reservoir elements to represent storage and release of water, lag-function elements to represent the transmission and delay of fluxes and junction elements to represent the splitting, merging and/or rescaling of fluxes can be combined in any way to create a rainfall-runoff model (Fenicia et al., 2011). Some examples of these elements are shown in Figure C—1.

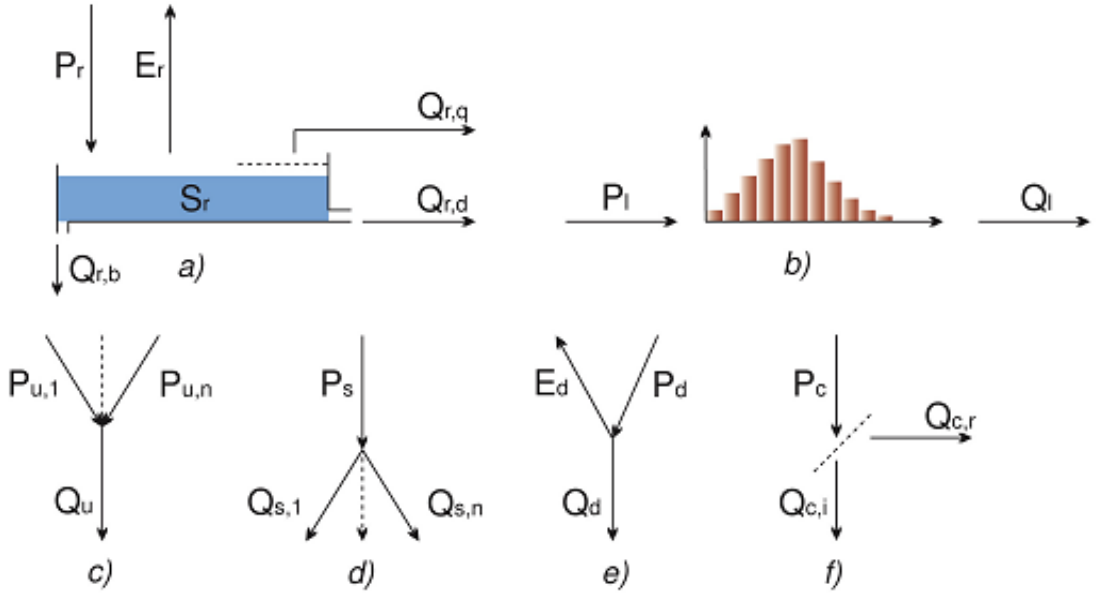

Figure C—1. Generic building block of the flexible framework: (a) generic reservoir and (b) lag-function. Junction elements: (c) union and (d) splitter. Splitters can be used to represent (e) the subtraction of potential evapotranspiration from rainfall and (f) the threshold type occurrence of Hortonian flow. After Fenicia et al. (2011).

The twelve used structures are exact copies of the twelve structures used in Fenicia et al. (2012), for convenience the details of the structures are repeated here (Table C—1 to Table C—5).

Table C—1. Components and parameters of model structures SF01-SF12, "✓" and "-" indicate presence or absence respectively. N_θ is the number of parameters and N_s is the number of states. UR, FR, SR and LF denote the unsaturated, fast, slow reservoirs and lag-function respectively.

Model			Components						Parameters											
	N_s	N_θ	IR	UR	FR	SR	RR	LF	C_e (-)	I_{max} (mm)	$S_{u,max}$ (mm)	β (-)	M (-)	K_r (1/h)	R_{max} (mm/h)	T_f (h)	K_f (mm$^{1-\alpha}$/h)	α (-)	D (-)	K_s (1/h)
SF01	1	3	-	-	✓	-	-	-	✓	-	-	-	-	-	-	-	✓	✓	-	-
SF02	1	4	-	✓	-	-	-	-	✓	-	✓	✓	-	-	✓	-	-	-	-	-
SF03	2	4	-	✓	✓	-	-	-	✓	-	✓	-	-	-	-	-	✓	✓	-	-
SF04	2	5	-	✓	✓	-	-	-	✓	-	✓	✓	-	-	-	-	✓	✓	-	-
SF05	3	6	-	✓	✓	-	-	✓	✓	-	✓	✓	-	-	-	✓	✓	✓	-	-
SF06	4	7	✓	✓	✓	-	-	✓	✓	✓	✓	✓	-	-	-	✓	✓	✓	-	-
SF07	4	8	-	✓	✓		✓	✓	✓	-	✓	✓	✓	✓	-	✓	✓	✓	-	-
SF08	2	4	-	-	✓	✓	-	-	✓	-	-	-	-	-	-	-	✓	-	✓	✓
SF09	3	5	-	✓	✓	✓	-	-	✓	-	✓	-	-	-	-	-	✓	-	✓	✓
SF10	4	6	-	✓	✓	✓	-	✓	✓	-	✓	-	-	-	-	✓	✓	-	✓	✓
SF11	4	7	-	✓	✓	✓	-	✓	✓	-	✓	✓	-	-	-	✓	✓	-	✓	✓
SF12	5	8	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	-	-	-	✓	✓	-	✓	✓

Table C—2. Calibration bounds of SUPERFLEX parameters.

SUPERFLEX	Lower Bound	Upper Bound
C_e [-]	0.1	3
I_{max} [mm]	$1 \cdot 10^{-2}$	10
$S_{u,max}$ [mm]	0.1	$1 \cdot 10^{4}$
β [-]	$1 \cdot 10^{-3}$	10
M [-]	0	0.2
K_r [1/h]	$5 \cdot 10^{-2}$	4
R_{max} [mm/h]	$1 \cdot 10^{-6}$	2
T_f [h]	1	100
K_f [1/h]	$1 \cdot 10^{-3}$	4
α [-]	1	10
D [-]	0	1
K_s [1/h]	$1 \cdot 10^{-7}$	$1 \cdot 10^{-2}$

Table C—3. Water balance equations of the SUPERFLEX models used in this study.

Water balance equations:	SF 01	SF 02	SF 03	SF 04	SF 05	SF 06	SF 07	SF 08	SF 09	SF 10	SF 11	SF 12
$\frac{dS_f}{dt} = P_f - Q_f - E_f$	✓	-	-	-	-	-	-	✓	-	-	-	-
$\frac{dS_f}{dt} = P_f - Q_f$	-	-	✓	✓	-	-	-	-	✓	-	-	-
$\frac{dS_f}{dt} = P_{fl} - Q_f$	-	-	-	-	✓	✓	✓	-	-	✓	✓	✓
$\frac{dS_u}{dt} = P_u - Q_q - Q_u - E_u$	-	✓	-	-	-	-	-	-	-	-	-	-
$\frac{dS_u}{dt} = P_u - Q_q - E_u$	-	-	✓	✓	✓	✓	✓	-	✓	✓	✓	✓
$\frac{dS_s}{dt} = P_s - Q_s$	-	-	-	-	-	-	-	✓	✓	✓	✓	✓
$\frac{dS_i}{dt} = P_t - P_u - E_i$	-	-	-	-	-	✓	-	-	-	-	-	✓
$\frac{dS_r}{dt} = P_r - Q_r$	-	-	-	-	-	-	✓	-	-	-	-	-
$P_t = P_u + P_r$	-	-	-	-	-	-	✓	-	-	-	-	-
$P_t = P_f + P_s$	-	-	-	-	-	-	-	✓	-	-	-	-
$P_t = P_f$	✓	-	-	-	-	-	-	-	-	-	-	-
$P_t = P_u$	-	✓	✓	✓	✓	-	-	-	✓	✓	✓	-
$Q_q = P_f + P_s$	-	-	-	-	-	-	-	-	✓	✓	✓	✓
$Q_t = Q_f$	✓	-	✓	✓	✓	✓	-	-	-	-	-	-
$Q_t = Q_f + Q_r$	-	-	-	-	-	-	✓	-	-	-	-	-
$Q_t = Q_q + Q_u$	-	✓	-	-	-	-	-	-	-	-	-	-
$Q_t = Q_f + Q_s$	-	-	-	-	-	-	-	✓	✓	✓	✓	✓

Table C—4. Constitutive functions of the SUPERFLEX models used in this study. The operator * in the equation for P_{fl} denotes the convolution operator[+].

Constitutive functions	SF 01	SF 02	SF 03	SF 04	SF 05	SF 06	SF 07	SF 08	SF 09	SF 10	SF 11	SF 12
$\bar{S}_i = S_i / S_{i,max}$	-	-	-	-	-	✓	-	-	-	-	-	✓
$P_u = P_t f_h(\bar{S}_i \mid m_1)$	-	-	-	-	-	✓	-	-	-	-	-	✓
$E_i = C_e E_p f_m(\bar{S}_i \mid m_2)$	-	-	-	-	-	✓	-	-	-	-	-	✓
$\bar{S}_u = S_u / S_{u,max}$	-	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓
$Q_q = P_u f_p(\bar{S}_u \mid \beta)$	-	✓	-	✓	✓	✓	✓	-	✓	✓	✓	✓
$Q_q = P_u f_h(\bar{S}_u \mid m_1)$	-	-	✓	-	-	-	-	-	-	-	-	-
$E_u = C_e E_p f_m(\bar{S}_u \mid m_2)$	-	✓	✓	✓	✓	-	✓	-	✓	✓	✓	-
$E_u = C_e E_p \left(1 - f_m(\bar{S}_i \mid m_2)\right) f_m(\bar{S}_u \mid m_2)$	-	-	-	-	-	✓	-	-	-	-	-	✓
$E_f = C_e E_p f_e(S_f \mid m_3)$	✓	-	-	-	-	-	-	✓	-	-	-	-
$P_{fl} = (P_f * h_f)(t)$	-	-	-	-	✓	✓	✓	-	-	✓	-	✓
$h_f = \begin{cases} 2t / T_f^2, & t < T_f \\ 0, & t > T_f \end{cases}$	-	-	-	-	✓	✓	✓	-	-	✓	-	✓
$P_r = MP_t$	-	-	-	-	-	-	✓	-	-	-	-	-
$P_s = DQ_q$	-	-	-	-	-	-	-	-	✓	✓	✓	✓
$P_s = DP_t$	-	-	-	-	-	-	-	✓	-	-	-	-
$Q_u = R_{max} \bar{S}_u$	-	✓	-	-	-	-	-	-	-	-	-	-
$Q_r = k_r S_r$	-	-	-	-	-	-	✓	-	-	-	-	-
$Q_f = k_f S_f$	-	-	-	-	-	-	-	✓	✓	✓	✓	✓
$Q_f = k_f S_f^{\alpha}$	✓	-	✓	✓	✓	✓	✓	-	-	-	-	-
$Q_s = k_s S_s$	-	-	-	-	-	-	-	✓	✓	✓	✓	✓

[+] **Lag-function smoothed using the method in Kavetski and Kuczera (2007).**

Table C—5. Constitutive functions used in the SUPERFLEX structures.

Functions	Name
$f_p(x \mid m) = x^m$	Power function
$f_r(x \mid m) = 1 - (1-x)^m$	"Reflected" power function
$f_m(x \mid m) = (1+m)\dfrac{x}{x+m}$	Monod-type kinetics, adjusted so that $f_m(1 \mid m)=1$
$f_h(x \mid m) = 1 - \dfrac{(1-x)(1+m)}{1-x+m}$	"Reflected" hyperbolic function, scaled to the unit square
$f_e(x \mid m) = 1 - \mathrm{e}^{-x/m}$	Tessier function (note that $f_e(x \mid m) \to 1$ as $x \to \infty$)

C.II. Detailed description of structure SF12

The most complex structure SF12, is shown in Figure C—2. It uses four reservoirs, one lag-function and several junction elements that require a total of eight parameters to be calibrated. Like the GR4H model it uses rainfall and potential evapotranspiration as input to generate discharge as output.

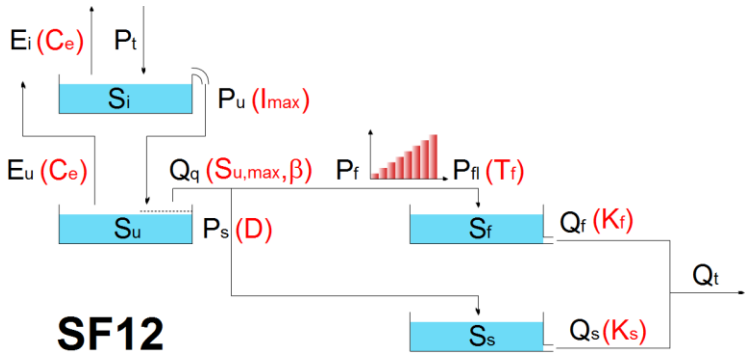

Figure C—2. SUPERFLEX structure SF12 (Fenicia et al., 2012).

Interception reservoir (IR)

The model structure starts by filling an interception reservoir (IR) with the observed rainfall P_t [mm]. From this reservoir evapotranspiration E_I [mm] takes place depending on the potential evapotranspiration E_P [mm], a «calibrated» multiplication factor C_e to correct the potential evapotranspiration, and a constitutive function f_m to relate E_I to the level in IR:

$$E_I = C_e \, E_P \, f_m(\overline{S_I}|m_1) \qquad \text{Eq. C—1}$$

The constitutive function is a Monod-type kinetics adjusted so $f_m(1|m_1) = 1$ and in this case dependent on the relative level in the reservoir $\overline{S_I}$ and m_1:

$$f_m(\overline{S_I}|m_1) = (1 + m_1)\frac{\overline{S_I}}{\overline{S_I} + m_1} \qquad \text{Eq. C—2}$$

This function ensures that there is no evapotranspiration from IR when the reservoir is empty while m_1 smoothes the behaviour. The relative level in IR is determined by the current level in the reservoir S_I [mm] and the maximum level of the reservoir $S_{I,max}$ [mm] «calibrated»:

$$\overline{S_I} = \frac{S_I}{S_{I,max}} \qquad \text{Eq. C—3}$$

The level in IR and P_t then determine how much water enters the unsaturated reservoir (UR), the reservoir representing an unsaturated soil layer. The amount of water entering UR P_u [mm] is determined as follows:

$$P_U = P_t \, f_h(\overline{\overline{S_I}}|m_1) \qquad \text{Eq. C—4}$$

Another constitutive function f_h is used, a reflected hyperbolic function scaled to the unit square:

$$f_h(\overline{S_I}|m_1) = 1 - \frac{(1 - \overline{S_I})(1 + \overline{S_I})}{1 - \overline{S_I} + m_1} \qquad \text{Eq. C—5}$$

This function stays close to zero for most values of $\overline{S_I}$ but rises to one very quickly when $\overline{S_I}$ becomes one, it therefore described a threshold kind behaviour of the interception reservoir.

Unsaturated reservoir (UR)

Like from IR, evapotranspiration E_U [mm] takes place from UR depending on the potential evapotranspiration E_P [mm], the multiplication factor C_e and the relative level in UR $\overline{S_U}$:

$$E_U = C_e\, E_P \left(1 - f_m(\overline{S_I}|m_2)\right) f_m(\overline{S_U}|m_2) \qquad \text{Eq. C—6}$$

Notice that the same constitutive function is used as in Eq. C—2 except with different input and that the amount of water already evaporated from IR is subtracted from E_U. $\overline{S_U}$ is calculated using the current level in UR (S_U [mm]) and the maximum level of UR ($S_{U,max}$ [mm] «calibrated») like in Eq. C—3. From UR, water flows to a slow reservoir (SR) and a fast reservoir (FR). The amount of water flowing from UR (Q_q [mm]) is dependent on the P_U and $\overline{S_U}$ through the use of another constitutive function f_p; a power function with «calibrated» power β:

$$Q_q = P_U\, f_p(\overline{S_U}|\beta) \qquad \text{Eq. C—7}$$

$$f_p(\overline{S_U}|\beta) = \overline{S_U}^{\beta} \qquad \text{Eq. C—8}$$

Q_q is the split between FR and SR by the use of «calibrated» coefficient D, P_F [mm] is sent to FR and P_S [mm] to SR:

$$P_F = (1 - D)\, Q_q \qquad \text{Eq. C—9}$$

$$P_S = D\, Q_q \qquad \text{Eq. C—10}$$

Fast reservoir (FR)

P_F coming from UR does not directly enter FR but is subject to a lag-function h_f resulting in the lagged fraction of water P_{FL} [mm]:

$$P_{FL} = \left(P_F * h_f\right)(t) \qquad \text{Eq. C—11}$$

$$h_f = \begin{cases} 2t/T_f^2, & t < T_f \\ 0, & t > T_f \end{cases} \qquad \text{Eq. C—12}$$

Here $*$ denotes the convolution of P_F and h_f and t the relative time step. This function spreads P_F from each time step over several time steps depending on T_f [mm], the «calibrated» time base of the lag-function. h_f gives a weight to each relative time step that is multiplied by P_F at the first time step j. The integration of h_f over one time step gives the weight given to this time step:

$$h_{f,j+1} = \frac{t_{j+1}^2}{T_f^2} - \frac{t_j^2}{T_f^2} \qquad \text{Eq. C—13}$$

$$P_{FL,j+1} = P_{F,j}\, h_{f,j+1} + P_{FL,i}\, h_{f,i+2} + \ldots \qquad \text{Eq. C—14}$$

Where i denotes the time step before j, depending on the size of T_F there will be more time steps contributing to $P_{FL,j}$. Integration of Eq. C—13 over the entire length T_f gives the combined weight of one, meaning all the water in $P_{F,j}$ is spread over P_{FL} in T_f time steps.

P_{FL} then fills FR, the level in FR is S_F [mm] and together with the «calibrated» retention time for the fast reservoir k_F [hr^{-1}] determines the outflow Q_F [mm/hr]:

$$Q_F = K_F\, S_F \qquad \text{Eq. C—15}$$

Slow reservoir (SR)

The slow reservoir is fed directly by P_S and the outflow Q_S [mm/hr] is determined by the level in the reservoir S_S [mm] and the «calibrated» retention time for the slow reservoir k_S [hr^{-1}]:

$$Q_S = K_S S_S \qquad \text{Eq. C—16}$$

The total outflow or discharge Q_t [mm/hr] is then:

$$Q_t = Q_F + Q_S \qquad \text{Eq. C—17}$$

D. Model performance

This appendix provides additional information on model performance in calibration and the performance of the approaches on CR1-CR4.

D.I. Model performance in calibration

Figure D—1 shows the calibration result of all models on the 237 catchments in boxplots. It shows that all the models perform higher in calibration than in validation. In all cases, increasing complexity stepwise gives a slightly higher performance. Only in case of SF02 the increased complexity (compared to SF01) does not give higher performance because SF02 uses a threshold type of function which apparently is less successful than the power relation used in SF01. This shows that a more complex conceptualisation can lead to lower performance.

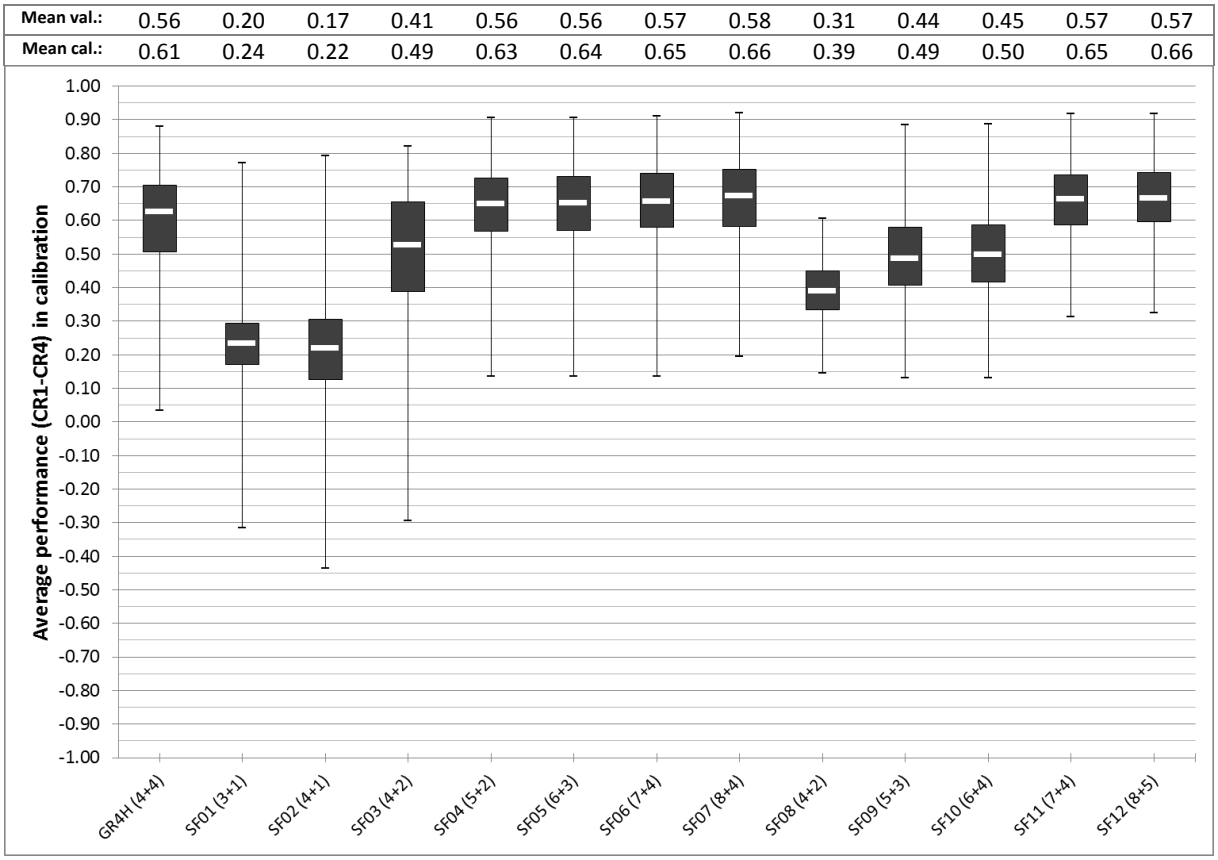

Figure D—1. Boxplots (maximum, 75th percentile, median, 25th percentile and minimum) of CR1-CR4 values obtained by all model structures in validation on the 237 catchments. The x-axis shows the twelve SUPERFLEX structures plus GR4H, the value between brackets denotes the complexity measure nr. of calibrated parameters + nr. of states. At the top of the figure the mean values for model performance in validation and calibration are given.

D.II. Performance of approaches on CR1-CR4

Figure D—2 shows the performance of the fixed and the flexible approach plus the distribution when both approaches are combined on the four criteria. These figures show that like in average performance for all models, scores on CR2 are low and on CR3 high compared to the others. The poor performance on low flow (CR2) is linked to general difficulties with simulating low flow and possible favouring of high flow during calibration. High water balance scores (CR3) shows that the water balance is relatively easy to simulate. All model structures contain specific parameter(s) to adapt for errors in the water balance. In all criteria the fixed approach performs better, but with less consistent catchments. The combination shows that the flexible approach performs poor on the inconsistent catchments for the fixed approach.

For cases where the water balance scores low, the Relative Volume Error (RVE) was calculated to make interpretation of CR3 easier. When CR3 is at 0.9 RVE is around 5%, when CR3=0.8 -> RVE=10%, CR3=0.4 -> RVE=35% and CR3=0.3 -> RVE=40%. CR3 stretches small changes in RVE while with larger errors it drops less. This rescaling was necessary to make all four criteria comparable.

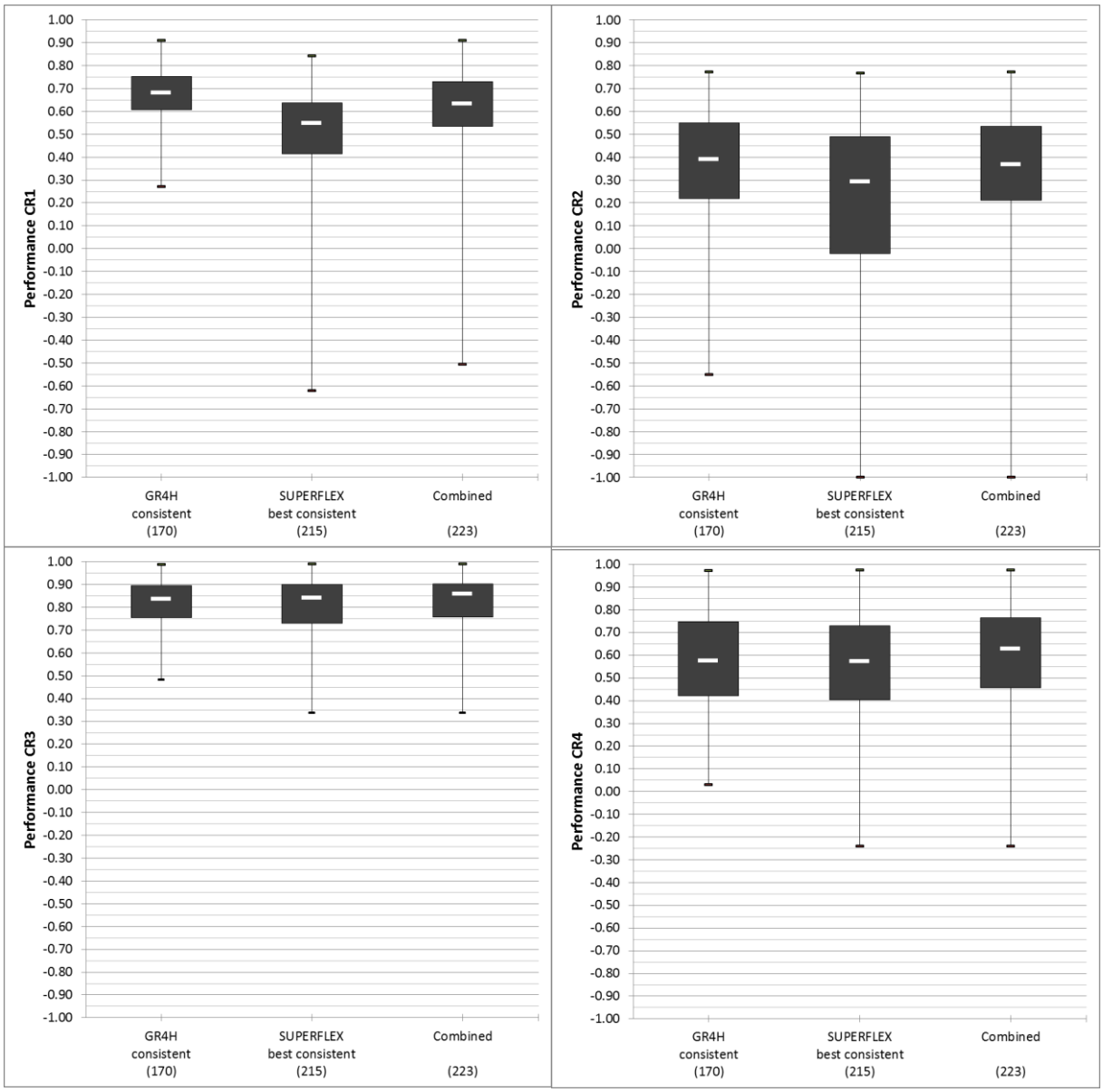

Figure D—2. Distribution of performance on CR1-CR4 of the fixed, flexible and a combined approach.

E. Hydrological monsters

This appendix contains descriptions of the performance of the modelling approaches on the monster catchments (E.I). It shows the difference in rainfall between the wet and the dry years and shows an example of the GR4H model where the roles of the stores are switched (E.II). Finally, lists of the monster catchments per group are given (E.III).

E.I. Performance of monster catchments

Figure E—1 shows the average performance of the modelling approaches on the selected monster catchments. The fixed GR4H model performs generally better than the flexible SUPERFLEX structures. However, the flexible approach is consistent on more catchments and the performance on these catchments is generally low. Table E—1 also shows the average performance on CR1-CR4 of the models on the monster catchments. These results show that CR2 (low flow) is the lowest scoring criterion for the monster catchments across all approaches and that CR3 (water balance) generally scores high. This is similar as observed in average performance. The robustness remains difficult to interpret, since monster catchments are not really less robust than all catchments together.

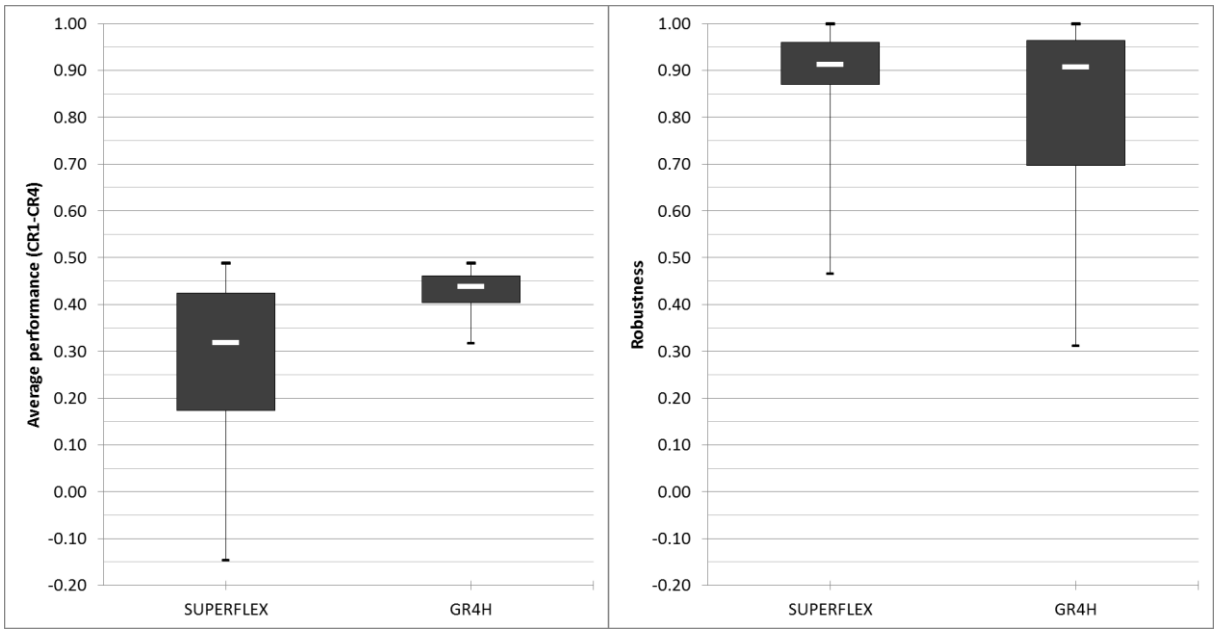

Figure E—1. Average performance (left) and robustness for the fixed GR4H model and the flexible SUPERFLEX approach on the monster catchments.

Table E—1. Average performance and robustness of both approaches on the monster catchments. Monster catchments are grouped according to the consistency of the opposing approach.

		Average performance	Robustness	CR1 (high flow)	CR2 (low flow)	CR3 (water balance)	CR4 (variability)
SF	GR4H inconsistent (29)	0.27	0.88	0.31	-0.34	0.70	0.40
	GR4H consistent (20)	0.30	0.88	0.31	-0.26	0.68	0.45
GR4H	SUPERFLEX inconsistent (6)	0.40	0.69	0.48	0.11	0.64	0.38
	SUPERFLEX consistent (20)	0.43	0.87	0.60	0.11	0.69	0.34

Table E—2 shows the performance of the monster catchments in four groups. The groups are based on the likely reasons for the poor performance of the approaches. For catchments with a strong base flow pattern (BFP), many models fail and the performance of the remaining models is very low. In the other groups SUPERFLEX finds a working model on more catchments but with low performance, as observed in the previous section.

Table E—2 Performance of monster catchments divided into four groups. Note that the groups are not of equal size and inconsistent have no score on the criteria.

		Inter-annual base flow pattern (13)	Climatic differences (24)	Flashy Flow (18)	Poor Data (14)
SUPERFLEX	**Performance**	0.18	0.31	0.26	0.38
	Robustness	0.9	0.87	0.91	0.88
	CR1	0.22	0.38	0.24	0.39
	CR2	-0.65	-0.28	-0.19	-0.12
	CR3	0.8	0.73	0.61	0.77
	CR4	0.34	0.42	0.37	0.48
	# Inconsistent	11	2	5	1
GR4H	**Performance**	0.4	0.4	0.44	0.48
	Robustness	0.43	0.82	0.89	0.69
	CR1	0.38	0.59	0.55	0.72
	CR2	0.01	0.03	0.25	-0.18
	CR3	0.55	0.65	0.68	0.62
	CR4	0.65	0.33	0.28	0.75
	# Inconsistent	11	12	13	7

E.II. Wet and dry years

Figure E—2 shows the difference between wet and dry years in different catchment groups. The difference between wet and dry years is larger in catchments that are selected as monsters and their neighbours. Note that the wet years are situated at the end of the first calibration period (1997-2001) and that this has a large impact on the difference in calibration conditions between the two periods. This high difference partially explains the difficulty of the models to select a consistent parameter set for the monster catchments. However, in the neighbouring catchments the differences in rainfall is very high as well. Closer examination of this group shows that in many of these catchments the GR4H model does fail but complex SUPERFLEX structures yield relatively high performance.

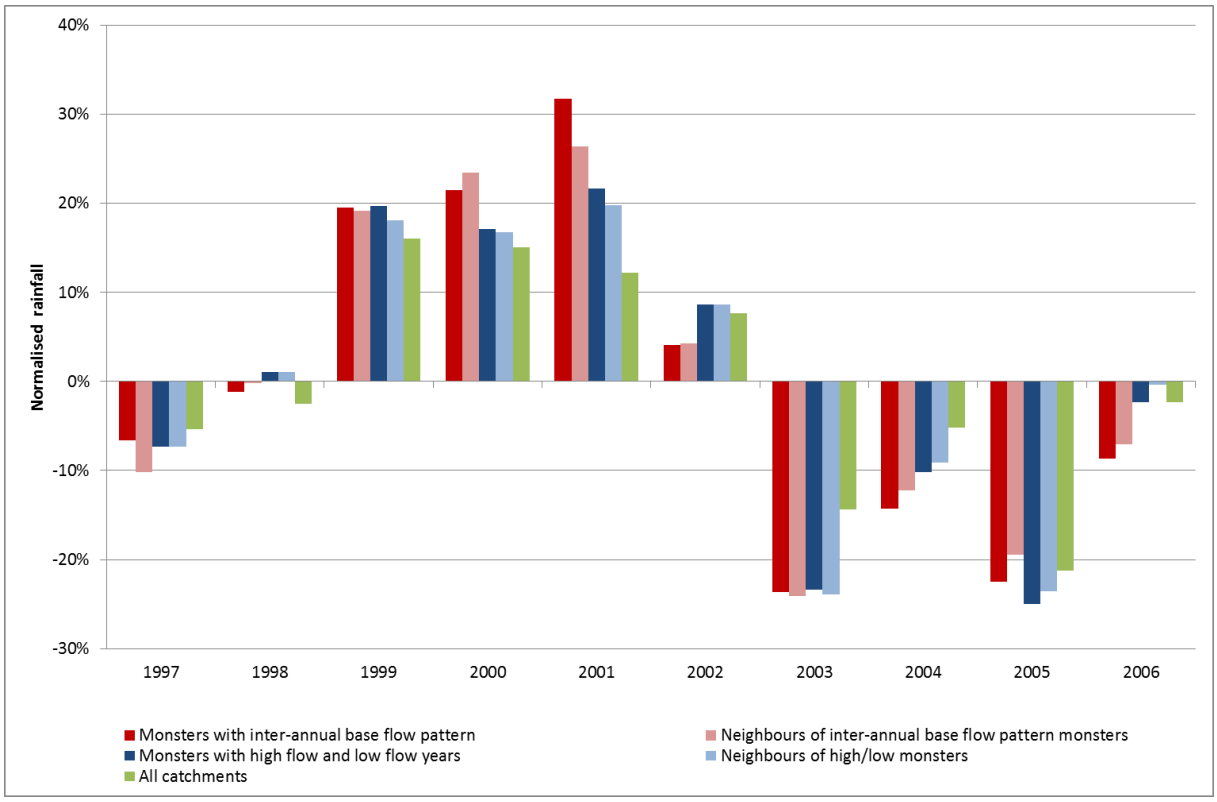

Figure E—2. Wet and dry years in monster catchments, their neighbours and all catchments together. Yearly rainfall was normalised with the 10-year average, i.e. rainfall of 20% means 20% more rain in this year than average over the 10 years of available data.

In the main text, catchment H8042010 –Epte at Fourges – 1386 km² is shown as an example. Here, the GR4H model gives two very different simulated hydrographs for the two calibration periods as a result of the large differences in calibration conditions. The function of the two stores in GR4H, the production and the routing store, appear to have switched role.

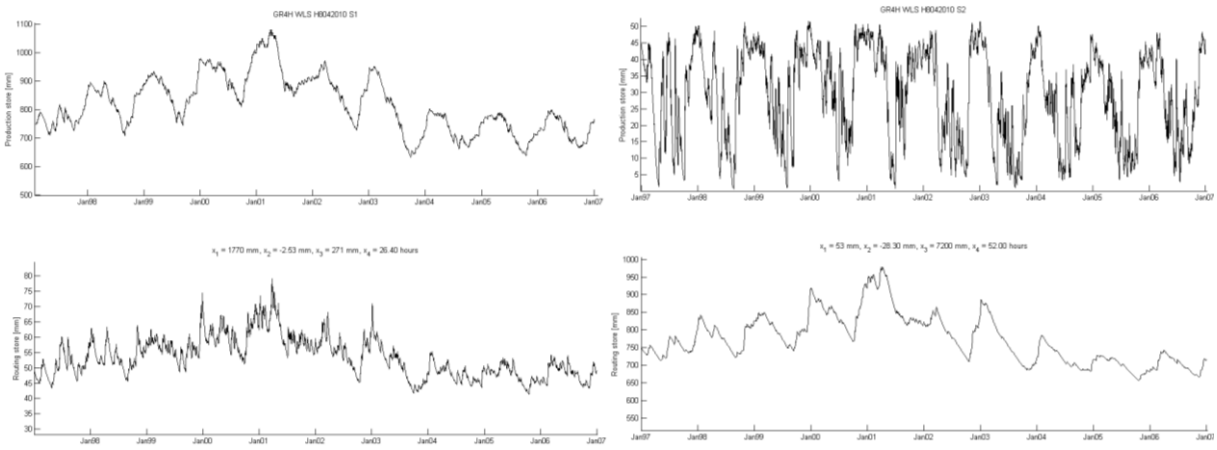

Figure E—3. Simulated level in the production store (upper two) and routing store (lower two) of the GR4H model of catchment H8042010 – L'Epte at Fourges – 1386 km². On the left the results of the calibration on period 1 (wet) and on the right the results of the calibration on period 2 (dry).

Figure E—3 shows the levels in the production and routing store of example catchment H80421010 calibrated on period 1 (S1) and period 2 (S2). The capacity of the production store in S1 is very large while in S2 it is very small, for the routing store the capacity is small in S1 and large in S2. This leads to simulation of flow components (direct response vs. slow base flow processes) by different stores in S1 and S2. This can be observed in the level of the stores, the function of the stores is switched between the periods.

E.III. Lists of monster catchments

Table E—3 to Table E—6 show the results of the individual monsters per group. The tables show the catchments description and the selection of the different models with the performance and robustness.

Table E—3. Model results of monster catchments where severe climatic differences between calibration and validation periods have led to inter-annual base flow patterns.

Code	Name	Area [km^2]	SUPERFLEX			GR4H			Best
			Structure	Performance	Robustness	GR4H?	Performance	Robustness	
E3518510	La Laquette à Witternesse	81	SF01	0.18	0.91				SF01
E6406010	L'Avre à Moreuil	621							
E6426010	La Selle à Plachy-Buyon	546							
H1713010	L'Ardusson à Saint-Aubin	158				GR4H	0.4	0.54	GR4H
H3613020	Le Lunain à Épisy	247							
H4022020	L'Essonne à Guigneville-sur-Essonne [La Mothe]	850							
H4022030	L'Essonne à Boulancourt	585							
H4042010	L'Essonne à Ballancourt-sur-Essonne	1858							
H5732010	Le Grand Morin à Pommeuse	762	SF02	0.18	0.9				SF02
H7602010	La Brêche à Nogent-sur-Oise	463				GR4H	0.4	0.31	GR4H
H7702010	Le Thérain à Bonnières	207							
H8042010	L'Epte à Fourges	1386							
H8043310	L'Aubette de Magny à Ambleville	101							

Table E—4. Remaining model results of monster catchments with severe climatic differences between calibration and validation periods.

Code	Name	Area [km^2]	SUPERFLEX			GR4H			Best
			Structure	Performance	Robustness	GR4H?	Performance	Robustness	
A5431010	Le Madon à Pulligny	950	SF03	0.32	0.75	GR4H	0.45	0.98	GR4H
A7642010	La Petite Seille à Château-Salins	153	SF09	0.46	0.91				SF09
A8322010	Le Woigot à Briey	71	SF03	0.38	0.84	GR4H	0.35	0.94	SF03
A8732010	La Canner à Koenigsmacker	109							

Code	Name	Area [km^2]	SUPERFLEX			GR4H			Best
			Structure	Performance	Robustness	GR4H?	Performance	Robustness	
A9021010	La Sarre à Sarrebourg	307	SF01	0.21	0.89				SF01
A9091060	La Sarre à Diedendorf	737	SF04	0.43	0.97				SF04
A9352050	L'Eichel à Oermingen	280	SF01	0.11	0.92				SF01
A9832010	La Nied Allemande à Faulquemont	203	SF01	0.17	0.71				SF01
B1092010	Le Mouzon à Circourt-sur-Mouzon [Villars]	401	SF03	0.4	0.96	GR4H	0.45	0.96	GR4H
B1322010	Le Vair à Belmont-sur-Vair	139	SF03	0.25	0.6	GR4H	0.33	0.68	GR4H
B2042010	L'Aroffe à Vannes-le-Châtel	197	SF09	0.16	0.89				SF09
H2073110	Le Sauzay à Corvol-l'Orgueilleux	89	SF09	0.42	0.91				SF09
H4322030	L'Yerres à Courtomer [Paradis]	426				GR4H	0.32	0.79	GR4H
H4333410	Le Réveillon à Férolles-Attilly [La Jonchère]	56	SF09	0.42	0.97				SF09
H5702010	Le Grand Morin à Meilleray	350	SF09	0.47	0.88				SF09
I0011010	La Risle à Rai	143	SF09	0.44	0.91				SF09
I0102010	La Charentonne à Bocquencé	67	SF04	0.45	0.92	GR4H	0.4	0.92	SF04
J7633010	Le Semnon à Bain-de-Bretagne [Rochereuil]	415	SF01	0.17	0.78	GR4H	0.44	0.79	GR4H
K7414010	La Tourmente à Villeloin-Coulangé [Coulangé]	106	SF05	0.24	0.91				SF05
M1041610	Le Loir à Saint-Maur-sur-le-Loir	1085	SF05	0.48	0.92	GR4H	0.33	0.97	SF05
M3711810	L'Oudon à Cossé-le-Vivien	135	SF01	0.3	0.93	GR4H	0.48	0.74	GR4H
M3771810	L'Oudon à Châtelais [Marcillé]	732	SF01	0.32	0.85	GR4H	0.42	0.72	GR4H
M3774010	Le Chéran à la Boissière	78	SF01	0.14	0.74	GR4H	0.42	0.68	GR4H
M6333020	L'Erdre à Nort-sur-Erdre [Moulin de Vault]	457	SF01	0.13	0.93	GR4H	0.41	0.68	GR4H

Table E—5. Model results of flashy flow monsters.

Code	Name	Area [km^2]	SUPERFLEX			GR4H			Best
			Structure	Performance	Robustness	GR4H?	Performance	Robustness	
K2821910	La Dore à Dore-l'Église	106	SF03	0.48	0.91	GR4H	0.49	0.91	SF03
K7777777	La Loire à Ardentes [station test bidon]	687	SF02	0.04	0.9	GR4H	0.46	1	GR4H
L4653010	La Bouzanne à Velles [Forges]	438	SF03	0.25	0.93	GR4H	0.46	0.94	GR4H
M5214020	L'Hyrome à Saint-Lambert-du-Lattay [Chauveau]	153	SF04	0.41	0.98	GR4H	0.42	0.9	SF04
O2215010	La Saune à Quint-Fonsegrives	109				GR4H	0.44	0.91	GR4H
O5754020	La Vère à Bruniquel [La Gauterie]	314	SF01	0.11	1	GR4H	0.48	0.98	GR4H
O8584010	La Lède à Casseneuil	413				GR4H	0.46	0.94	GR4H
O9685310	La Pimpine à Cénac	52	SF04	0.34	0.95				SF04
O9785310	La Jalle de Ludon au Pian-Médoc	32				GR4H	0.41	0.65	GR4H
P2484010	Le Céou à Saint-Cybranet	573	SF01	0.15	0.98	GR4H	0.45	1	GR4H
P5715010	L'Engranne à Baigneaux	28	SF01	-0.15	0.47				SF01
V4225010	Le Bez à Châtillon-en-Diois	225	SF09	0.41	0.99				SF09
V4264010	La Drôme à Saillans	1131	SF01	0.19	0.91				SF01
V4414010	Le Roubion à Soyans	186	SF09	0.32	0.97	GR4H	0.46	0.99	GR4H
V6035010	Le Toulourenc à Malaucène [Veaux]	157							
V7124010	Le Gardon de Mialet à Générargues [Roucan]	240	SF09	0.32	0.98	GR4H	0.47	0.99	GR4H
V7135010	Le Gardon de Saint-Jean à Corbès [Roc Courbe]	262	SF01	0.27	0.94				SF01
V7216510	Le Vigueirat à Tarascon [Saint-Gabriel]	247	SF08	0.39	0.96				SF08
Y0624020	L'Agly à Saint-Paul-de-Fenouillet [Clue de la Fou]	225							
Y1345010	Le Lampy à Raissac-sur-Lampy	58	SF01	0.1	0.9				SF01
Y1415020	L'Orbiel à Bouilhonnac [Villedubert]	239	SF02	0.21	0.92				SF02

Code	Name	Area [km²]	SUPERFLEX			GR4H			Best
			Structure	Performance	Robustness	GR4H?	Performance	Robustness	
Y2035010	La Vis à Saint-Laurent-le-Minier	309							
Y2102010	L'Hérault à Laroque	918	SF03	0.47	0.99				SF03
Y2142010	L'Hérault à Gignac	1432	SF08	0.46	1				SF08
Y3315080	Le Salaison à Mauguio	55	SF03	0.04	0.79				SF03

Table E—6. Model results of catchments with disturbances on observed data.

Code	Name	Area [km²]	SUPERFLEX			GR4H			Best
			Structure	Performance	Robustness	GR4H?	Performance	Robustness	
A3422010	La Zorn à Saverne [Schinderthal]	183	SF09	0.49	0.84				SF09
H4202020	L'Orge à Saint-Chéron [Saint-Évroult]	111							
H4223110	La Remarde à Saint-Cyr-sous-Dourdan	151	SF07	0.36	0.87				SF07
I2213610	L'Ancre à Cricqueville-en-Auge	60	SF04	0.4	0.92				SF04
K0614010	Le Furan à Andrézieux-Bouthéon	178	SF08	0.45	0.99				SF08
K2514010	L'Allanche à Joursac [Pont du Vernet]	156	SF01	0.08	0.74				SF01
K2821910	La Dore à Dore-l'Église	106	SF03	0.48	0.91	GR4H	0.49	0.91	SF03
K7777777	La Loire à Ardentes [station test bidon]	687	SF02	0.04	0.9	GR4H	0.46	1	GR4H
L4653010	La Bouzanne à Velles [Forges]	438	SF03	0.25	0.93	GR4H	0.46	0.94	GR4H
O5754020	La Vère à Bruniquel [La Gauterie]	314	SF01	0.11	1	GR4H	0.48	0.98	GR4H
O8584010	La Lède à Casseneuil	413				GR4H	0.46	0.94	GR4H
P2484010	Le Céou à Saint-Cybranet	573	SF01	0.15	0.98	GR4H	0.45	1	GR4H
P7261510	L'Isle à Abzac	3758	SF04	0.48	0.83	GR4H	0.48	0.69	SF04
V7216510	Le Vigueirat à Tarascon [Saint-Gabriel]	247	SF08	0.39	0.96				SF08