Imperial College London
Department of Civil and Environmental Engineering

Exploring data mining for hydrological modelling

Claudia Vitolo

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy from Imperial College London
2015

Abstract

Technological advances in computer science, namely cloud computing and data mining, are reshaping the way the world looks at data. Data are becoming the drivers of discoveries and strategic developments. In environmental sciences, for instance, big volumes of information are produced by monitoring networks, satellites and model simulations and are processed to uncover hidden patterns, correlations and trends to, ultimately, support policy and decision making.

Hydrologists, in particular, use models to simulate river discharges and estimate the concentration of pollutants as well as the risk of floods and droughts. The very first step of any hydrological modelling exercise consists of selecting an appropriate model. However, the choice is often made by the modeller based on his/her expertise rather than on the model's suitability to reproduce the most important processes for the area under study. Since this approach defeats the "scientific method" for its lack of reproducibility and consistency across experts as well as locations, a shift towards a data-driven selection process is deemed necessary.

This work presents the design, development and testing results of a completely novel data mining algorithm, called AMCA, able to automatically identify the most suitable model configurations for a given catchment, using minimum data requirements and an inventory of model structures. In the design phase a transdisciplinary approach was adopted, borrowing techniques from the fields of machine learning, signal processing and marketing.

The algorithm was tested on the Severn at Plynlimon flume catchment, in the Plynlimon study area (Wales, UK). This area was selected because of its reliable measurements and the homogeneity of its soils and vegetation. The Framework for Understanding Structural Errors (FUSE) was used as sample model inventory, but the methodology can easily be adapted to others, including more sophisticated model structures.

The model configuration problem, that the AMCA attempts to solve, can be categorised as "fully unsupervised" if there is no prior knowledge of interactions and relationships amongst observed data at a certain location and available model structures and parameters. Therefore, the first set of tests was run on a synthetic dataset to evaluate the algorithm's performance against known

outcomes. Most of the component of the synthetic model structure were clearly identified by the AMCA, which allowed to proceed with further testing using observed data.

Using real observations, the AMCA efficiently selected the most suitable model structures and, when coupled with association rule mining techniques, could also identify optimal parameter ranges. The performance of the ensemble suggested by the combination of AMCA and association rules was calibrated and validated against four widely used models (Topmodel, ARNOVIC, PRMS and Sacramento). The ensemble configuration always returned the best average efficiency, characterised by the narrowest spread and, therefore, lowest uncertainty.

As final application, the full set of FUSE models was used to predict the effect of land use changes on catchment flows. The predictive uncertainty improved significantly when the prior distributions of model structures and parameters were conditioned using the AMCA approach. It was also noticed that such improvement is due to constrains applied to both model and parameter space, however the parameter space seems to contribute more.

These results confirm that a considerable part of the uncertainty in prediction is due to the definition of the prior choice of the model configuration and that more objective ways to constrain the prior using formal data-driven techniques are needed. AMCA is, however, a procedure that can only be applied to gauged catchment. Future experiments could test whether AMCA configurations could be regionalised or transferred to ungauged catchments on the basis of catchment characteristics.

To my husband Saber

Declaration of Originality

The work presented in this thesis is my own except where otherwise acknowledged.

Material from Chapter 5 regarding the data mining of hydrological model performances was presented at the EGU General Assembly 2013, in Vienna, Austria.

Copyright Declaration

Acknowledgements

I would like to express my sincere gratitude to my current and former supervisors: Dr. Wouter Buytaert and Prof. Neil McIntyre. I have been greatly inspired by their scientific enthusiasm and I cannot thank them enough for giving me the opportunity to carry out my research and providing the framework and resources to develop and bring forward the ideas underlying this thesis. I thank Dr. Nataliya Le Vine for co-supervising the last part of this work, for the fruitful discussions and for giving me a hand taking my first steps towards Bayesian thinking. Particular thanks belong to Dr. Adrian Butler and Dr. Gwyn Rees, my examiners, for reviewing and providing thoughtful comments on the manuscript.

I also thank Hilary McMillan and Martyn Clark for answering my numerous emails on the FUSE work, Dominik Reusser for the brainstorming that helped me improve the FUSE code and the data mining algorithm. Thanks also to Matthew Fry and the Centre for Ecology and Hydrology for making Plynlimon data available and Simon Burbridge for his support using the High Performance Computing service at Imperial College, without which this work could not have been carried out.

I also thank all colleagues and friends at the Environmental and Water Resources Section of the Department of Civil and Environmental Engineering for providing such a motivating atmosphere that has brought a great deal of critical considerations. Particularly, I want to thank Zed Zulkafli, Susana Ochoa, Lipen Wang, Bastian Manz, Simon Parker and Barbara Orellana Bobadilla.

Finally, I would like to thank my husband Saber, my parents Edoardo and Franca, my sister Valentina, my brothers Carlo and Fabio, and my friend Nicole Kalas for having supported and encouraged me during these years.

Acronyms

A AMCA result space. 62

AI Artificial Intelligence. 3

AMCA Automatic Model Configuration Algorithm. 6

DM Data Mining. 3

DTW Dynamic Time Warping. 67

I Array containing performance indices. 62

I' Array I after preliminary model selection. 64

I" Array I after Pareto filtering. 65

I"' Array I after redundancy reduction. 68

IE Initial Ensemble. 62

IR Information Retrieval. 3

ML Machine Learning. 3

MPI Model Performance Indices. 60

NR Net Radiation. 41

P Precipitation. 41

RE Reduced Ensemble. 68

SOM Self Organising Maps. 65

SR Solar Radiation. 41

T Array containing simulated Time series. 62

T” Array T after Pareto frontier. 65

T”’ Array T after redundancy reduction. 68

TD Temperature Dry bulb. 41

TW Temperature Wet bulb. 41

WD Wind Direction. 41

WS Wind Speed. 41

Glossary

Algorithm A set of steps to accomplish a task. xi, 19

Calibration This is an operation that aims to find the optimal model parameters in a multidimensional space. It can be carried out at random or optimised employing combinatorial, evolutionary or stochastic methods. 12

Data Mining Process of designing the sequence and mode of combining machine learning techniques in order to discover information from a large amount of data. 52, 57, 62, 112

Machine Learning Systems [or algorithms] to automate decision making and classification of data. xi, 20

Supervised Learning Algorithm used to make predictions based on a set of examples. 20

Unsupervised Learning Algorithm used to explore possible interactions amongst the features of the dataset. 20

Workflow Computer program in which multiple algorithms are used in a modular fashion. xi, 19

Contents

List of Tables

List of Figures

Chapter 1

Introduction

River hydrology is a complex science that uses physical, empirical or statistical models to learn hydrological processes occurring in a river catchment and infer predictions in time and space. Hydrological modelling has a multitude of applications, from hazard risk management (e.g. flood and droughts) to water quality planning (e.g. European Water Framework Directive[1]) and water supply and sanitation[2] programmes. Regardless of the specific objectives, modelling is an important tool to support policy and decision making because it allows to understand and compare the effects of policy options and assess consequences on population, properties and infrastructures.

In hydrology, modelling consists of an analysis of local climate data (such as precipitation and potential evapotranspiration) and their transformation into streamflow discharge using a set of equations, also called "hydrological model". Depending on the phenomena taken into consideration the equations involved may vary, changing the so called "model structure". Over the last two centuries a wide variety of models have been implemented, some of which are designed to fit particular climatic-hydrologic behaviours. As an example, TOPMODEL is a physically based distributed watershed model designed by Beven and Kirkby (1979) to simu-

[1] `http://ec.europa.eu/environment/water/water-framework/`

[2] European Water Supply and Sanitation Technology Platform: `http://wsstp.eu/`

late the rainfall-discharge relationship in mountainous catchments under humid climates where surface runoff is mainly generated by surface saturation.

The different steps of the modelling process are formally described by Beven (Figure 1.2, 2001*b*). Indeed, Beven's modelling process has become, over the years, an important reference for hydrologists and practitioners but relies heavily on the role of the local expert. This is usually the modeller, who identifies relevant hydrological processes and translates this perception of the reality in a mathematical model and then into a computer code that can be used for generating simulations.

Although the expert's role is fundamentally important, especially when there is high uncertainty due to insufficient data, it also makes the modelling exercise inherently subjective because the modelling strategy is influenced by the individual training and previous modelling experience. Considering these limitations, Beven (2007) introduces the need for more evidence-based modelling approaches. He envisages the use of a "modelling system" in which hydrological predictions are made taking into account the relationship among process representations, site characterisations and boundary conditions. The implementation of such a system, called "Environmental Models of Everywhere", could theoretically allow more consistent and reproducible modelling, but has never been implemented in practice.

This research project presents a way to implement Beven's modelling procedure reducing subjectivity and hence improving reproducibility. The main idea is that, with the advent of new technologies and mathematical methods (e.g. as cloud computing and data mining) the role of the local expert could be replaced by an instance of artificial intelligence, such as an algorithm. In other words, modellers could avoid to make a subjective model selection by using all the model structures available and simulate all the possible hydrological behaviours. Data mining of model performances should, then, allow to discern between suitable and unsuitable configurations.

1.1 Problem statement

In computational sciences it is important to rely on reproducible workflows to be able to audit, examine and review suggested procedures. Over the years, reproducibility has become a major concern for managers and policy makers committed to consistent and transparent modelling applications. Reproducibility should span from the discovery of relevant data, through the selection and use of models and algorithms (including pre/post-processing) to communication and dissemination of results.

Data-driven reproducible protocols are needed to extract evidence from data and dictate the most suitable choices amongst those available to the modeller. Various branches of mathematics and computer science focus on the automatic extraction of trends and patterns from data, examples are: Information Retrieval (IR), Machine Learning (ML) and Artificial Intelligence (AI). These disciplines provide a plethora of tools and techniques for supervised (e.g. classification) and unsupervised learning (e.g. clustering, similarity search and pattern matching). Designing the sequence and mode of combining these techniques together, in order to discover information from a large amount of data, is generally called a Data Mining (DM) procedure.

In the hydrology domain, reproducibility is often compromised by the widespread use of expert elicitation. The model selection problem, for instance, would certainly benefit from a less subjective approach to the analysis. Currently, there are relatively few attempts to define DM procedures relevant for hydrological purposes and it is deemed to be plenty of scope for researchers to design new ones.

Based on the problems mentioned above, the following research objectives are identified:

- To design simulation experiments and data mining procedures to reveal trends and patterns leading to a more objective model configuration process.
- To explore how model components and parameters are expected to interact based on evidence from data, and consequently improve reliability of predictions.

- To extend findings to non-stationary conditions and data-poor environments, exploring the relation between model configuration and time-dependent catchment characteristics so that modellers can learn and infer information on hydrological processes.

1.2 Research questions

The primary objective of this thesis is to answer the following research questions:

1. Can a data mining procedure be designed for automatic model configuration? How can its performance be determined?

2. What determines an optimal model configuration? Is it the interaction amongst model components and/or parameters? Can this improve parameter identifiability and predictions and reduce uncertainty?

3. When land use changes occur, how can model configurations be adapted accordingly? How to condition model predictions for data-rich as well as data-poor environments?

1.3 Contributions

This dissertation contributes to several aspects of environmental and hydrological modelling, data science and uncertainty estimation using a trans-disciplinary approach. It explores the limitations of current modelling tools for hydrologists and borrows techniques from data science, signal processing and marketing to trigger novel thinking.

It explores the use of these techniques to make the modelling exercise less subjective and more transparent and reproducible. This is done by proposing a novel data mining procedure for automatic configuration of hydrological models. The proposed algorithms, based mainly on unsupervised machine learning techniques, can be used as learning tool to uncover non-trivial

catchment/processes information. Its plotting functionalities provide a systematic approach to summarise multi-dimensional model results while exploring data patterns through parameter and model similarities and interactions. Finally, it shows how to combine various sources of information to understand whether and how anthropogenic changes to the environment can be detected by model simulations and re-used as what-if scenario tools. The Table 1.1 maps the objectives to the specific research questions, case studies, main contributions and chapter references.

1.4 Development effort

The work carried out entails the development of computer code to automate tasks. Amongst the different programming languages available, I have chosen to use a combination of R and Bash because of their suitability and the excellent support of the open source communities. R is used for data discovery, data preparation, data mining and visualisation. Bash is used for string manipulation, and the automatic dispatch of jobs on High Performance Computing systems for basic process parallelisation. Computer specifications of the system used to develop and test the abovementioned code are described in appendix A.

1.5 Dissemination

As a part of this thesis, various attempts have been made to disseminate the work done within the community of hydrologists. All the contributions have been implemented as computer code and made available through a public repository under an open licence. For the major libraries, web pages have been set up to facilitate discovery, retrieval and installation. All the libraries are accompanied by in-depth code documentation, video demonstrations, tutorials and online surveys to collect user feedbacks. The release of the information on libraries and any development update is advertised in real-time via a dedicated social network community which already

Table 1.1: Mapping objectives to specific research questions and contributions.

Chapter	Objectives	Research questions	Case studies	Contributions
5	To develop tools for the analysis of a wide spectrum of model simulations to uncover trends and patterns that may lead to a more objective model configuration.	Can a data mining procedure be designed for automatic model configuration? How can its performance be determined?	Synthetically generated datasets	Development of an Automatic Model Configuration Algorithm (AMCA), implemented as an R package.

Continued on next page

Table1.1 Mapping objectives to specific research questions and contributions (continued from previous page).

Chapter	Objectives	Research questions	Case studies	Contributions
6	To explore how model components and parameters are expected to interact based on evidence from data, and consequently improve reliability of predictions.	What determines an optimal model configuration? Is it the interaction amongst model components and/or parameters? Can this improve parameter identifiability and predictions and reduce uncertainty?	Plynlimon (UK)	Guidelines to identify how model components and parameters interact.
				Continued on next page

Table1.1 Mapping objectives to specific research questions and contributions (continued from previous page).

Chapter	Objectives	Research questions	Case studies	Contributions
7	To extend findings to non-stationary conditions and data-poor environments, exploring the relation between model configuration and time-dependent catchment characteristics so that modellers can learn and infer information on hydrological processes.	When land use changes occur, how can model configurations be adapted accordingly? How to condition model predictions for data-rich as well as data-poor environments?	Plynlimon (UK)	Guidelines to understand whether and how anthropogenic changes to the environment (e.g. deforestation) can be detected by model simulations.

counts dozens of members.

1.6 Structure of the thesis

Chapter 2 provides the intellectual context for the research, reviewing the state-of-the-art in hydrological modelling and flexible model formulation, algorithm design and data mining techniques. This is vital to identify opposing views, avoid duplicated efforts and also to identify methods, information and ideas relevant to this research project.

Chapter 3 describes the Framework for Understanding Structural Errors (FUSE) (Clark et al., 2008), the underlying concepts and its re-implementation into an R package. This framework is used as an example model inventory.

Chapter 4 describes the Plynlimon catchments which are used as study areas for testing the proposed algorithms. This Chapter also explores data availability and the use of a pre-processing workfow for data quality assessment and pre-modelling data manipulations.

Chapter 5 presents the AMCA open source project: the design and implementation of a novel data mining algorithm developed to identify the most suitable model configurations for a catchment of interest. This is tested on a synthetic dataset and accompanied by a series of sensitivity tests.

Chapter 6 presents a way of using the AMCA algorithm as a learning tool, exploring parameters and model component similarities and interactions to improve parameter identifiability, predictions and reduce uncertainties.

Chapter 7 proposes a methodology to understand whether and how anthropogenic changes to the environment can be detected by model simulations. Model simulations are conditioned based on regionalised information, enabling predictions in data-rich as well as data-poor environments and the implementation of what-if scenario tools.

Finally, Chapter 8 presents a summary and a critical assessment of the work that has been carried out, the most important conclusions and possible future research activities that could stem from it.

Chapter 2

Hydrological modelling and (Big) Data analytics: a brief literature review

Observed data is used to run model simulations which, in turn, help design development strategies and facilitate decision making. In hydrology, the understanding of processes is tightly linked to the way models are conceptualised and these tools have evolved significantly over less than 200 years of dedicated scientific research.

From simple mathematical equations to complex distributed physically-based models and ensembles, modellers have a plethora of options to choose from. Single model structures are usually designed *ad hoc* to model particular regions/processes. Multi-model ensembles and physically based models, instead, have wider ranges of applications. The choice can be based on data availability and suitability to reproduce known hydrological behaviours, on the project's objectives, or more pragmatically, based on the computer system in use and familiarity with a certain model.

Once suitable models are identified, their behaviour can be manipulated configuring the parameter space. Calibration is the operation that aims to find optimal model parameters. It consists of a search in a multidimensional space that can be carried out at random or optimised employ-

ing combinatorial, evolutionary or stochastic methods. The search identifies areas for which the combination of parameters is more likely to generate a response similar to the observations.

This review covers the main literature on hydrological models, their calibration and related uncertainties as well as the analytical implications of having to interpret large volumes of complex data.

2.1 Hydrological models

Hydrological models initially consisted of empirical formulas or modified forms of the rational method (Mulvany, 1851; Dooge, 1957; Todini, 1988), whose development was largely driven by the need to address particular engineering problems. With the advent of digital computers, it became possible to numerically simulate processes occurring in a watershed and hence many complex hydrological models were developed. An extensive overview of available hydrological models is available in Beven (2001*b*).

The simplest type of model is conceptually lumped, treating the basin as a single homogeneous element. These models consider geological, climate and land use information averaged over the catchemnt area and develop a single outflow hydrograph (Jones, 1997). They are typically used for water resources assessment and management including real-time forecasting (Blackie and Eeles, 1985; Paudel, 2010). Their main disadvantage is that they may not capture all of the important processes occurring in a catchment.

Distributed models, on the other hand are usually physically based and take into account spatial heterogeneity of parameters, topographic features, geologic and land cover variability (Kampf and Burges, 2007). They are often used for forecasting the effects of land-use changes or the movements of pollutants and sediments (Beven and OConnell, 1982; Paudel, 2010). Some widely used distributed models are: GBHM model (Yang et al., 1998), IHDM model (Morris, 1980), SHE model (Beven et al., 1980; Abbott et al., 1986), SWAT model (Arnold et al., 1998),

TOPMODEL (Beven and Kirkby, 1979), VIC (Liang, 1994).

However, in order to deal with the complexity of hydrological systems, many distributed physically based (hydrological) models have become too complex to understand and too data-hungry to be widely applied (Ma and Yang, 2010). There are widely divergent points of view as to whether they offer a significant improvement in actual performance when compared to the well-proven lumped conceptual model type (Refsgaard and Knudsen, 1996).

A more recent approach, developed in the 1970s, consists of a model structure able to adapt itself to the inputs and outputs of the system without taking into account underlying physics. Such a model is known as "empirical black box" and it can be either a simple autoregressive model, based on time series analysis, or incorporate information based on newer data (e.g. artificial neural network models).

Although each type of model has its advantages and disadvantages, this research does not deal with empirical black box models because it is intended to be an opportunity to understand more about the hydrological processes occurring in the investigated catchments. This research is concerned with providing an objective methodology to approach the hydrological modelling exercise in a location-agnostic manner. Considering the limited availability of distributed information on a global scale, only lumped models are considered hereafter.

2.1.1 Model calibration approaches and uncertainty

Model calibration is a process in which model parameters are modified to minimise the difference between model output and measurement data or maximise some goodness-of-fit criteria, e.g. root mean square error. Manual calibration can be performed, for instance, by changing one parameter at the time while the others are kept constant or at random with a trial and error approach (Boyle et al., 2000). This approach is also used to explore model sensitivity to parameter variations. However, even for models with a low number of parameters, exploring manually the range of responses over the entire parameter space can be rather tedious.

Calibration can be made automatic taking advantage of computer power. A number of parameter sets can be generated by randomly sampling the parameter space, then the model outputs corresponding to these parameter sets (Monte Carlo simulations) are compared to observed data using a fitting criteria. Local search optimization methods assume that there is a unique optimum in the parameter space, therefore the optimum parameter set is the one that corresponds to the maximum value of a given criteria (Dawdy and O'Donnell, 1965; Nash and Sutcliffe, 1970; Gupta and Sorooshian, 1985). Duan et al. (1992), however, showed that in case of multiple maxima, global optimization is preferable because does not incur in a premature convergence to a local optimum. Examples of global optimization approaches are given by Genetic Algorithms (Wang, 1991) and the Shuffled Complex Evolution algorithm (Duan et al., 1992). However, even with global searches parameters may be poorly identified generating a high level of uncertainty in the model results. Methods like the likelihood ratios (Beven and Binley, 1992), parameteric bootstrapping (Tarantola, 1987) and Markov Chain Monte Carlo (e.g. Metropolis-Hastings samplers: Kuczera and Parent, 1998; Vrugt et al., 2003) have introduced approaches to estimate parameter uncertainty.

Modellers can incur in four sources of uncertainty: random and systematic errors in the inputs/boundary conditions, random and systematic errors in the observed outputs, identification of sub-optimal parameter values (as mentioned above) and utilisation of incomplete/biased model structures (Butts et al., 2004). While issues related to observed variables and parameter estimation have been widely investigated, the impact of model structure error remains relatively less explored (Butts et al., 2004).

Depending on how a model structure is defined, the predictive uncertainty may vary. This is demonstrated, for instance, by a number of streamflow simulation studies and climate change impact studies. Butts et al. (2004) tried to identify a trade-off between model complexity and predictive ability and compared the magnitude of model structure uncertainty to the other sources of uncertainty. They showed high level of dependency between model performance and model structure and that "the sensitivity to variations in acceptable model structures are of the

same magnitude as uncertainties arising from the other evaluated sources". Jiang et al. (2007) used six monthly water balance models and 29-year long records of monthly streamflow and climate in the Dongjiang basin in China and showed that the simulated runoff change differed up to 20% under an increasing temperature of $4°C$ and decreasing precipitation of 20%. Najafi et al. (2011) conducted a study over the Tualatin River Basin in Oregon (USA) and showed that the hydrologic model uncertainty is considerably smaller than GCM uncertainty, except during the dry season, suggesting that the hydrologic model selection-combination is critical when assessing the hydrologic climate change impact. These results seem also to suggest that, in order to improve the reliability of streamflow predictions, modellers should combine, or at least compare, several model structures.

2.1.2 Modelling inventories

In theory, any model can be applied to any catchment but some can fails to provide satisfactory performance in terms of predictability of results/uncertainty. A failure can be due to problems in parameter identifiability, errors in data, structural inadequacy of the model, etc. Andréassian et al. (2010) suggest a focus on diagnosing model failures in order to progress in hydrological modelling.

Modern research is, however, moving towards "modelling frameworks, software tools combining single aspects of the catchment representation which, under rigorous constraints, enables flexible and creative problem solving" (Andrews et al., 2011). In the last decade several frameworks have been developed, some examples are: the Community Hydrologic Modeling Platform (CHyMP by Murdoch et al., 2008), the Framework for Understanding Structural Errors[1] (FUSE by Clark et al., 2008), the Hydrological Model Assessment and Development (HydroMAD by Andrews et al., 2011), the Object Modelling System (OMS by Leavesley et al., 2006), the PREcipitation Runoff EVApotranspiration HRU (PREVAH by Viviroli et al., 2009), the

[1] Based on BATEA, software for calibration and uncertainty analysis

Rainfall-Runoff Modelling Toolbox (RRMT by Wagener et al., 2001), the Representative Elementary Watershed (REW by Reggiani et al., 1998) and SUPERFLEX (Kavetski et al., 2010).

Within the class of frameworks using lumped models, FUSE can be considered the most complete inventory of conceptual model currently available. This modelling framework was re-implemented by translating the original FORTRAN code into the R language and released as open source project[2]. The following section reviews how model selection and calibration have been carried out utilising single and multi-model frameworks.

2.1.3 Approaches for model selection and calibration in a multi-model framework

The advantage of using an inventory of models rather than a single model stands in the possibility to compare different structure and assess how structural variability may affect predictions. Given a hydrological model inventory, various strategies can be employed to perform model selection and calibration.

Neural networks can be used to identify statistical relationships between inputs and outputs. However, the relation (or statistical model) does not provide means to understand the physical processes involved in the system (Lees, 2000). These black-box models can also be used within a Data-Based Mechanistic (DBM) modelling framework to generated a physical understanding (Young, 1998). Working with conceptual rainfall-runoff models, Coxon et al. (2014) built a diagnostic approach testing multiple hypotheses of hydrological behaviour using hydrologic signatures and time step-based metrics within the limits of acceptability uncertainty analysis approach (Beven, 2006). Coxon et al. (2014) used the FUSE framework (Clark et al., 2008) and stated that "the importance of selecting an appropriate model structure varies by catchment, and in some catchments, the model-structural choice is relatively unimportant in comparison with the selection of model parameters".

[2]Available from the public repository: http://ichydro.github.io/r_fuse/

Marshall et al. (2005) suggest various options, ranging from single and multi-criteria approaches for assessing model accuracy to Bayesian methods to describe parameter and model uncertainty probabilistically. They also highlight that the former methods tend to favour complex models, while Bayesian methods are usually biased towards simple models and the evaluation of various statistics from the posterior distribution in multidimensional functions can be challenging to evaluate analytically. Approximated methods exist and employ the use of Monte Carlo sampling.

More recently, Clark et al. (2008) simulated streamflow discharges for two rivers in the United States (the Guadalupe River in Texas and the French Broad River in North Carolina) comparing 79 model structures within the FUSE framework. Models were calibrated both independently and as an ensemble and results were compared in terms of Nash-Sutcliffe efficiency. The ensemble was always performing better than the single model structures. In a similar context, McMillan et al. (2010) suggested to use FUSE's model identification number as an additional parameter and proceed with calibration using traditional methods.

Although the methods presented by Clark et al. (2008) and McMillan et al. (2010) are simple to implement, they have a number of limitations. They do not allow to learn information about the selection of a particular model structure nor provide insight on the interaction between model components and parameter values. If this information becomes available, it could be used to learn about the catchment of interest in terms of dominant processes.

2.2 Analytical implications of working with large volumes of complex data

Calibrating multiple models in a multi-objective framework generates a large volume of complex data which are difficult to analyse and interpret. In machine learning, data are analysed building statistical models.

From an analytical perspective, when problems involve complex domains, such as the simulation of hydrological processes, these are not solvable by concise and neat formulas, and imply the introduction of simplifications and assumptions. Wu (2013) suggests that "having more data allows the data to speak for itself, instead of relying on unproven assumptions and weak correlations". This seems to be backed up by a number of machine learning studies that show the data set size to be more important than the statistical model being trained (Halevy et al., 2009; Brill, 2003). Shotton et al. (2013), for instance, as part of a Microsoft Research project related to their Kinect gaming device, tried to recognise human poses from single-depth images. They stated that the key factors in their success was the large volume of data they generated using computer graphics and a simple classifier[3].

However, there are also cases in which a complex statistical model allows to use less data (Pilászy and Tikk, 2009). In reality, it all depends on finding the right model, the statistical model that works best for a given problem. The performance of a model is often assessed in terms of bias and variance of its results. If a statistical model is too complicated for the amount of available data, this leads to model overfitting which is due to its high variance (Amatriain, 2015). Conversely, if the statistical model is too simple for the number and types of features to be analysed, then the model has a high bias (Amatriain, 2015). It becomes clear that the best statistical model is the one providing an optimal trade-off between bias and variance.

2.3 Algorithms and Workflows

An algorithm is "a set of steps to accomplish a task" (Cormen, 2013) and algorithm design is an important area of scientific research. Designing a novel algorithm does not necessarily involve new operations, but finding "efficient ways of doing something that requires a higher level of intuition than the most apparent solution" (Cutrell, 2012).

Methodologically, there are a number of steps in developing an algorithm (Goodrich and Tamas-

[3] A classifier is a statistical model used for supervised learning.

sia, 2002):

- Problem definition
- Development of a model
- Specification of Algorithm
- Designing an Algorithm
- Checking the correctness of Algorithm
- Analysis of Algorithm
- Implementation of Algorithm
- Program testing
- Documentation Preparation

Based on the list of steps above, an algorithm is first designed in terms of mathematical equations and then converted into a computer code. When a task is particularly complex, it is good practice to separate the functionality of a program into independent modules (Brogi et al., 1994). A workflow is defined, hereafter, as a computer program in which multiple algorithms are used in a modular fashion.

2.4 Machine Learning

Warden (2011) defines machine learning as the "systems [or algorithms] to automate decision making and classification of data". In machine learning there are two major types of algorithms: supervised and unsupervised learning (Mohri et al., 2012). The first is used to make predictions based on a set of examples. The dataset is usually divided into two parts, the training and test

sets. The first set typically contains labels that are missing in the second set. The scope of these algorithms is to classify the cause-effect relationship based on observed data and use these classes to make prediction on new data. Supervised Learning algorithms are widely used in environmental modelling: Ireland et al. (2015) use them to extract flooded areas from Landsat TM imagery, Wohlfahrt et al. (2010) for assessing the impact of the spatial arrangement of agricultural practices on pesticide runoff in small catchments, Chandramouli and Raman (2001) for training neural networks in multireservoir modeling, Chang and Chen (2003) for water-stage forecasting in an estuary under high flood and tidal effects and Rogers and Dowla (1994) uses them to set up a nonlinear groundwater management methodology that optimizes aquifer remediation.

In Unsupervised Learning, instead, the dependent variable is unknown and the algorithm explores possible interactions amongst the features of the dataset. The aim is to "provide structure to relatively unstructured data" (Perry, 2013). This type of algorithm is also widely used in hydrology: Lin and Chen (2006) use them to identify the homogeneous regions for regional frequency analysis, Lin and Wu (2011) for developing a reservoir inflow forecasting model, Einax et al. (1998) for the evaluation and interpretation of river pollution data.

The combination of supervised and unsupervised learning seems far less common and, to the best knowledge of the author, there have been no attempt in literature to combine these techniques to automate model selection and calibration.

2.5 Conclusions

This research is concerned with providing an objective methodology to guide the model selection and configuration process in a location-agnostic manner. As a consequence, the procedure is designed to rely on minimal data requirements and, therefore, only lumped models are considered hereafter.

Amongst the many inventories of available conceptual rainfall-runoff models, FUSE was selected to test the algorithms designed in the following chapters because it provides the largest number of model structures. Next chapter illustrates in detail the implementation of the FUSE modules and the design of model building options.

To the best knowledge of the author, there seem to be no attempt in literature to combine supervised and unsupervised learning algorithms to implement an automatic model selection and configuration. This PhD work attempts to cover this gap.

Chapter 3

The FUSE modelling framework

The Framework for Understanding Structural Errors (FUSE) was developed by Clark et al. (2008) and is a state-of-the-art modelling toolbox which includes well established models for rainfall-runoff simulations (i.e. PRMS, SACRAMENTO, TOPMODEL and ARNO/VIC, also defined as parent models). Each model is characterised by a different architecture of the upper and lower soil layers and parameterisation of processes such as: evaporation, vertical percolation, interflow, base flow and surface runoff. FUSE can combine elements from different models to obtain several structures. Rainfall and potential evapotranspiration observations are used as inputs to simulate a streamflow discharge time series. As part of this PhD work, the original FORTRAN code was re-implemented in the R programming language to facilitate interoperability with other pre/post processing algorithms. The code is available from a public repository[1]. This chapter illustrates the peculiarity of this modelling framework, its input requirements and outputs.

[1]FUSE R package: `http://ichydro.github.io/r_fuse/`

3.1 Introduction

FUSE is a modular framework of conceptual rainfall-runoff models. The catchment is considered a closed system in which the precipitation is the only input[2]. The potential evapotranspiration is a loss and it is generally derived from temperature, humidity and other climate variables. The streamflow discharge is the output that closes the long-term water balance equation:

$$Q = P - E \tag{3.1}$$

The framework consists of four widely used models, called *parent models*: PRMS, SACRAMENTO, TOPMODEL and ARNO/VIC. Each model structure consists of a list of *building decisions* describing the type of rainfall error (if included in the inference), the structure of upper and lower soil layers and the parameterisation of processes such as: evaporation, vertical percolation between soil layers, interflow, base flow, surface runoff and routing scheme. Each building decision can be parameterised using different *modelling options*. In FUSE, modelling options have been implemented as separate modules using a consistent set of parameters. The major advantage in doing so is that additional model structures can be generated by shuffling the parent models' options. For instance, a model structure can be generated by combining the upper soil layer characterised by a single state variable (as in TOPMODEL) with the lower soil layer characterised by a combination of three storages (two for tension and one for free water, as in SACRAMENTO) with the PRMS runoff mechanism and no interflow (as in ARNO/VIC). Figure 3.1 shows FUSE's parent models, state variables and fluxes, as defined by Clark et al. (2008).

The default list of FUSE's model structures is represented by a table made of 1248 rows and 9 columns. Each row represents a model structure which is identified by a Model ID number (first column, "mid"). Table 3.1 lists the first ten model structures in the list.

[2]If relevant, the snow melt should be modelled and added to the precipitation.

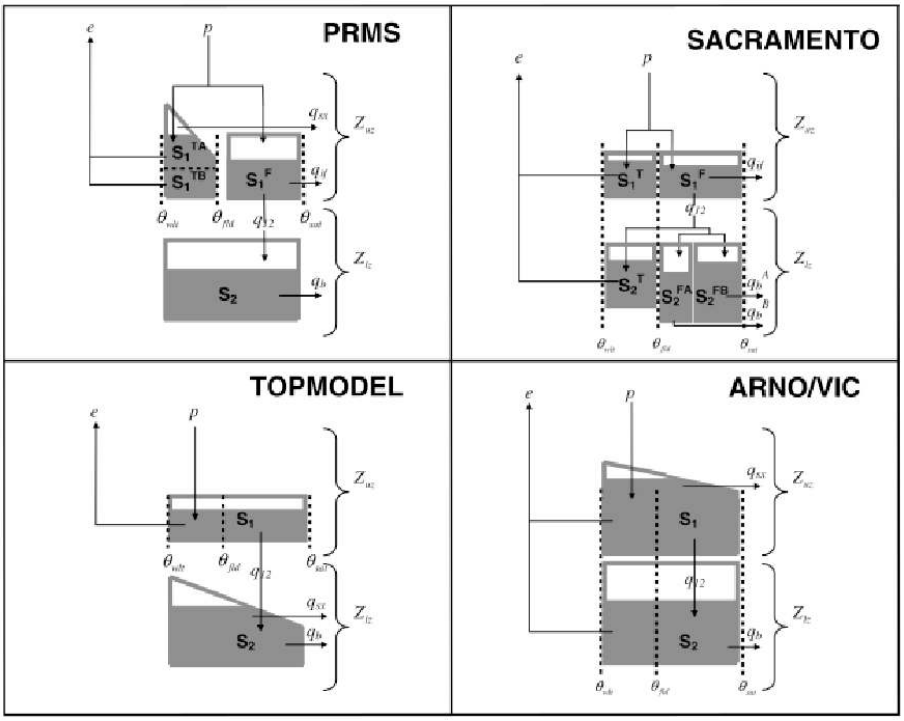

Figure 3.1: Parent models used in FUSE.
source by: Clark *et al.* Clark et al. (2008)

Table 3.1: First 10 model structures in the FUSE's model list.

mid	rferr	arch1	arch2	qsurf	qperc	esoil	qintf	q_tdh
1	additive_e	tension1_1	tens2pll_2	arno_x_vic	perc_f2sat	sequential	intflwnone	rout_gamma
2	multiplc_e	tension1_1	tens2pll_2	arno_x_vic	perc_f2sat	sequential	intflwnone	rout_gamma
3	additive_e	onestate_1	tens2pll_2	arno_x_vic	perc_f2sat	sequential	intflwnone	rout_gamma
4	multiplc_e	onestate_1	tens2pll_2	arno_x_vic	perc_f2sat	sequential	intflwnone	rout_gamma
5	additive_e	tension1_1	unlimfrc_2	arno_x_vic	perc_f2sat	sequential	intflwnone	rout_gamma
6	multiplc_e	tension1_1	unlimfrc_2	arno_x_vic	perc_f2sat	sequential	intflwnone	rout_gamma
7	additive_e	tension2_1	unlimfrc_2	arno_x_vic	perc_f2sat	sequential	intflwnone	rout_gamma
8	multiplc_e	tension2_1	unlimfrc_2	arno_x_vic	perc_f2sat	sequential	intflwnone	rout_gamma
9	additive_e	onestate_1	unlimfrc_2	arno_x_vic	perc_f2sat	sequential	intflwnone	rout_gamma
10	multiplc_e	onestate_1	unlimfrc_2	arno_x_vic	perc_f2sat	sequential	intflwnone	rout_gamma

The first column contains the model identification number (mid), an integer between 1 and 1248. The second column is the rainfall error type (rferr), which can be either additive or multiplicative. The third column contains the architecture of the upper soil layer (arch1), this can be characterised by a single state variable, broken up into tension and free storage or characterised by a tension storage sub-divided into recharge and excess. The fourth column contains the architecture of the lower soil layer (arch2), which can be characterised by either a limited or unlimited size reservoir. In case of a reservoir of limited size, this can be defined by a single

state variable or as a tension reservoir plus two parallel tanks (in this case the upper soil layer cannot be split into two tension and one free storage). In case of an unlimited reservoir, the depletion rate can be either a power function or a fraction. In the fifth column is the surface runoff scheme (qsurf), based only on the saturation-excess mechanism. The saturated area is calculated using one of the following parameterisations: the ARNO/VIC (upper zone control), the PRMS variant (fraction of upper tension storage) or the TOPMODEL. In the sixth column is the percolation scheme (qperc), in which water availability for percolation is limited by the moisture content in lower layer, field capacity or wilting point (in this case the upper soil layer can only be single state). The evaporation scheme (esoil) is in the seventh column and can be either sequential or depending on the relative root fractions in each of the soil layers. In the former case the potential evaporative demand is first satisfied by evaporation from the upper layer, then the residual from the lower layer. The interflow (qintf) (column 8) can be allowed or not allowed (e.g. for TOPMODEL and ARNO/VIC). Similarly the routing (q_tdh) can be allowed or not allowed (column 9). If allowed, this uses a two-parameters gamma distribution to empirically route runoff to the basin outlet.

In the original FUSE's model list the variables that identify the options are expressed in terms of character strings. In the R implementation, instead, all the strings are converted into numerical factors to speed up processing[3]. All the possible modelling option names/IDs and dependent parameters are summarised in Table 3.2.

[3]The code using numerical values was benchmarched against the code using strings and it was found to run about 100 times faster.

Table 3.2: FUSE model building decisions, options (name and ID number) and depending parameters.

Model building decision	Option name	Option ID	Depending parameters
Rainfall Error	additive_e = additive rainfall error	11	$rferr_{add}$
	multiplc_e = multiplicative rainfall error	12	$rferr_{mlt}$
Upper layer architecture	onestate_1 = single state variable	21	$S_{1,max}$, ϕ_{tens}
	tension1_1 = tension and free storage	22	$S_{1,max}$, ϕ_{tens}
	tension2_1 = 2 tension and 1 free storage	23	$S_{1,max}$, ϕ_{tens}, ϕ_{rchr}
Lower layer architecture	fixedsiz_2 = fixed size	31	$S_{2,max}$, k_s, n
	tens2pll_2 = tension reservoir and two parallel tanks	32	$S_{2,max}$, κ, ϕ_{base}, ν_A, ν_B
	unlimfrc_2 = unlimited (frac. rate)	33	$S_{2,max}$, ν
	unlimpow_2 = unlimited (power recession)	34	$S_{2,max}$, k_s, n, λ, χ
Runoff	arno_x_vic = ARNO-Xzang-VIC (Unsaturated zone Pareto)	41	b
	prms_varnt = PRMS (Unsaturated zone linear)	42	$A_{c,max}$
	tmdl_param = TOPMODEL (Saturated zone topographic)	43	n, λ, χ
Percolation	perc_f2sat = water availability from field capacity to saturation	51	k_u, c
	perc_lower = percolation defined by moisture content in lower layer (gravity drainage)	52	α, ψ
	perc_w2sat = water availability from wilt point to saturation (Saturated zone control)	53	k_u, c
Evaporation	rootweight = root weighting	61	r_1
	sequential = sequential evaporation model	62	-
Interflow	intflwnone = interflow denied	71	-
	intflwsome = interflow allowed	72	k_i
Routing	no_routing = routing denied	81	-
	rout_gamma = routing allowed	82	μ_τ

3.1.1 Distribution of modelling options

Each building decision can have from a minimum of 2 to a maximum of 4 options. Table 3.3 shows the frequency of each option within the default list of FUSE's model structures. According to this list, the options for rainfall error, runoff, evaporation, interflow and routing are

evenly distributed. The upper and lower soil layer along with the percolation parameterisation, instead, have preferential options that are more frequent than the others. For instance, 46% of the model structures have an upper soil layer made of a single state variable (onestate_1 = 21), 31% of the model structures is characterised by one tension and one free storage (tension1_1 = 22) and only the remaining 23% has two tension and one free storage (tension2_1 = 23). The lower soil layer made of one tension reservoir and two parallel tanks (tens2pll_2 = 32) is found in only 19% of the model structures while the remaining options are evenly distributed (27% each). Similarly the percolation scheme dependent on the wilt point (perc_w2sat = 53) is only found in 16% of the structures, while each of the other options covers 42% of the structures.

Table 3.3: Distribution of FUSE modelling options.

Model building decision	Option name	Option ID	Frequency (%)
Rainfall Error	additive_e = additive rainfall error	11	50
	multiplc_e = multiplicative rainfall error	12	50
Upper layer architecture	onestate_1 = single state variable	21	46
	tension1_1 = tension and free storage	22	31
	tension2_1 = 2 tension and 1 free storage	23	23
Lower layer architecture	fixedsiz_2 = fixed size	31	27
	tens2pll_2 = tension reservoir and two parallel tanks	32	19
	unlimfrc_2 = unlimited (frac. rate)	33	27
	unlimpow_2 = unlimited (power recession)	34	27
Runoff	arno_x_vic = ARNO-Xzang-VIC (Unsaturated zone Pareto)	41	$3\overline{3}$
	prms_varnt = PRMS (Unsaturated zone linear)	42	$3\overline{3}$
	tmdl_param = TOPMODEL (Saturated zone topographic)	43	$3\overline{3}$
Percolation	perc_f2sat = water availability from field capacity to saturation	51	42
	perc_lower = percolation defined by moisture content in lower layer (gravity drainage)	52	42
	perc_w2sat = water availability from wilt point to saturation (Saturated zone control)	53	16
Evaporation	rootweight = root weighting	61	50
	sequential = sequential evaporation model	62	50
Interflow	intflwnone = interflow denied	71	50
	intflwsome = interflow allowed	72	50
Routing	no_routing = routing denied	81	50
	rout_gamma = routing allowed	82	50

3.2 Parameters, state variables and internal fluxes

The FUSE framework uses 24 parameters. Each model structure, however, only uses a subset between 7 and 15 parameters. The full list of parameters with description, units and naming convention is available in Table 3.4.

FUSE's models are based on a set of ordinary differential equations (ODEs) that are solved with respect to a list of state variables describing the capacity of the water storages in each soil layer

(see Table 3.5). Depending on the complexity of the model, the ODEs have from a minimum of 4 to a maximum of 10 state variables.

Beside instantaneous and routed runoff, FUSE can calculate a total of 18 internal fluxes. The *percolation*, for instance, is the flow of water that moves from the upper to the lower soil layer. The *interflow* is a lateral loss for the system and occurs in the unsaturated zone of the upper soil layer. The *surface runoff* is the volume of water that does not enter the soil due to excess of saturation. The soil is schematised as a combination of water storages (or buckets). When a bucket reaches its capacity the volume in excess moves to another bucket generating an *overflow*. Generally, water moves from one bucket to another in the same soil layer. If the upper layer reaches saturation then the overflowing water becomes additional runoff. If the free storage in the lower layer reaches saturation then the overflowing water becomes additional baseflow. Base flow is the volume of water in the lower layer which was accumulated due to precedent rainfall events. This can be a constant or decreasing function. Each of these fluxes is described in Table 3.6.

Table 3.4: FUSE's parameters

Description	Name	Symbol
Additive rainfall error (mm day-1)	rferr_add	$rferr_{add}$
Multiplicative rainfall error (-)	rferr_mlt	$rferr_{mlt}$
Fraction of tension storage in recharge zone (-)	frchzne	ϕ_{rchr}
Fraction total storage as tension storage (-)	fracten	ϕ_{tens}
Maximum total storage in upper soil layer (mm)	maxwatr_1	$S_{1,max}$
Fraction of percolation to tension storage in the lower layer (-)	percfrac	κ
Fraction of storage in the first baseflow reservoir (-)	fprimqb	ϕ_{base}
Baseflow depletion rate in the first reservoir (day-1)	qbrate_2a	ν_A
Baseflow depletion rate in the second reservoir (day-1)	qbrate_2b	ν_B
Baseflow depletion rate (day-1)	qb_prms	ν
Maximum total storage in lower soil layer (mm)	maxwatr_2	$S_{2,max}$
Baseflow rate (mm day-1)	baserte	k_s
Fraction of roots in the upper layer (-)	rtfrac1	r_1
Percolation rate (mm day-1)	percrte	k_u
Percolation exponent (-)	percexp	c
Sacramento model percolation multiplier for dry soil layer (-)	sacpmlt	α
Sacramento model percolation exponent for dry soil layer (-)	sacpexp	ψ
Interflow rate (mm day-1)	iflwrte	k_i
ARNO/VIC "b" exponent (-)	axv_bexp	b
Maximum saturated area (-)	sareamax	$A_{c,max}$
Mean value of the log-transformed topographic index (m)	loglamb	λ
Shape parameter for the topo index gamma distribution (-)	tishape	χ
Baseflow exponent (-)	qb_powr	n
Time delay (days)	timedelay	μ_τ

Table 3.5: FUSE's state variables

Description	Name	Symbol
Total water content in the upper soil layer (mm)	watr_1	S_1
Tension water content in the upper soil layer (mm)	tens_1	S_1^T
Primary tension water content in the upper soil layer (mm)	tens_1a	S_1^{TA}
Secondary tension water content in the upper soil layer (mm)	tens_1b	S_1^{TB}
Free water content in the upper soil layer (mm)	free_1	S_1^F
Total water content in the lower soil layer (mm)	watr_2	S_2
Tension water content in the lower soil layer (mm)	tens_2	S_2^T
Free water content in the lower soil layer (mm)	free_2	S_2^F
Free water content in the primary base flow reservoir (mm)	free_2a	S_2^{FA}
Free water content in the secondary base flow reservoir (mm)	free_2b	S_2^{FB}

Table 3.6: FUSE's internal fluxes

Description	**Name**	**Symbol**
Evaporation from the upper soil layer	evap_1	e_1
Evaporation from the lower soil layer	evap_2	e_2
Evaporation from the primary tension store	evap_1a	e_1^A
Evaporation from the secondary tension store	evap_1b	e_1^B
Surface runoff	qrunoff	q_{sx}
Percolation of water from the upper to the lower layer	qperc_12	q_{12}
Interflow	qintf_1	q_{if}
Base flow	qbase_2	q_b
Base flow from the primary reservoir	qbase_2a	q_b^A
Base flow from the secondary reservoir	qbase_2b	q_b^B
Overflow of water from the primary tension store in the upper soil layer	oflow_1	q_{urof}
Overflow of water from tension storage in the upper soil layer	tens2free_1	q_{utof}
Overflow of water from free storage in the upper soil layer	rchr2excs	q_{ufof}
Overflow of water from tension storage in the lower soil layer	tens2free_2	q_{stof}
Overflow of water from free storage in the lower soil layer	oflow_2	q_{sfof}
Overflow of water from primary base flow storage in the lower soil layer	oflow_2a	q_{sfofa}
Overflow of water from secondary base flow storage in the lower soil layer	oflow_2b	q_{sfofb}
Instantaneous runoff	U	

3.3 FUSE modules

The FUSE framework is divided into two modules: a soil moisture accounting module and a routing module. In the *fuse* R package, each module can be run separately using dedicated functions. Forcing inputs are in tabular format (e.g. data.frame) with fixed headers. Two columns are required: P (precipitation time series) and E (potential evapo-transpiration time series). A third column, named Q (discharge time series) is used for calibration only. Precipitation, potential evapotranspiration and streamflow discharge are measured in mm/day. The parameter set is provided as a named list (for suggested parameter ranges see Clark et al., 2008).

As already mentioned before, within FUSE it is possible to modify the precipitation input to take into account measurement errors of the following forms: additive and multiplicative. The former modifies the precipitation input by a fixed volume measured in mm/day and it is usually used to correct shifts in the recordings. The latter is a number between 0 and 1 and modifies the precipitation input by a given percentage. The new input is called *effective precipitation*. For the purpose of this work, the rainfall error is always excluded from the inference, therefore the effective precipitation is always equal to the input precipitation.

3.3.1 The Soil Moisture Accounting module

The Soil Moisture Accounting (SMA) module is used to determine catchment wetness in terms of moisture storage volumes and rainfall losses. The SMA function takes as inputs:

- the table containing time series of precipitation and potential evapo-transpiration;
- the model id number (mid, e.g. 5);
- the model list (modlist);
- observation time step in days (deltim, e.g. 1/24 for hourly time steps);

- states output options (TRUE outputs all the state variables);
- fluxes output options (TRUE outputs all the internal fluxes);
- initial water content (*fracstate0*, by default this is equal to 25% of the maximum storage capacity);
- a parameter set.

3.3.2 Routing module

The routing module is used to simulate the delay in runoff (in days). This would normally require an hydraulic model to calculate the wave propagation based on topography, condition of the channel and any hydraulic structure along the river. For small catchments, where man-made modifications are negligible, this can be simplified by applying an empirical formula. FUSE uses the gamma function.

The routing function takes as inputs: the instantaneous runoff computed using the SMA module (U), the model id number, the model list, daily delay in runoff (timedelay) and the observation time step.

3.4 Concluding remarks

FUSE is a modular framework of conceptual rainfall-runoff models in which modelling options have been implemented as separate modules using a consistent set of parameters. By shuffling the default modelling options, FUSE generates 1248 model structures which can be used to simulate a wide range of hydrological behaviours.

The R implementation of the FUSE framework presented in this chapter was designed to work not only as a stand alone package for the R environment but also as an openly available build-

ing block to compose modelling workflows that can be re-used and re-purposed to generate information useful across different projects, fields and research domains.

Chapter 4

Study area

A case study site was selected to test the algorithms proposed in Chapters 5 to 7: the Plynlimon catchments in the United Kingdom. In this chapter, the site is first described in relation to its geographical location and climatic characteristics. The available information is collated exploring various data sources. A screening of the available datasets is carried out to highlight the presence of missing and unrealistic values, fill gaps in the records as well as preparing the datasets to be used within the FUSE modelling framework.

4.1 Site description and data sources

Plynlimon area is located in mid-Wales (see Figure 4.1) and its climate is classified as "warm temperate climate, fully humid with warm summer" according to the Koeppen-Geiger classification system (Kottek et al., 2006; Peel et al., 2007).

The Global Runoff Data Centre's catalogue identifies two main catchments in the area: the Wye at Cefn Brwyn and the Severn at Plynlimon flume. The GRDC also points to the National River Flow Archive (NRFA), hosted by the UK Centre for Ecology and Hydrology (CEH), as the official data provider.

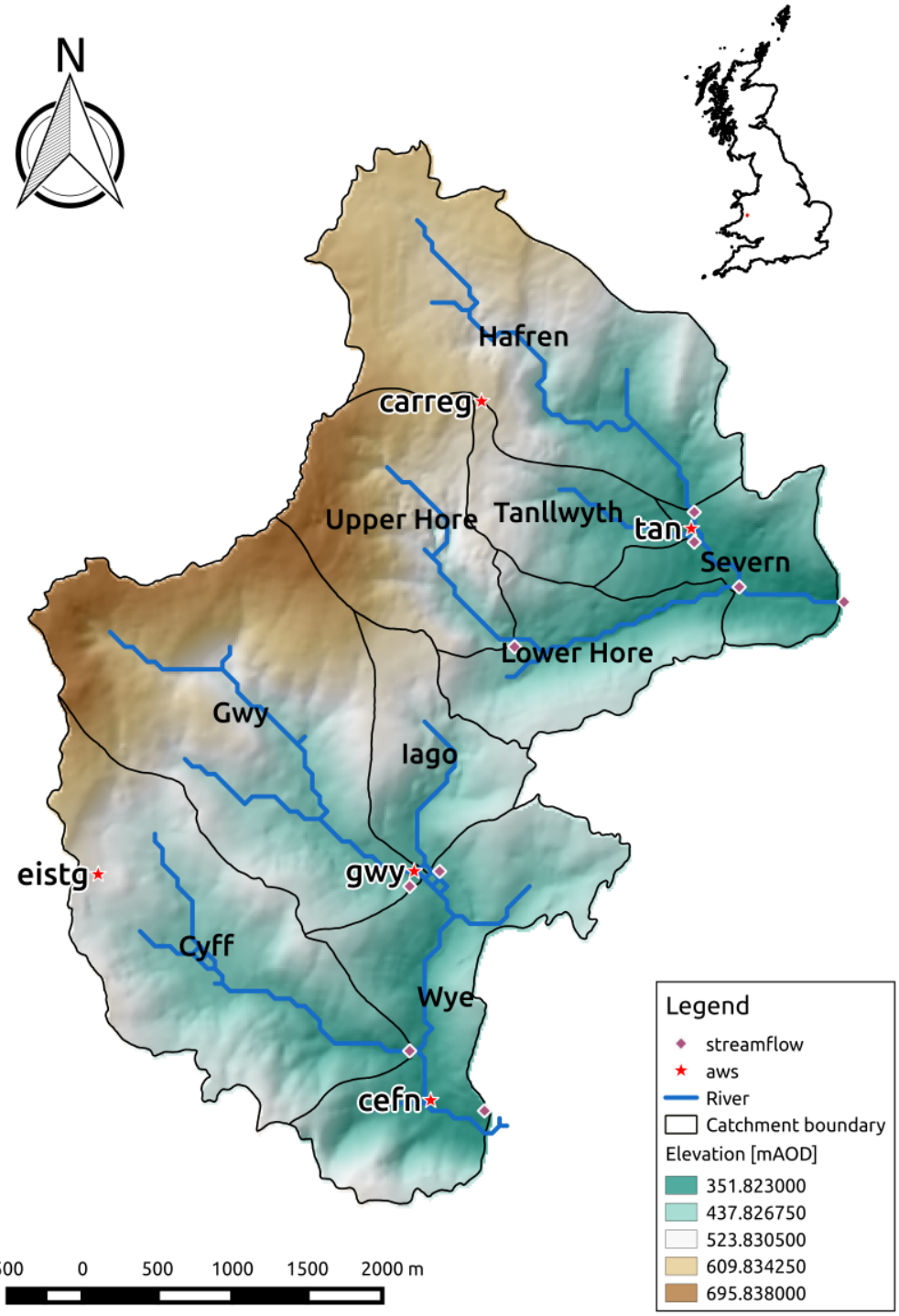

Figure 4.1: Location of Plynlimon Experimental Catchments. Streamflow gauges are shown as purple diamonds, automatic weather stations are shown as red stars. The shading illustrates the elevation.

According to the NRFA records, the *Severn at Plynlimon flume* is a small upland catchment ($8.7\,Km^2$) mainly covered with coniferous forest (67%). This is divided into four sub-catchments:

Tanllwyth, *Hafren*, *Upper Hore* and *Lower Hore*. The Severn at Plynlimon flume and its upstream tributaries have steep profiles, rapid mean flow and flashy hydrographs. The mean precipitation is around 1700-2500 mm/yr.

The *Wye at Cefn Brwyn* is also quite small ($10.55\,Km^2$) mainly covered with grazed grassland. It is divided into three sub-catchments: *Gwy*, *Cyff* and *Iago*. The total annual average rainfall is about 2500 mm/yr. On the high plateau, the predominant vegetation is grassland and heath, whereas mires dominate on the valley floors. Soils are rather shallow (thickness is generally less than $1\,m$) and dominated by peat, peaty peazols and peaty gleys overlying shales and mudstone (Storm and Jensen, 1984).

Catchments and sub-catchments are shown in Figure 4.1, while their main characteristics are summarised in Table 4.1. The Information Gateway portal (CEHIG[1]), also hosted by the CEH, provides the most detailed datasets for Plynlimon catchments.

Table 4.1: Plynlimon stations. Area (A), Length of main channel (L) and average slope (S) are calculated from GIS layers.

ID	**Name**	**River**	**A** [Km^2]	**L** [Km]	**S** [%]	**Z** [$mAOD$]
54022	Severn at Plynlimon flume	Severn	8.70	4.601	6.3	331
54090	Tanllwyth at Tanllwyth flume	Tanllwyth	0.89	0.974	10.9	357
54091	Severn at Hafren flume	Severn	3.67	3.263	5.94	357
54092	Hore at Hore flume	Hore	3.08	3.350	7.05	336
54097	Hore at Upper Hore flume	Hore	1.60	1.668	10.6	412
55008	Wye at Cefn Brwyn	Wye	10.55	5.363	3.63	341
55033	Wye at Gwy flume	Wye	3.98	2.989	2.03	405
55034	Cyff at Cyff flume	Cyff	3.13	2.703	2.76	356
55035	Iago at Iago flume	Iago	1.02	1.228	3.07	386

4.1.1 Geospatial information

The CEHIG datasets contain spatial data as well as time series data. Topographic information is represented by a Digital Terrain Model with a resolution of $15\,m$ and spot heights. In regard

[1] `https://gateway.ceh.ac.uk/`, accessed 15th September 2014.

to the quality of the DTM, the CEHIG metadata catalogue states:

"Plynlimon digital terrain model was derived from digitised elevation data. The elevation data had been digitised from scanned topographic maps (reference below). Once the elevation data was captured (spot heights and contour lines) a TIN was created which was then used to derived a hydrologically corrected, grid-based DTM. All the processing was done in ARCINFO however it was not documented therefore no further details can be provided. Topographic maps reference: Plynlimon Catchment Areas - Severn Catchment. Sheet No. 1. Scale 1:5000. Wallingford, Berkshire. Institute of Hydrology, Natural Environmental Research Council, 1968. Plynlimon Catchment Areas - Wye Catchment. Sheet No. 2. Scale 1:5000. Wallingford, Berkshire. Institute of Hydrology, Natural Environmental Research Council, 1968."

The same source provides maps of vegetation classes (Figure 4.2) and soil types (Figure 4.3). The former has a spatial resolution of 25 m resolution and is derived by digitizing aerial photography of the area from 2009 and was published in 2013. CEHIG states that:

"...the digitised map was verified on the ground and amended as needed".

Over the period 1990-2007 Land Cover Maps were also produced but CEH states that the methodologies used are fundamentally different, which implies that it is not possible to derive assessment of land use changes over time.

Soil related information can be derived from the Hydrolody of soil types (Boorman et al., 1995) or from the map of soil materials. The latter was digitised from a scanned paper map. CEHIG states:

"...most likely the soil type boundaries had been hand drawn on a paper map as a result of direct observations on site".

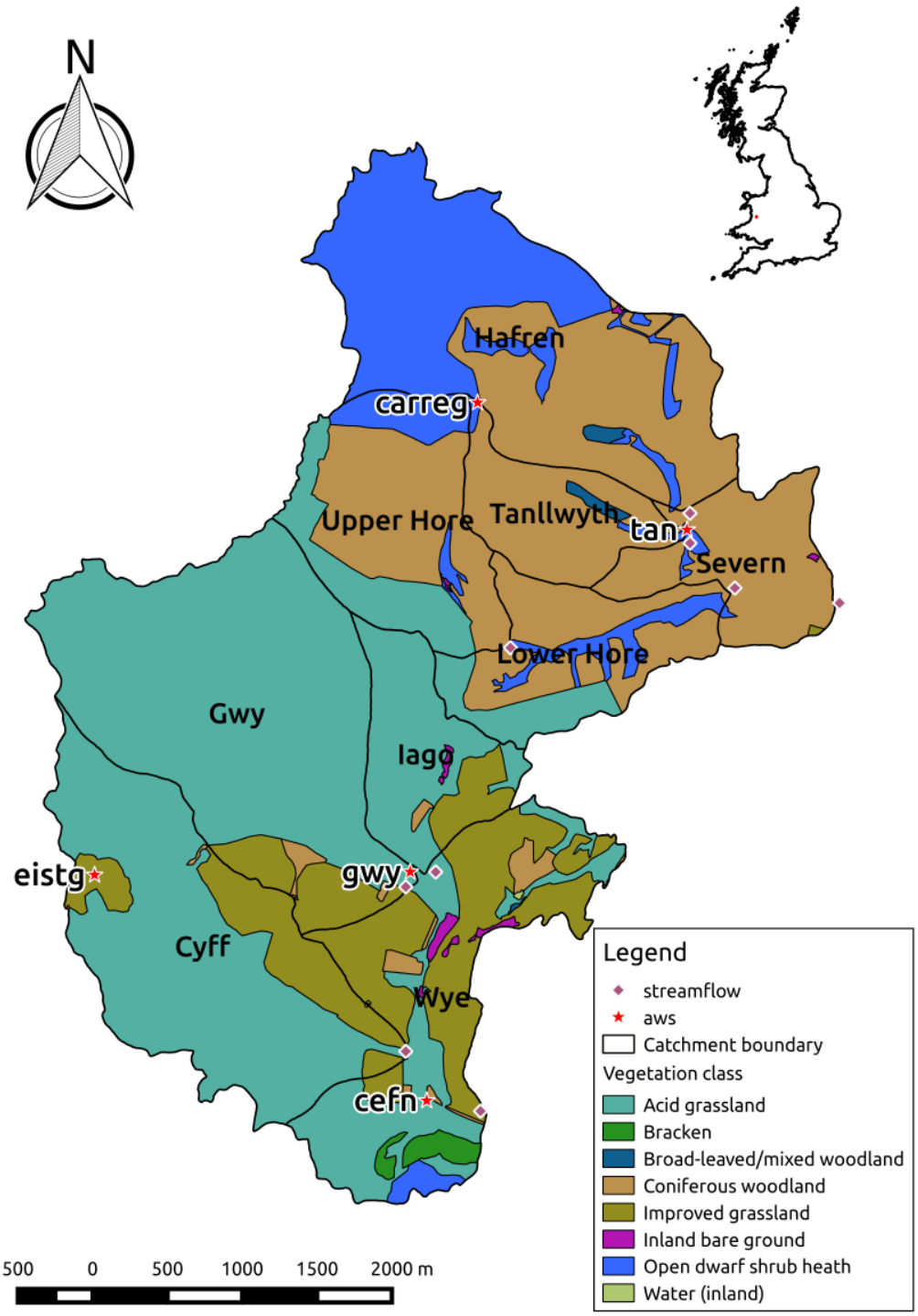

Figure 4.2: Vegetation map of the study catchment (source: Centre for Ecology and Hydrology Information Gateway). The Severn at Plynlimon flume is generally covered with forest while the Wye at Cefn flume area is mainly covered with grassland.

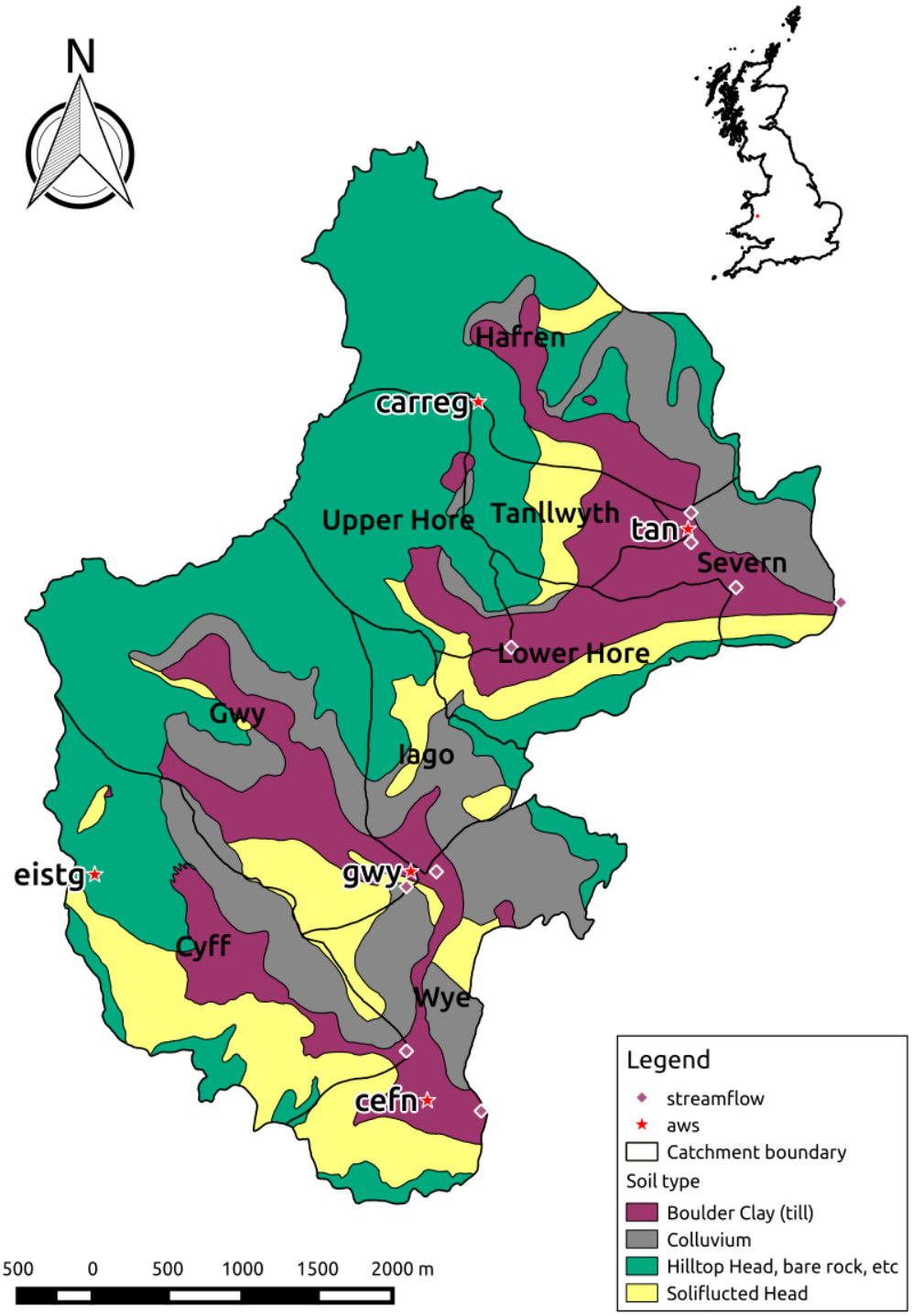

Figure 4.3: Soil map of the study catchment (source: Centre for Ecology and Hydrology Information Gateway). The Severn at Plynlimon flume and the Wye at Cefn flume have similar soil types. The upstream areas are characterised by bare rocks while the downstream areas by clays and peaty soils.

4.1.2 Time series information

Both Plynlimon catchments are heavily instrumented and used in numerous studies Storm and Jensen (1984); Kirby et al. (1991); Beven and Binley (1992); Hudson et al. (1997); Bulygina et al. (2011). Five weather stations and nine gauging stations are currently installed, these are represented in figures4.2 and 4.3 as stars and diamonds respectively. Streamflow time series were available only for eight gauging stations (the dataset for Iago at Iago flume is known to be recorded but not available). Streamflow (Q) is recorded in m^3/s while weather stations provide the following variables:

- incoming Solar Radiation (SR) averaged over the recording interval and measured in W/m^2 (Watts per square metre);
- Net Radiation (NR), the difference between the incoming and outgoing radiation (i.e. reflected by the ground) averaged over the recording interval and measured in W/m^2 (Watts per square metre);
- Temperature Wet bulb (TW), air temperature as a function of the moisture content of the air, averaged over the recording interval and measured in °C (degree Celsius);
- Temperature Dry bulb (TD), temperature of the air averaged over the measuring interval, measured in °C (degree Celsius);
- Wind Speed (WS) averaged over the recording interval and measured in m/s (metres per second);
- Wind Direction (WD) averaged over the recording interval and measured in degrees;
- Precipitation (P) accumulated over the recording interval and measured in *mm* (millimeters).

4.2 Data pre-processing

In order to prepare Plynlimon data for feeding the modelling framework FUSE (used in the following chapters), few pre-processing steps were needed. FUSE requires as forcing inputs: rainfall (plus snowmelt, if relevant), potential evapotranspiration and streamflow (for calibration only) time series. The inputs should be made of gap-free areal averaged time series, measured in mm/day, for consistency with the default parameter ranges. It is assumed that, in the Plynlimon study area, the contribution of snowmelt to runoff is negligible while potential evapotranspiration is expected to have a small but not negligible impact.

The computer code to reproduce the data preparation steps is available as an R package (called *pure*[2]). The pure package, was developed as part of this PhD work and is currently being tested on various case studies within the NERC-funded Probability, Uncertainty and Risk in the Environment (PURE) project[3]. The package consists of a collection of utility functions used in a pre-defined sequence to improve the reproducibility of the pre-processing task. This is achieved in four steps, according to Figure 4.4:

- initial screening to identify missing and unrealistic values,
- calculating derived variables,
- averaging time series over a catchment area,
- filling gaps in the records and converting to common units.

[2] `http://cvitolo.github.io/r_pure/`, accessed 27th September 2014.
[3] `http://www.nerc.ac.uk/research/funded/programmes/pure/`, accessed 27th September 2014.

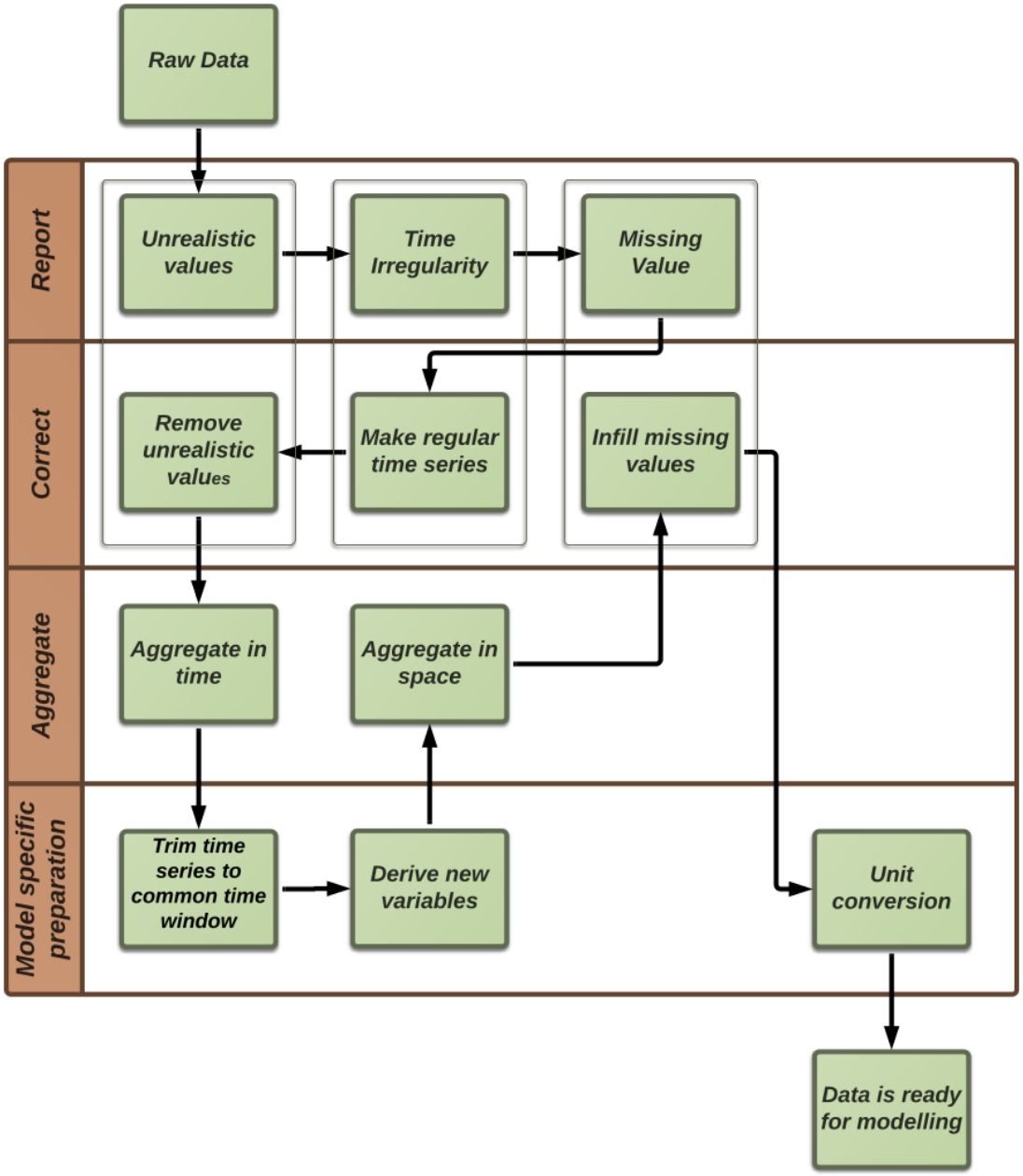

Figure 4.4: The PURE data preparation workflow. Raw data is made ready for modelling through a series of processing steps schematised as green rectangles. Each process belongs to one of four operations: report, correct, aggregate and model specific preparation. The first operation consists of scanning the time series for unrealistic values (e.g. negative rainfall), records irregurarly spaced in time and missing value. A summary report is produced, based on which, unevenly spaced time series are transformed to evenly spaced ones by linearly interpolating between data points. Unrealistic values are removed generating additional missing values. These are not infilled straight away but after some additional steps. The time series are first aggregated in time, if the modelling time step is longer than the one of the regular time series. The length of the time series is trimmed to take into account only simultaneous recordings. Any derived variable is calculated, then gaps are infilled and finally units are converted according to the model's requirements.

4.2.1 Screening

During screening, time series are first scanned to identify start and end date of the dataset, percentage of missing values, presence of any unrealistic values or obvious outliers. According to this initial screening, Plynlimon time series are characterised by regular recordings from 1968 to 2010 with various degrees of coverage. There are neither unrealistic values, nor clear outliers. Weather variables are characterised by time series with less then 15% of missing values, while streamflow datasets have generally less than 2% missing values. The detailed results of the screening are summarised in Tables 4.2 and 4.3.

Table 4.2: Summary of available 15-minute datasets from streamflow stations, where the variable flow is measured in m^3/s. A line separates the first three stations falling within the Wye at Cefn Brwyn catchment from the remaining five stations falling within the Severn at Plynlimon flume catchment.

ID	SiteName	Start of record	End of record	% missing values
55033	Wye at Gwy flume	1973-10-10	2008-12-31	1.96
55034	Cyff at Cyff flume	1973-10-02	2008-12-31	0.84
55033	Wye at Cefn Brwyn	1968-10-01	2008-12-31	0.23
54097	Hore at Upper Hore flume	1985-11-08	2008-12-31	1.91
54092	Hore at Hore flume	1973-10-01	2008-12-31	0.32
54091	Severn at Hafren flume	1976-01-01	2008-12-31	0.15
54090	Tanllwyth at Tanllwyth flume	1973-10-01	2008-12-31	0.48
54022	Severn at Plynlimon flume	1971-10-01	2008-12-31	0.14

Table 4.3: Summary of available hourly datasets from automatic weather stations.

ID	SiteName	Start of record	End of record	Variable	% missing values
cefn	Cefn Brwyn	1975-04-02	2003-06-10	SR	9.27
				NR	10.84
				TW	11.22
				TD	10.17
				WS	9.88
				WD	10.34
				P	8.99
eistg	Eisteddfa Gurig	1976-01-07	2010-12-31	SR	13.93
				NR	11.26
				TW	10.87
				TD	11.09
				WS	10.19
				WD	10.53
				P	14.46
gwy	Gwy	1999-07-06	2010-12-31	SR	0.13
				NR	0.06
				TW	5.03
				TD	0.13
				WS	1.99
				WD	3.72
				P	7.51
carreg	Carreg Wen	1976-01-03	2010-12-31	SR	8.25
				NR	10.66
				TW	8.03
				TD	7.84
				WS	9.94
				WD	7.43
				P	8.06
tan	Tanllwyth	1975-04-25	2010-12-31	SR	1.65
				NR	12.84
				TW	9.49
				TD	8.38
				WS	12.43
				WD	7.67
				P	10.17

4.2.2 Deriving potential evapotranspiration from weather variables

The Penman-Monteith method was used to generate potential evapotranspiration values from weather information. The methodology used to transform hourly time series is suggested by the Food and Agriculture Organization (FAO) of the United Nations (Allen et al., 1998).

The FAO method assumes grass as reference crop, therefore the potential evapotranspiration is equal to the reference evapotranspiration [mm/h]. This can be calculated as follows:

$$E = \frac{0.408\Delta(NR - G) + \gamma \frac{37}{T+273} WS(e^o - e^a)}{\Delta + \gamma(1 + 0.34WS)} \tag{4.1}$$

In equation 4.1, NR is the net radiation at the grass surface measured in MJ/(m^2 hour). As the observed net radiation is measured in W/m^2, the following conversion was applied:

$$1[\frac{MJ}{m^2 hour}] = 1000000/3600[\frac{J}{m^2 s}] = 277.778[\frac{W}{m^2}]. \tag{4.2}$$

G is the soil heat flux density measured in MJ/(m^2 hour). Hourly G can be approximated during daylight periods as

$$Gday = 0.1NR, \tag{4.3}$$

and during night time periods as

$$Gnight = 0.5NR. \tag{4.4}$$

T is the mean hourly air temperature, which can be approximated to the dry bulb temperature TD [$^\circ C$].

Δ is the saturation slope vapour pressure curve at T $[kPa/^{\circ}C]$.

$$\Delta = \frac{4098(0.6108 \exp \frac{17.27TD}{TD+237.3})}{(TD+237.3)^2} \tag{4.5}$$

Also, γ is the psychrometric constant measured in $kPa/^{\circ}C$ and available as tabulated values for different altitudes (z) (Allen et al., 1998, p. 214).

The saturation vapour pressure at air temperature T, called e^o, is measured in kPa:

$$e^o = 0.6108 \exp \frac{17.27 * TW}{TW + 237.3} \tag{4.6}$$

Finally, WS is the average hourly wind speed measured in m/s, while e^a is the average hourly actual vapour pressure measured in kPa:

$$e^a = e^o - \gamma(TD - TW) \tag{4.7}$$

However, FAO's is only one of the available methods to calculate the potential evapotranspiration. Depending on the granularity of the weather information and any missing inputs, other methods may be more appropriate to use. Grace and Quick (1988) and Lu et al. (2005) proved that the application of different methods may induce very large discrepancies in the PE estimate with serious consequences in decision making applications such as irrigation scheduling.

4.2.3 Areal averaging

Although areal averaging can be achieved using a number of interpolation methods, Singh and Chowdhury (1986) demonstrated that often simpler methods are preferable. In this work, precipitation and evapotranspiration for each subcatchment are averaged over the relative area using the Voronoi tessellation method (Voronoi, 1908), also known as the Theissen's method. This

method assumes that each weather station is representative of an area. This area can be drawn following a three-step procedure:

1. stations are connected to each other with a straight line segment,
2. a perpendicular line is drawn through the middle point of each segment,
3. these lines split the catchment in smaller polygons and the area of influence of station X is given by the polygon that contains such a station.

Figure 4.5 shows areas of influence, while tables 4.4 and 4.5 summarise the areal weights assigned to each station, based on which the average precipitation and potential evapotranspiration were calculated. In case one or more stations have missing values, an arithmetic mean from the remaining stations is calculated. This procedure reduces significantly the number of missing values.

Table 4.4: Wye at Cefn Brwyn catchment, areal weights from Voronoi polygons. The weights are calculated as the ratio between the area of a polygon and the total area of a subcatchment.

ID	**Name**	**cefn**	**eistg**	**gwy**	**carreg**
55034	Cyff at Cyff flume	0.30	0.60	0.10	0
55033	Wye at Gwy flume	0	0.48	0.38	0.14
55008	Wye at Cefn Brwyn	0.20	0.34	0.41	0.05

Table 4.5: Severn catchment, areal weights from Voronoi polygons. The weights are calculated as the ratio between the area of a polygon and the total area of a subcatchment.

ID	**Name**	**tan**	**gwy**	**carreg**
54097	Hore at Upper Hore flume	0.03	0.02	0.95
54092	Hore at Hore flume	0.32	0.12	0.56
54091	Severn at Hafren flume	0.21	0	0.79
54090	Tanllwyth at Tanllwyth flume	0.53	0	0.47
54022	Severn at Plynlimon flume	0.38	0.04	0.58

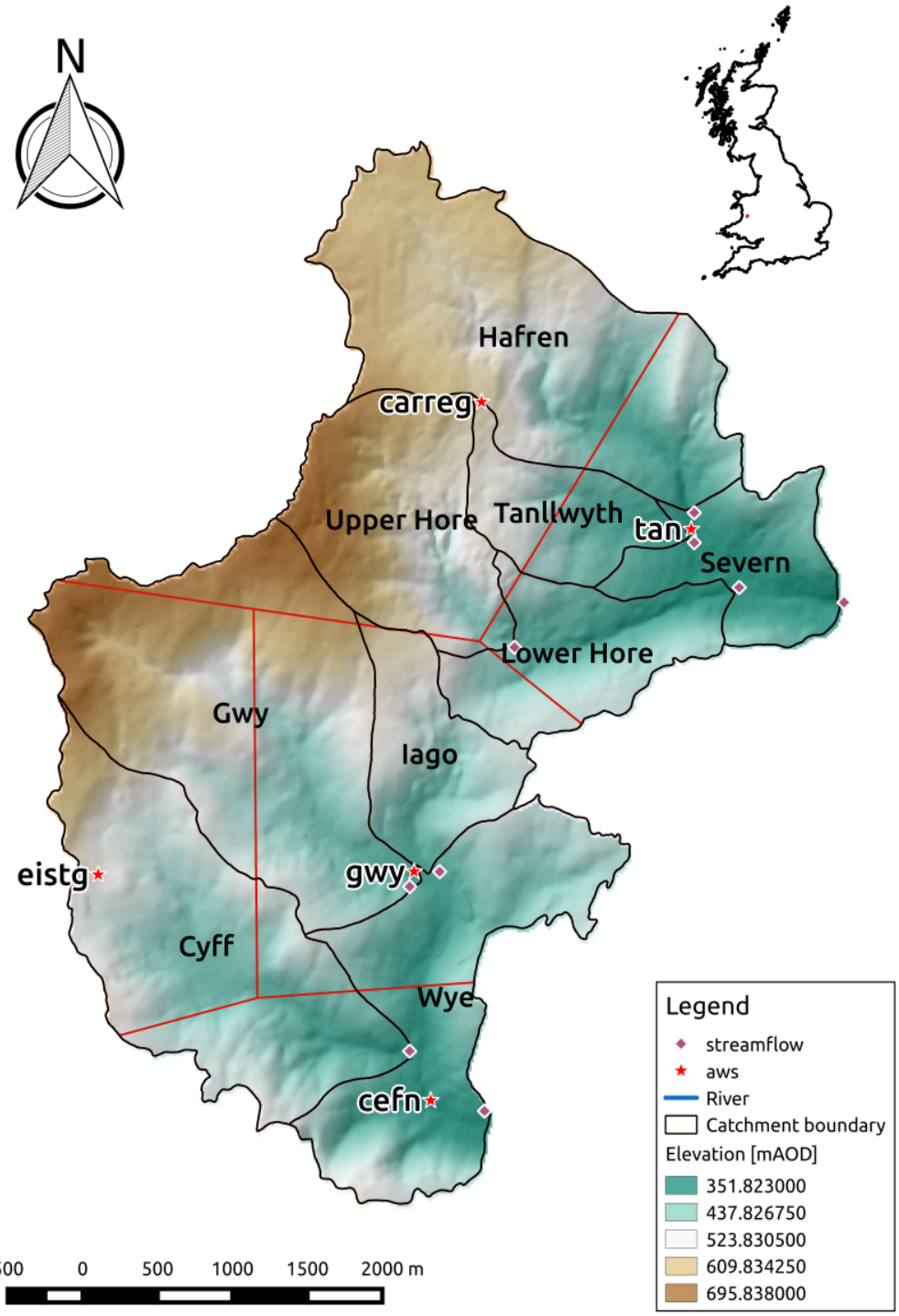

Figure 4.5: Voronoi polygons of the Plynlimon catchments. Each subcatchment is partitioned into regions (polygons) based on their closeness to a certain recording station. For instance, the Severn at Hafren flume (54091) is divided into 2 polygons: the upstream polygon delineates the area closer to the *carreg* weather station, the downstream polygon delineates the area closer to the *tan* weather station. The ratio between the area of a polygon and the total area of a subcatchment returns a weight (see Tables 4.4 and 4.5) that is used to calculate the rainfall contribution from each region.

4.2.4 Filling gaps and converting to common units

Remaining gaps (up to ten consecutive time steps) are infilled using a linear interpolation. Although quadratic and higher order functions could interpolate the ascending/descending limb of the hydrograph better than linear function, at low flows, they generate unrealistic negative values. Therefore, linear interpolation was used. As alternative methods arithmetic mean and infilling methods using donor stations could also be used.

Finally, time series are converted to mm/day. In order to be able to compare observed discharges and simulated values, the streamflow time series were first converted in depth of water per unit area (m/h, dividing the observed values by the relative catchment area), then converted to mm/day. Figure 4.6 shows discharge, precipitation and potential evapotranspiration time series for Severn at Plynlimon flume from the 6th October 1981 to the 2nd November 1981.

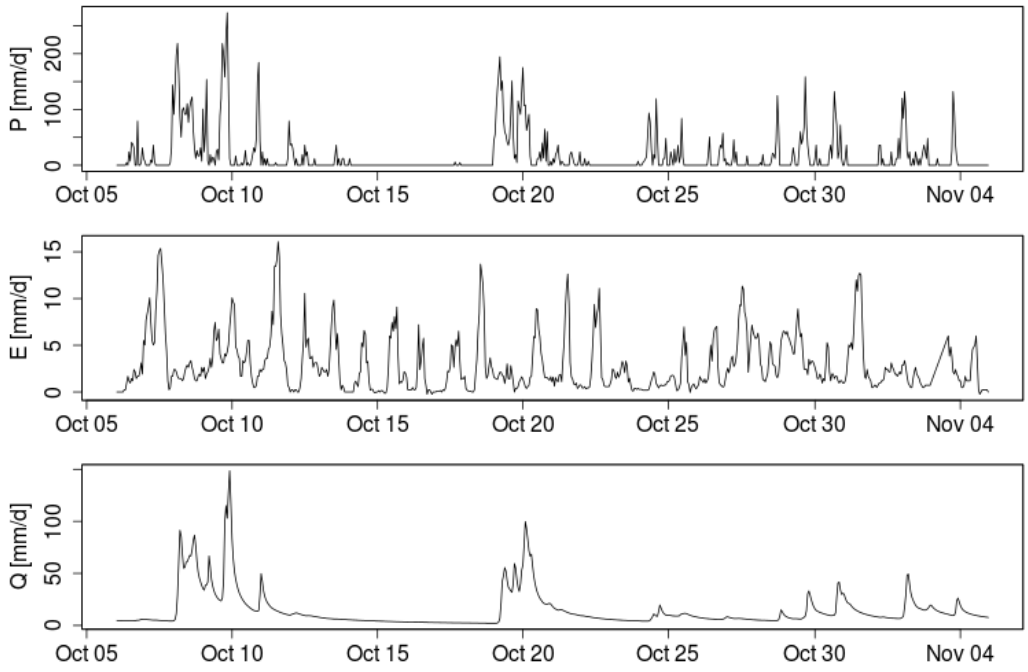

Figure 4.6: Manipulated time series for Severn at Plynlimon flume. The top panel shows precipitation records (P), the middle panel potential evapotranspiration (E) and the bottom panel streamflow discharge (Q). All the variables are measured in mm/d.

4.3 Concluding remarks

Two major catchments are identifiable in the Plynlimon area: the Severn at Plynlimon flume and the Wye at Cefn Brwyn. They are both highly instrumented and data is managed by the Centre for Ecology and Hydrology via the Information Gateway web service. The available information was collated and manipulated to convert the time series in consistent gap-free forcing inputs for the FUSE framework that will be used in the following chapters.

Data, at this location, is highly reliable thus considered suitable to investigate the variability of simulated hydrological responses using different model structure (see Chapters 5 and 6) because the noise due to input data errors is expected to be negligible.

Both catchments are characterized by steep profiles and responsive regime. The major difference between them is in the predominant land cover. The Severn at Plynlimon flume is mainly afforested while the Wye at Cefn Brwyn is under extensively grazed grassland. These characteristics make this location suitable to investigate the effects of land use changes on catchment flows (see Chapter 7).

Chapter 5

A data mining algorithm for automatic hydrological model selection and parameterisation

Hydrological modelling practice is characterised by a large degree of subjectivity, especially in the selection of model structures and parameter ranges. When making such decisions, a wide range of options exists and a particular choice is often hard to justify on the basis of model assumptions and knowledge about the hydrological system that is to be modelled. In this chapter, a novel Data Mining workflow is presented to facilitate an explicit and transparent approach to model structure selection, parameter identifiability and redundancy reduction combining Pareto filtering, clustering techniques and time series matching algorithms in a multi-objective framework. The result of the algorithm is a set of suggested model configurations. The selection is less modeller dependent and more consistent and reproducible than traditional approaches. A series of experiments are presented, using the FUSE modelling framework as model inventory. The approach is tested using a synthetic dataset to validate the results against the known model configuration and investigate the sensitivity of the proposed approach to factors such as the size of the sampled space and parameter variability.

5.1 Introduction

Hydrological modelling on the catchment scale, although based on scientific procedures and techniques, still remains subjective and is strongly influenced by the modelling experience and hydrological judgement of the investigator for the selection of the most suitable model structures, parameter ranges and objective functions (Pechlivanidis et al., 2011).

Given the high variability of catchment properties and processes, it would be ideal if modellers could construct tailored model structures that correspond as closely as possible to their expert perception of the local hydrology. In this context, Beven (2001*b*) suggests a standard theoretical approach to the modelling process by defining a number of consecutive steps. Modellers make use of local knowledge, experience and expertise to translate their perception of the hydrological processes occurring in a catchment (the so-called "perceptual model") into a "conceptual model" structure: a set of mathematical equations used to simulate the hydrological behaviour at a given scale. The perceptual model evolves into the conceptual model through a series of assumptions and simplifications, thus potentially increasing the expected uncertainty. When the equations are translated into computer code, these uncertainties are inflated even further by numerical approximation. At this stage, the model becomes "procedural" and can finally be calibrated and validated.

In practice, modellers do not produce tailored model structures on a project-by-project basis but often refer to inventories of existing models, such as those suggested by Nemeč (1993) and Singh et al. (2006). Modellers' preferences and familiarity for specific models can influence model selection and evaluation but typically the most important driving force remains the scope of the project (Cunderlik, 2003). One model is usually preferred over another based on criteria such as required model outputs, hydrological processes that need to be modelled, input data availability and costs (World Meteorological Organization, 1975; Nemeč, 1993; Klok et al., 2001). A model selection based primarily on such pragmatic criteria has important consequences for the interpretation of model simulations and can lead to "disappointing results",

as recognised by (CRCCH, 2005) who suggest guidelines to select models as trade-off between model complexity and predictive performance. Very often the lack of benchmarks for model adequacy causes the need to evaluate models in relative terms (Oreskes et al., 1994; Schaefli and Gupta, 2007). A model structure, for instance, can be compared to itself under different scenarios, including variable initial/boundary conditions, and climatic/topographic conditions. Andréassian et al. (2009) suggests a focus on model improvement using "crash tests for a standardised evaluation of hydrological models", while Reusser and Zehe (2011) use temporal disaggregation of model performance to identify model deficiencies. However, there is a large body of experimental evidence that suggests the existence of limits to the single model structure improvement. This is due to the difficulty of describing all the aspects of the response hydrograph by simply acting on the parameter space (Gupta et al., 1998; Wagener et al., 2001).

An alternative to improving a single model structure is to work with multiple model components and objective functions, which can reveal useful insights into structural adequacy (Wagener et al., 2001). Abbott et al. (1986) give one of the first examples of multi-model approach in the introduction to the European Hydrological System "SHE". This distributed model was developed to flexibly adapt its structure to more or less information, according to the availability of data. Leavesley et al. (1996) introduce the concept of "Modular Modeling System" expanded over many years to collect functions taking into account a "variety of constraints that include the types of data available and the spatial and temporal scales of application" (Leavesley et al., 2002).

The model flexibility pioneered by Abbott et al. (1986) and Leavesley et al. (1996, 2002) was taken to a next level with the development of modelling frameworks (and toolboxes) such as the Rainfall-Runoff Modelling Toolbox (RRMT) (Wagener et al., 2001) and HYDROMAD (Andrews et al., 2011). In this context, the Framework for Understanding Structural Error (FUSE) (Clark et al., 2008) deserves a particular mention because it contains an unprecedented number of model structures and has led to the introduction of the highly customisable SUPERFLEX representation (Fenicia et al., 2008, 2011).

Beven (2000*b*) highlighted the problem that every catchment is unique in relation to model representation of flow processes ("uniqueness of places"). If there were a model that perfectly represented these processes and a set of observations that could describe the full "distribution of characteristics that may be important in controlling storm runoff generation" (Beven, 2001*a*), model parameters could be deterministically identifiable. In the real world, neither the perfect model nor the perfect set of observations exist. Therefore, given a model, there could be several "optimal" parameter sets and, given multiple models, there would be even more "optimal" model configurations[1], which Beven (2006) calls "equifinality problem".

Based on the availability of multi-model frameworks, many suggest tailor-made solutions combining model components and existing libraries (Buytaert et al., 2008; David et al., 2002; Castronova, Goodall and Ercan, 2013; Castronova, Goodall and Elag, 2013; Fenicia et al., 2011; Ames et al., 2012). Systems like CUAHSI HIS (Ames et al., 2012) integrate smart ways of combining models and heterogeneous data sources (Beran and Piasecki, 2009) but still rely on the modeller's input for model selection and parameter definition. Similarly, the Australian "eWater CRC toolkit" (Jordan et al., 2007) provides various water and catchment management utility tools and a series of documents, the "Model Choice series", to help modellers decide the right catchment model for their needs. Others propose flexible approaches to support a wide range of hydrological behaviours. Argent et al. (2005), for instance, introduced the "E2 framework" which is based on hierarchical model selection, while Young (1998) proposed a data-based mechanistic (DBM) modelling approach. The latter uses statistical procedures to identify the most appropriate structure from a class of models.

Unlike other approaches, this work builds upon the schematic modelling process outlined by Beven (2001*b*) to provide a transparent, data-based procedure for the selection of an acceptable ensemble of model configurations for a given catchment using locally available data. The selection is made using a combination of machine learning techniques, multiple model structures and multiple objective functions. This algorithm is tested on a sample inventory, but the

[1] A model configuration is defined here as the combination of one model structure and one parameter set.

methodology can easily be adapted to others, including more sophisticated model structures.

5.2 Method

This work presents an Automatic Model Configuration Algorithm, called AMCA hereafter, based on the selection process suggested by Beven (2001*b*). The process is illustrated in Figure 5.1 and consists of the following 4 steps:

1. *Step I: defining the procedural model space.* In line with a rejectionist modelling approach, modellers should initially consider any model structure that may represent a plausible conceptualisation and define parameters in the most conservative range. The output of this step is the ensemble equivalent of Beven's procedural model, for which a further reduction of model structures is only made on practical considerations, due, for example, to the availability of code implementations and compatibility with the available data.

2. *Step II: generating the configuration space.* Performance indices are defined to compare simulated results to the observations. A number of parameter sets are uniformly sampled and Monte Carlo simulations are generated for each configuration[2].

3. *Step III: filtering the configuration space.* This step consists of a Data Mining procedure to progressively reduce the model structure and parameter spaces. The algorithm takes into account model performances, Pareto efficiency and structural redundancies. It is designed to make model selection and evaluation more transparent, and amenable to comparative studies, sensitivity analysis and further refinement in general.

4. *Evaluating the simulated results.* The output of the algorithm is an ensemble of configurations for which accuracy, precision and statistical reliability are calculated.

[2]A model configuration is defined here as the combination of a model structure and a parameter set.

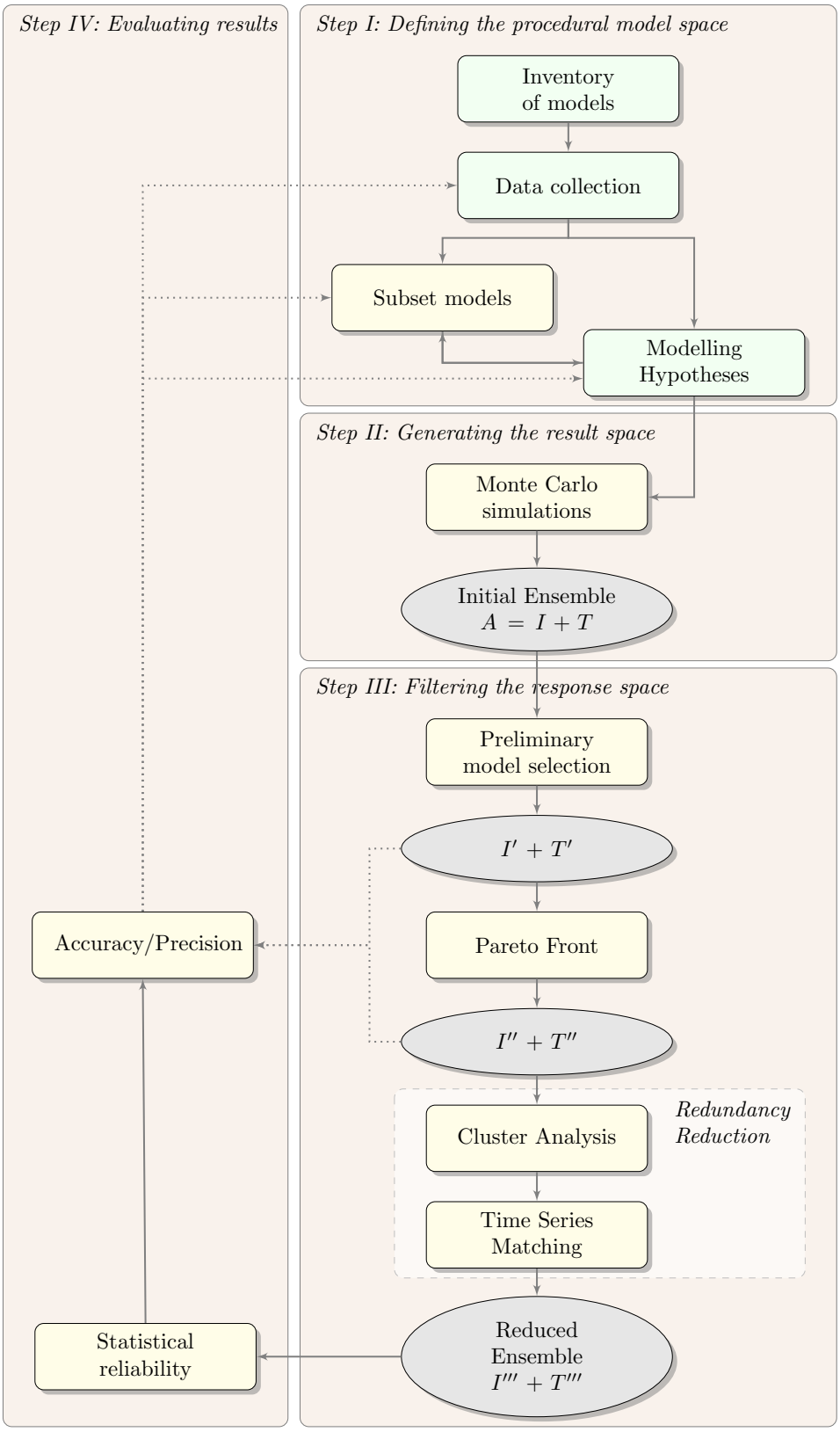

Figure 5.1: The Automatic Model Configuration Algorithm (AMCA). Green rectangles are user inputs, yellow rectangles are automatic procedures. Ellipses are generated data objects containing: MPIs (I), and simulated discharges (T). A is the 3-dimensional array containing both I and T.

5.2.1 Step I: Defining the procedural model space

In theory, the model space can be imagined as made of an infinite number of model structures. In practice, only a subset of these models are available. This is either because not all perceptual models can be formulated as conceptual models and implemented as procedural models, or for more practical reasons related to incompatibilities of computer platforms and licensing issues. The subset of available models is called here *Inventory of models*.

A first selection criterion for procedural modelling is data availability. After *data collection*, the inventory of models may be reduced to accommodate only structures compatible with the available data at the location under study (e.g. if only lumped information is available, distributed models can usually be discarded). In this study, the multi-model framework FUSE developed by Clark et al. (2008), and extensively described in Chapter 3, is adopted as model inventory. FUSE's models require as inputs hydrological and meteorological observations lumped over the region of interest (precipitation plus snowmelt and potential evapotranspiration, expressed in mm/day regardless of the recording time step) and return the simulated streamflow discharge at the outlet. Any FUSE model can be used in regions where this information is available as direct or derived observation. Therefore, the inventory of models is not subsetted at this stage.

Finally, some *modelling hypotheses* are defined. These are related to the model's initial conditions, the length of the warmup period[3], the number of parameter sets to take into account and the type of model performance measures to adopt.

Model Performance Indices

The goodness of fit between simulated and observed discharges is usually evaluated via measures of model performance, for which a wide variety is available in literature (Legates and McCabe, 1999; Krause et al., 2005; Dawson et al., 2007; Reusser et al., 2009). The choice of

[3]The warmup period is defined here as the number of time steps needed by the model to eliminate the bias generated by initial conditions.

performance indices should relate to aspects of the simulated hydrograph that can be controlled by the model. For example, in flood related projects is very important to match the timing and magnitude of observed and simulated hydrograph peaks. Similarly, for drought related projects, it is important to match the volumes of water available over a given period. By contrast, an index that measures the amount of snow that becomes runoff would not be useful using FUSE because its models have no power to change this proportion.

A list of 5 Model Performance Indices (MPIs) are used here:

1. the shift resulting in the maximum cross correlation of the observed and simulated time series. According to the implementation used by Reusser (2014), this has a bounded range $[-36, +36]$ days. The absolute value of this shift is the index *LAGTIME* used hereafter to measure timing errors and varies in the range $[0, 36]$ days.

2. the module of the mean of the difference between the observed and simulated discharge (Van Den Boogaart et al., 2014; Jachner et al., 2007), with a range of variability is $[0, +\infty[$. This index is called *MAE* and is used to detect volume errors.

3. the Nash-Sutcliffe efficiency, this objective function is particularly sensitive to high flows (Nash and Sutcliffe, 1970). The range of variability goes from $[-\infty, 1[$, however, the index *NSHF* considered hereafter is calculated as 1 minus the efficiency, which varies in the range $[0, +\infty[$.

4. the Nash-Sutcliffe efficiency calculated using the logarithm of observed and simulated discharges and it is particularly sensitive to low flows (Krause et al., 2005). The index *NSLF* considered hereafter is calculated as 1 minus the efficiency, which varies in the range $[0, +\infty[$.

5. Rainfall-Runoff coefficient, defined as the volume of runoff divided by the corresponding rainfall (Sawicz et al., 2011). The *RR* index considered hereafter is the difference between observed and simulated RR, varying in the range $[0, +\infty[$. It is used to match the percentage of runoff generated relative to incoming rainfall.

All the indices are rescaled to an interval between 0 and 1, with 0 being the best performance and 1 the worst.

5.2.2 Step II: Generating the result space

Model simulations

Monte Carlo experiments are employed to investigate how model performances are affected by different combinations of parameter sets and model structures. A number of parameter sets are generated using the Latin Hypercube Sampling (LHS) method (McKay et al., 1979).

The LHS method is used here only as screening tool. This is based on the hypothesis that a limited number of parameter samples could be enough to highlight the model structures that are more suitable to simulate a given response. Once these model structures are identified, however, modellers are required to carry out more rigorous searches of the parameter space using, for instance, optimization algorithms.

The Initial Ensemble

The generated result space consists of a three-dimensional array A containing the simulated time series for each combination of model structure and parameter set plus the corresponding set of MPIs. In order to apply the Data Mining algorithm described in the next section and produce graphical representations of the result space, the array A is split into two sub arrays: T and I. T contains one simulated discharge time series for each parameter set and model structure. I contains only the MPIs corresponding to each realisation.

The most conservative result in terms of simulated discharge, is obtained by combining all the simulated time series contained in T. This composite result will hereafter be referred to as the Initial Ensemble (IE). This is expected to be characterised by a wide spread of the simulated

flows, due to the large variability of structures and parameters. If observations do not fall within the IE's bounds, the modeller should assess whether this is due to input errors or the relevant model configurations are missing from the inventory.

5.2.3 Step III: Filtering the ensemble response space

This step aims to filter out the realisations which are clearly unsuitable to predict the given response and those which are redundant. It consists of 3 tasks:

1. Preliminary Selection;
2. Identification of non-dominated points (Pareto Front);
3. Redundancy Reduction, which can be subdivided into
   (a) Cluster Analysis and
   (b) Time Series Matching.

The majority of the filtering procedure makes use of the array I, while array T is used only in the time series matching algorithm.

Preliminary Selection

Amongst the numerous FUSE model structures, it is expected that only some of them contain parameterisations of processes that are relevant for the catchment under study. Therefore some of them will be more suitable than others in reproducing the observed behaviour. In order to identify the most suitable model structures with regards to a certain MPI, the variability of the index over the parameter space can be analysed for each model structure and quantified in terms of a basic statistics, the median. The median of each MPI is calculated for each model structure across the parameter space. It is assumed that a set of suitable models can be identified

by looking for a median lower than a certain threshold. Such approach is vaguely similar to that used by Hornberger and Spear (1981) and within the Generalised Likelihood Uncertainty Estimation by Beven and Binley (1992). In these methodologies, the threshold is set by the modeller making the result highly subjective.

The algorithm presented here, however, has an internal mechanism to automatically set a threshold which consists of the following steps:

- define a sequence of thresholds, e.g. from 0.1 to 1 with 0.1 step,
- for each model structure calculate the median of each MPI across all the parameter sets,
- for each threshold, the model structures with median below the threshold for all the MPIs are selected as suitable and collected in a vector called *nS*,
- in case all the performances are above the threshold, the entire set of models is considered suitable for the next step, at the cost of a slower processing,
- the auto-generated threshold is the one that correspond to the minimum length of *nS* (red dot in Figure 5.2).

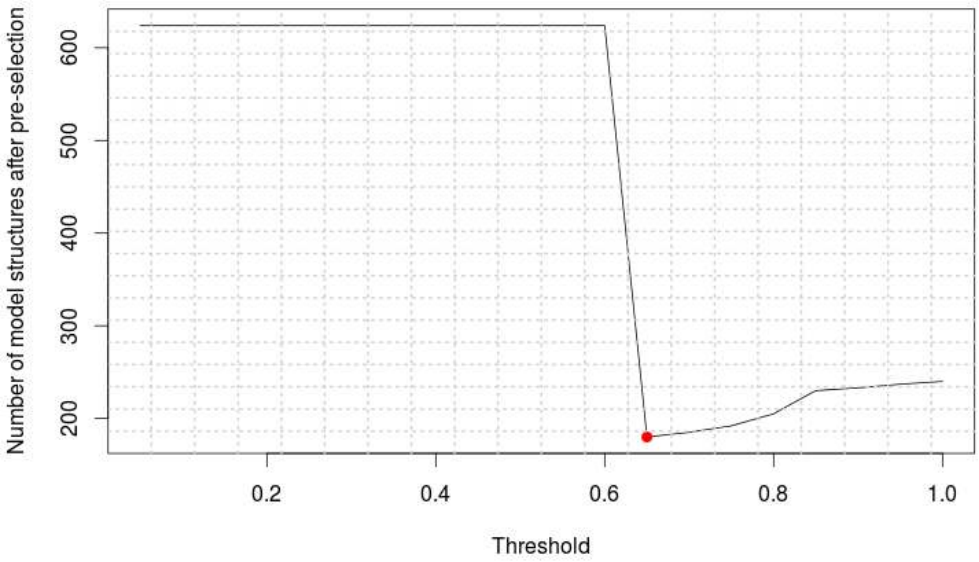

Figure 5.2: Various thresholds (on the x-axis) are plotted against the number of selected model structures (on the y-axis) calculated in the pre-selection step. The red dot shows the threshold selected by the algorithm.

It should be noted that this method only identifies the overall behaviour of a model structure over the parameter space but does not take into consideration isolated well-performing realisations. The array containing only the pre-selected models is called I' and its related ensemble T'.

Pareto Front

Given the (reduced) number of combinations of model structures, the best realisation minimises a given index (maximises performance). However, this may not correspond to the best possible scenario with respect to other indices.

To obtain a set of optimal solutions over a multi-dimensional performance space, the Pareto Front is extracted using the related function in the R package "emoa" (Mersmann, 2015). This leads to the partitioning of I' into dominated (to be discarded) and non-dominated realisations I''. T'' is the ensemble corresponding to I''.

The filtering efficiency of a Pareto front algorithm is known to decrease with an increase in the

number of performance indices taken into account (Corne and Knowles, 2007). This means there will probably be groups of realisations characterised by similar combinations of performance indices and that the Pareto front cannot discern amongst them. The redundancy reduction task is set up to identify and remove redundant realisations from the Pareto Front utilising a clustering technique to group them and a time series matching procedure to filter one representative configuration from each cluster.

Redundancy Reduction

The clustering is based on Self Organising Maps (SOM), non-linear representations of multi-dimensional data already widely used in hydrological modelling (Herbst et al., 2008; Reusser et al., 2009; Ley et al., 2011; Toth, 2013). A SOM is represented as a map of a pre-defined size showing interconnected nodes (see examples in Figure 5.3). Each node represents a cluster centroid. The dimension of the node is proportional to the number of elements contained in the cluster and the distance between nodes is a measure of the associated weight (Kohonen et al., 1996; Yan, 2010). It is assumed a Gaussian neighbourhood function and rectangular topology, whose dimensions are defined based on the total number of realisations selected up to this point. For instance, if 50 realisations have been selected, the first dimension is the square root of 50 (rounded up to the nearest integer). The second dimension is the square root of 50 divided by the first dimension and rounded up. In this way, if there are no similarities, realisations will not be forced to cluster.

The Pareto front is used as input of the SOM algorithm, and the weight vector is initialised with the linear grid obtained from the first and second principle component directions. When the SOMs are trained, the realisations are grouped in a number of clusters which does not necessarily equal the number of nodes (some nodes can be empty). Similarities between time series belonging to a certain cluster cannot be evaluated using MPIs as clustering, by definition, tries to minimise within-cluster variability. Therefore, observations and simulations are compared using an additional similarity score, called the Dynamic Time Warping distance.

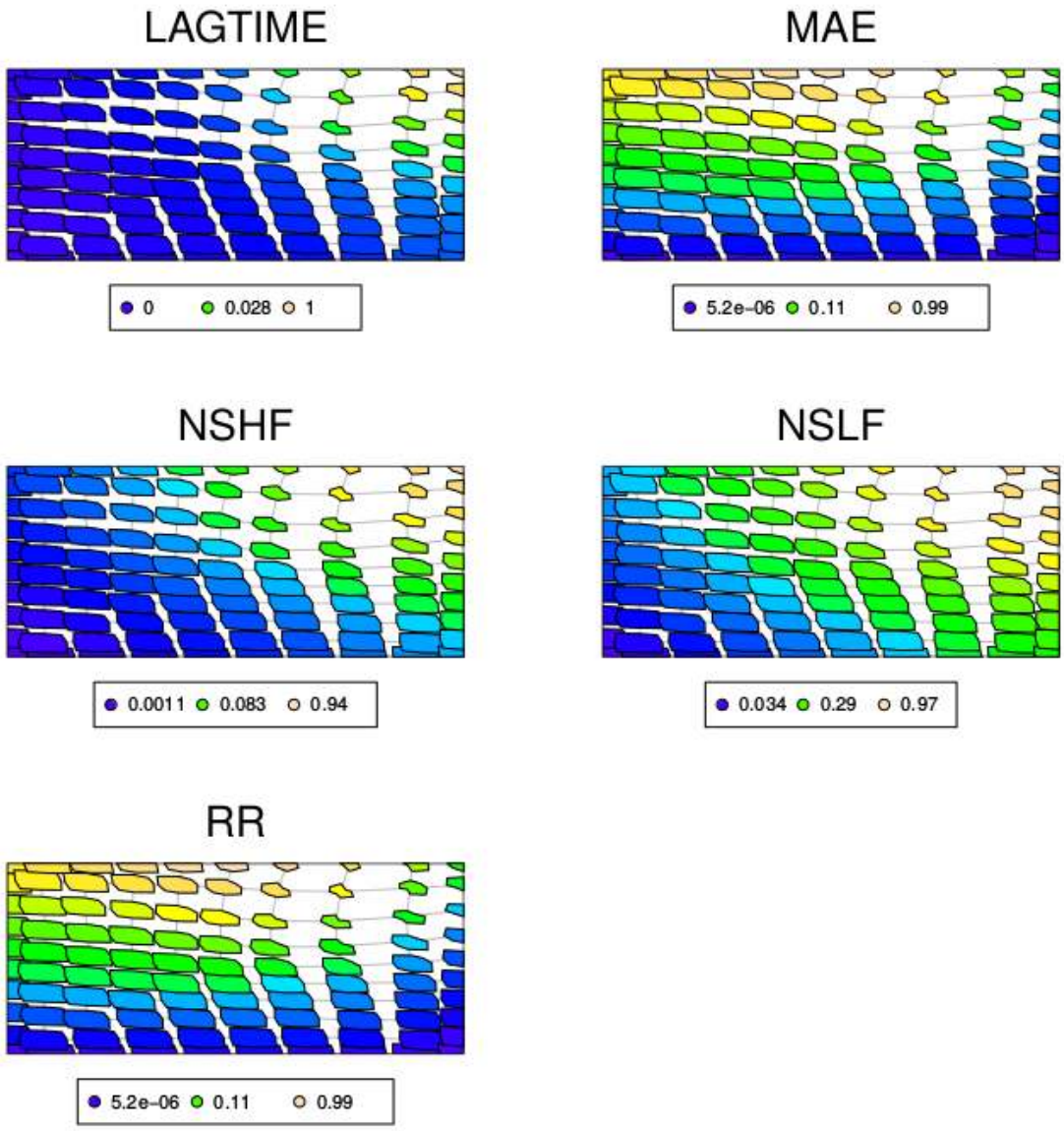

Figure 5.3: Example of Self Organizing Maps for 5 MPIs and dimension 10 x 10. The nodes are colour coded based on the performance of the realisations contained. The range goes from 0 (blue, best performances) to 1 (yellow, worst performances). There is no particular meaning associated with the x and y axes.

The time series matching algorithm called Dynamic Time Warping (DTW) was developed by Sakoe and Chiba (1978) for speech recognition to align two time series by warping iteratively the time axis until a suitable metrics is satisfied. The example matrix in Figure 5.4 shows two time series: A and B. These time series are made of a discrete number of observations. At the first time step (starting from the bottom left cell), the distances between the first point of A and

all the points of B are calculated. The element in the matrix that corresponds to the minimum distance is highlighted with a red dot. The same procedure is repeated for the remaining time steps and the sequence of dots determines the minimum distance path (or warping path). Being based on matrix operations, this method becomes computationally very expensive for long time series.

In the AMCA algorithm, time series A and B are the observed and simulated streamflow respectively. DTW assumes that simulations are non-linear time-stretched modifications of the observations and the similarity score depends on how the simulation is stretched in respect to the observation (Berndt and Clifford, 1994). For this reason the similarity score is expressed in terms of "distance". The AMCA only compares time series with the same length but in case time series of different length need to be compared, a normalised distance should be adopted. This is because "longer timeseries have naturally higher distances, making comparisons impossible" (Giorgino, 2009; Tormene et al., 2009). From each cluster identified by SOMs, only one member is selected, the one with the lowest DTW distance.

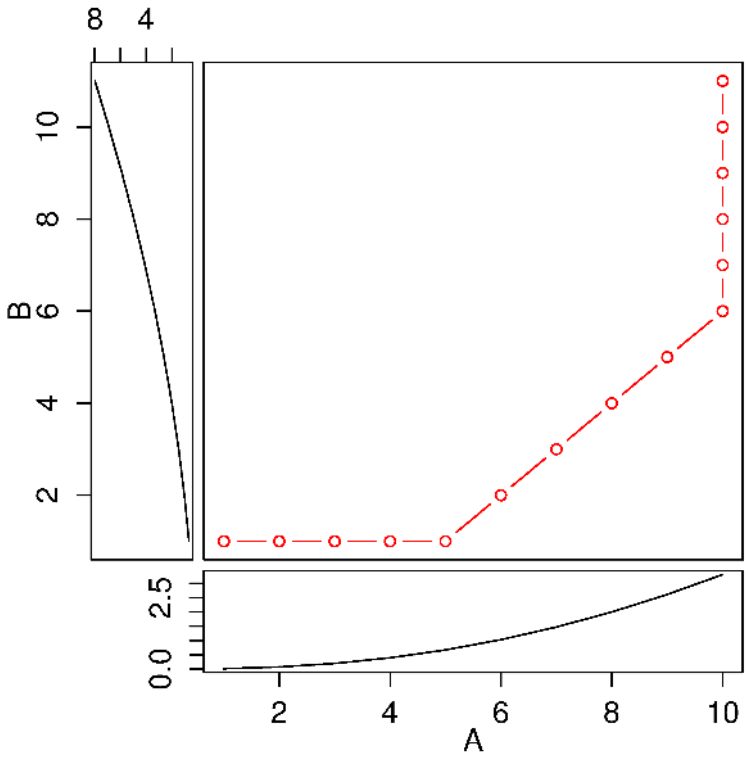

Figure 5.4: Schematic example of Dynamic Time Warping path. A and B are two numeric vectors

The result of the redundancy reduction task is a set of realisations (I''' and T'''), which composite result is referred to as the Reduced Ensemble (RE). Although the techniques illustrated here are widely used, the main novelty of this work is to combine them in a reproducible workflow, which have not been previously done.

5.2.4 Step IV: Evaluating the simulated results

The result of the suggested procedure is assessed taking the *Initial Ensemble* as baseline and scoring the *Reduced Ensemble* based on: the accuracy, precision and statistical reliability.

Although, there are more complex ways to assess the *accuracy* of an ensemble coverage (Garaud and Mallet, 2011), here this is calculated as the percentage of time steps in which the observations fall into the predicted interval. According to Christoffersen (1998) an indicator

variable can be defined as follows:

$$c_t = \begin{cases} 1 & \text{if } d_t \in [L_t(q), U_t(q)] \\ 0 & \text{if } d_t \notin [L_t(q), U_t(q)] \end{cases} \tag{5.1}$$

where $L_t(q)$ and $U_t(q)$ are the lower and upper bounds of the ensemble, d_t is the observation and q is the coverage probability. The actual coverage gives a measure of the accuracy of the interval (expressed in percentage):

$$\text{Actual Coverage} = \frac{1}{N}\sum_{t=1}^{N} c_t \tag{5.2}$$

It is desirable for an ensemble result to be accurate but also precise (as close to a deterministic output as possible). The reason for this is two-fold. On one hand, the "best (deterministic) estimate" is often a requirement for many practical applications such as flood forecasting systems (Vaughan and McIntyre, 2012). On the other hand, the frequency and distribution of the ensemble's parameters and model components could be highly informative (if there is a good level of consistency). Yadav et al. (2007) suggest to measure precision as the relative spread between two ensembles, calculated as the average distance between U_t and L_t for the *Reduced Ensemble* divided by the average distance between U_t and L_t for the *Initial Ensemble*. This measure is expressed in percentage with 0% meaning no reduction compared to the *Initial Ensemble* and 100% meaning the *Reduced Ensemble* is a line.

Laio and Tamea (2007) suggest to evaluate the statistical reliability of an ensemble using a graphical approach known as Quantile-Quantile plots. In this study a more-quantitive approach is adopted, summarising the "overall statistical reliability" (α) based on the shape and distance of the result from the bisector of the plot, as suggested by Renard et al. (2010).

5.2.5 Experimental setup

The procedure is tested on one hydrological year of hourly data at the Severn at Plynlimon flume catchment in the Plynlimon area, United Kindgom (see Chapter 4). Data comprises of precipitation, potential evapotranspiration and streamflow discharge time series recorder between October 2005 and September 2006. The period between October 2004 and September 2005 is used to warmup the models. This period was chosen due to the low number of missing values.

A synthetic experiment is set up to analyse the efficiency of every intermediate step. The synthetic dataset was generated using the forcing inputs described above along with model structure and parameters listed in Table 5.1.

The experiments include a sensitivity analysis investigating how the *Reduced Ensemble*'s performances are affected by changes in the following factors:

- Size of the sampled parameter space,
- Routing parameter range,
- Preliminary model selection step,
- Warmup period.

5.3 Results

5.3.1 Stepping in the algorithm using synthetic data

In order to clarify the implications of each part of the workflow, this section shows the outcome of each step of the AMCA algorithm applied to the synthetic dataset. The results are summarised in Table 5.2.

Table 5.1: Model structure (first eight lines) and parameters (remaining 24 lines) used to generate the synthetic dataset. Parameters with no values are not used by the selected model structure.

Description	**Name**	**Value**
Model structure IDs	mid	55
Upper layer architecture made of tension and free storages	tension1_1	22
Lower layer architecture made of a reservoir size characterised by a unlimited power recession	unlimpow_2	34
Topmodel-like Runoff parameterisation	tmdl_param	43
Percolation scheme with water availability from field capacity to saturation	perc_f2sat	51
Sequential evaporation scheme	sequential	62
Interflow denied	intflwnome	71
Routing allowed	rout_gamma	82
Additive rainfall error (mm/day)	rferr_add	0
Multiplicative rainfall error (-)	rferr_mlt	-
Fraction of tension storage in recharge zone (-)	frchzne	-
Fraction total storage as tension storage (-)	fracten	0.57
Maximum total storage in upper soil layer (mm)	maxwatr_1	220
Fraction of percolation to tension storage in the lower layer (-)	percfrac	-
Fraction of storage in the first baseflow reservoir (-)	fprimqb	-
Baseflow depletion rate in the first reservoir (day^{-1})	qbrate_2a	-
Baseflow depletion rate in the second reservoir (day^{-1})	qbrate_2b	-
Baseflow depletion rate (day^{-1})	qb_prms	-
Maximum total storage in lower soil layer (mm)	maxwatr_2	3385
Baseflow rate (mm/day)	baserte	95
Fraction of roots in the upper layer (-)	rtfrac1	-
Percolation rate (mm/day)	percrte	41
Percolation exponent (-)	percexp	17
Sacramento model percolation multiplier for dry soil layer (-)	sacpmlt	-
Sacramento model percolation exponent for dry soil layer (-)	sacpexp	-
Interflow rate (mm/day)	iflwrte	-
ARNO/VIC ”b” exponent (-)	axv_bexp	-
Maximum saturated area (-)	sareamax	-
Mean value of the log-transformed topographic index (m)	loglamb	5.54
Shape parameter for the topo index gamma distribution (-)	tishape	2.19
Baseflow exponent (-)	qb_powr	3.53
Time delay ($days$)	timedelay	2.7

At step I, the algorithm uses the full set of FUSE's model structures (1248) as inventory. Based on the available data, there is no need to subset the initial model inventory while the modelling hypotheses are based on the default FUSE settings listed below:

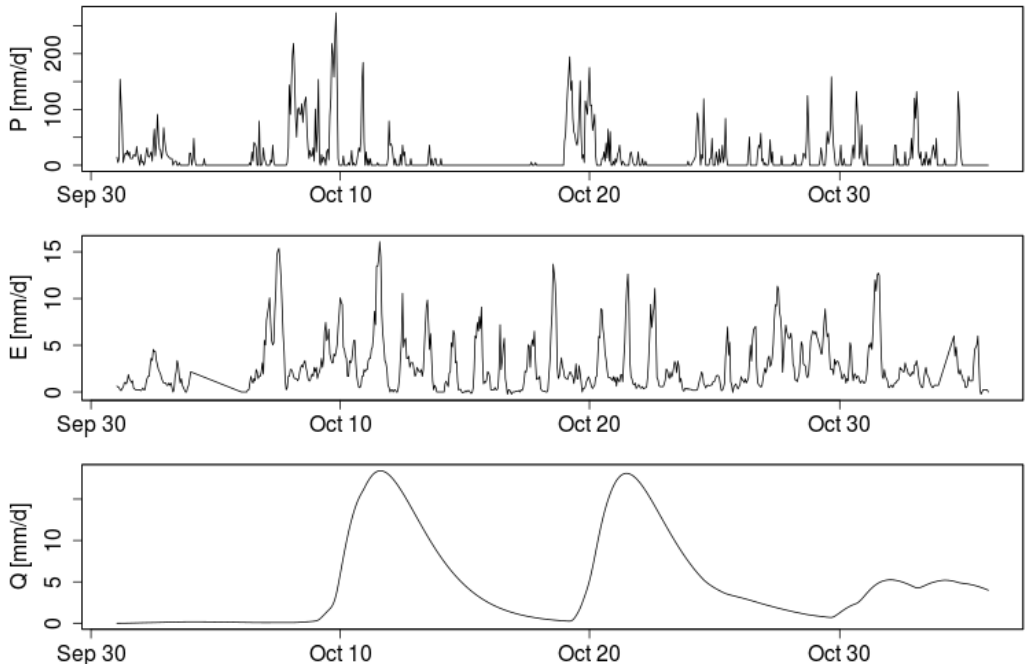

Figure 5.5: Synthetic time series of hourly precipitation (P in the top panel), potential evapotranspiration (E in the middle panel) and streamflow discharge (Q in the bottom panel). All the units are in mm/day.

- The extent of the parameter space is set to the default ranges suggested by Clark et al. (2011).

- 10000 parameter sets are sampled using the Latin Hypercube Sampling method. The same hypercube is used with all models.

At step II, the result space is generated by running each of the 1248 models using the same 10000 sampled parameter sets. The simulated discharges are collected in the array T while MPIs are collected in the array I. Due to the large number of simulations taken into account at this stage (12,480,000), there is a plethora of factors (e.g. sensitivity to particular model components and parameters) which effects overlap and impede a clear distinction between suitable and unsuitable model structures. For this reason, the filtering process is designed to act in a sequential fashion. At first, it performs a screening of inputs to remove equivalent settings. This happens, for instance, when the rainfall error is not included in the inference and the additive and multiplicative error factors are set to 0 and 1 respectively. If this initial screening is not

performed, the result space is populated with pairs of identical performances. The algorithm, therefore, checks whether the rainfall errors are set to the default values. If so, model structures with multiplicative rainfall errors are automatically filtered out reducing the model space from 1248 to 624 model structures.

The preliminary model selection in the filtering process (step III in Figure 5.1), sets an optimal performance threshold using simple statistics and set theory operations to summarise the suitability of each model to reproduce the desired response. The array I can be sliced along the dimension representing an MPI to extract and inspect the resulting 2D array, as in Figure 5.6.

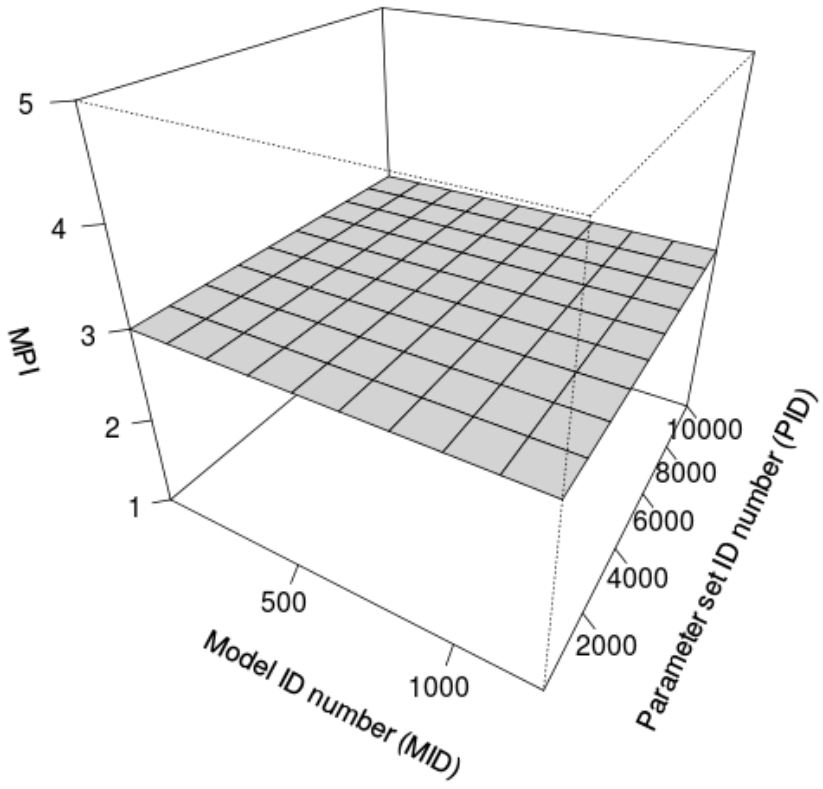

Figure 5.6: Array I, sliced along one of the axes to extract 2D array containing the NSHF index. MID is the model ID number, PID is the parameter set ID number and MPI is the Model Performance Index.

Figure 5.7 shows a 2-dimensional representation of the array I sliced along each MPI, which gives a first insight into the variability of performances across model structures and parameter sets. This plot is divided in 5 panels (one for each MPI). Each panel shows 1248 FUSE models on the x-axis (labelled by the original FUSE Model ID number, MID) and the first 1000 param-

eter sets on the y-axis (labelled by the parameter set ID number, PID). Plotting only a portion of the result space is necessary for practical reasons, however the overall pattern is only negligibly different.

Each cell is colour coded depending on the value of the MPI, which is represented by a gradient of colours going from green (best performance) to red (worst performance). Each column represents the performance of a model structure, across all the sampled parameter sets, and each row represents the performance of a parameter set, across all the model structures. Prevalence of red on vertical bands suggests that the correspondent model structure is not likely to be able to simulate the given response, and vice-versa in the case of prevalence of green. Similarly, inspecting horizontal bands, can help identifying problems with parameter ranges.

In this setting, the need for a multi-objective evaluation framework becomes immediately evident. The prevalence of green in the panels related to NSHF and NSLF suggests that the range of magnitude of the observations are captured by the vast majority of model-parameter set combinations, as over 98.7% of realisations are below 0.5. These measures, if used on their own, are not able to identify the true differences between realisations and observations which, as is discussed below, are due to timing errors.

Timing errors are captured by the LAGTIME index (first panel from left in Figure 5.7) which shows less than 33% of the values below 1 and that the first half of the model structures performed better than the second half (see also Figure 5.8). The first half is characterised by model structures in which there is a component that delays the instantaneous runoff (routing allowed), while for the second half routing is not allowed.

The less restrictive threshold that can be applied is 1, the upper limit of the performance range. This means that models whose median of any MPI is equal to 1 are always discarded. Amongst the remaining model structures, many are still performing better then others as some columns show a prevalence of red. This can also be summarised by calculating the median of performances over each column and plotting it against the MID as shown in Figure 5.8. This figure

is made of 5 horizontal panels (one for each MPI) and shows the MIDs on the x-axis and the median of performances for each model structure on the y-axis. First, third and fourth panels (from the top) confirm that performances tend to worsen using the second half of the model structures (MID from 624 to 1248), for which no routing is allowed. The algorithm filters out these models reducing the number from 624 to 312.

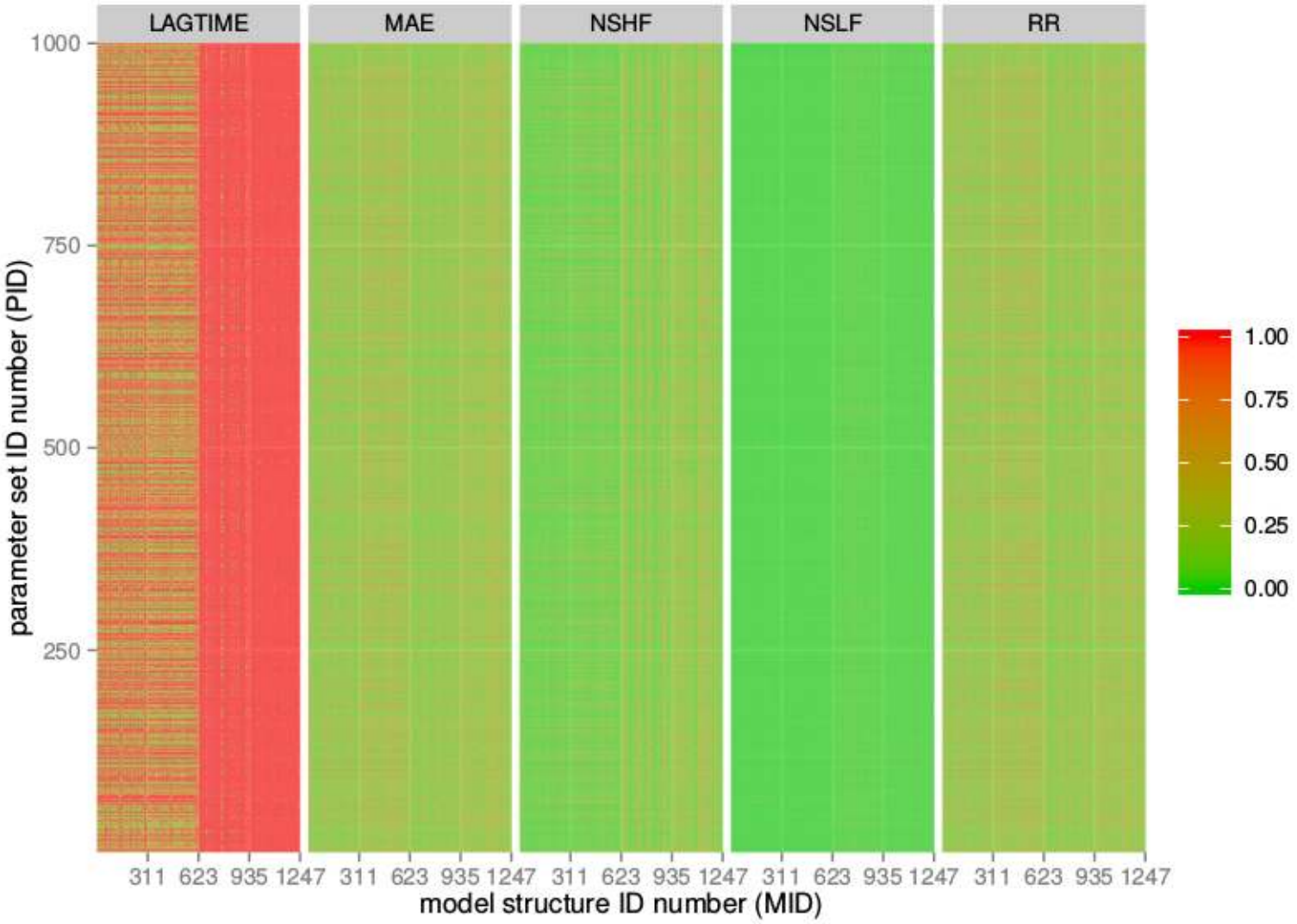

Figure 5.7: The 5 vertical panels show the array *I* (*Initial Ensemble*), sliced along each of the MPIs. The x-axis shows the model structure ID number (MID). For practical reasons, the y-axis shows only the first 1000 parameter set ID numbers (PID). Each cell is colour coded from green (best performance) to red (worst performance).

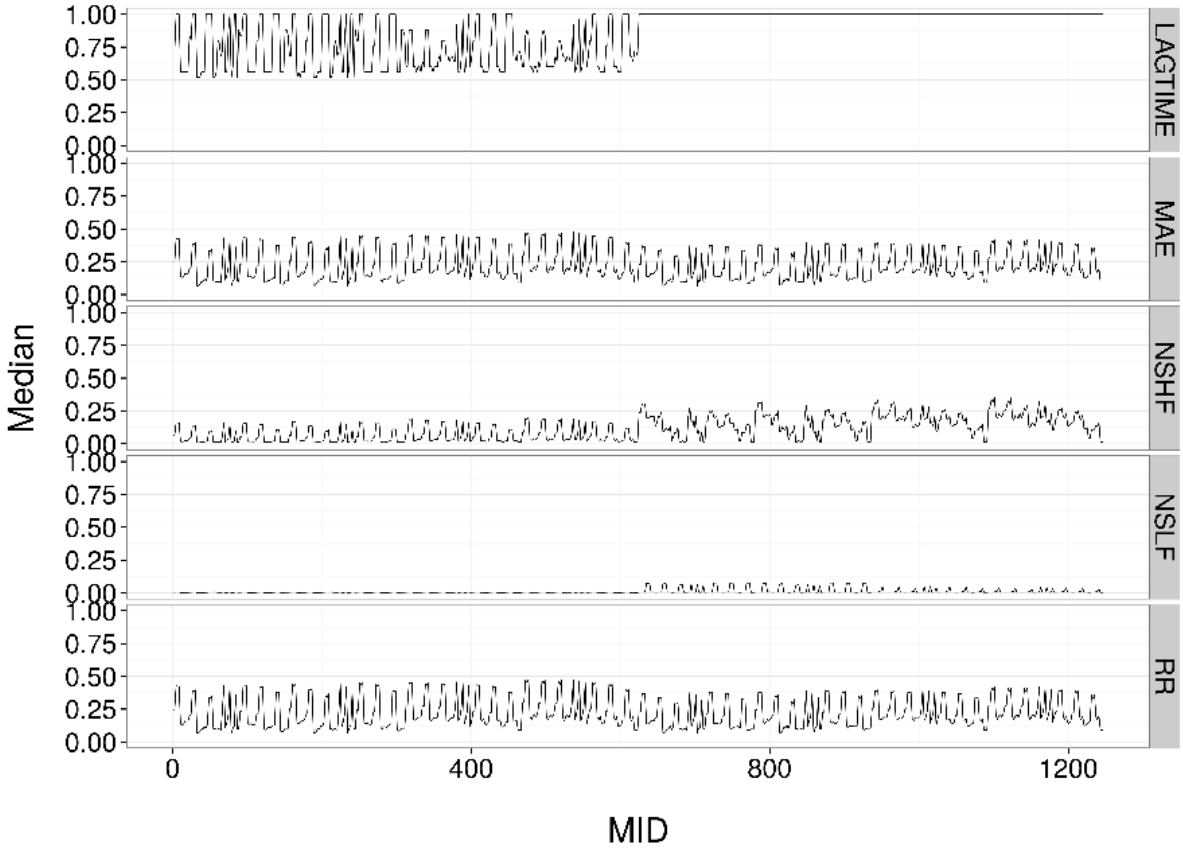

Figure 5.8: The 5 horizontal panels show the median of performances calculated for each model structure. The most concerning performance relates to timing (top panel) as all the model structures have a median performance above 0.5. The first half of the model structures (routing allowed) is generally performing better then the second half (routing not allowed). The oscillating pattern suggests there are also other model components to be discarded.

Distribution of performances

In order to understand better the multi-objective nature of the results for a single model, Figure 5.9 shows the variability of the performances of model 55 as a scatterplot. On the x-axis are the PID numbers (in order to make the plot easier to read, only the first 1000 are visualised). The values on the y-axis vary in the range [0,1] and shows the rescaled performances. these are colour coded: red (LAGTIME), dark green (MAE), light green (NSHF), blue (NSLF), pink (RR). With the exception of the LAGTIME, the majority of the other performances does not exceed 0.2. The different appearance of the performance distributions is due to the fact that LAGTIME is the only measure with a bounded upper limit prior to rescaling. This means that LAGTIME does not have outliers but all the extremes are forced to fall on the upper bound. Because the AMCA focuses on the lower bound of the range (best performance), this effect could be considered negligible in relation to the selection process. However future experiments could be performed to verify whether the use of a timing performance measure with unbounded upper limit would affect the selection.

Nevertheless, timing errors are dominant and this is also evident for other model structures. Figure 5.10, for instance, compares the results obtained for model 55 (first horizontal panel) to the results obtained for FUSE's parent models: Topmodel (MID = 59), Arno/Vic (MID = 229), PRMS (MID = 341), Sacramento (MID = 425). Performances are distributed differently across models, timing errors are dominant for the 5 model structures, however, the RR and NSHF errors for models 341 and 425 span wider ranges compared to the other models. The different distribution of performances is probably due to the interflow model component. Assuming that in the catchment there is no interflow, when this option is switched off (as in models 55, 59 and 229) the model performs generally better. Switching on the interflow option (as in models 341 and 425), instead, reduces the simulated streamflow incrementing the volume errors (NSLF) and errors related to the ratio between generated runoff and incoming rainfall (RR).

Referring back to Figure 5.8, this shows peculiar oscillating patterns on the left hand side of each panel. In order to understand which model configuration causes the spikes, the frequency

of model components can be analysed with regards to the highest MPI values. In this case, the bad-performing configurations have in common a combination of two model components: percolation due to gravity drainage (qperc = perc_lower) and an architecture of the lower soil layer (arch2) characterised by a baseflow reservoir of unlimited size with a fractional depletion rate (arch2 = unlimfrc_2). The percolation due to gravity depends on the moisture in the lower soil layer through a fraction where the denominator is infinite due to the unlimited size of the reservoir. The percolation from the upper to the lower soil layer, therefore, is always zero causing a consistent runoff overestimation and low performances. Removing these structures reduces the total number of models from 312 to 276.

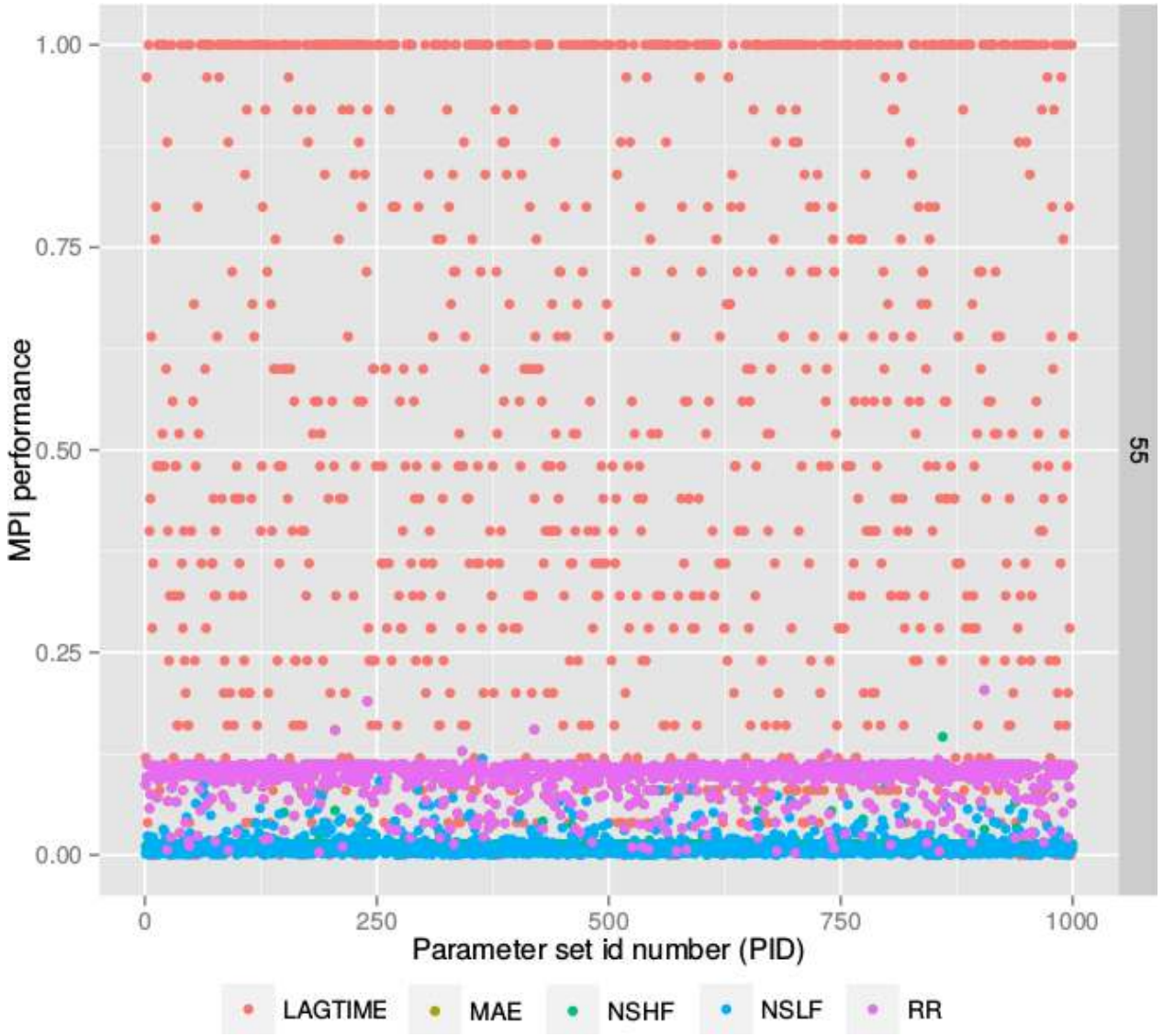

Figure 5.9: Performances of model id 55. On the x-axis are the first 1000 parameter sets id numbers (PID). On the y-axis is the performance (value in the range [0,1]). The performances are colour coded: red (LAGTIME), dark green (MAE), light green (NSHF), blue (NSLF), pink (RR).

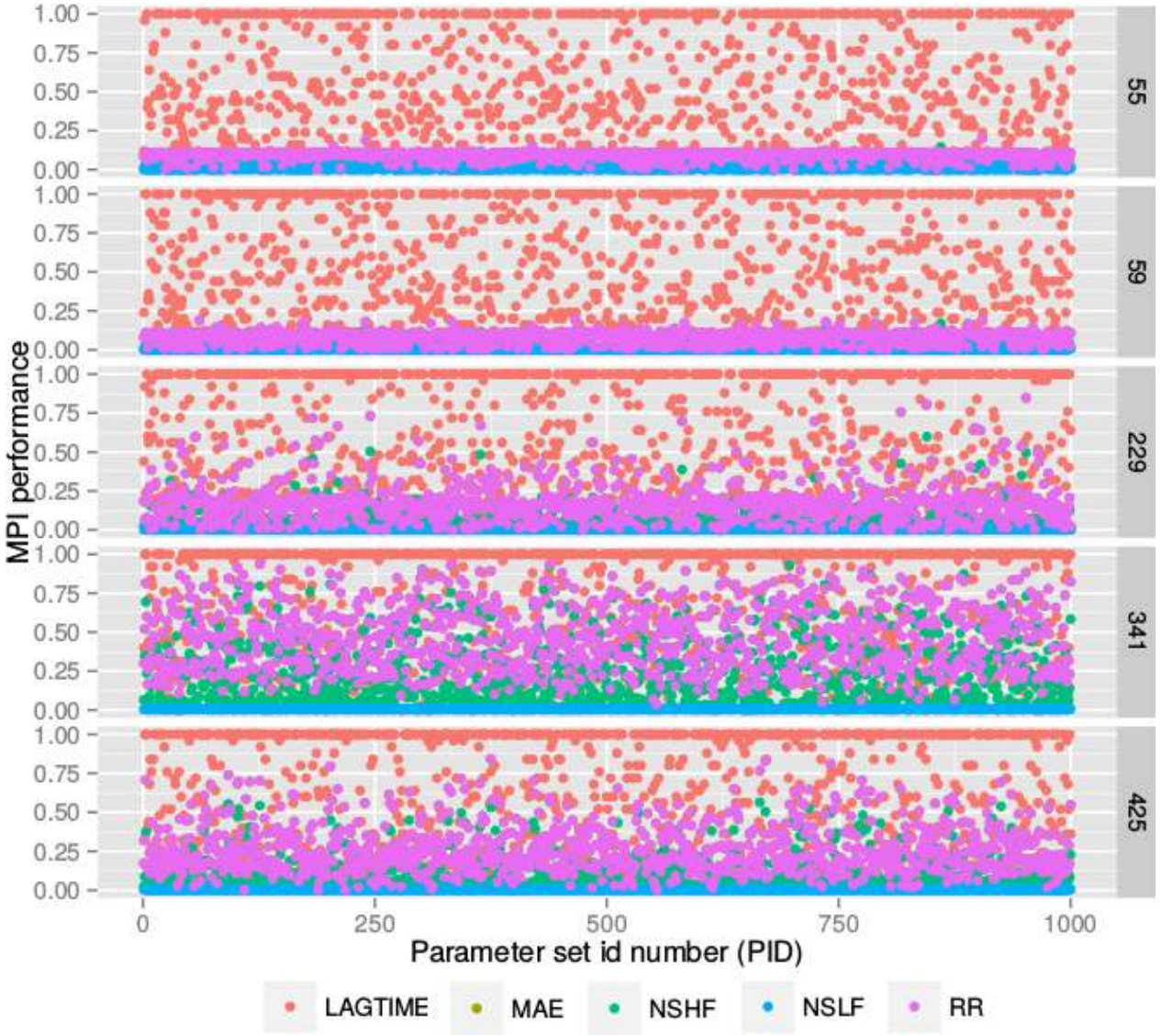

Figure 5.10: The five horizontal panels show the comparison of performances of five models (MID): 55, 59, 229, 341, 425. On the x-axis are the first 1000 parameter set id numbers (PID). On the y-axis is the performance (value in the range [0,1]). The performances are colour coded: red (LAGTIME), dark green (MAE), light green (NSHF), blue (NSLF), pink (RR).

Preliminary Selection (PS) Ensemble

The threshold identified by the algorithm is 0.6, therefore model structures with any median MPI in the range [0.6,1] are also discarded. The pre-selection retains 94 model structures with a dimensionality reduction of 92% (the *Initial Ensemble* is used as baseline).

Figure 5.11 shows the minimum and maximum bounds of the *Initial Ensemble* (T) and the ensemble generated from the pre-selected realisations (T'). A dimensionality reduction of about 92% translates in filtering out configurations that overestimate discharges at peaks and initial time steps. The accuracy of the 5th and 95th percentiles (area between red lines) is still 100% while the precision increases to 70%. Although the spread of T' is still significantly wide, the majority of the selected simulations lay in the dark blue area. This shows that the wide spread is due to very few realisations with high magnitude errors that have not been successfully filtered out in this preliminary stage.

Pareto Front (PF) Ensemble

Figure 5.12 compares the 5th and 95th percentiles of T' (black dotted lines) with the distributions percentiles over time of T'' generated from the Pareto front (grey-blue area). T" is still 100% accurate and slightly more precise (81%), simulating better timing and magnitude of the peaks.

Reduced Ensemble (RE)

Figure 5.13 compares the 5th and 95th percentiles of T'' (black dotted lines) with the distribution percentiles over time of T''' generated from the redundancy reduction steps (grey-blue area). The difference between the two ensembles is extremely subtle demonstrating the configurations removed at this last step do not add much to the overall results and can be considered as redundancies.

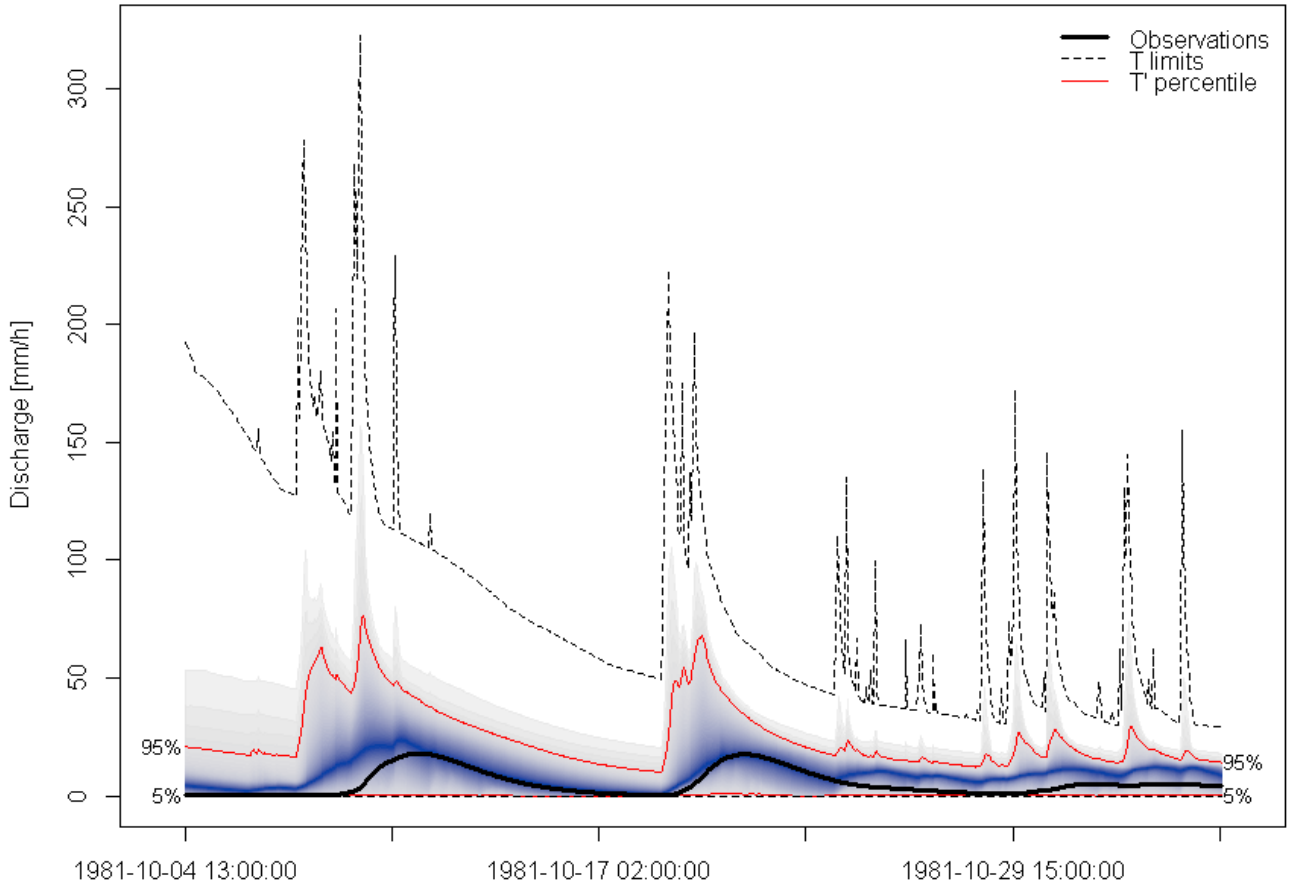

Figure 5.11: Comparison between the *Initial Ensemble* limits (T, black dotted line) and the intermediate ensemble obtained after the preliminary selection step (T'). The distribution percentiles over time of T' are shown as a gradient of colours from light grey to dark blue. The 5th and 95th percentiles are highlighted in red.

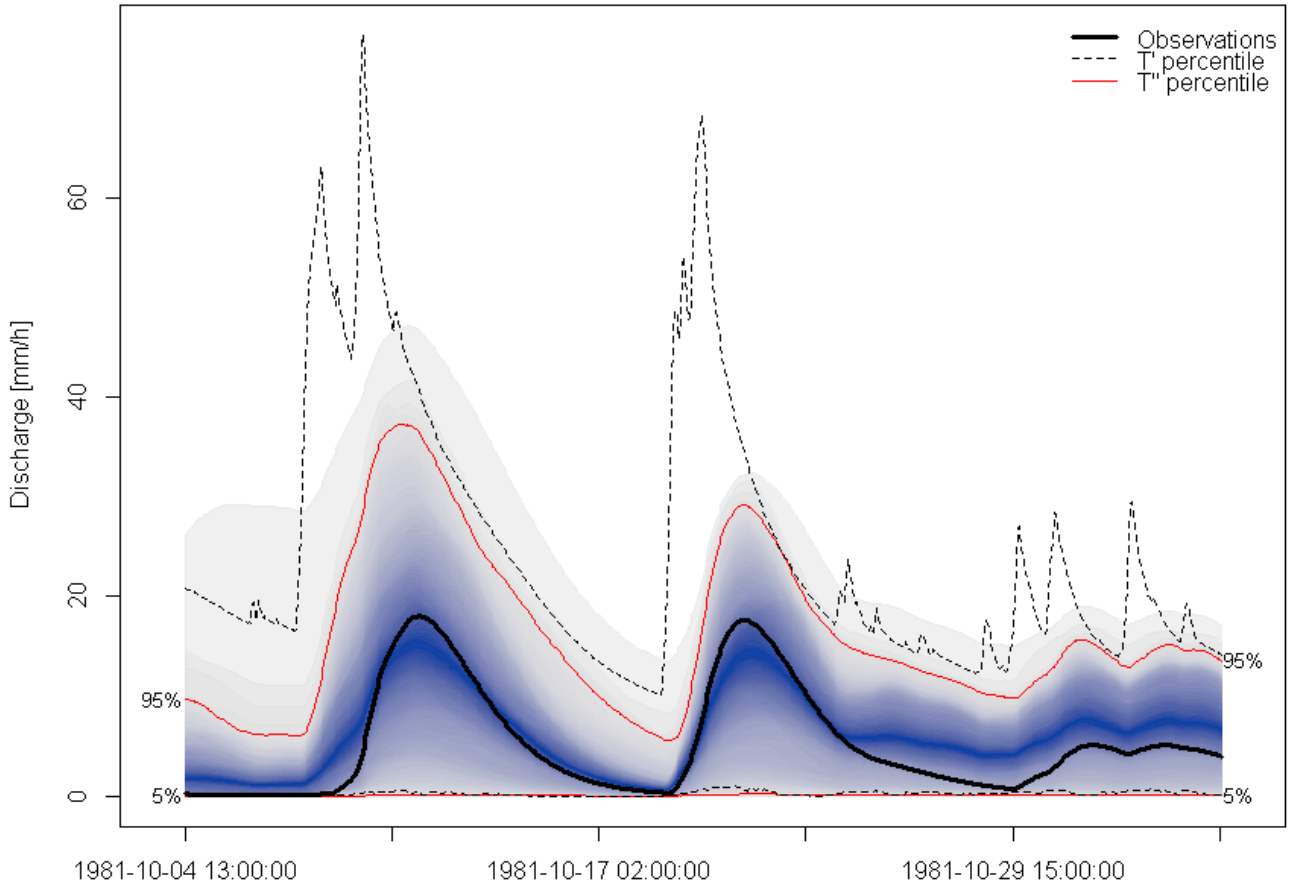

Figure 5.12: Comparison between the ensembles obtained from the pre-selection and the Pareto front steps. The 5th and 95th percentiles of T' are shown as black dotted line. The distribution percentiles over time of T'' are shown as a gradient of colours from light grey to dark blue. The 5th and 95th percentiles are highlighted in red.

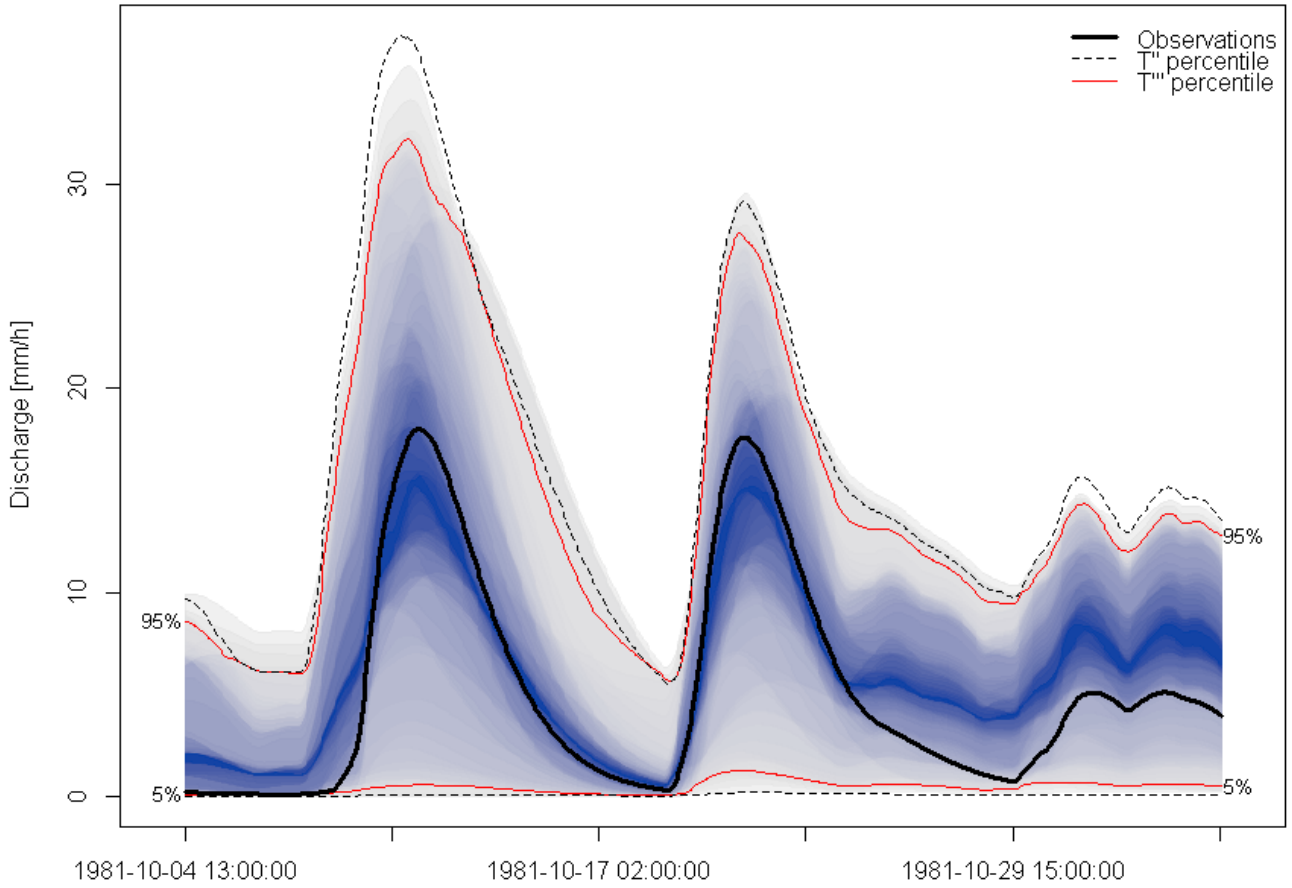

Figure 5.13: Comparison between the ensembles obtained from the Pareto front and redundancy reduction steps. The 5th and 95th percentiles of T'' are shown as black dotted line. The distribution percentiles over time of T''' are shown as a gradient of colours from light grey to dark blue. The 5th and 95th percentiles are highlighted in red.

Compare cumulative probabilities

Figures 5.14 to 5.18 show the cumulative probability of each MPI, a black line for the *Initial Ensemble*, a red line for the ensemble derived from the Pre-selection (PS), a blue line for the one derived from the Pareto front (PF) and a green line for the *Reduced Ensemble*. Blue and green lines have very similar ranges but the green one is generally performing better which demonstrates, again, the success of the redundancy reduction step.

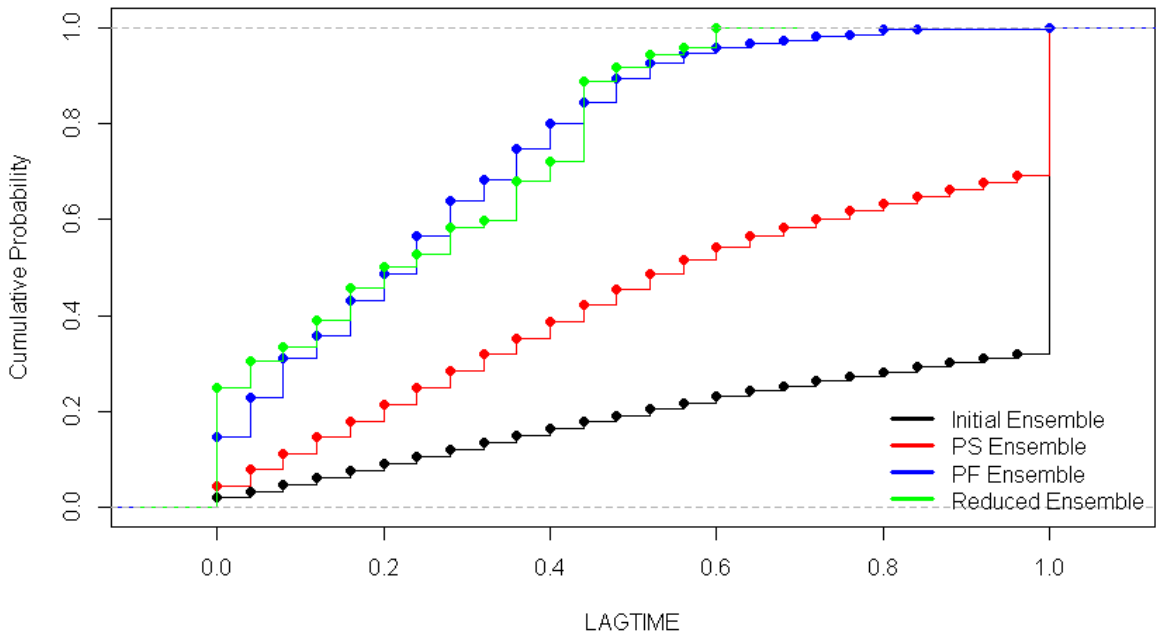

Figure 5.14: Cumulative probability distribution of the performance indicator LAGTIME.

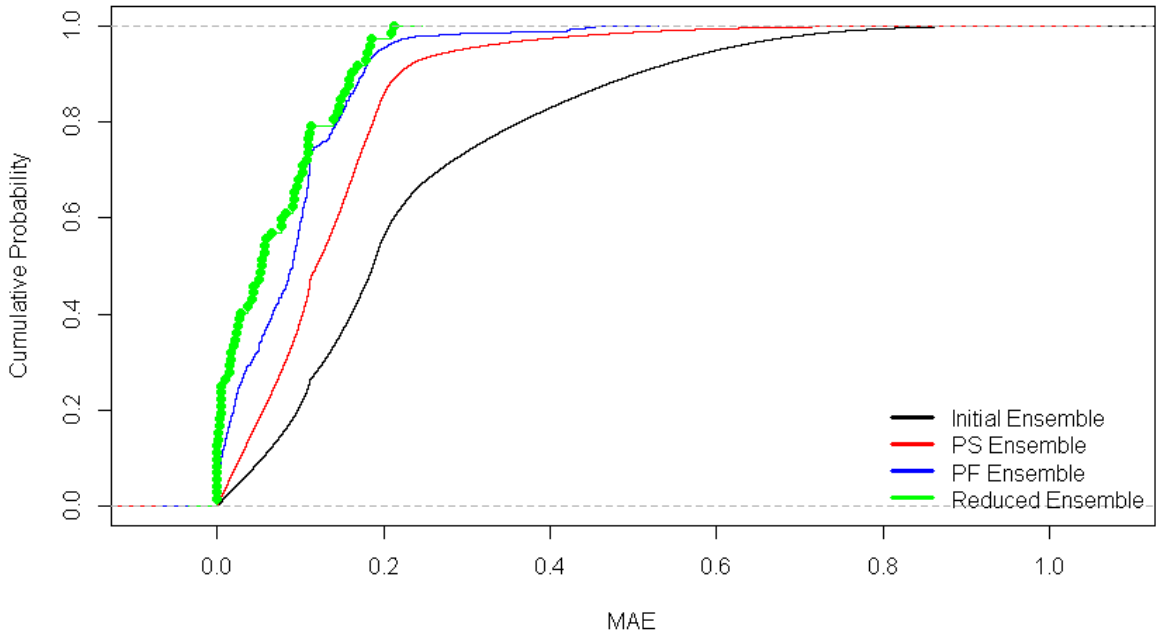

Figure 5.15: Cumulative probability distribution of the performance indicator MAE.

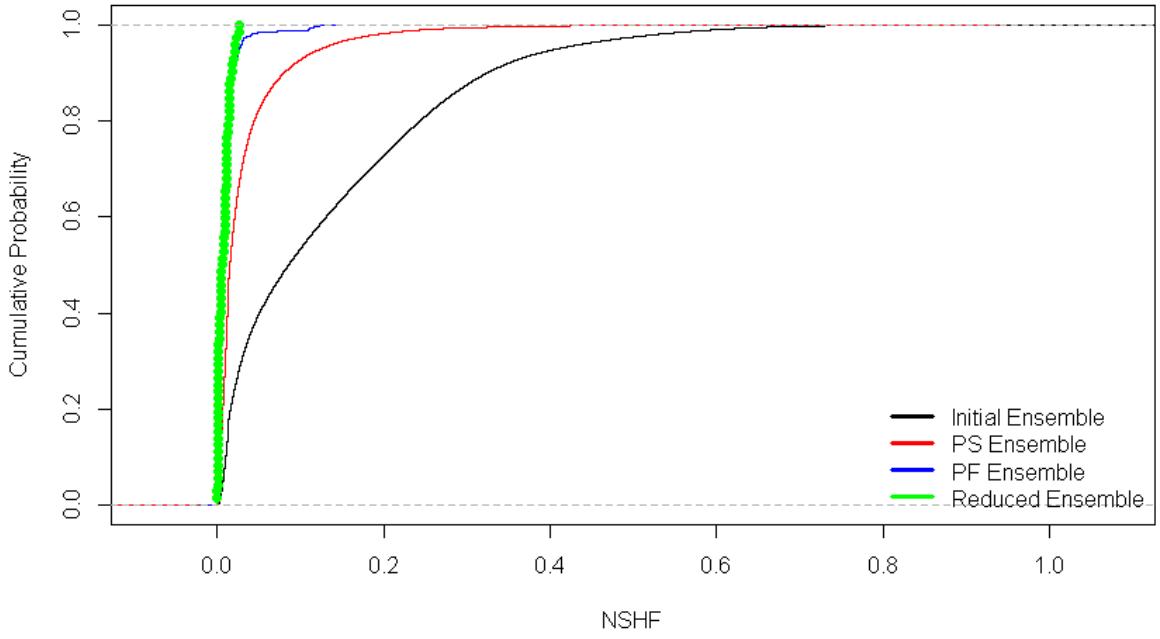

Figure 5.16: Cumulative probability distribution of the performance indicator NSHF.

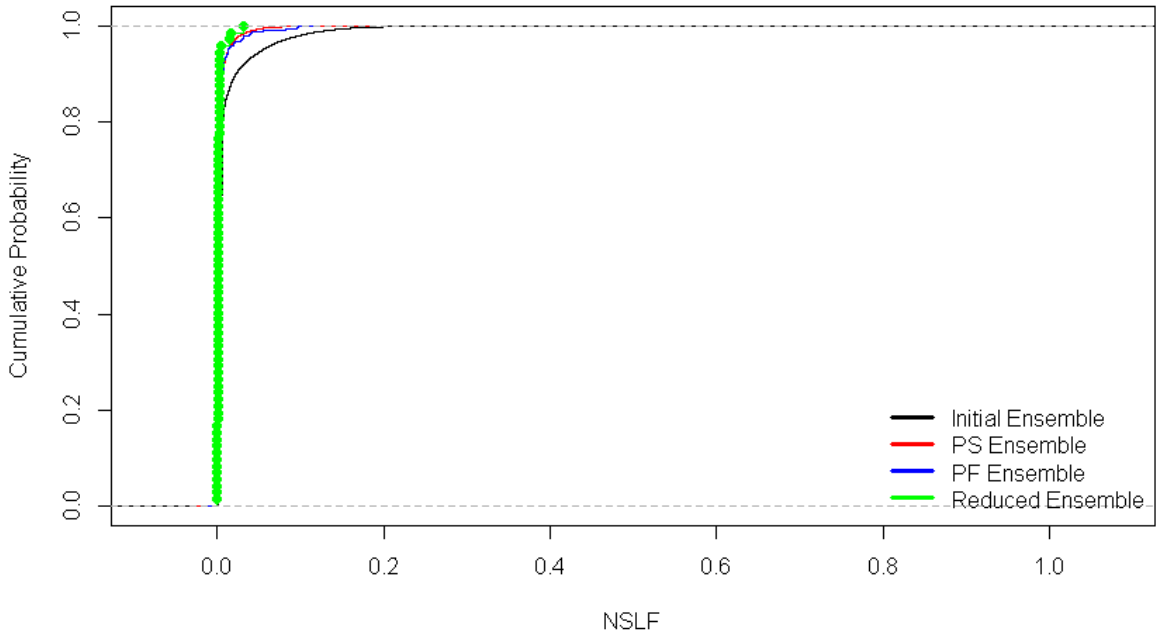

Figure 5.17: Cumulative probability distribution of the performance indicator NSLF.

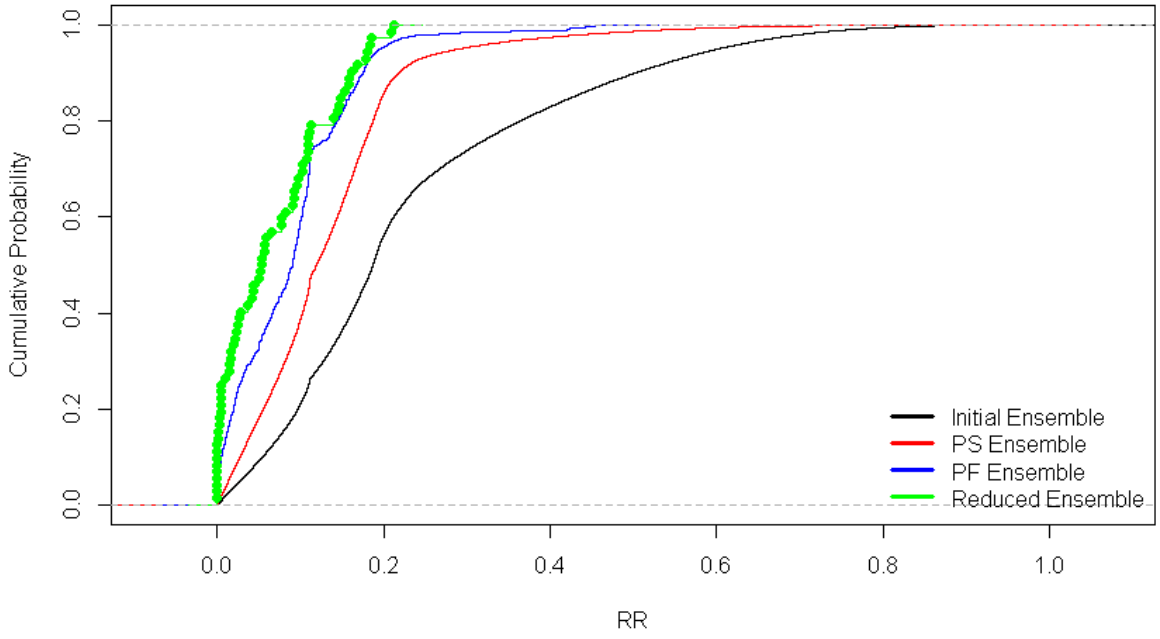

Figure 5.18: Cumulative probability distribution of the performance indicator RR.

Algorithm's suggestions in terms of model components and parameters

To check whether the algorithm selects the correct modelling options, the frequency distribution of the selected model components is compared to the synthetic model structure. Figure 5.19 shows seven vertical panels (one per model building decision), the relative frequency[4] on the y-axis and modelling options on the x-axis. The options that correspond to the synthetic model structure are green, while the others are red. The algorithm performs a correct selection if the most frequent options coincide with the synthetic model structure. In other words, the highest bar in each panel should be green.

The algorithm suggests a model structure with an upper soil layer made of two storages (one tension and one free), a lower soil layer of unlimited size and power recession law, a Topmodel runoff mechanism, sequential parameterisation for the evaporation scheme, no interflow and routing allowed. All these options coincide with the synthetic model components. However, the algorithm failed to identify the correct percolation parametrisation.

Similarly, the posterior distribution of parameters can be compared to the synthetic values to check whether the algorithm can identify optimal parameter ranges. Figure 5.20 illustrates a comparison between the parameter distributions of the *Initial Ensemble* (grey) and the *Reduced Ensemble* (yellow). On the x-axis of each plot is the parameter range, while on the y-axis is the kernel density estimate. Only parameters used to generate the synthetic data are shown, so that the distributions can be compared with the true values (red dots). The procedure did not noticeably reduce the initial parameter ranges, except for the "timedelay" parameter, which narrowed from [0,5] to [2.1,4.7] bracketing closer the synthetic value of 2.7 days.

[4]The relative frequency is defined here as the ratio between the frequency count of a modelling option and the total number of ensemble components.

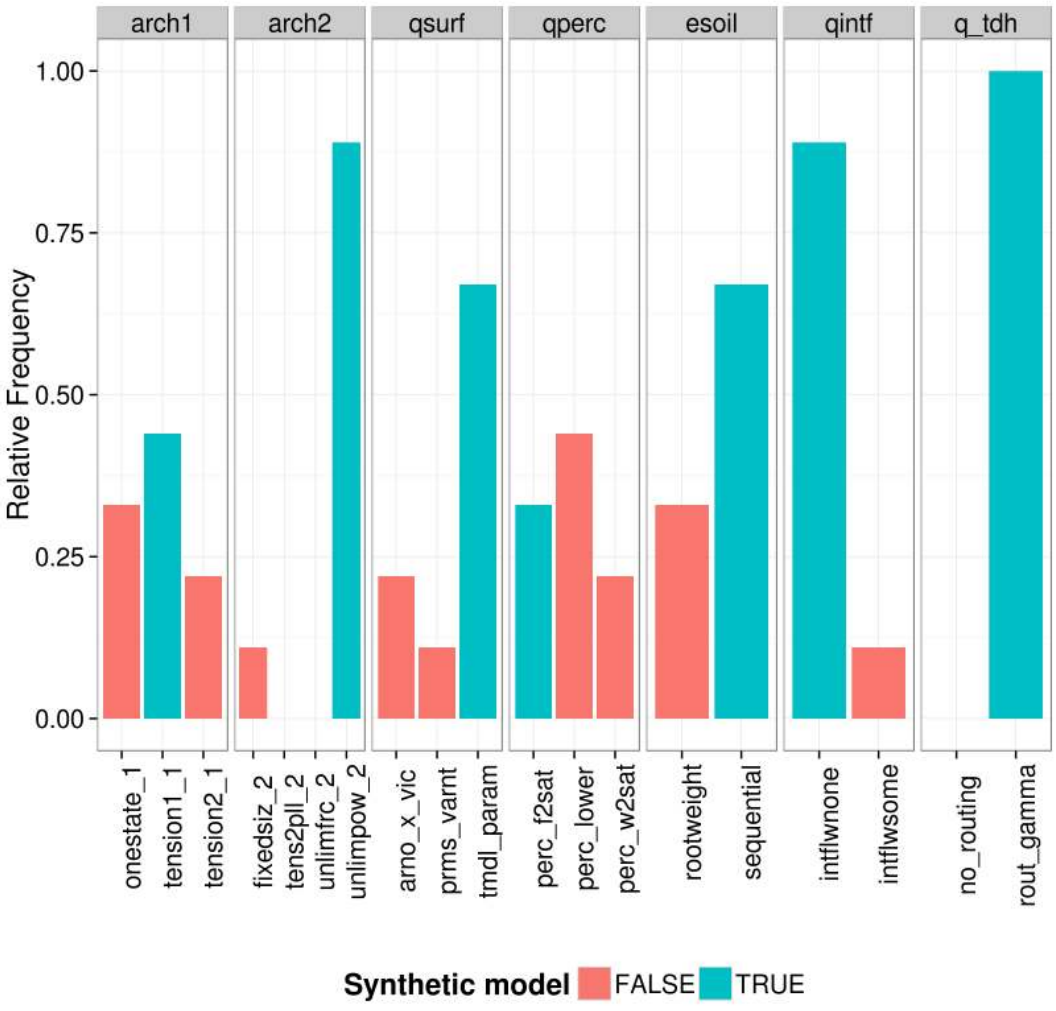

Figure 5.19: Relative frequency of the selected model components. Components used in the synthetic model are shown in green, the others are in red. The majority of the most frequent components coincide with the synthetic ones.

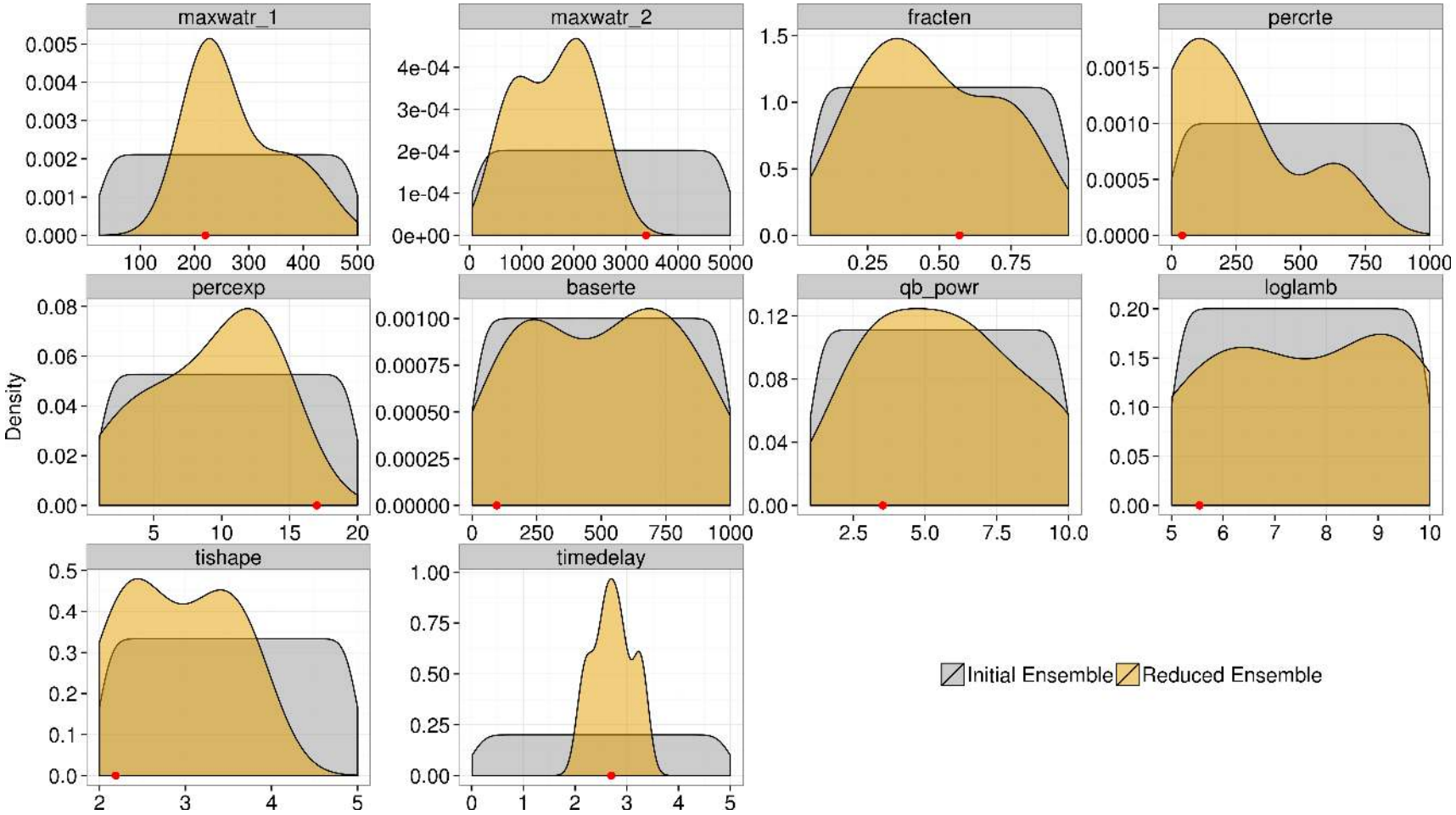

Figure 5.20: Comparison between the parameter distributions of the *Initial Ensemble* (grey) and the *Reduced Ensemble* (yellow). The parameter values are on the x-axis, while the kernel density estimate is on the y-axis. The parameters *maxwatr_1* and *maxwatr_2* are measured in mm, *baserte* and *percrte* are in mm/day, *loglamb* is in m, *timedelay* is in days, the remaining parameters are pure numbers. The red dots show the synthetic parameter values.

While the prior (*Initial Ensemble*) parameter distributions were designed to be uniform, a χ^2 test showed that most posterior (*Reduced Ensemble*) distributions were not (e.g. timedelay, percrte, maxwatr_1 and maxwatr_2). This is also evident by visually inspecting Figure 5.20. If both prior and posterior distributions of a generic parameter X are uniform, it means that the AMCA filtering process is not affected by its variations and therefore X is considered a not sensitive parameter. Conversely, if the posterior distribution spans only a subrange and/or is not equally spread across the range of possible values, X is considered a sensitive parameter. Amongst all the parameters, the timedelay appears to be the most sensitive as it spans a subrange much smaller than the original one and within this subrange the majority of the values are concentrated around 2.7 days (the synthetic value). This is a sign that timedelay was sampled from a too wide range. The sampling error generated significantly poor timing performances that dominated the overall result space hiding the effect of the model structure variability. Section 5.3.2 investigates whether the algorithm's performances can be improved by sampling the timedelay from a more realistic range.

Table 5.2: Summary table of the filtering algorithm's results. At each step *i*, a number of outputs are reported: utilised model structures (# Models), parameter sets (# Parameter sets), number of realisations (# Realisations), dimensionality reduction (D-reduction, calculated as the ratio of the number of realisation at step *i* over the number of realisations of the *Initial Ensemble*) and accuracy and precision of derived ensembles. In addition, at the third step is also reported the auto-generated threshold and at the last step the statistical reliability of the *Reduced Ensemble*.

Step	Description	Type of result	Value
1	$Result\ Space \rightarrow I/T$	# Models	1,248
		# Parameter sets	10,000
		# Realisations	12,480,000
		Accuracy (%)	100
		Precision (%)	0
2	$Rainfall\ error\ screening$	# Models	624
		# Parameter sets	10,000
		# Realisations	6,240,000
		D-reduction (%)	50
		Accuracy (%)	100
		Precision (%)	0
3	$Pre.model\ selection \rightarrow I/T'$	Auto-generated threshold	0.6
		# Models	94
		# Parameter sets	10000
		# Realisations	940000
		D-reduction (%)	92
		Accuracy (%)	100
		Precision (%)	70
4	$Pareto\ Front \rightarrow I/T''$	# Models	32
		# Parameter sets	39
		# Realisations	1248
		D-reduction (%)	+99
		Accuracy (%)	100
		Precision (%)	81
5	$Redundancy\ Reduction \rightarrow I/T'''$	# Models	8
		# Parameter sets	9
		# Realisations	72
		D-reduction (%)	+99
		Accuracy (%)	100
		Precision (%)	84
		Reliability (%)	84

5.3.2 Sensitivity analysis

Based on the synthetic test illustrated in the previous section, the AMCA algorithm seems to be fairly accurate in filtering out unsuitable model structures and narrowing parameter ranges. However, it is difficult to determine the potentials of such an approach without further testing. This section collects a series of experiments to analyse the sensitivity of the algorithm to the size of the sampled parameter space, to the variability of the routing parameter, to the use of the preliminary model selection step and to the length of the warmup period.

Size of the sampled parameter space

The AMCA algorithm uses random sampling to define the parameter space. The number of runs required could be determined based on the number of parameters forming a set. Beven (2001*b*, p. 219) suggests that a typical dependency can be formulated as follows:

$$nS = 10^{nP} \tag{5.3}$$

where nS is the number of required runs to sample and nP is the number of parameters in a set.

FUSE's models use a total of 22 parameters, leading to an ideal number of runs being simply unmanageable. Therefore, it is important to identify an ideal trade-off between a satisfactory representation of the sampled space, the processing time and the use of computer resources.

There are few factors that could justify a reduction in the required number of runs, one being that each model in the FUSE framework only utilises a subset from a minimum of 8 to a maximum of 16 parameters.

The HPC facilities available at Imperial College are based on a queueing system that only allows the use of a maximum of 40 cores at a time. The AMCA algorithm itself is highly demanding

in terms of resources. Using the synthethic case and limiting the processing time to 1 day, a maximum of 10000 parameter sets could be sampled.

It is expected that the more parameter sets are sampled, the more accurate and precise the *Reduced Ensemble* is going to be. Experiments were carried out to understand how accuracy and precision of the *Reduced Ensemble* worsen when reducing the number of samples, and it was extrapolate what results could be expected if a higher number of samples could be used.

The performance of the algorithm is valued in terms of the accuracy, precision and reliability of the *Reduced Ensemble*. Performances were collected for samples of the following sizes: 1000, 2500, 5000, 7500, 10000.

Figure 5.21 summarises the results of these experiments. On the x-axis is the number of samples, while on the y-axis percentages from 0 to 100. The red line shows that the accuracy of the *Initial Ensemble* is always 100%, regardless of the sample size. This suggests that the variability over the model structure space could compensate for a sparsely sampled parameter space.

The yellow, green and light blue lines show the accuracy, precision and reliability of the *Reduced Ensemble*, respectively. these are always above 68%. The smallest set of samples (1000), is characterised by the lowest accuracy. However this rapidly increases to 100% with as few as 5000 samples. An increase in accuracy generally corresponds to a loss in precision and reliability. However, results become fairly stable using as few as 5000 parameter sets suggesting that the AMCA does not need a more densely sampled parameter space to converge, at least for the dataset used for testing.

The next section illustrates an experiment to understand whether narrowing the range of what appears to be the most sensitive parameter (timedelay) allows the use of an even more sparsely sampled parameter space.

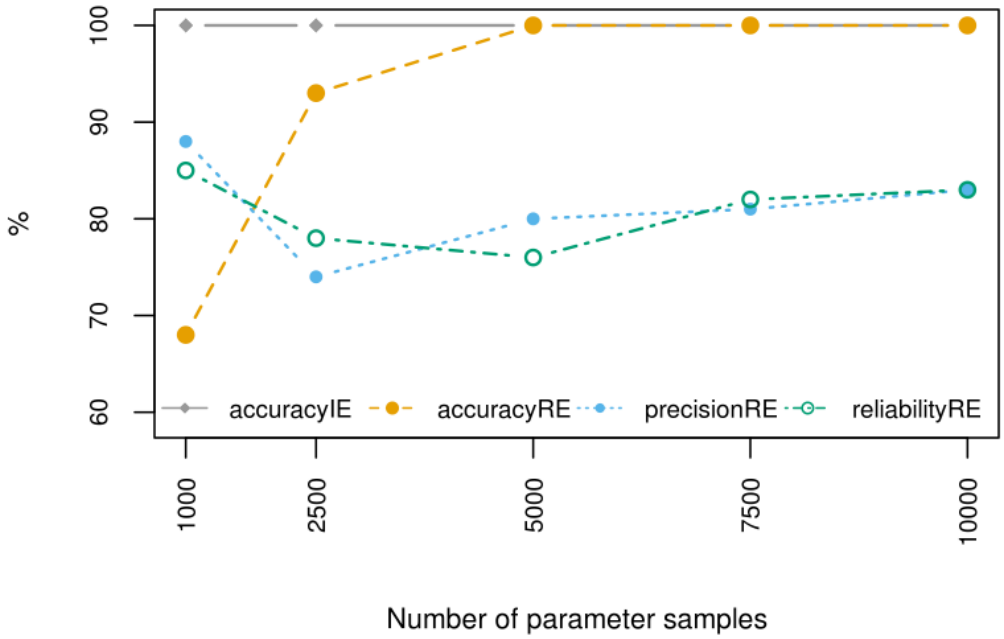

Figure 5.21: Algorithm performances based on parameter sample size. The x-axis shows the number of samples and the y-axis the percentage value. IE = *Initial Ensemble*, RE = *Reduced Ensemble*.

Routing parameter range

The timedelay has been found to be the most sensitive parameter of the FUSE framework. It is expected that narrowing the range of this parameter could improve the accuracy, precision and reliability of more sparsely sampled parameter spaces.

The true timedelay of the synthetic example is known to be 2.7 days, therefore the results of the algorithm were compared in the following cases:

- parameter sets are sampled setting the timedelay equal to the true value,
- parameter sets are sampled from the default range [0.1,5] days.

Figure 5.22 shows the bounds of the *Reduced Ensemble* for the default parameter ranges as black dotted lines, while fixing the timedelay to 2.7 days returns the bounds shown as red

lines. Imposing the timedelay equal to the true value significantly improves the ensemble results which are highly accurate but more precise (87%) and reliable (92%). The identification of model components also looks more clearly defined, as shown in Figure 5.23, suggesting a lower layer of unlimited size and power recession law, a Topmodel-like runoff parameterisation, a percolation scheme dependent of the field capacity and no interflow. However, the algorithm failed to identify the correct upper soil layer architecture.

The experiment in the previous section was also repeated imposing the timedelay equal to 2.7 days and the results illustrated in Figure 5.24. Even though there is a slight descrease in accuracy, the performances are generally higher than in the previous case. This suggests that, in order to obtain better ensemble predictions and well identified model components, parameter ranges should be as close as possible to realistic values.

Using real datasets, the timedelay could be calculated as the time of concentration, derived from the geospatial characteristics of the catchment, such as length of the longest river reach, mean slope and area of the catchment. When this information is not available, the algorithm itself can be used recursively, running a first time to narrow the parameter ranges and a second time to analyse model structure variability.

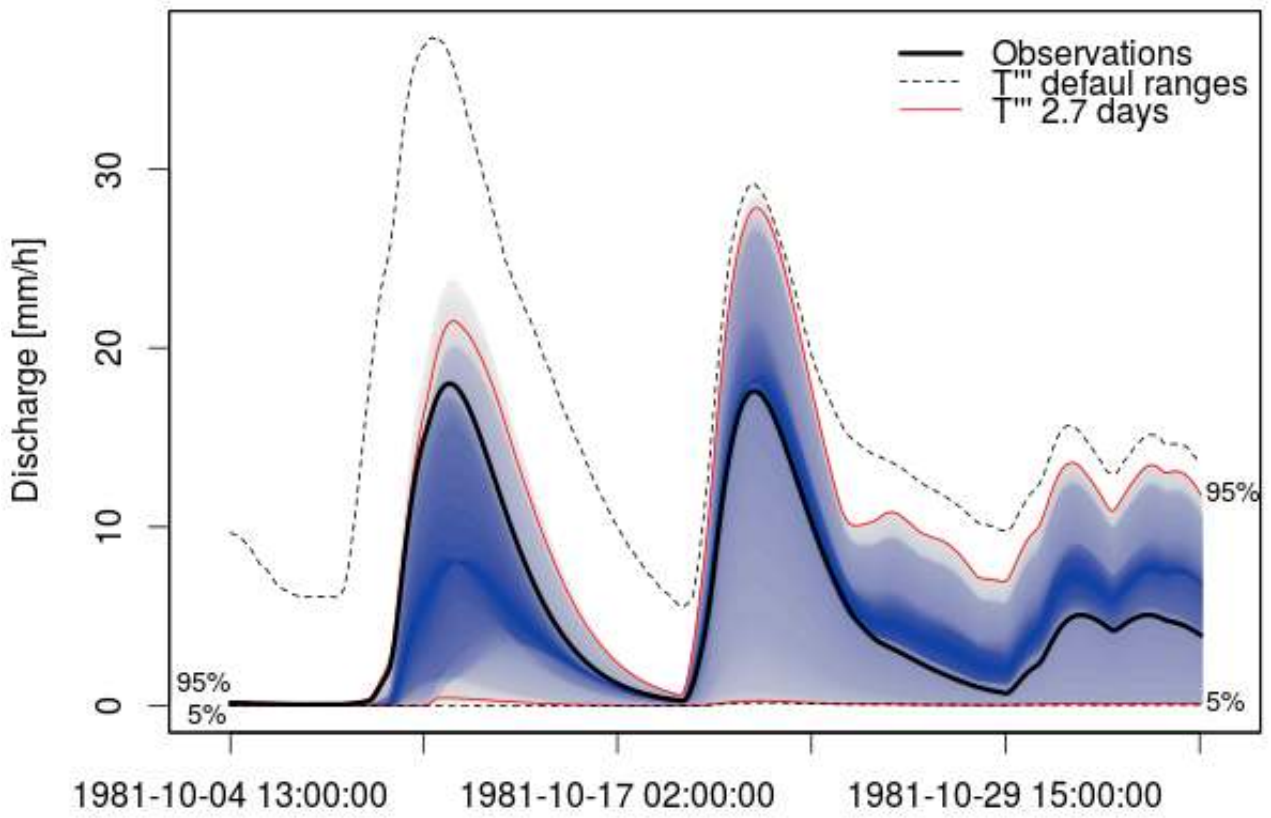

Figure 5.22: Improved ensemble obtained by imposing a timedelay equal to its true value (2.7 days). The dotted black lines show the 5^{th} and 95^{th} percentiles of the ensemble from default ranges while the red lines show the 5^{th} and 95^{th} of the ensemble obtained by imposing the timedelay equal to the true value. The grey-blue area shows the latter's distribution percentiles over time.

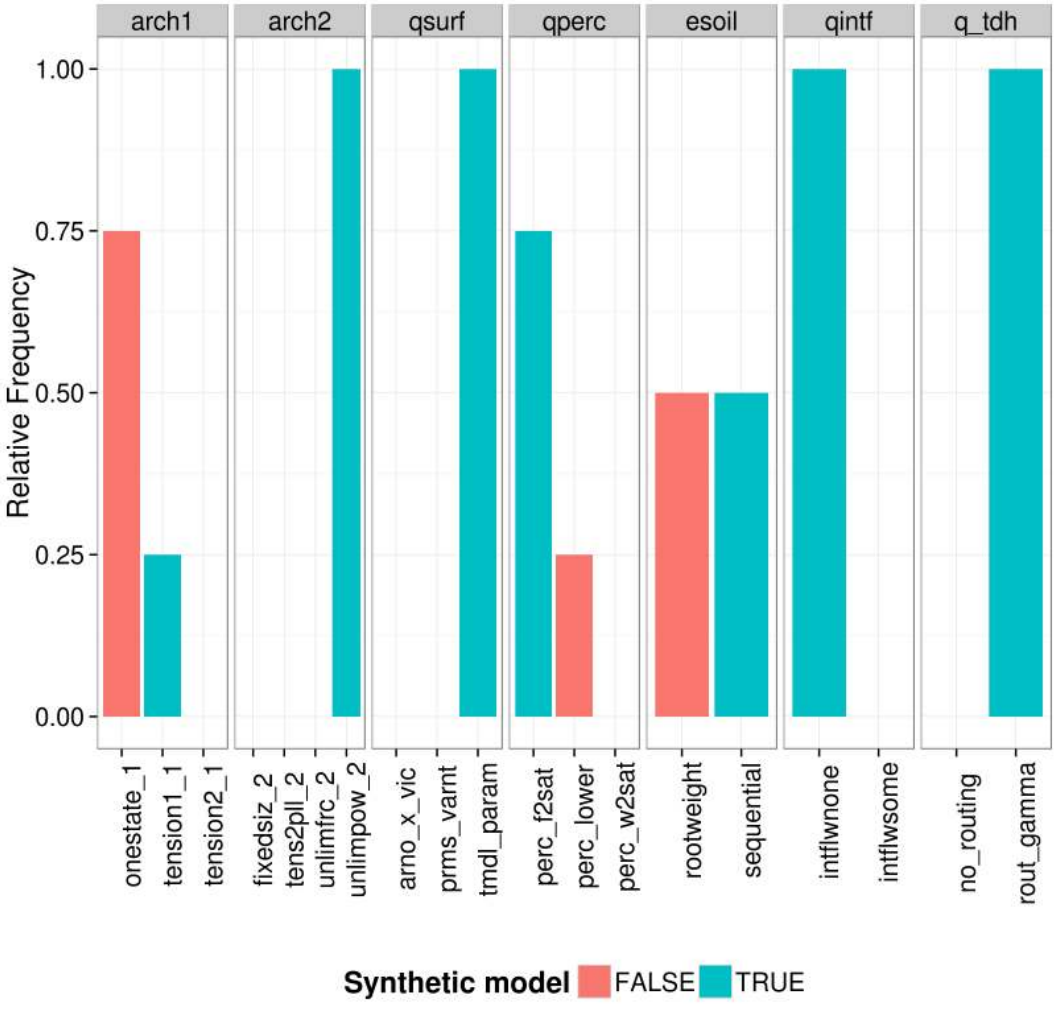

Figure 5.23: Improved identification of model components obtained by imposing a timedelay equal to its true value (2.7 days). The plot shows the relative frequency of the selected model components. Components used in the synthetic model are shown in green, the others are in red.

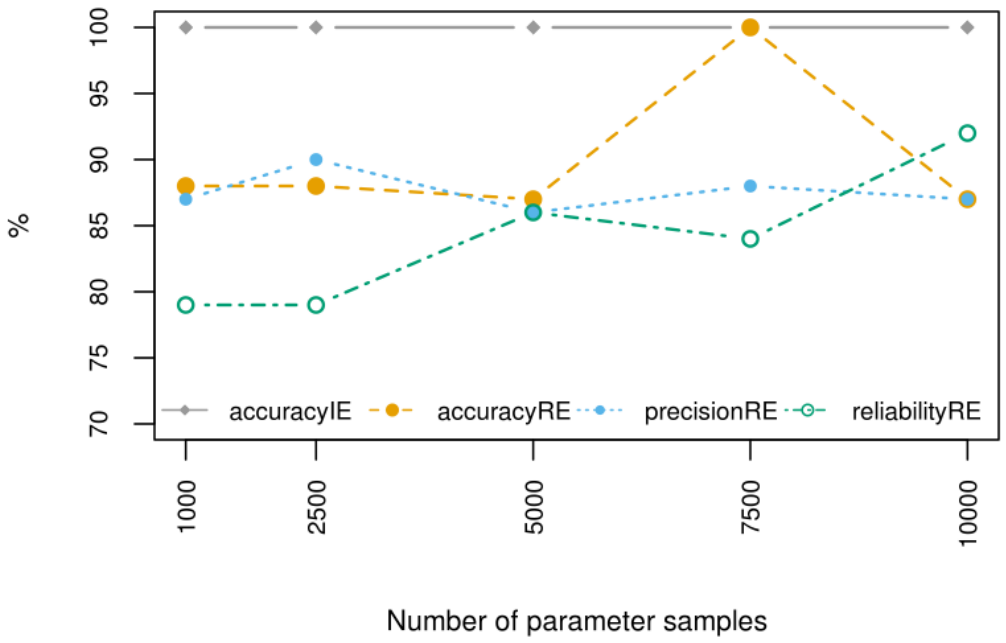

Figure 5.24: Algorithm performances based on parameter sample size. The x-axis shows the number of samples and the y-axis the percentage value. IE = *Initial Ensemble*, RE = *Reduced Ensemble*. In the above simulations the time delay is always equal to 2.7 days.

Avoiding the preliminary model selection step

The preliminary model selection was designed to expedite processing, with fewer realisations to process, runs are performed faster but this may translate into a loss of information as filtering out model structures based on their "overall performance" could lead to ignore isolated well performing realisations. The convenience of having a preliminary selection was evaluated by comparing the results in Table 5.2 to the case in which the preliminary selection step is skipped.

Skipping the preliminary selection step can be considered a more conservative approach, as the new RE is characterised by an increase in reliability (from 84% to 85%) and a decrease in precision (from 83% to 66%) while still 100% accurate. Figure 5.25 shows that the RE medians with (red dotted line) and without (blue dashed line) pre-selection are very similar. However the two approaches differ significantly in the spread of the results at the beginning of the simulation period (see Figure 5.26). This is a sign that the warmup period of 3.5 days (for hourly datasets) is not adequate and a longer warmup should be allowed (this is further discusses in Section 5.3.2). The model structures causing the wider spread are all characterised by a runoff generation mechanism different from Topmodel (see all the selected model components in Figure 5.27). These model structures are more sensitive to the initial conditions but the Pareto algorithm seems not being able to detect the difference in terms of MPIs. The use of the preliminary selection is, therefore, always recommended as this not only speeds up the processing time but also complements the filtering capabilities of the Pareto algorithm when using multiple performance indices and short time series for which the warmup period may not be adequate.

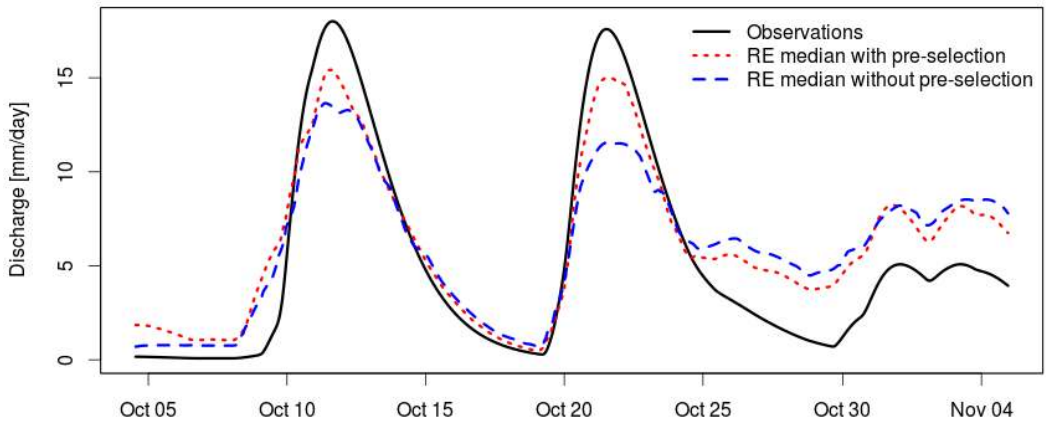

Figure 5.25: Comparison of observed results with the median of the *Reduced Ensembles* obtained with (red dotted line) and without (blue dashed line) pre-selection step.

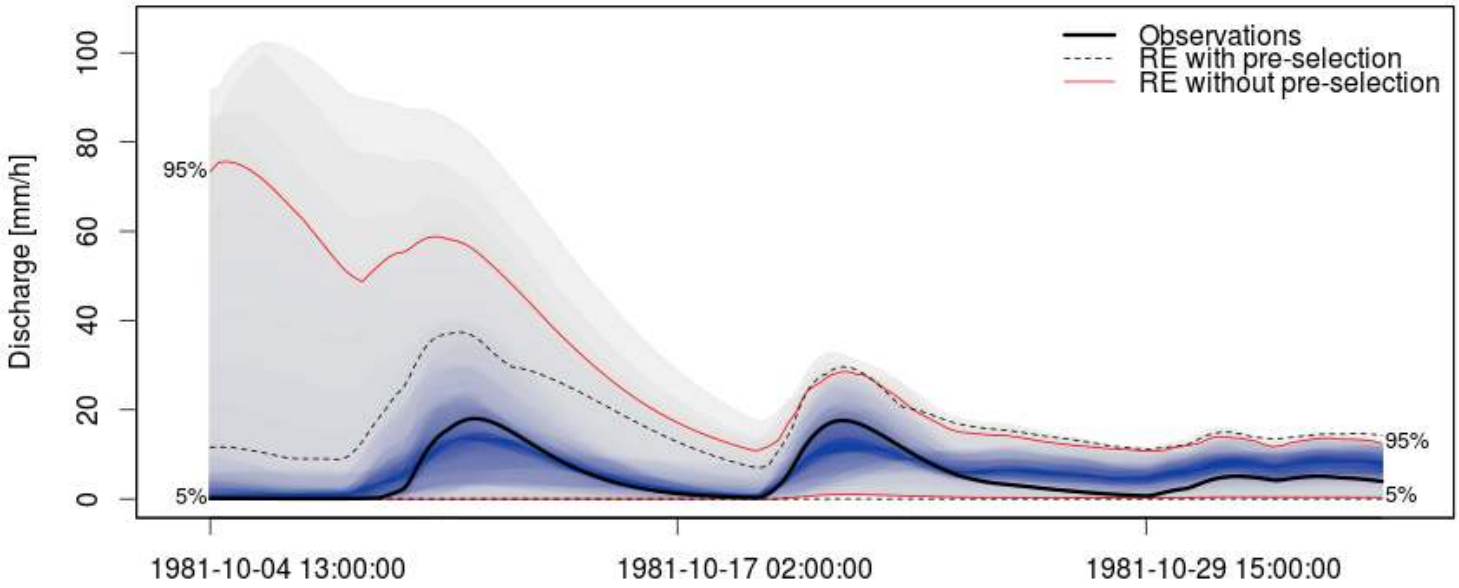

Figure 5.26: Comparison of the spread of the *Reduced Ensembles*, represented by the bound between the 5^{th} and 95^{th} percentile, obtained with and without pre-selection step. The former bound is the area between the black dotted lines, while the latter is the area between the red lines. The grey-blue colour gradient is the latter's distribution percentiles over time.

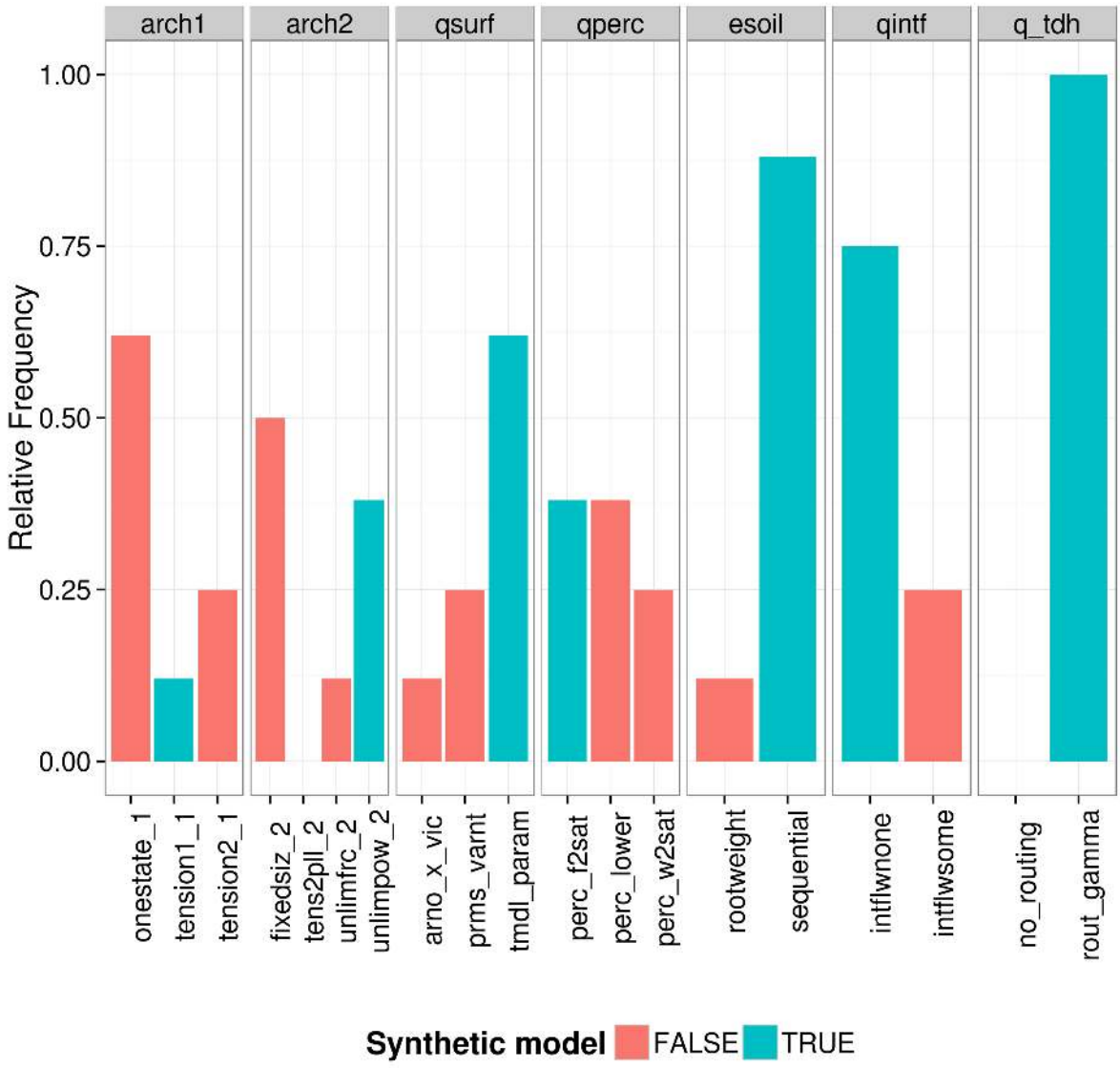

Figure 5.27: Relative frequency of the model components obtained by skipping the preliminary selection step. Components used in the synthetic model are shown in green, the others are in red.

Warmup period

Hydrological models can be very sensitive to initial and boundary conditions (Stephenson and Freeze, 1974; Beven, 2001*b*). This is also true for FUSE, which core is a set of ordinary differential equations to be resolved with regards to the state variables once initial conditions are defined. These state variables represent the water storages in the upper and lower soil layers, defining the structure of the Soil Moisture Accounting module. By default, the state variables are initialised as containing 25% of the maximum storage capacity (Clark et al., 2008). This is a generic assumption that tends to overestimate the soil moisture in dry periods and underestimate it in wet periods. In order to limit the effects of such arbitrary assumption, model performances are generally calculated after running the model for a certain number of time steps which length is known as "warmup period" (or spin-up). Western and Grayson (1998) suggest to assign the warmup period based on the response time of the system, however, there is no general rule to quantitatively determine an "adequate" number.

The sensitivity to the initial conditions is not part of the model behaviour as it tends to disappear over time. However, choosing an inadequate warmup period can cause the model behaviour to be masked behind the model sensitivity to the initial conditions and during the selection process model structures could be filtered out for the "wrong reason". This section describes an experiment to relate the time-response of a catchment to the warmup period, for models within the FUSE framework. This test is useful to assess whether the 3.5 days warmup period used in the previous experiments was an adequate choice. If not, it could help quantifying the performance loss.

Using a synthetic example, the response time of the system corresponds to the timedelay parameter (known) and the warmup period can be set a number of times greater than this parameter. The first experiment compares the results of the AMCA using a 2 year synthetic dataset and four different warmup periods: 0%, 10%, 30% and 50% of the dataset length. The last year is used for calculating the MPIs, while different portions of the first year are used to warm up the models. It is expected that for adequate warmup periods the model configuration converges in

terms of ensemble accuracy, precision and reliability.

The dataset in the first experiment is generated using the same model and parameters listed in Table 5.1 with the exception of the timedelay which is set to 1 hour. Other three datasets are generated setting the timedelay equal to: 6 hours, 1 day and 3 days. The previous experiment is then repeated. From the parameter space, 2500 sets are sampled and used with all the combinations of timedelay/warmup periods.

Figure 5.28 shows the combined results of these experiments. For fast responses (timedelay set to 1 and 6 hours) the accuracy of simulations with no warmup is generally lower than in other cases, this is probably because too many models have been erroneously eliminated at the preselection step. For slow responses (timedelay set to 1 and 3 days), instead, all the performances seem not to be significantly affected by the warmup period. Therefore, increasing the warmup period in the previous sections' experiments would not have had a noticeable effect on the algorithm's performances.

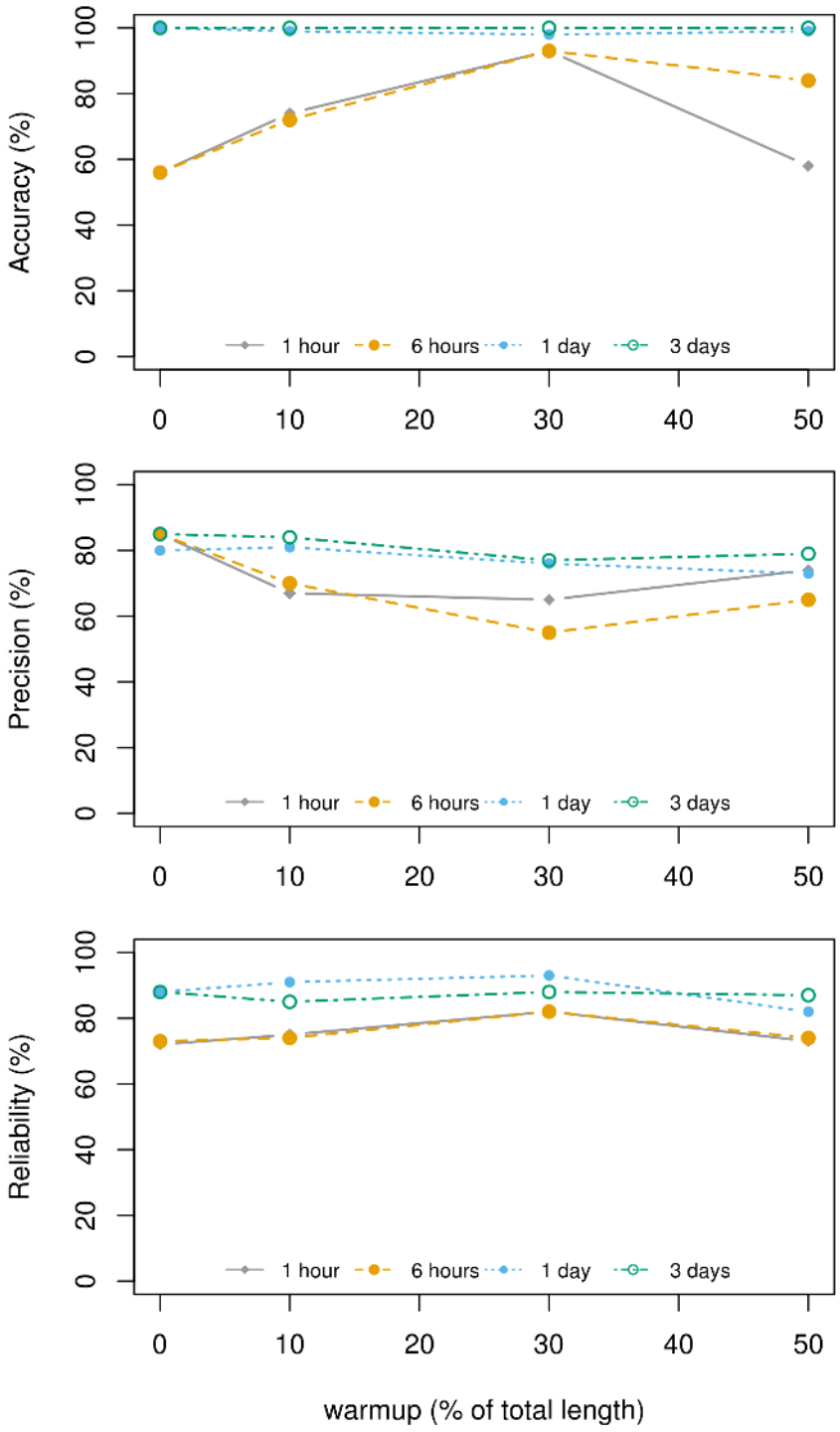

Figure 5.28: Multipanel plot to illustrate sensitivity to warmup period, the number of time steps needed by the model to eliminate the initial bias. The experiments were running on 3 synthetic time series: the first one was generated using a timedelay parameter of 1 day (top panel), the second using a timedelay of 3 days (middle panel) and the third using a time delay of 5 days (bottom panel). On the x-axis is the warmup period, expressed as percentage of the total length (1 year). On the y-axis are the performances (accuracy = pink line, precision = green line and reliability = blue line), in the range between 0 and 100 %.

5.4 Discussion

The definition of a suitable model configurations to be used for simulating scenarios, is often modeller dependent. In this chapter an automated algorithm, called AMCA, was developed with the aim to reduce subjectivity in the analysis or at least make the modelling hypotheses more explicit and transparent. This algorithm cannot be considered an alternative to calibration and optimization procedures but, rather, an initial screening tool that compares possible configurations and selects the most adequate subset among those available to the modeller. Its main purpose is to identify the basic and most suitable model configuration, in terms of modelling options and parameter ranges. However, these configurations should then be calibrated and validated in subsequent steps.

The core of the AMCA is a filtering process that attempts to discard as many unsuitable model configurations (combinations of modelling options and parameter sets) as possible without compromising the representation of model structure uncertainty. The filtering process is based on the analysis of errors related to different attributes of the response (timing, volume, magnitude of the peaks and low flows). The algorithm can easily be adapted to incorporate different/additional indices, as the suggested list is only a selection of the most widely used in literature and should not be considered exhaustive. There are, for instance, no attempts to identify errors in the baseflow rate which would have required modelling separately groundwater flows.

The first step of the algorithm filters model structures based on their overall performance, discarding isolated well performing configurations. The concept of equifinality (Beven, 2006) suggests that there might be multiple model configurations that can simulate equally well a certain hydrological response. Therefore, are model configurations interchangeable? Implicitly, the preliminary selection step assumes they are. It removes model structures based on their overall performance and ignores isolated well performing configurations assuming that there will probably equifinal configurations amongst those not discarded. This assumption was tested comparing the AMCA results with and without the preliminary selection step. Results showed

that the two ensembles are not significantly different and that the selected configurations compensated for those filtered out.

In the second part of the filtering process, the Pareto Front generally identifies less than 0.01% of the realisations as non-dominated. However, the Pareto efficiency is expected to decline increasing the number of competing criteria (Corne and Knowles, 2007) and this could result in a much larger number of realisations to consider at the redundancy reduction stage. If some of the Pareto non-dominated realisations are redundant, these are removed by the redundancy reduction step which uses Self-Organizing Maps and time series matching methods. These are powerful tools to identify (also visually) clusters and inspect time series similarities. In a SOM, similar realisations arrange themselves into clusters on a 2-dimensional map projecting not only multiple performances but also the relationships amongst them into weights, node sizes and connections. One of the advantages of using SOMs over other clustering techniques is that there is no need to assign *a priori* the number of clusters. Working with SOMs, however, is not a completely objective procedure. The clustering is quite variable with the dimension of the map (Reusser et al., 2009). Using a small map size forces the data to cluster in few groups without necessarily mirroring real similarities. Conversely, increasing the map size could show the data as too dispersed. In this study, an adaptive map size was adopted, dependent on the dimension of the Pareto Front. This was necessary to make sure that each realisation could fall in a different cluster, if necessary. Further studies are needed to understand how results are affected.

From each cluster only one 'representative' realisation is selected. The selection is made via the DTW method which is based on matrix operations and therefore suitable only for relatively short-length time series. It could be interesting to investigate the DTW's limit of applicability in terms of length of time series to be processed. Some preliminary tests showed that, based on the computing capabilities of the machine used, the process would fail to allocate memory for datasets longer than 10000 time steps. There might also be a lower limit, as the effect of the initial conditions could be difficult to minimize for very short datasets. Finally, the results of

the redundancy reduction step are combined in a model ensemble. The fact that clustering and DTW only remove redundancies is confirmed by the high degree of similarity of the cumulative distribution functions calculated for the Pareto step and the RE.

The performance of the algorithm is measured in terms of accuracy, precision and statistical reliability. The information that can be derived depends on how consistent the ensemble components are. The synthetic case obtained by imposing default parameter ranges shows well defined model components which often coincide with the true modelling options used for generating the synthetic data.

The algorithm generally converges towards stable values of accuracy, precision and reliability with less than 5000 samples. Using a smaller number of parameter samples and more realistic parameter ranges considerably reduces the number of simulations and related processing time. For instance, if a realistic range of the time delay is identifiable, the number of model structures to take into consideration halves because structures not allowing routing can be removed. If the parameter sets to take into account are also halved (from 10000 to 5000), the processing is generally reduced to 25% of the original time.

It was also noticed that model component identifiability can be further improved by narrowing parameter ranges to more realistic values. One of the experiments showed that there is much less uncertainty in the definition of model components when the timedelay is set to its synthetic value. However, in this particular case, the algorithm failed to identify the correct architecture for the upper soil layers. The one suggested by the algorithm consist of a single storage unit, while the true one is made of two storages. This inconsistency may be due to the fact that modelling options are not evenly distributed across the various model structures. In fact, 576 model structures are characterised by a single storage, 384 are characterised by an upper layer split into tension and free storage, and the remaining 288 model structures have the tension storage sub-divided into recharge and excess. Therefore, as the majority of the structures contain a single upper storage option, it is likely that this option is selected as the most suitable one.

For the purpose of testing the AMCA, a short dataset (about one month) was used, even though FUSE models are generally used with multiple years of daily data (Clark et al., 2008; McMillan et al., 2010; Clark et al., 2011; McMillan et al., 2011). The reason is that the Severn at Plynlimon catchment is relatively small (below 10 Km^2) and characterised by fast responses, therefore, in order to properly describe the hydrological behaviour observations are recorded hourly. With this temporal resolution, only a short dataset would have allowed to run many million simulations in a feasible time frame: about 12.5 million simulation were generated for the synthetic test illustrated in Section 5.3.1 and 132 millions for the sensitivity tests in Section 5.3.2. However, in the next two chapters the AMCA is tested on much longer real observations.

More experiments are needed to understand whether parameter identifiability can be improved by narrowing model component options. Is a low degree of parameter identifiability linked to parameter interaction? How can this interactions be investigated? Next chapter tries to answer these questions engaging association rule mining.

5.5 Concluding remarks

This chapter describes the design, development and testing of an automated algorithm to select an adequate set of model structures and parameter ranges to simulate the dominant processes occurring in a catchment while properly accounting for model structural uncertainty. This is achieved by using a Data Mining approach and combining various techniques (e.g. set theory operations, clustering, time series matching) into one single algorithm which is able to uncover non-trivial data patterns.

The algorithm was tested using synthetic events, demonstrating that the result space is affected by many overlapping effects ranging from redundancies due to default framework settings, to systematic discharge overestimations/underestimations due to arbitrary initial conditions and sub-optimal parameter ranges.

Results are analysed in terms of precision, accuracy and reliability. It is important to highlight that the best configurations are not those corresponding to the most precise ensemble, nor to the most accurate or reliable one but rather those identifying the best trade-off. Next chapter illustrates further experiments carried out to understand whether the identifiability of model components and parameters can be improved mining their interactions.

Chapter 6

Coupling the AMCA with association rule mining to improve the identifiability of optimal model configurations

In the previous chapter, an automatic data mining procedure was introduced to identify the most suitable model configurations for a catchment of interest. The algorithm was applied to a synthetic dataset in a multi-objective and multi-model framework, returning model configurations characterised by high level of consistency with the expected results. However, the algorithm performed poorly when trying to narrow parameter ranges, probably due to high degree of parameter interaction. The aim of this chapter is to investigate how model components and parameters interact. This is done by coupling the AMCA with association rule mining, a machine learning technique that can efficiently identify the simultaneous occurrence of a set of variables within multivariate distributions. The methodology is tested on the Plynlimon area in the UK. Results show that a significant degree of interaction can be identified amongst model components and parameters. The combined use of AMCA and association rules-based filtering allowed to identify an optimum set of 2 out of 312 model structures and constrain 7 out of 15 parameter ranges. The ensemble generated by the suggested configurations was found 100%

accurate and 37% more precise than the default FUSE ensemble. The same configurations were calibrated using an optimisation algorithm and generated the best average efficiency and smallest uncertainty (for both calibration and validation periods) when compared with the calibrated FUSE's parent models.

6.1 Background

Conceptual lumped rainfall-runoff models are widely used for hydrological modelling due to their parsimonious use of parameters and fast computation time, when compared to physically based models. The parameters of a conceptual model cannot be measured *in situ*, they are evaluated by inferential procedures using historical data. For models base on non-linear ordinary differential equations, such as most FUSE model structures (Clark et al., 2008), parameter estimation becomes notoriously difficult (Strebel, 2013) because more than one set of parameters may generate the same distribution of simulations, known as equifinality problem (Beven and Freer, 2001). Parameter identification problems and sensitivity analysis often consider model parameters independent, and therefore uncorrelated. The relative importance of individual parameters in determining model performance can be investigated using regional sensitivity analysis (RSA) based on Monte Carlo simulations (Hornberger, 1980; Spear, 1980). In RSA, model simulations are evaluated using an objective function (e.g. Nash-Sutcliffe efficiency) and divided in behavioural and non-behavioural simulations depending on whether the perfomance falls above or below a given threshold. For each parameter, the empirical cumulative distribution function (CDFs) from behavioural simulations is compared to the CDF computed from non-behavioural simulations via either visual inspection or using formal statistical tests such as the Kolmogorov-Smirnov test (Spear, 1980). If the CDFs are significantly different, the parameter is considered sensitive. Pianosi and Wagener (2015) use CDFs to assess how the simulated distributions vary altering inputs. However, Mo et al. (2006) noticed that the model performance becomes behavioural (or non-behavioural) depending on the interaction between the individual parameter values, therefore, the entire parameter set is important to achieve a behavioural simu-

lation. Relaxing the hypothesis of independence, it becomes important to define the structure of a covariance matrix in order to implement a Monte Carlo procedure. Beck (1987) pointed out that, as the number of parameters increases, the posterior parameter distributions become more difficult to interpret. Multivariate analysis can be used to take into account interactions. Spear (1980) suggested to use a principal components analysis of the covariance matrix to explore the problem over multiple dimensions. According to Kuczera (1997), interacting parameters lie in small narrow regions of the parameter space and suggested a subspace probabilistic search strategy. Spear et al. (1994) explored the connectedness of these small regions of interaction using nearest neighbour metrics and concluded that "this is generally a single connected region". They suggested to describe this region using a tree-structured density estimation approach.

However, using multi-model frameworks, the complexity increases due to the number of combinations of models and parameter sets able to produce the same distribution of simulations. In the previous chapter, an ensemble of suitable configurations for a catchment of interest was generated using the AMCA approach. The ensemble was highly accurate and combined many configurations from which suggestions on optimal model components could be derived. However, excluding the routing coefficient (timedelay), optimal parameter ranges could not be identified. Assuming that the lack of identifiability is due to parameter interaction, if significant interactions are uncovered these could be used to contain the size of the ensemble improving its precision and without compromising its accuracy. As it is formulated, this is an ensemble pruning problem. This kind of problem is well known in machine learning, for which unnecessarily large ensemble can lead to a number of issues such as extra memory usage and computational costs (Zhang et al., 2006). Ensemble pruning can be approached in numerous ways, involving various degrees of complexity. For instance, statistical tests can be used to subset models that differ significantly from the others in terms of accuracy (Tsoumakas et al., 2004, 2005). If each ensemble member is given a weight, optimization can improve the performance of the ensemble by tuning these weights (Mason et al., 1998). The effectiveness of this approach depend on the number of zeros in the weights (Demiriz et al., 2002) and there is no control over the dimension of the final ensemble. In this regard, Zhang et al. (2006) proposed to constrain the weights to be

binary and set the size of the final ensemble using a cardinality constraint. All these approaches, however, assume parameters and model components to be independent of each other. In Chapter 5, frequency-based approaches that assume configuration variables independent were found to be of limited use. The true model configuration was checked, for instance, against the *Reduced Ensemble* by plotting the relative frequency of each model component (i.e. see Figure 5.19). This type of plot is useful to highlight model components occurring more often than the competing ones but did not inform on the frequency of occurrence of a particular combination of these components.

The possibility to formulate the ensemble pruning problem to enforce the independence of its variables is an active field of research. This is based on the assumption that an ensemble member should be preserved based on whether its variables show a significant level of correlation. This is a combinatorial problem that, in other disciplines, is solved using association rule mining. This technique works by identifying all possible combinations of variables within discrete multivariate distributions and sorts them based on their joint frequency of occurrence. Association rule mining was first introduced by Agrawal et al. (1993) and has been used since in computer science for mining web usage and traffic (Mobasher et al., 2001), in finance for market basket analysis[1] and in the field of bioinformatics for the analysis of gene expression (Carmona-Saez et al., 2006). Hydrological applications are relatively limited and focus mainly on finding rules for clustering hydrological time series and spatial characteristics(McGuire and Gangopadhyay, 2006; Wan et al., 2007; Malhotra and Venugopal, 2011).

This chapter explores the possibility to combine the AMCA to unsupervised learning techniques called association rules as a multi-model alternative to Kuczera (1997) and Spear et al. (1994) approaches to identify the region in the configuration space where parameters and model components interact. The ensemble members for which variables fall in the region of high interactions will be preserved, while the remaining will be pruned. Being a technique that can

[1] `http://machinelearningmastery.com/market-basket-analysis-with-association\ -rule-learning/`, accessed 27th November 2014.
`http://www.xlstat.com/en/learning-center/tutorials/how-associations-rules-can-help-\ for-market-basket-analysis.html`, accessed 27th November 2014.

be adopted for data as well as text mining Wong et al. (1999); Manimaran and Velmurugan (2013), this is particularly suitable in the context of an AMCA ensemble, for which parameters are numerical while model components are categorical variables.

The chapter is organised as follows: Section 6.2 describes the proposed methodology, Section 6.3 illustrate the modelling setup, 6.4 illustrates results for a case study and explores the limitations of the use of association rules mining, while Section 6.6 draws the concluding remarks.

6.2 Methodology

Given a catchment of interest, FUSE (Clark et al., 2008) is used as model inventory to generate simulated discharge time series combining 1248 model structures and 10000 parameter sets. The AMCA algorithm, then, is used to select from the available 12480000 configurations those that best suit the observations. The result of this procedure, as seen in the previous chapter, is the *Reduced Ensemble*, which is a table that contains all the variables for all the selected configurations.

In the context of association rules, the *Reduced Ensemble* is called *transaction table*. In a transaction table, configurations are presented such that each row of the table represents a configuration, called *transaction*, and each column represents a variable, called *item*. A group of items occurring together is called *itemset*. Each transaction is made of 33 items: 9 model components (see Tables 3.1 and 3.2) and 24 parameters (see Table 3.4).

Association rules expect all the items to be categorical variables[2], therefore, each parameter range is first divided into a number of bins of equal size, then each parameter value is converted to a categorical variable through binning. As an example, if the timedelay is defined in the range [0,5], and I divide this range into 5 bins, each parameter value belongs to one of these 5

[2]A categorical variable is defined as a variable that can assume a limited number of values so that it can be assigned to a particular group or category.

categories: (0,1], (1,2], (2,3], (3,4] and (4,5]. A timedelay of 0.2 days is converted into (0,1], 1.5 days is converted into (1,2] and so forth.

The itemsets occurring in two or more transactions are identified, collected and analysed against each other to determine whether any if-then relationships, also called *rules*, can be identified. For instance, if for any itemset the timedelay is always (0,1], this generates the following rule:

$$\{\} \implies timedelay = (0,1]$$

where the empty curly brackets stand for "any itemset". This rule has length 1, because only one item is involved. Another example involving more than one item is the following: if in a certain number of transactions the timedelay is (0,1], then the maximum water storage capacity of the upper soil layer is (120,215]. This rule can be written as follows:

$$timedelay = (0,1] \implies maxwatr\_1 = (120,215]$$

This rule has length 2, because there are two items involved: one on the left hand side (lhs) and one on the right hand side (rhs).

In general mathematical terms, rules are expressed as follows:

$$X \implies Y \tag{6.1}$$

where X is called *antecedent itemset* (lhs) and can be made of 1 or multiple items. Y is called *consequent itemset* (rhs) and it is made of only 1 item. X and Y are mutually exclusive and the number of identifiable rules depends on the number of observed itemsets. The more variables are in a transaction, the more rules may be generated. However some rules are more significant than others and Agrawal and Srikant (1994) designed the *apriori* algorithm to define a way

to choose the most relevant rules amongst all the possible ones, based on three measures of significance: support, confidence and lift.

The support is the proportion of transactions containing a certain itemset.

$$Support = supp(X) = f_i/n \tag{6.2}$$

where f_i is the frequency of a certain itemset and n is the total number of transactions.

The confidence is the ratio of the observed support to the support of X and is an indication of the number of times the if-then relationship was true in the transaction table:

$$Confidence = conf(X \implies Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{6.3}$$

The lift is the ratio of the observed support to that expected if X and Y were independent:

$$Lift = lift(X \implies Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)} \tag{6.4}$$

The apriori algorithm assumes, by default, that a rule is significant if the support is above 10% and confidence is above 80%. In this work, I used a slightly more restrictive threshold for the support (30%) to be able to work with a small and highly significant set of rules. The implications of such subjective decision are discussed further in Section 6.5. It is important to note that this algorithm does not distinguish the direction of the implication (i.e. whether X implies Y or viceversa), to remove any redundancies and avoid double counting, the list of rules should be pruned before visualisation or further manipulation. All the possible rules can be ordered in a sparse symmetrical matrix, therefore, rules pruning simply consists of retaining the elements above (or below) the diagonal.

There are several visual aids that can help exploring pruned rules. They can be visualised schematically as in the examples in Figures 6.1 to 6.3. The scatterplot (Figure 6.1) is used to determine how the lift changes based on support and confidence. Generally, rules with high lift

have a relatively low support and the most interesting rules reside on the boundary (Bayardo and Agrawal, 1999).

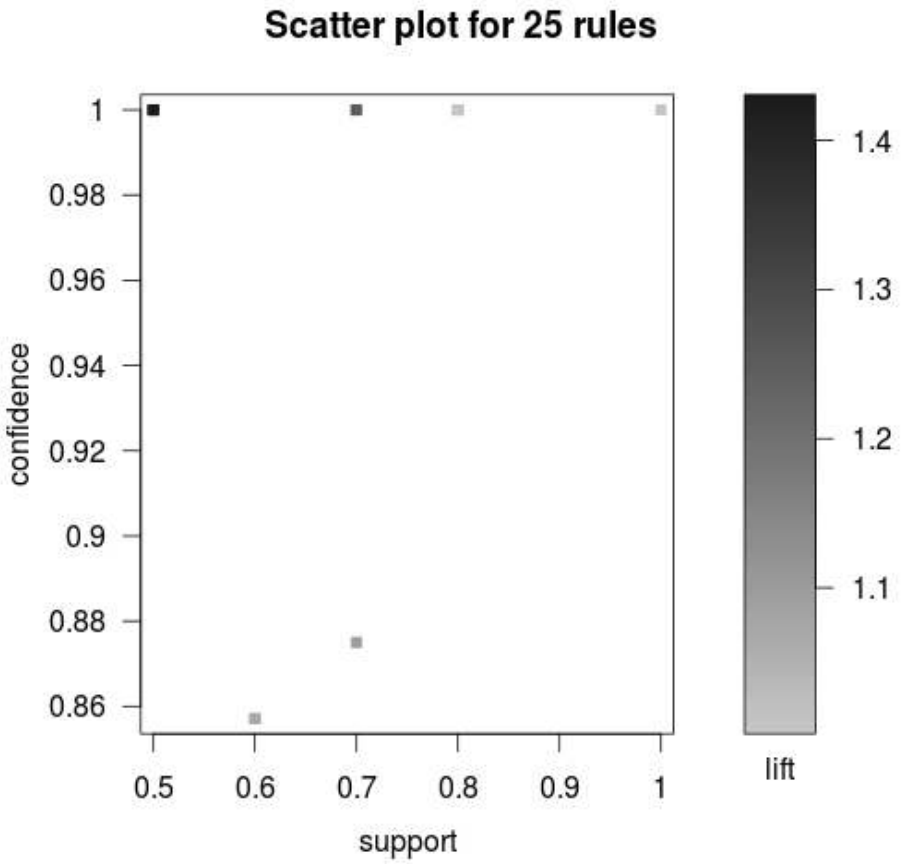

Figure 6.1: Example scatter plot showing support level on the x-axis, confidence level on the y-axis. Each point is colour coded from light to dark grey to show the degree of lift.

Figure 6.2 illustrates the circle-connection graph which shows items as text, rules as circles (colour coded based on the lift and with size dependent on the support) and implications as arrows. This plot is useful to highlight items of particular interest (arrows converging to a certain item).

The parallel coordinates plot (bottom) shows on the x-axis the number of items forming a rule. In this particular example, the consequent itemset (rhs) is made of 1 item, while the antecedent itemset (lhs) is made of a maximum of two items (position 1 and 2). On the y-axis are the items involved in the rules. The arrows are colour coded based on the lift. The parallel coordinates plot is useful because it detects the most significant links amongst variables. In the example

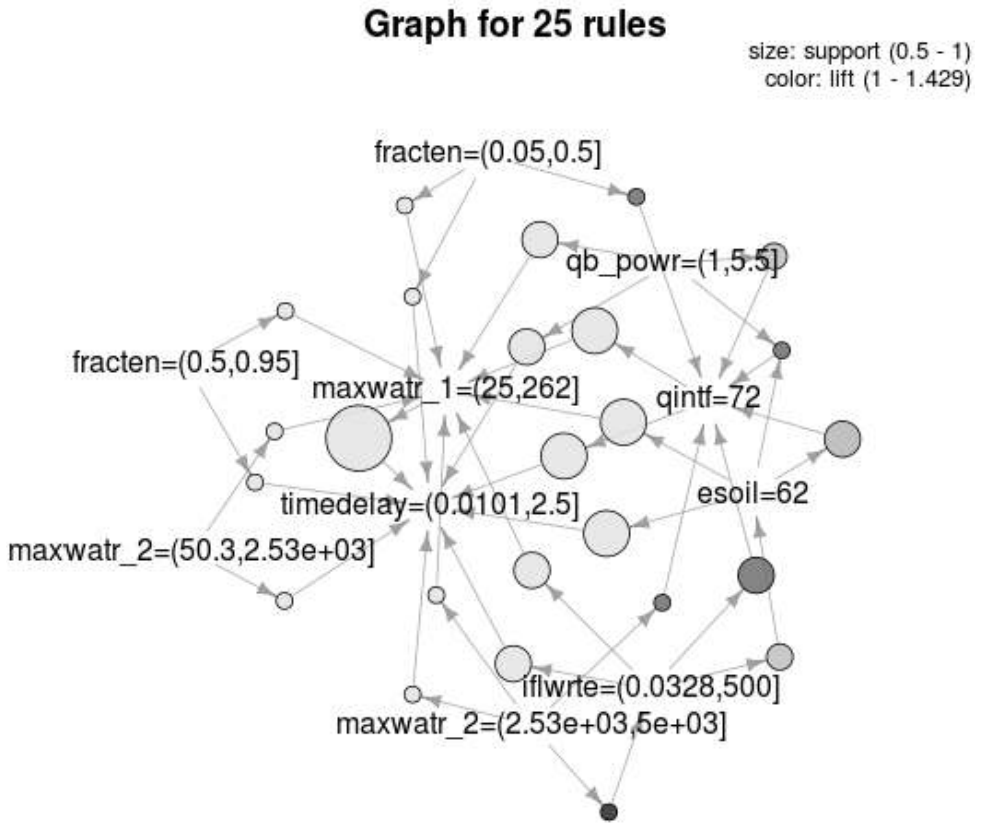

Figure 6.2: Example circle-connection graph which shows items as text, implications as arrows and rules as circles (size and colour depend on the support and lift, respectively).

one of the arrows is coloured in red to demonstrate how to read a rule from the plot. The rule is made of 2 antecedent items: qb_powr in the range (1.5,5] is in position 2 and maxwatr_1 in the range (25,262] is in position 1 of the lhs. The simultaneous occurrence of these two items causes the selection of possible interflow (qintf = 72), which is the rhs item.

The scatterplot is considered the preferred plotting option in the case of a high number of rules, as the cloud of points and connections would make the other plots unreadable. The circle-connection graph and the parallel coordinate plot are, instead, preferable when dealing with a low number of rules.

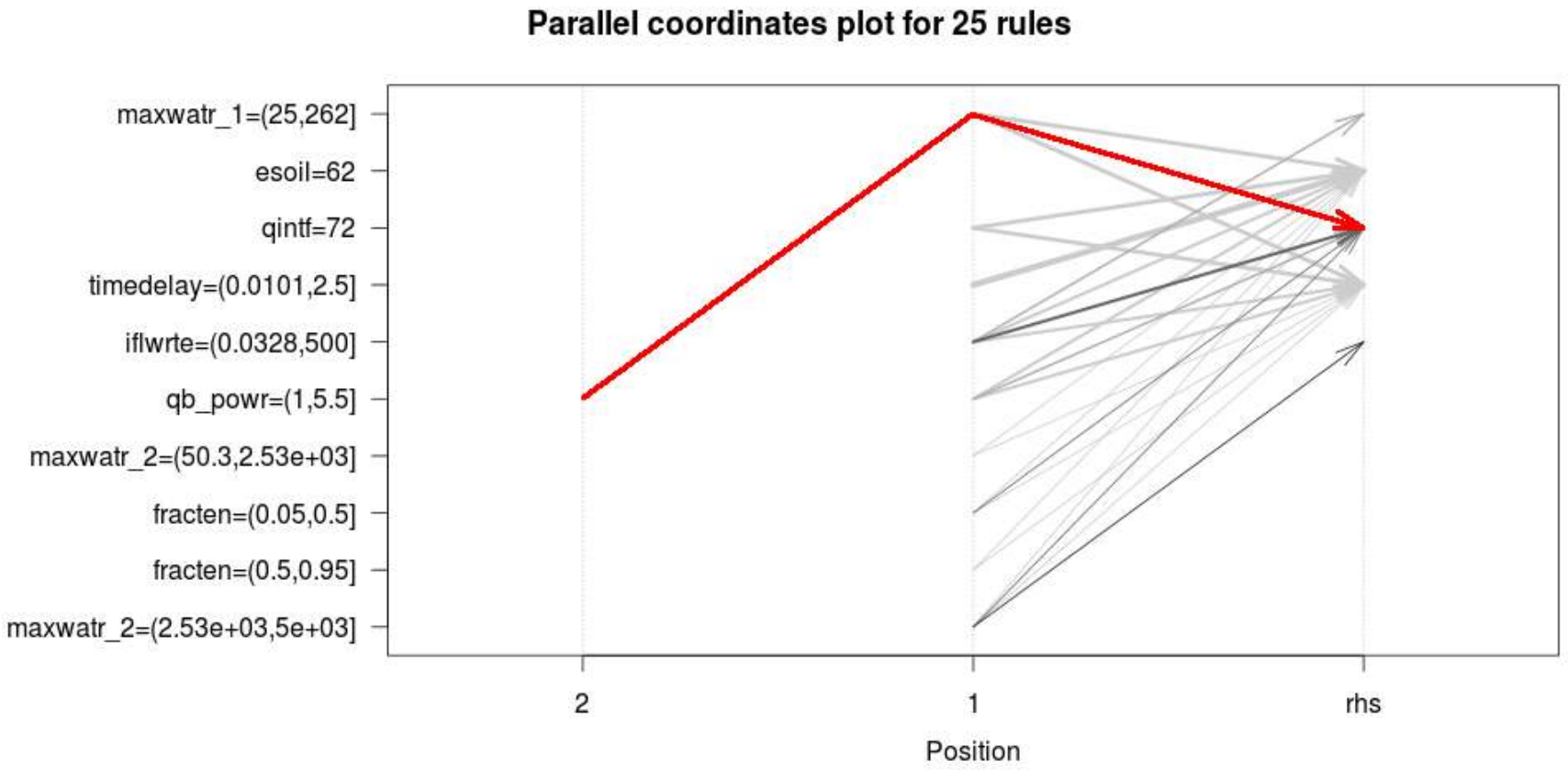

Figure 6.3: Example of parallel coordinates plot which shows on the x-axis the number of items forming a rule and on the y-axis the items involved in the rules. The arrows are colour coded from light to dark grey based on the lift. The red arrow is used to illustrate the following example: the simultaneous occurrence of qb_powr in the range (1, 5.5] and maxwatr_1 in the range (25,262] is likely associated with interflow mechanisms.

In order to interpret the results of association rules plots, it is suggested to refer to the lookup tables 3.2 and 3.4, that list the FUSE model building decisions, options and depending parameters.

Interactions amongst model components and parameters are identified in terms of association rules using a recursive approach which consists of the following steps:

1. run the AMCA using default settings,
2. identify interactions using association rules,
3. narrow the AMCA model structures/parameters based on the identified rules and re-run the AMCA,
4. repeat points 2 and 3 until no additional interactions can be identified.

At each loop new parameter sets are sampled, using the Latin Hypercube Sampling method. As the number of model structure reduces, the sampling becomes more efficient because takes into account only the parameters used in the selected model structures.

Lastly, the configuration obtained pruning the original ensemble using association rules, called AR hereafter, is compared to the default FUSE configurations to test whether the use of association rules leads to an ensemble more precise than the original one, without compromising its accuracy.

Increasing the precision of an ensemble should allow to reduce the expected uncertainty in predictions. To test this hypothesis, the configurations in AR are calibrated and compared with the calibrated results obtained from the 4 FUSE parent models.

6.3 Case study and modelling set up

An experiment is set up to determine how model configuration components tend to interact and whether these can be used to improve parameter identifiability. The proposed methodology is tested on the Severn at Plynlimon flume catchment (UK) in the period from 1975 and 1984. During this 10 year period, there is no record of human activities that could have altered the hydrological regime. The first five years (from 1975 to 1979) are used as training period, while the period from 1980 to 1984 is used as validation period.

The AMCA algorithm was run using 312 model structures (the rainfall error is not accounted for and the routing is always allowed) and 10000 parameter sets sampled, using a Latin Hypercube, from the default ranges. The RE tables for each year were appended one after the other to generate the transaction table and the levels of support and confidence are set to 30% and 80% respectively. The minimum length of the rules is set to 2 when looking for simultaneous occurrences, it is set to 1 when the search is extended to also independent frequent variables. Parameters are converted splitting the ranges into 5 categories.

The AR model configurations and the FUSE parent models (Topmodel, ARNOVIC,PRMS and Sacramento) were calibrated separately over the period from 1975 to 1979 using the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt and Ter Braak, 2011), limiting the calibration runs to a maximum of 10000 function evaluations per model structure. The DREAM algorithm was used to identify the parameter set in each model that minimises the Nash-Sutcliffe efficiency factor (Nash and Sutcliffe, 1970). The calibrated parameters were then used to validate the results over the period from October 1979 to April 1985.

6.4 Results

Parameters and model components are first analysed separately. Figure 6.4 shows that, at the first loop, only two rules emerged, each made of 1 item. This means that these parameter ranges

are frequent but that one does not necessarily implies the other. The first rule suggests that the timedelay should be in the lowest sub-range ([0.01,1]), which is consistent with the small size of the catchment and its typical flashy response. This rule has strong significance, with support and confidence of about 0.8 and a lift of 1. The second rule suggests that the maximum storage in the lower soil layer should also be narrowed to the lowest sub-range ([50,1000]). Although characterised by a lower support and confidence (0.3), this rule is consistent with the low permeability soils characterising the area.

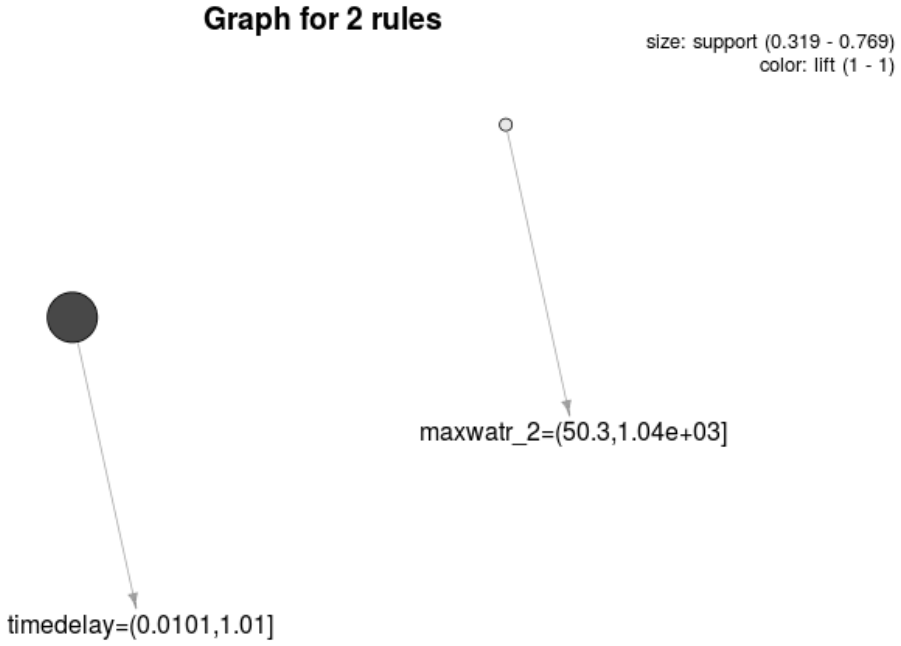

Figure 6.4: Association rules for parameters in the period 1975-1979, loop n. 1. At this initial stage there are only two rules identifiable: regardless between 30% and 70% of the selected realisations are characterised by a timedelay and maxwatr_2 parameters in the lowest range. As the circles are not connected, it can be derived that these two rules do not necessarily occur within the same model structure.

The process is repeated by restricting the analysis to model components only. Figure 6.5 shows 38 pruned rules emerged. The rules are made of 1 consequent item (rhs on the x-axis) and a maximum of 3 antecedent items. The rule with highest lift (darkest grey), for instance, converges to a lower soil layer architecture made of a single state reservoir with no evaporation (based on a power recession law, arch2=34). This item is selected as a consequence of the simultaneous occurrence of the ARNOVIC model (qsurf=41), an upper soil layer made of a single

reservoir (arch1=21) and no interflow (qintf=71). Other suggestions can be derived following the trajectory of the other arrows. In summary, the 38 rules suggest:

- to limit the upper soil layer to either a single state (21) or separate tension storage (22);
- to limit the lower soil layer to a single state reservoir with no evaporation (based on a power recession law) (34);
- to limit the runoff mechanism to those defined in the ARNOVIC model (41) and the PRMS model (42);
- drainage above field capacity (51);
- rootweight evaporation (61);
- and absence of interflow (71).

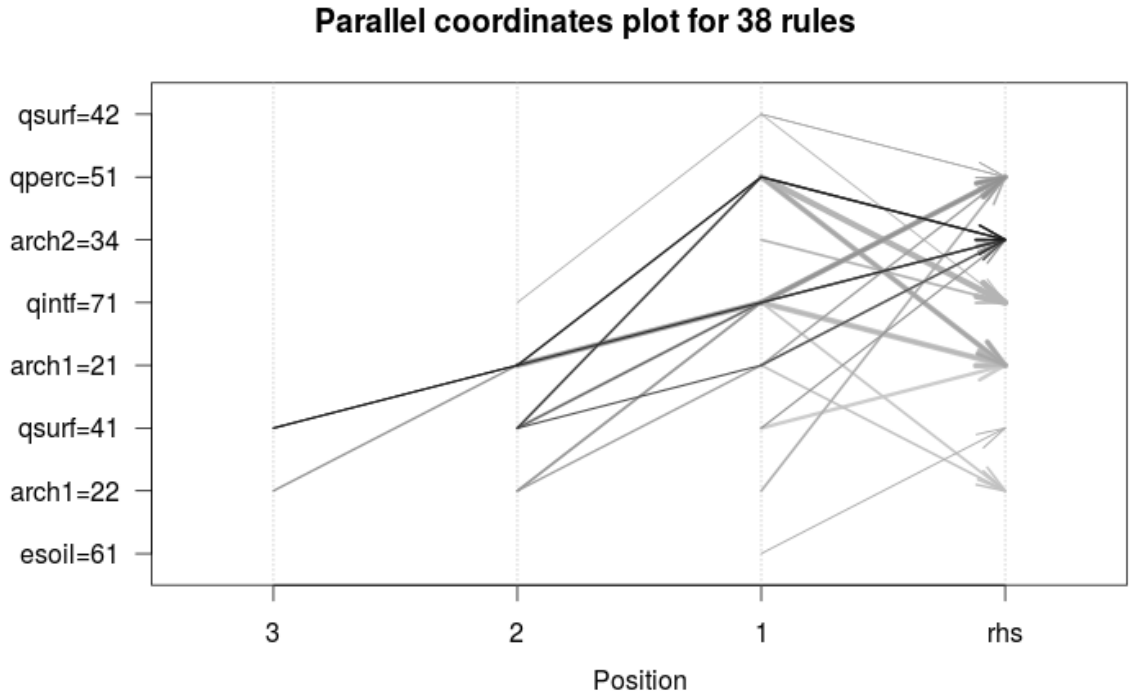

Figure 6.5: Association rules for model components in the period 1975-1979, loop n. 1. The strongest rule (darkest arrow) seems to suggest that if the runoff mechanism at this catchment uses an ARNOVIC parameterisation, then the baseflow model component should be schematised as an unlimited reservoir with power recession law.

Finally, model components and parameters are combined and the minimum length of the rules is set to 2, so that only rules showing evidence of interactions are considered. A total of 45 rules emerged (see Figure 6.6) confirming interactions among the same model components and parameters identified above with the addition of a weaker but still significant possibility that the lower layer could be described by a tension reservoir plus two parallel tanks (arch2=32). Although the parameters timedelay and maxwatr_2 do not necessarily interact with each other, there is evidence of interaction with certain model components. For instance, the timedelay in the lowest range is always linked to the absence of interflow. The maximum lower storage in the lower range, instead, shows a direct implication with the rootweighted evaporation mechanism and runoff mechanisms depending on the saturation rate of the upper soil layer. The strongest rule (darkest arrow) shows that the combination of a ARNOVIC runoff mechanism and a lower soil layer made of a tension reservoir plus two parallel tanks (arch2=32) implies the frequent occurrence of an upper soil layer made of a single reservoir. These rules represent all the most

frequent model configurations and significant interactions, they are used to constrain the AMCA default settings and start a new loop.

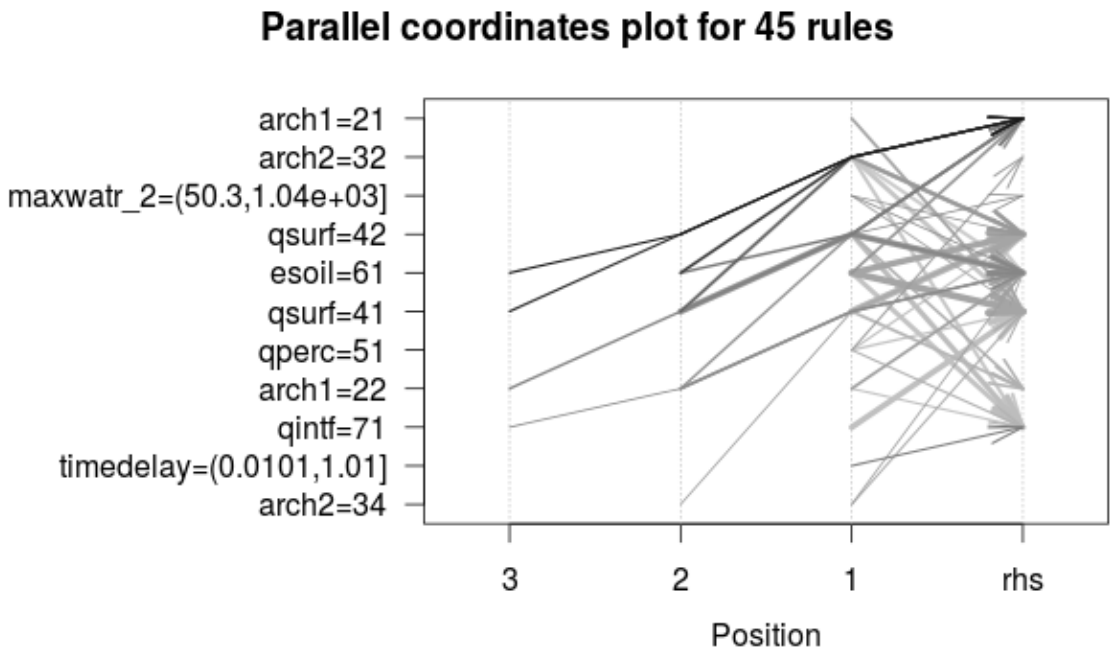

Figure 6.6: Association rules identified for parameters and model components in the period 1975-1979, loop n. 1. Here the strongest rule (darkest arrow) seems to suggest that if the runoff mechanism at this catchment uses an ARNOVIC parameterisation, then the baseflow model component could be schematised as a tension reservoir and two parallel tanks combined with single storage in the upper soil layer.

The second loop is characterised by a slightly smaller parameter space, as only two parameter ranges were narrowed (timedelay and maxwatr_2). However, the model space significantly reduced, from 312 to only 8 model structures. The new set of rules identified confirms a strong interaction with a lower soil layer made of a tension reservoir plus two parallel tanks (line converging to arch2=32 in Figure 6.7). Interactions involving the upper soil layer architecture do not add additional information to the previously identified configuration, while the PRMS mechanism seems to be preferred to the ARNOVIC. These considerations reduce the number of model structures to 2. The rule with highest support (largest circle), for instance, assumes that if the upper soil layer is made of a single reservoir, the runoff mechanism is PRMS and the maximum saturated area (sareamax) is in the range [0.05,0.41], then the lower layer is likely to be characterised by a tension reservoir plus two parallel tanks (arch2=32). Similar suggestions

can also be made in regard to the following parameters:

- baseflow depletion rates for primary and secondary reservoirs are limited to the lowest ranges (qbrate_2a and qbrate_2b in range [0.001,0.05]);
- fraction of percolation to tension storage in the lower layer (percfrac) in the range [0.59,0.77];
- maximum total storage in upper soil layer in the lowest range (maxwatr_1 in [25,120]);
- lastly, the timedelay could be further narrowed to the range [0.2,0.4].

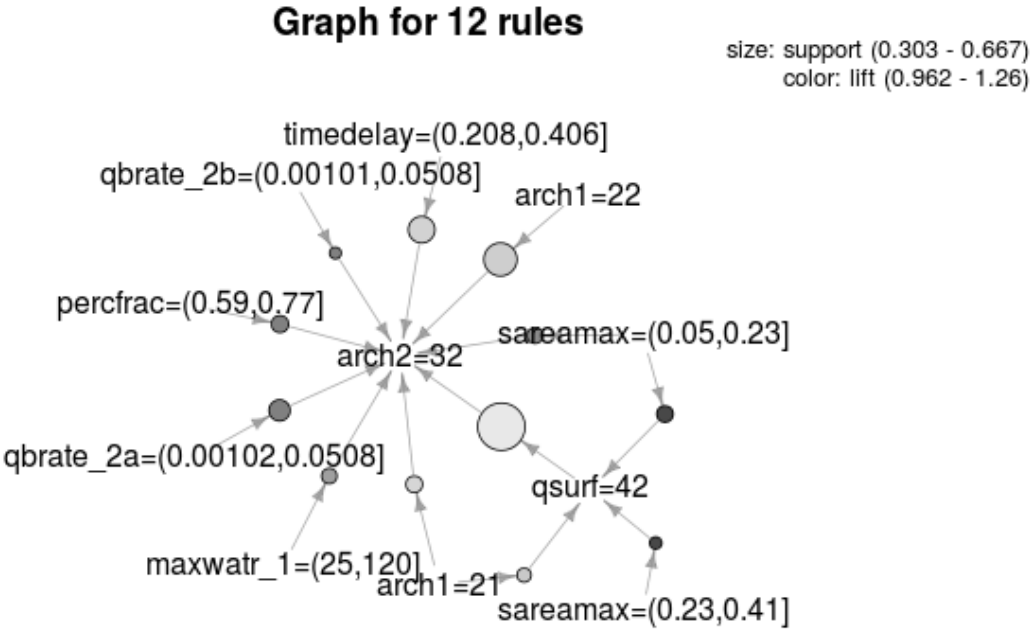

Figure 6.7: Association rules identified for the REs in the period 1975-1979, loop n. 2.

One more loop followed until no more rules where identified for the pre-defined threshold of support and confidence. Table 6.1 summarises the final suggested configuration.

Figure 6.8 shows a large event occurred between February and March 1982 (black line). The ensemble generated using the default FUSE configurations (yellow polygon) is compared with the ensemble generated by the AR configurations in Table 6.1 (green polygon). The event is bracketed with 100% accuracy by both FUSE and AR but the latter is 37% more precise.

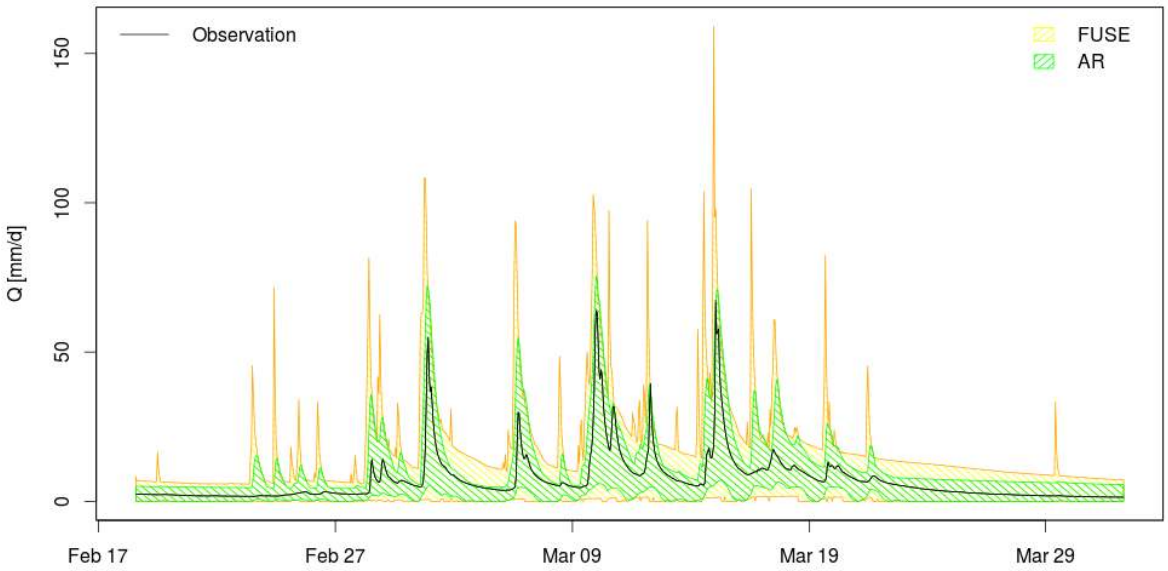

Figure 6.8: Simulated ensembles for a large event in February/March 1982. The observation is shown as black line, while the yellow and green polygons show the maximum extents of the ensembles generated using the default FUSE and AR (from Table 6.1) configurations, respectively.

To test whether the reduction in precision determines lower prediction uncertainty, the AR configurations is calibrated and compared with the calibrated results using the FUSE parent models (FUSE60 = Topmodel, FUSE230 = ARNOVIC, FUSE342 = PRMS, FUSE426 = Sacramento). Figures 6.9 and 6.10 show the boxplots of function evaluations after convergence for the calibration and validation periods respectively. As expected, the AR configuration returns the best average efficiency. This is also characterised by the narrowest spread with a standard deviation of 0.01 for the calibration and 0.02 for the validation period. Although FUSE230 and FUSE426 show the potential to reach the highest performances in the validation period, they are also characterised by a much wider spread, with a standard deviation of 0.09 and 0.16 respectively.

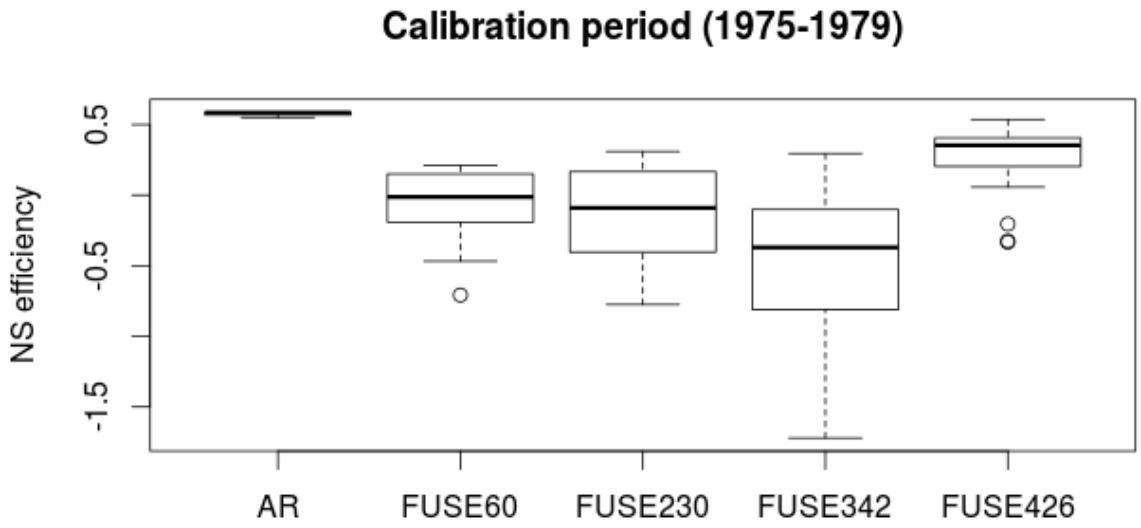

Figure 6.9: Boxplots of NS efficiencies for FUSE models 60 (Topmodel), 185 (configuration identified by the association rules), 230 (ARNOVIC), 342 (PRMS), 426 (Sacramento) over the calibration period.

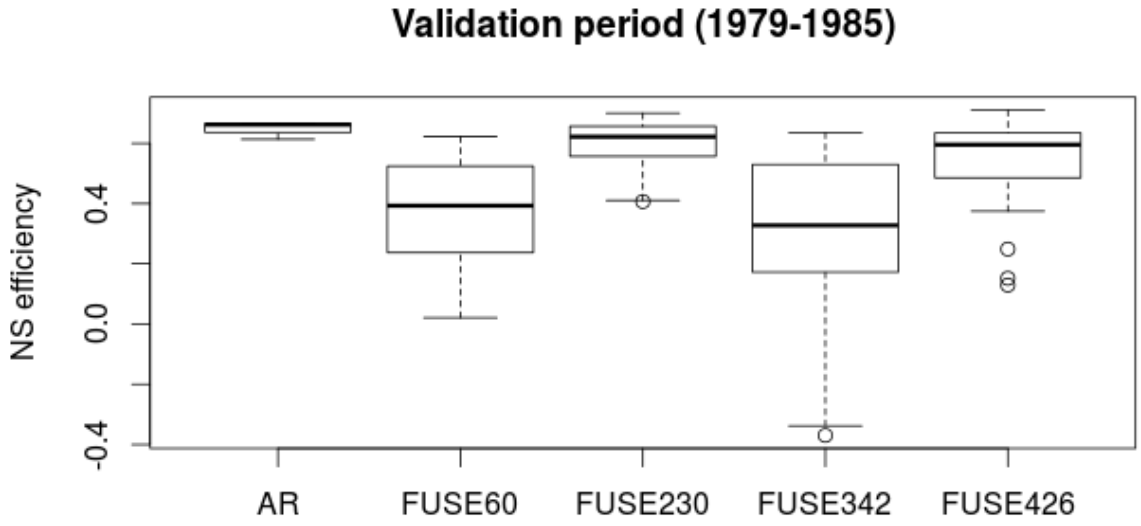

Figure 6.10: Boxplots of NS efficiencies for FUSE models 60 (Topmodel), 185 (configuration identified by the association rules), 230 (ARNOVIC), 342 (PRMS), 426 (Sacramento) over the validation period.

Table 6.1: Suggested model configuration obtained by coupling AMCA and association rule mining techniques for the period 1975-1984 (pre-fell). The line divides the first six rows, describing the suggested model structures, from the list of optimal parameter ranges. Ranges in bold are narrower than the default ones.

Description	**Name**	**Range**
Upper layer architecture made of either a single state variable or tension and free storages	onestate_1 tension1_1	21 - 22
Lower layer architecture made of a tension reservoir and two parallel tanks	tens2pll_2	32
Runoff PRMS-like, unsaturated zone linear	prms_varnt	42
Percolation scheme with water availability from field capacity to saturation	perc_f2sat	51
Evaporation, rootweighted	rootweight	61
Interflow not allowed	intflwnone	71
Additive rainfall error (mm/day)	rferr_add	0
Multiplicative rainfall error (-)	rferr_mlt	1
Fraction total storage as tension storage (-)	fracten	[0.05, 0.95]
Maximum total storage in upper soil layer (mm)	maxwatr_1	**[25, 120]**
Fraction of percolation to tension storage in the lower layer (-)	percfrac	**[0.59, 0.77]**
Fraction of storage in the first baseflow reservoir (-)	fprimqb	[0.05, 0.95]
Baseflow depletion rate in the first reservoir (day^{-1})	qbrate_2a	**[0.001, 0.05]**
Baseflow depletion rate in the second reservoir (day^{-1})	qbrate_2b	**[0.001, 0.05]**
Maximum total storage in lower soil layer (mm)	maxwatr_2	**[50, 1000]**
Fraction of roots in the upper layer (-)	rtfrac1	[0.05, 0.95]
Percolation rate (mm/day)	percrte	[0.01, 1000]
Percolation exponent (-)	percexp	[1, 20]
Maximum saturated area (-)	sareamax	**[0.05, 0.41]**
Mean value of the log-transformed topographic index (m)	loglamb	[5, 10]
Shape parameter for the topo index gamma distribution (-)	tishape	[2, 5]
Baseflow exponent (-)	qb_powr	[1, 10]
Time delay ($days$)	timedelay	**[0.2, 0.4]**

6.5 Discussion

There are various subjective decisions related to association rule mining. The first occurs when numerical variables are converted into categorical ones. The conversion assumes that the ranges are split into bins but, the ideal number is not well defined. This certainly depends on the extent of the range and on the models' sensitivity to parameter variation. If a low number is chosen (e.g. 2 or 3) and the ideal range falls across two or more bins, the identifiability of the parameter could be hindered. Choosing a higher number, multiple bins can be merged together

if necessary, but there is also the risk that the occurrences in each single bin do not reach the minimum level of support and confidence.

The second subjective decision involved in association rule mining: what is the ideal threshold for support and confidence levels? The apriori algorithm uses as default threshold 10% for support and 80% for confidence level. Adopting these thresholds, thousands of rules where identified and they did not highlight any particular interaction. On the contrary, rising the level of support above 0.5 did not return any significant rule. More experiments should be carried out to identify the ideal trade-off, however a support threshold to 0.3 was deemed to be reasonable as it allowed to identify configurations highly consistent with the observed level of soil permeability and type of expected hydrological regime.

The minimum length of items per rule is another factor controlling the number of identifiable rules. In the experiments carried out in this chapter, the minimum length was set to 2. Rules made of 1 item have no implications to other items, meaning that they are frequent but do not show significant evidence of interactions.

The recursive nature of this approach allows to discover interactions over multiple layers of complexities. In the first loop parameter interactions were not visible, probably due to the overwhelming timing errors (in Chapter 5 was already noticed that ensemble results are highly sensitive to the timedelay). Once this first layer of complexity was revealed, new rules became identifiable allowing to constrain 7 out of 15 parameter ranges relevant for the selected model structures. Unconstrained parameters are probably the less sensitive.

Identifying optimal model components and parameter ranges allows to considerably increase the precision of the ensemble without jeopardising its accuracy. This reduces the expected uncertainty, improving the usability of the ensemble for problems related to design.

The approach to combine AMCA and association rules, to select suitable models and identify dependencies between model components and parameters, could be easily transferred to other modelling framework, assuming that the required processing time is similar to FUSE's.

Conversely, frameworks made of large distributed models, instead, would not be compatible because the processing would be too computationally expensive. However, association rule mining could be used on its own, as machine learning tool, to learn about variable interdependencies for any model settings. An interesting application, for instance, could be to set up a regionalisation experiments in which a spatially distributed model (e.g. the UK MetOffice's JULES model) is calibrated on a number of catchments where association rules are used to derive dependencies between model parameters and catchment characteristics and automatically construct spatial regression laws.

6.6 Concluding remarks

This chapter suggests to combine the AMCA to association rule techniques to guide the interpretation of ensemble results. Interactions among parameters and model components are revealed over multiple layers of complexities using associations rules to constrain the AMCA settings through a recursive approach.

The Severn catchment closed at Plynlimon flume was used as study area because of its limited extension and the homogeneity of its soils and vegetation. Results show that the most significant rules can be used to constrain the default model configuration increasing the average simulated performance. According the suggested configurations, the Severn at Plynlimon flume appears to be characterized by an upper soil layer characterised by either a single storage or divided into free and tension storages, with a maximum capacity below 120mm and a fraction of percolation to the tension storage in the lower layer in the range [0.59,0.77]. The lower soil layer is made of a tension reservoir plus two parallel tanks with a maximum storage capacity of 1000mm and depletion rates below 0.05 day^{-1} for both primary and secondary reservoirs. The runoff generation mechanism depends on the excess of saturation in the tension storage of the upper soil layer (PRMS) with a maximum saturated area of 41%. The percolation depends on the field capacity, the evaporation on the fraction of soil in which there are roots, while lateral flows are

negligible. Finally, the delay in runoff is expected to be in the range [0.2,0.4] days.

Association rules appear to be a powerful tool to identify model interactions and guide the configuration of the ideal model ensemble. A similar approach could be used to investigate whether significant changes in the hydrological regime can be mapped in terms of model configuration shifts. This will be further explored in the next chapter.

Chapter 7

Using a probabilistic multi-model framework to predict the effects of land use changes on catchment flows

Land use and cover changes are altering the Earth's surface at an unprecedented pace. Assessing their impact is still problematic, due to the limited availability of documented cause-effect records. In this chapter, a multi-model probabilistic framework is presented to predict the effects of land use changes on catchment flows. The framework uses regionalised information on soil and vegetation summarised in two catchment signatures: the Base Flow Index and the Curve Number. The former is an indicator of the runoff properties for a particular area under extreme precipitation events and therefore high flow response, while the latter is an indicator of low flow response. As these signatures can theoretically classify a wide range of catchment responses, in this work they are used to condition model configurations and make predictions for deforestation/afforestation scenarios. Performances are assessed using the Nash-Sutcliffe efficiency analog for probabilistic predictions. The methodology is tested on the upland area of Plynlimon in the United Kingdom. Results show that flow peaks increase and become more sensitive to changes as the land use condition worsen. Prediction bounds from regionalised information are

characterised by high accuracy but low sharpness, probably due to the signatures's low constraining power. However, the same signatures allow to discern amongst modelling options and identify configurations similar to those suggested by the AMCA in the previous chapter. When the AMCA approach is used to condition the prior, performances improve significantly. This is due to constrains applied to both model and parameter space, however the latter contributed to 80% of the improvement.

7.1 Introduction

Human activities can significantly shape the hydrology of a catchment, affecting hydrological processes in a complex way that makes it difficult to predict their impact on streamflow variability. Upland agricultural management practises such as deforestation and intense grazing, for instance, can reduce water interception and infiltration to the deeper soil layers determining faster surface runoff, and consequently increasing the frequency and magnitude of floods (Marshall et al., 2009; Boardman et al., 1994; Burt, 2001).

A number of studies have attempted to assess quantitatively the impact of deforestation and tree plantation on local hydrology (McCulloch and Robinson, 1993; Bird et al., 2003). Main challenges are related to limited data availability, lack of modelling methods and tools (McIntyre et al., 2013). For only few experimental catchments there is a historical record of land use changes and information and methodologies available to characterise the impact of these non-stationarities are still limited (Beven, 2001*b*) and characterised by major uncertainties. There is, therefore, scope for further investigations that could provide reliable guidelines for policy and land management strategies (Bird et al., 2003).

The effects of land use changes on streamflow generation can be assessed using methods based on either paired catchment analysis or hydrological models (Zégre et al., 2010). The former method investigates the statistical relationships between the flow records at a control catchment (A) and at a treated catchment (B), where A and B have similar characteristics in terms of soil,

vegetation, topography and climate. The relationship is often expressed in terms of regression model that is calibrated on the pre-change period. When a change in land use occurs on catchment B, this is detected if the residuals exceed the model prediction limits. A comprehensive list of paired catchment studies can be found in Brown et al. (2005). Changes occurring in the same catchment can also be analysed. This is usually done using a first time window as *reference period* and observing the response to a change in vegetation over a second time window. In this approach, however, the predictive uncertainty is inflated by the effect of weather variability.

An alternative method to paired catchment analysis is based on hydrological modelling. Within this category, various approaches are possible. The classical bottom-up approach defined by Blöschl and Sivapalan (1995) assumes that the model structure is defined a-priori by the modeller, the model is physically-based and predictions are made through detailed simulation of hydrological processes at small temporal and spatial scales. On one side, this type of models allows to characterise explicitly the spatial heterogeneity of catchment properties and investigate the corresponding physical changes. On the other side, they are computationally expensive and the interpretation of local processes is not straightforward to scale up to the catchment level (Ballard, 2011). Dominant processes are scale dependent and some of them can be considered negligible at a small scale but more and more relevant at bigger scale (e.g. large-scale preferential flow paths in the subsurface, Sivapalan et al., 2003). Furthermore, in case of limited data availability the risk of over-parameterisation becomes larger thus leading to large uncertainties in predictions, parameter non-identifiability and equifinality (Beven, 2000*a*; Grayson and Blöschl, 2001). Therefore, in many studies, the rigid bottom-up methodology, that relies on complex physically based models, has been set aside preferring a more generic and flexible top-down approach. This approach consists of defining statistical relationships between climate forcing (e.g. rainfall) and hydrological responses (e.g. streamflow) without attempting to interpret the underlying physical processes (Sivapalan et al., 2003). Bowling et al. (2000), for instance, used a mechanistic model to assess the effect of logging on peak flow events. Although less complex, these models return results that cannot be safely extrapolated to predict, for instance, scenarios that have not been previously observed (Ballard, 2011). Nowadays, such

unobserved scenarios are designed using agro-economic models which can predict feasible land use systems based on political, economic, social and ecological conditions (Breuer et al., 2009). Some examples with variable degrees of complexity are the ProLand model (Weinmann et al., 2006), the CLUE model (Veldkamp and Fresco, 1996) and four basic conceptual models designed by Hersperger et al. (2010). There is, therefore, the need to interpret future scenarios using parsimonious and computationally efficient models. This need is met by conceptual models that treat the basin as a single homogeneous entity in the sense that all the inputs are lumped over the catchment area (e.g areal average rainfall) and develop a single outflow hydrograph (Jones, 1997). The parameters are constant over the catchment area but can vary significantly from one catchment to another.

To be able to apply this type of models in data poor as well as data rich areas, it is important to estimate the spatial variability of model parameters. In data-poor environments, for instance, Buytaert and Beven (2009) used TOPMODEL to predict streamflow discharge in a set of catchments in the Ecuadorian Andes. In this experiment, parameters are first generated for a donor catchment and then transformed using a probabilistic transformation function, to migrate the model properties from one catchment to another. According to McIntyre et al. (2005), another common approach is to consider the regression of model parameters against catchment descriptors. In other words, if there is a clear correlation between a catchment descriptor (e.g. area) and a model parameter, it is possible to define a statistical model to infer the parameter value given the descriptor with a certain level of confidence. They reviewed a number of regionalisation studies and concluded that this type of regression methods have a limited applicability because of parameters interdependencies and the weak relationships between model parameters and catchment descriptors. They suggested to improve flow prediction using ensemble modelling and a probabilistic approach called Similarity Weighted Averaging (SWA). Bulygina et al. (2011), in the same context, proposed a Bayesian conditioning method that uses regionalised indices to investigate the effect of land use changes in the UK. These experiments utilised a single model structure, the Probability-Distributed Model (PDM), and predicted flows in various soil/vegetation conditions. McIntyre et al. (2005) identified as primary source of prediction

bias the model structure error and suggested to integrate the results of a wider range of model types in further work.

This chapter builds upon the methodology proposed by Bulygina et al. (2011) to predict the impact of replacement of forest by pasture (and viceversa) in areas where these changes have not been observed yet. The novelty of this work stands in the integration of multiple models in a probabilistic framework, as was suggested by McIntyre et al. (2005), and shows whether the effect of land use changes can be estimated more precisely allowing to switch amongst different models.

Section 7.2 describes the proposed methodology, Section 7.3 the data sources selected for the Plynlimon study area and the land use scenarios and 7.4 the modelled set up. Appendices C and B provide additional details on the procedures carried out to calculate catchment signatures from soil and vegetation data and time series data of precipitation and streamflow discharge. Section 7.5 illustrate the results obtained for the Severn at Plynlimon flume and the Wye at Gwy flume, over two different periods, while Section 7.6 draws the main conclusions.

Functions and scripts needed to reproduce the results of this work are publicly available under the *ad-hoc* implemented R-package "CurveNumber"(Vitolo and Le Vine, 2015).

7.2 Methodology

The objective of this work is to provide a methodology to estimate the streamflow discharge under unobserved deforestation/afforestation scenarios, in gauged as well as in ungauged catchments. Given a study area and the regionalisation model underlying the soil and vegetation classifications (e.g. USDA and HOST), the proposed methodology consists of the following steps:

1. Obtain regionalised catchment signatures based on information about soil and vegetation.

2. Using a set of rainfall-runoff model structures and a prior parameter distribution calculate signatures for different combinations of model structures and parameters.

3. Calculate the corresponding posterior parameter distribution based on the likelihood of simulated signatures.

4. Based on the considered land use change scenario, modify the information about vegetation and repeat steps 1 to 3.

7.2.1 Catchment signatures

Bulygina et al. (2011) suggests to summarise the hydrological properties of a catchment using two catchment signatures: the Curve Number from the U.S. Department of Agriculture's (USDA) Soil Conservation Service soil and land use classification and the Base Flow Index from the UK Hydrology of Soil Types (HOST) classification.

According to the USDA, the CN is an empirical measure that characterizes the runoff properties for a particular area. It is an integer between 0 and 100, where higher values mean more surface runoff and less infiltration to deeper soil layers (USDA, 1986; Reynolds et al., 1988). It can be calculated from an event-based analysis of observed precipitation and streamflow discharge time series using, for instance, the asymptotic method suggested by Hawkins (1993). Over the years, the CN has also been experimentally measured for a wide variety of catchments, allowing the tabulation of the results based on the combination of soil and land cover classes (USDA, 1986). The availability of tabulated values makes it possible to estimate CN for ungauged catchments from soil and vegetation maps, and use it as catchment signature in regionalisation problems (Bulygina et al., 2011; Blöschl et al., 2013).

Bulygina et al. (2011) showed that the CN method can also be used to identify the hydrological properties of catchments in the UK, and propose a mapping between the USDA classes and the HOST classification developed by the UK Institute of Hydrology[1] (Boorman et al., 1995).

[1]The UK Institute of Hydrology is now part of the Centre for Ecology and Hydrology.

A similar mapping was proposed by Halcrow and University of Stirling (2011) for the Allan Water at Kinbuck catchment in Scotland (UK). They also proposed a mapping between the runoff Curve Number for HOST hydrologic soil group and dominant Land Cover Map 2000 classification (LCM2000 code).

There are a number of location-dependent factors that may significantly affect the mapping between runoff Curve Number and the Land Cover Map code, such as topography, vegetation, soil's hydrological condition, drainage density, depth to water table and unclassified vegetation types. Some of the above mentioned factors are stable in the long term (e.g. soil and topography), while others may change in relatively short time (e.g. land use, land management practices, presence of drainage systems). It is widely recognised, for instance, that steep slopes are associated with increased surface runoff and the Curve Number should be consistently adjusted (Huang et al., 2006; Ebrahimian et al., 2012).

The HOST classification also provides, for catchments in the UK, theoretical values of BFI associated with each HOST class, which were derived from multiple regression analyses Boorman et al. (1995). In this work, CN and BFI are calculated as spatially weighted averages of the distributed soil and vegetation information (see detailed procedure in appendix C) and from precipitation and flow records using the asymptotic method suggested by Hawkins (1993) and illustrated in appendix B. These signatures provide a low-dimensional objective framework able to assess similarities in the magnitude of observed extreme events as well as low flows characteristics.

7.2.2 Likelihood of model configurations

In this chapter, I use Bayesian statistics as a probabilistic approach to uncertainty analysis. In the context of a multi-model probabilistic framework for simulating streamflow predictions, Bayesian statistics assumes that the degree of belief in a single configuration[2] is non-negative

[2] A configuration is defined here as the combination of one model structure and one parameter set.

and the total belief in all possible configurations is fixed to be one. To reflect the lack of prior knowledge it is common to assume that, initially, probabilities are uniformly distributed. This prior distribution can then be updated on the basis of evidence (comparing regionalised signatures to the simulated ones) to obtain posterior beliefs, which may be used as the basis for inferential decisions.

In mathematical terms, the posterior probability of a certain configuration c is:

$$p(c|D) = \frac{p(c)p(D|c)}{p(D)} \tag{7.1}$$

where c is the configuration (combination of one model structure and one parameter set), D is the data (or signature), $p(c)$ is a prior configuration distribution, $p(D|c)$ is the likelihood that a certain configuration is suitable to simulate the target signature and $p(D)$ is a normalising coefficient. Assuming that the residuals between observed and simulated signatures (ε_s) are normally distributed, their probability density function is as follows:

$$f(\varepsilon_s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left[-\frac{1}{2}\left(\frac{\varepsilon_s - \mu}{\sigma_s}\right)^2\right] \tag{7.2}$$

In equation 7.2, s can be one of the two signatures CN and BFI, while μ and σ_s are the mean and standard deviation of the residuals, respectively. For unbiased residuals the mean is equal to zero and equation 7.2 becomes equivalent to:

$$f(\varepsilon_s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{\varepsilon_s^2}{2\sigma_s^2}\right) \tag{7.3}$$

Assuming also that the residuals are independent, the likelihood is calculated as the product of their probabilities:

$$P(D|c) = \prod_{s=1}^{2} \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{\varepsilon_s^2}{2\sigma_s^2}\right) \tag{7.4}$$

Bulygina et al. (2011) suggest to assume the standard deviation of regionalisation residuals of

CN equal to 3, based on the inter-separation of CN values across land use and management classes. The standard deviation of regionalisation residuals of BFI, instead, is tabulated for each HOST soil class and can be calculated as follows:

$$\sigma_{BFI_{HOST}} = \sqrt{\sum_{i=1}^{n} a_i^2 \cdot \sigma_i^2} \tag{7.5}$$

where a_i is the fraction of catchment area that falls in the HOST/USDA class i and σ_i is the related standard deviation reported in table S1 of Bulygina et al. (2011).

The simulated ensembles will be compared in terms of Nash-Sutcliffe analog for probabilistic predictions (Bulygina et al., 2009):

$$NS = \left\{ 1 - \frac{\sum_{t=1}^{T} (E[s_t] - o_t^0)^2}{\sum_{t=1}^{T} (o_t^0)^2} \right\} - \frac{\sum_{t=1}^{T} Var[s_t]}{\sum_{t=1}^{T} (o_t^0 - o^0)^2} \tag{7.6}$$

where o_t^0 is the observed discharge at time t, o^0 is the average of o_t^0, s_t is the simulated discharge at time t, while $Var[]$ and $E[]$ denote the variance and the expectation, respectively.

7.3 Study area and land use scenarios

The proposed methodology is tested on the Severn at Plynlimon flume and the Wye at Gwy flume catchments (United Kingdom). Catchments, flow and input data are described in Chapter 4. Although more than 30 years of hourly records are available, predictions were made only for the period 1979-1981 (using the first year as warmup) to limit computation time and data storage. This particular period was selected because it does not contain missing records. The information related to soil, vegetation, topography and drainage network was retrieved from the CEH Information Gateway portal and summarised in Table 7.1.

Table 7.1: Data sources.

Type	Source	Notes
Soil	Hydrology of Soil Types	The area is characterised by low permeability bedrock.
Vegetation	Plynlimon vegetation (2013) which classes correspond to those defined in CEH Land Cover Map 2000	Wye at Cefn Brwyn is mainly covered with grassland, while the Severn at Plynlimon flume is mostly afforested.
Topography	The hydrologically corrected digital terrain model (DTM) of Plynlimon catchments published in June 2010	Slopes are generally below 11%.
River network	The digital river network of the natural and artificial streams within the Plynlimon catchments was published in June 2010	In the Wye at Cefn Brwyn area the natural drainage system is combined to a dense network of artificial ditches. The ratio between artificial and natural stream length is in the range 28-44%. Although there is no record of artificial drainage in the river network map for the Severn at Plynlimon flume area, Kirby et al. (1991) mentioned that part of the Severn at Plynlimon flume was hand-dug for turf planting and sparsely drained. Therefore the entire Plynlimon area is considered artificially drained, with a ratio below 50%.

The entire Plynlimon area was pasture for sheep grazing until 1930s, when a significant decrease in land prices led to government purchase of the Severn at Plynlimon flume area (Kirby et al., 1991). Various interventions of conifer planting started in late 1930s and continued until mid 1960s. In the summer 1985, the Forestry Commission initiated a major tree-felling in the Severn side of Plynlimon. In the first instance, the programme lasted four years and slopes were replanted within two years of felling. Few other deforestation and afforestation cycles followed. Each time, only a relatively small portion of the catchment would be affected by the change, making difficult to detect the effect of changes on the streamflow signal.

In order to provide few simple examples of how to analyse the impact of land use changes, in this work two scenarios are taken into account: 1) a forested catchment is converted to pasture and 2) viceversa, an area used for pasture is planted with forest. The first scenario is assessed on the Severn at Plynlimon flume catchment, while the second on the Wye at Gwy flume catchment. The proposed methodology, however, can be applied to any change in land use and hydrological

condition.

CN and BFI have been calculated with the following methods: a) using observed precipitation and streamflow discharge time series (observed signatures, CN_O and BFI_O) and b) using soil and vegetation maps (regionalised signatures, CN_R and BFI_R). In the former case, the observed BFI is calculated using the baseflow separation filter suggested by Gustard et al. (1992) while the observed CN is obtained using the asymptotic method suggested by Hawkins (1993) and described in detail in Appendix B. Results for all the Plynlimon subcatchments are listed in Tables B.1 and B.2, in particular, BFI_O is 0.34 for Severn at Plynlimon flume and 0.33 for Wye at Gwy flume, while CN_O is 81 for Severn at Plynlimon flume and 88 for Wye at Gwy flume.

The method to calculate regionalised signatures, instead, is described in Appendix C. The workflow schematised in Figure C.1 shows that it is possible to calculate CN and BFI given soil, vegetation, slope and hydrological condition of the soil (which may depend on the drainage level). Soil compaction problems due to overgrazing in the Plynlimon area have been reported by Thomas (1965). According to USDA (1986) guidelines, this was interpreted classifying the area as pasture in poor condition. Also, the dense network of ditches in the Plynlimon soils causes an excess of drainage and therefore an increase in the runoff potential, which was interpreted classifying afforested areas as woods in fair condition. Based on the above considerations, the regionalised values of CN and BFI were calculated for all the Plynlimon subcatchments and listed in Tables C.4 and C.3, in particular, BFI_R is 0.33 for Severn at Plynlimon flume and 0.34 for Wye at Gwy flume. Based on this method CN can be calculated for any land use scenario by creating fictitious vegetation maps. CN can increase/descrease depending on the hydrological condition of the soil which are tabulated values. For instance, CN_R the for both Severn at Plynlimon flume and Wye at Gwy flume is 77 in case of forest coverage and 88 for pasture.

In this study, the likelihood calculated using observed and regionalised indices uses the same standard deviations: 3 for CN and about 0.02 for BFI. The detailed procedures to calculate regionalised and observed flow indices are reported in appendices C and B. Based on the re-

gionalised and observed indices, the two catchments look very similar. However, being characterised by different existing land use (forest for the Severn at Plynlimon flume and pasture for the Wye at Gwy flume), they will be used for the assessment of prediction quality for the different land uses.

7.4 Modelling set up

The FUSE multi-model ensemble (Clark et al., 2008) was used to model the transformation of rainfall into runoff. A total of 1000 parameter sets were sampled uniformly, using a Latin Hypercube, from the default ranges suggested by Clark et al. (2011). Excluding the rainfall error from the inference and allowing the routing scheme, the maximum number of model structures available within the framework is 312. On one hand, this choice allows a wide range of possible responses but at a considerable computational cost[3]. More computationally efficient alternatives are also considered. Bulygina et al. (2011), for instance, uses expert knowledge to identify the PDM as suitable model structure. Unfortunately, the PDM model used by Bulygina et al. (2011) is not embedded in the FUSE framework and direct comparison with previous results could not be drawn. Beside expert knowledge, for gauged catchments the model configuration space can also be constrained based on evidence from data. In this cases, the prior configuration could be that suggested by the AMCA algorithm. In the previous chapters the ideal model configurations for the Severn at Plynlimon flume catchment were indentified for the period 1975-1985 and summarised in Table 6.1. The AMCA configurations are, therefore, retained for comparison to the default FUSE configurations.

The prior distribution of simulated flows is approximated by two different discrete distributions:

1. FUSE, in this case values are defined by 312000 model configurations (312 models are

[3]Running the suggested workflow for a single model structures takes about 18 hours. The slow computation is due to a non-optimised code to calculate the Curve Number using the event-based methodology suggested by Hawkins (1993).

combined with 1000 parameter sets sampled, from default ranges, using the Latin Hypercube Sampling method),

2. AMCA, in this case values are defined by 2000 model configurations (the two models suggested by AMCA are combined with 1000 parameter sets sampled, from AMCA-narrowed parameter ranges, using the Latin Hypercube Sampling method, for details on how these configurations were generated see Chapters 5 and 6).

For each model configuration (c), two indices are calculated: CN_c and BFI_c. The residuals obtained as difference between the regionalised (CN_R and BFI_R) and empirical signatures (CN_c and BFI_c) are then used to calculate the likelihood of each configuration, based on equation 7.4. The prior distributions have all equal weights, while the posterior weights are calculated by normalising the related likelihood (probability).

In order to characterize the frequency of modelling options for each probability-bin, a new measure called Persistence Rate (PR) is defined as the ratio between the percentage frequency of one option in the posterior distribution and the percentage frequency of the same option in the prior. If the persistence is 0, it means that the option was not selected in the posterior. If the persistence is 1, the option is as frequent in the posterior as it is in the prior. If the persistence is greater than 1 the option is more frequent in the posterior than in the prior.

The reader is reminded that FUSE modelling options are not uniformly distributed (see Table 3.3). The single reservoir in the upper soil architecture, for instance, is used by over 46% of the configurations and it is expected that, being more frequent, it is also more likely to be selected. In order to take into account the non-uniformity in the prior configurations, an additional measure is introduced. This is called Normalised Frequency (NF) and it is defined as the probability of each modelling option to belong to one of the probability bins when all options are sampled uniformly a priori. According to probability rules, this can be calculated as follows:

$$p_u(o|D) = \frac{p_u(o)p_{nu}(o|D)}{p_{nu}(o)} \tag{7.7}$$

where o is a model option, D is data, u stands for uniform, nu stands for non-uniform, $p_u(o)$ is the prior uniform probability of modelling options, $p_{nu}(o|D)$ is the posterior non-uniform probability of modelling options, $p_{nu}(o)$ is the prior non-uniform probability of modelling options. NF is the above probability after normalisation, which allows to characterize the configuration frequency over each bin, after accounting for configuration non-uniformity in the initial population of models (when some configurations appear more often than others). The performances of the various scenarios analysed are summarised in terms of Nash-Sutcliffe analog for probabilistic predictions (Bulygina et al., 2009), while the accuracy and precision of the ensembles are assessed as described in Chapter 5.

7.5 Results and discussion

The aim of this section is to predict the effects of land use changes, considering the uncertainty related to the soil's hydrological conditions and model structure variability.

7.5.1 Comparing predictions from observed and regionalised indices

The first experiment consists of comparing flow predictions generated using observed[4] and regionalised[5] indices, to quantify the uncertainty due to the regionalisation process. For the Severn at Plynlimon flume catchment (covered with forest in fair hydrological condition) and the FUSE prior configurations, results show that the prior 95% confidence interval is characterised by 90% accuracy. Low and medium events are accurately encompassed but major high flows are underestimated. An example of inaccurately simulated event is shown in Figure 7.1. The event occurred between the 6th and 8th October 1980. The grey-blue shaded area is the prior's distribution percentiles over time (95% probability mass).

[4] CN_O and BFI_O are calculated from precipitation and streamflow time series using the asymptotic method suggested by Hawkins (1993), see appendix B.

[5] CN_R and BFI_R are calculated from soil and vegetation information, see appendix C.

Constraining the predictions using flow indices improves significantly the accuracy and reduces uncertainty. For instance, the posterior from observed indices (green dotted lines) encompasses 98% of the observations. The regionalised posterior (red dotted lines) is also highly accurate (96%) and brackets more closely the observations, with an increase in precision of 22%. The NS efficiencies are also similar, 0.54 for the posterior constrained by observed indices and 0.51 for the posterior constrained by regionalised indices. Based on this comparison, the regionalisation process is not expected to add a significant layer of uncertainty to the predictions.

Although regionalised indices improve the predictions, CN and BFI seem to have a low constraining power as the confidence intervals still remain wide. This is also evident for smaller events. Figure 7.2 shows predictions for July 1981 for which low flows appear over-estimated and slightly shifted in time.

Table 7.2: Nash-Sutcliffe efficiency (NS) analog for probabilistic predictions for the Severn at Plynlimon flume in the period 1979-1981.

Predictions (FUSE configurations)	**NS**
Prior	0.19
Posterior constrained by observed indices	0.54
Posterior constrained by regionalised indices	0.51

7.5.2 Modelling options

The indices' constraining power was assessed in relation to the distribution of modelling options for different probability-bins. For about 1% of the simulations it was not possible to calculate the likelihood due to the lack of asymptotic behaviour in the relationship CN-P (non-standard behaviour). The posterior probabilities of the remaining configurations were divided into probability bins. 99% of the probability mass is concentrated in zero while the remaining 1% is below 0.004.

Nonetheless, the regionalised CN and BFI allow to clearly discern amongst modelling options. Figure 7.3 shows the persistence of each option on the y-axis, grouped based on the related

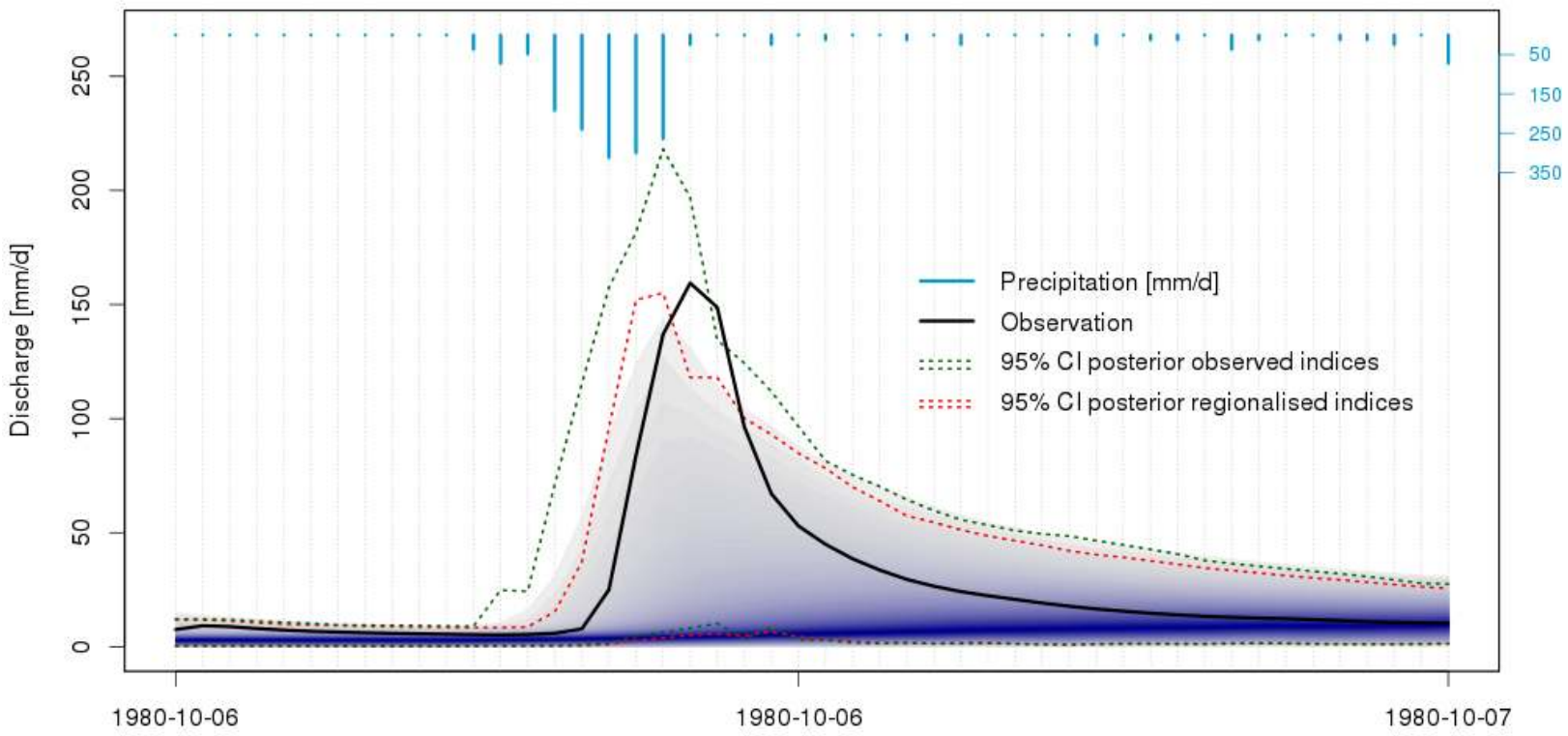

Figure 7.1: Prediction uncertainty bounds for the event occurred in the Severn at Plynlimon flume catchment between the 6th and the 8th October 1980. The precipitation is shown as blue bars on top, values are in mm/day on the right hand axis. The grey-blue shaded area shows the prior's distribution percentiles over time (95% probability mass). The green and red dotted lines show the posterior's 95% confidence intervals obtained from observed and regionalised indices, respectively. All the bounds are generated using 312 FUSE model structures.

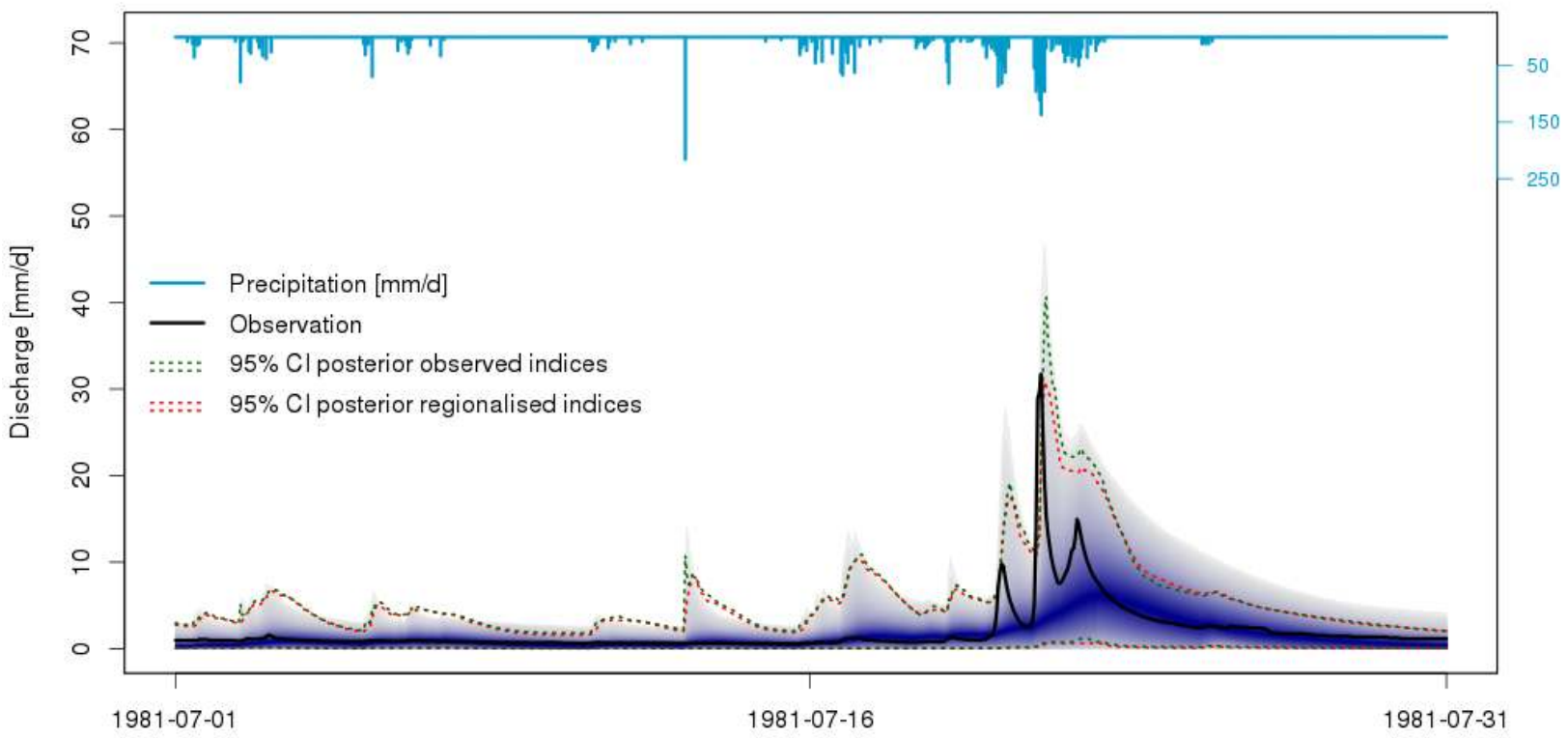

Figure 7.2: Prediction uncertainty bounds for July 1981. The precipitation is shown as blue bars on top, values are in mm/day on the right hand axis. The grey-blue shaded area shows the prior's distribution percentiles over time (95% probability mass). The green and red dotted lines show the posterior's 95% confidence intervals obtained from observed and regionalised indices, respectively. All the bounds are generated using 312 FUSE model structures.

model building decision, and 6 ranges of probability values (bins) on the x-axis. Discarded model components appear uniformly spread, as the persistence at zero-probability is always equal to one. Options with highest probabilities, instead, suggest the following configuration: a single reservoir for the upper soil layer, a baseflow reservoir of unlimited size and fraction recession law, a runoff mechanism based on ARNOVIC parameterisation, a percolation scheme controlled by the satured zone, evaporation scheme dependent on the fraction of roots and no interflow. The upper soil architecture, percolation and interlow schemes coincide with those identified by the AMCA algorithm in Chapter 6.

Figure 7.4 shows the normalised frequency, which is the option frequency normalised over a probability-bin, after accounting for configuration non-uniformity. From this plot, the distribution of options for the upper soil layer, evaporation and interflow schemes do not look significantly different from the initial population of models, while a clear diversion from the original distribution is observed for the lower soil architecture, percolation and runoff schemes.

7.5.3 Predictions based on different land use conditions

The USDA classification mentions that good hydrological conditions correspond to low runoff potential but do not attempt to set quantitative ranges to discern between good, fair and poor conditions. Therefore the assumptions on the hydrological conditions made earlier, on the basis of the drainage system and current grazing practice, can be considered subjective. In this section, the predictions for the Severn at Plynlimon flume (forest) and Wye at Gwy flume (pasture) are compared in case of good, fair and poor conditions to quantify the impact of this decision as well as assess the prediction quality based on the existing land uses.

Forest in fair condition is considered the reference scenario for the Severn at Plynlimon flume catchment. Results are shown with regard to the largest event in the period taken into account, occurred in the October 1980. As already mentioned before, CN varies depending on the hydrological condition of the soil and the mapping is available as tabulated values (see an example

of lookup table in Appendix C, Table C.1), therefore changing soil condition can be simulated by simply reading the tabulated values from a different row in the lookup table. Switching from fair to poor conditions, for instance, leads to a 28% weighted mean increase in the peak flow, while a 12% weighted mean decrease is expected switching to good hydrological conditions (see Figure 7.5). Therefore, high flows are sensitive to the hydrological condition selected.

Looking at the same event in October 1980 for the Wye at Gwy flume catchment, results show that the prediction quality changes for different land uses. Assuming pasture in fair condition as the reference scenario, 52% weighted mean increase in the peak flow is expected switching to poor conditions, while 29% weighted mean decrease is expected switching to good hydrological conditions (see Figure 7.6). Therefore, high flows sensitivity depends on both land use and hydrological condition selected.

7.5.4 Forest to pasture scenario

Changes in land use have a direct effect on CN. For the Plynlimon soils, in various conditions, 100% forest coverage leads to a Curve Number in the range [75,81] while for pasture the Curve Number is expected in the range [78,88].

Figure 7.7 compares predictions, at the Severn at Plynlimon flume catchment for the October 1980 event, for the hypothesised current land use condition (forest in fair conditions, red area) and for the scenario in which forest becomes pasture in poor condition (green area). The regionalised CN for fair woods is 77 while for poor pasture is 88. This scenario leads to 60% mean increase in the peak flow. The opposite case is shown at the Wye at Gwy flume catchment (Figure 7.8) where the current land cover is pasture that becomes forest. In this case, it is expected a 56% mean decrease in the peak flow.

The CN ranges for forest and pasture overlap and the difference in CN values increases as the soil hydrological conditions worsen. This means that the increase/decrease in flow peak becomes more subtle in case of both land uses are in good hydrological conditions. An interesting

case arises when the CN ranges overlap, as in the case in which forest is in poor condition and pasture in good condition. Figure 7.9 shows this case for the event in October 1980 for the Severn at Plynlimon flume. Predictions are very similar, but a 17% mean decrease in the peak flow is expected converting forest to pasture. This is the only case in which, the conversion of forest to pasture generates a reduced runoff.

7.5.5 Comparing predictions from different prior configurations

In this section, predictions generated from different prior configurations are compared to quantify the uncertainty related to model structure variability. Figures 7.10 and 7.11 compare the regionalised posterior generated using FUSE configurations (red polygon) to the posterior generated by AMCA configurations (green polygon) for high and low events.

In order to keep the computation time for each catchment manageable on the available computer cluster, the number of parameter samples was limited to 1000, which led to poorly sampled parameter space. Without time constraints, or using a larger cluster, it could be possible to explore better the parameter space and therefore have a more representative prior. Keeping the same number of parameter samples and conditioning the procedure using the AMCA prior, the performance increases significantly (from 0.51 to 0.71). Simply conditioning the prior using regionalised indices and restricting the model space to the two model structures identified by the AMCA, the posterior's performance increased from 0.51 to 0.55 (20% of the total improvement). By also adding the restrictions imposed on the parameter space, the posterior's performance raises to 0.71. It is, therefore, clear that the improvement is due to constrains applied to both model and parameter space, but the parameter space seems to contribute the most (80% of the total improvement).

Analysing a various high and low flow events, however, the AMCA predictions in correspondence of the lowest flows were found to have a rather low accuracy (68%). This loss in performance could be due, again, to the BFI's low constraining power.

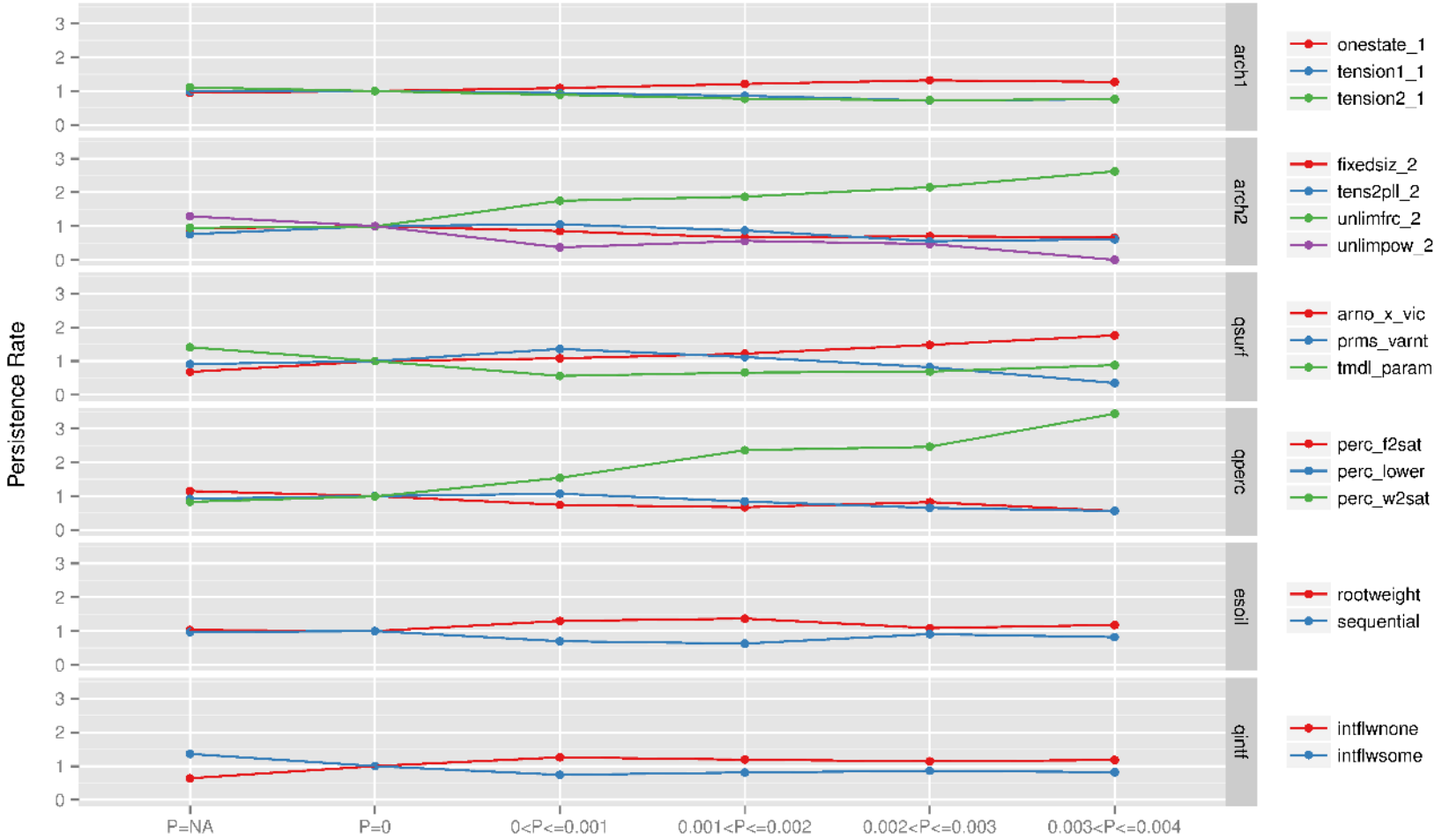

Figure 7.3: Persistence of modelling options (y-axis), based on 6 grouped probability ranges (P, on the x-axis). Discarded model components appear uniformly spread (persistence rate at zero is equal to one). The more the persistance rate increases the more significant the modelling option is. For instance, a baseflow reservoir of unlimited size and fraction recession law combined with a percolation scheme controlled by the satured zone become more and more significant as the probability increases.

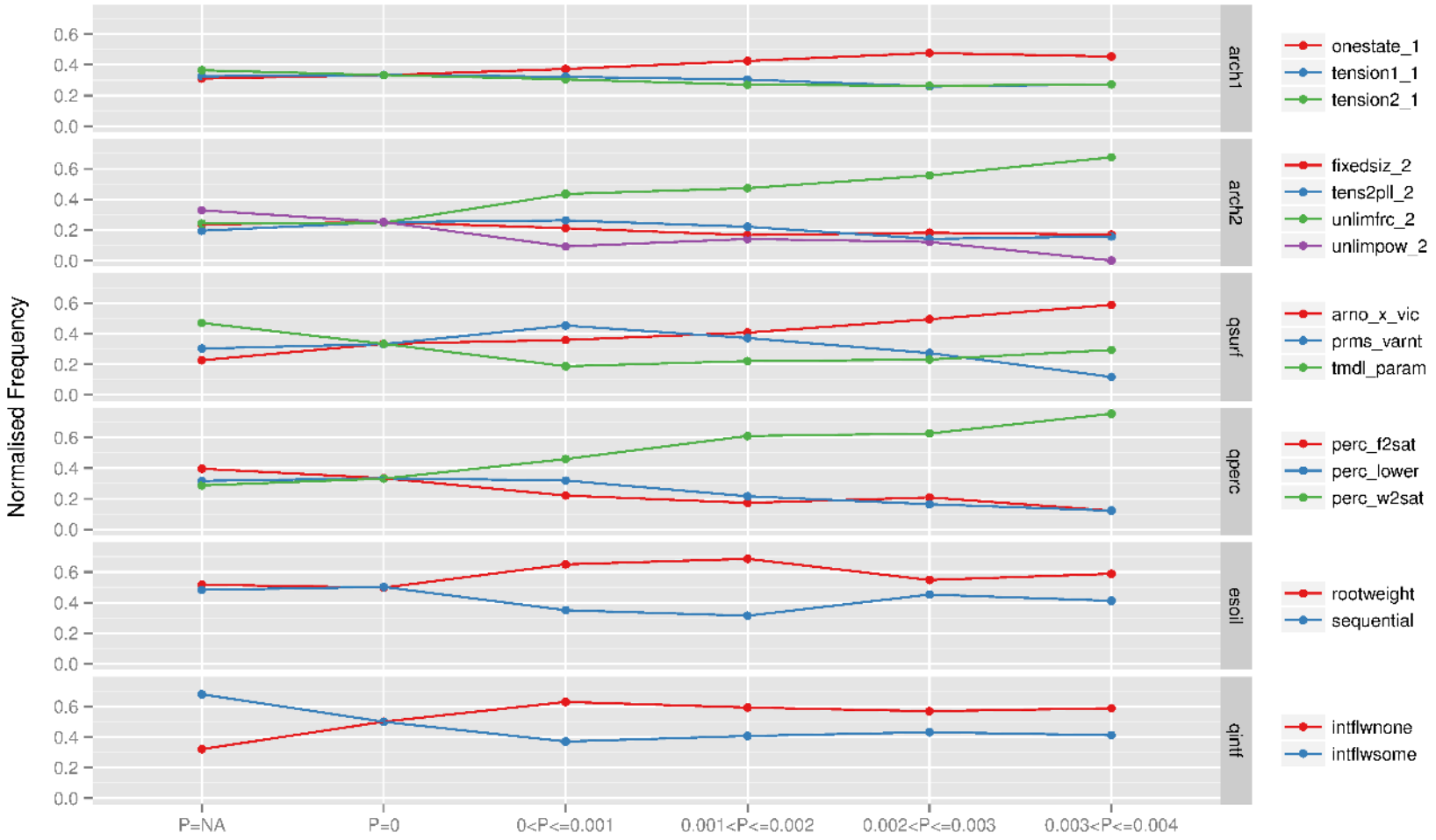

Figure 7.4: Normalised frequency of modelling options (y-axis), based on 6 grouped probability ranges (P, on the x-axis). The distribution of options for the upper soil layer, evaporation and interflow schemes do not look significantly different from the initial population of models, while a clear diversion from the original distribution is observed for the lower soil architecture, percolation and runoff schemes.

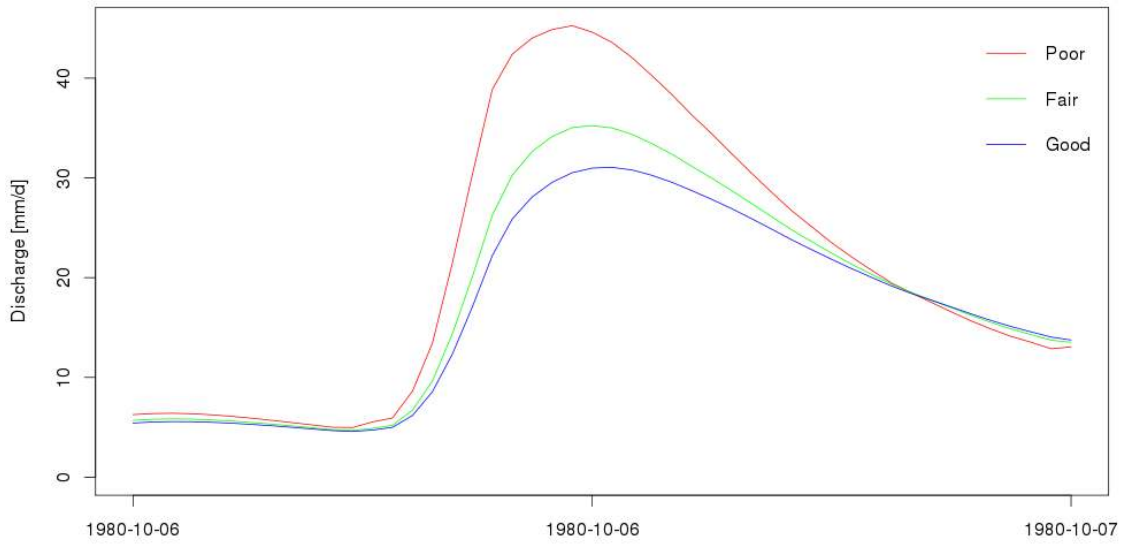

Figure 7.5: Weighted mean of flow predictions for Severn at Plynlimon flume covered with forest in good (red), fair (green), poor (blue) conditions.

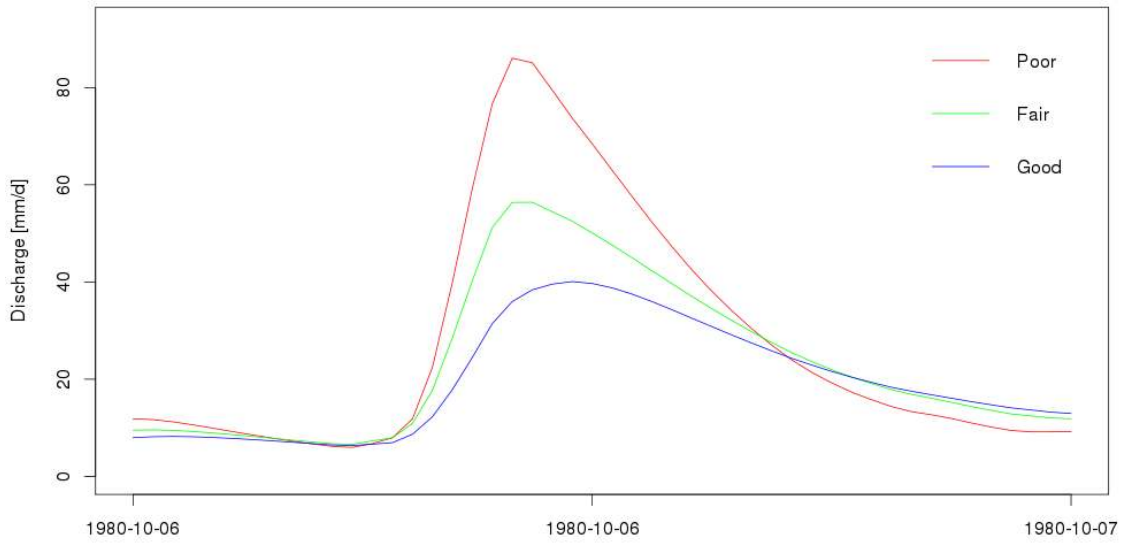

Figure 7.6: Weighted mean of flow predictions for Wye at Gwy flume covered with pasture in good (red), fair (green), poor (blue) conditions.

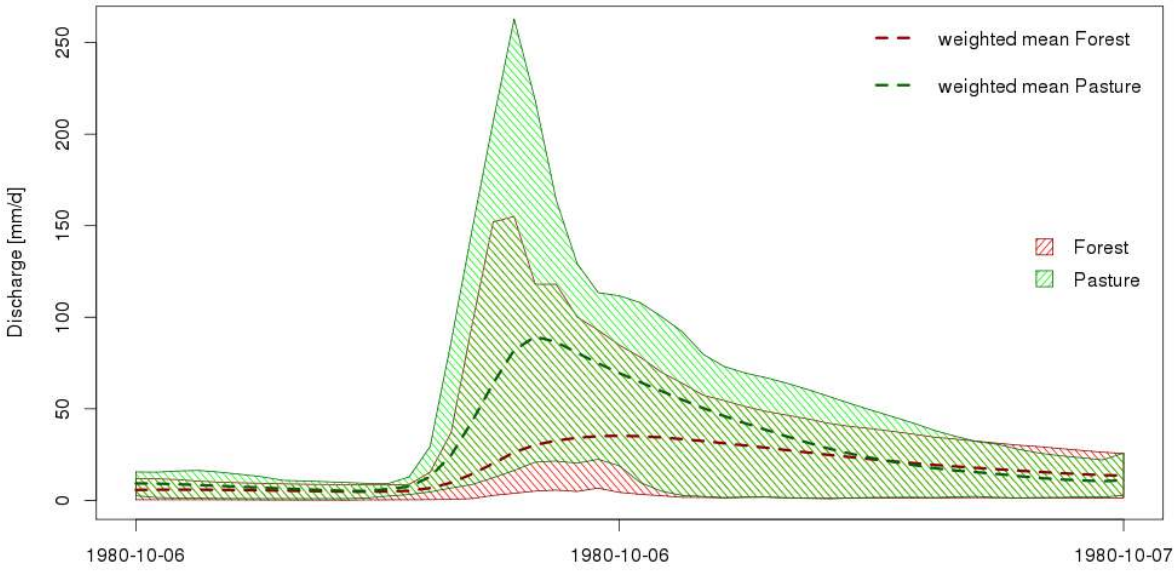

Figure 7.7: Posterior's 95% confidence intervals of streamflow predictions for the Severn at Plynlimon flume covered with forest in fair condition (red polygon) and pasture in poor condition (green polygon). Dashed lines show the weighted mean for forest (red) and pasture (green).

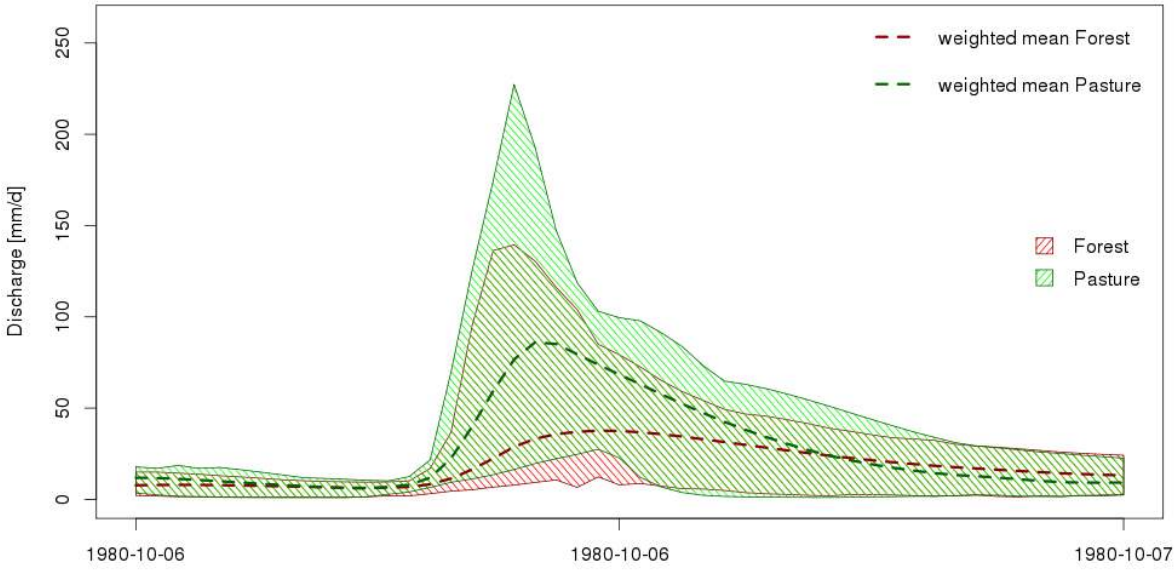

Figure 7.8: Posterior's 95% confidence intervals of streamflow predictions for the Wye at Gwy flume covered with pasture in poor condition (red polygon) and forest in fair condition (green polygon). Dashed lines show the weighted mean for a forest cover (red) and pasture (green).

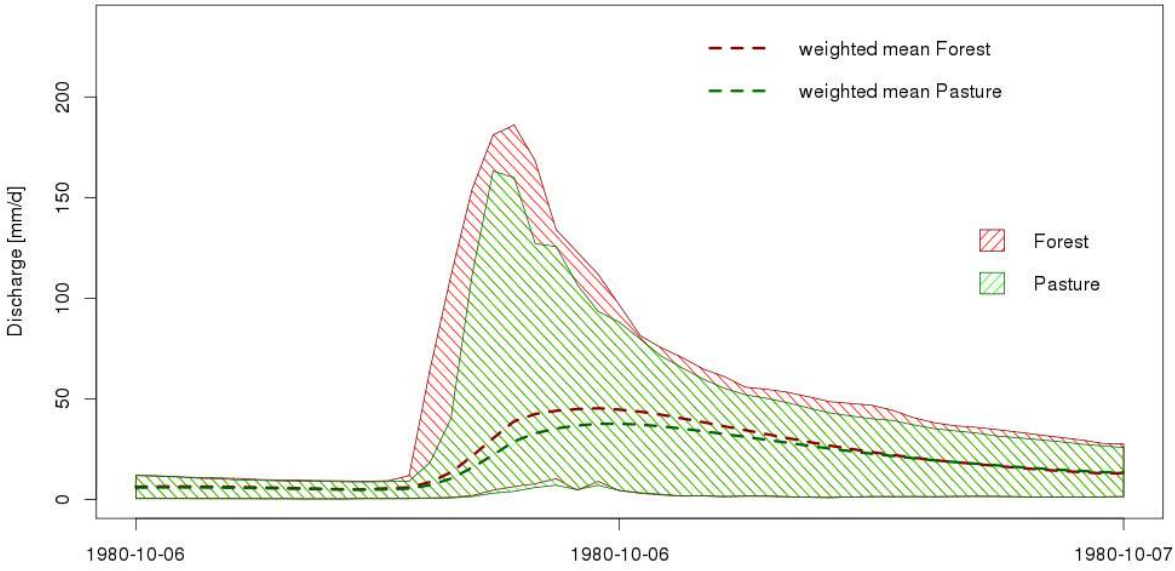

Figure 7.9: Posterior's 95% confidence intervals of streamflow predictions for the Severn at Plynlimon flume covered with forest in poor condition (red polygon) and pasture in good condition (green polygon). Dashed lines show the weighted mean for forest (red) and pasture (green).

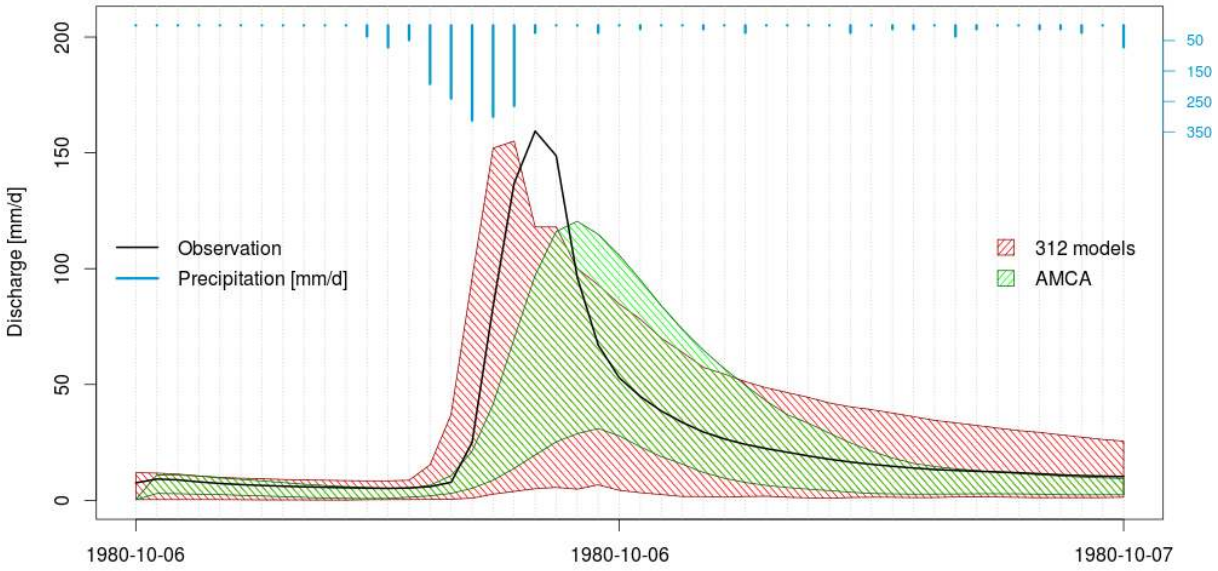

Figure 7.10: Posterior's 95% confidence intervals of streamflow predictions generated using 312 models (red polygon) and the AMCA configuration (green polygon) for event in October 1980 (high flows).

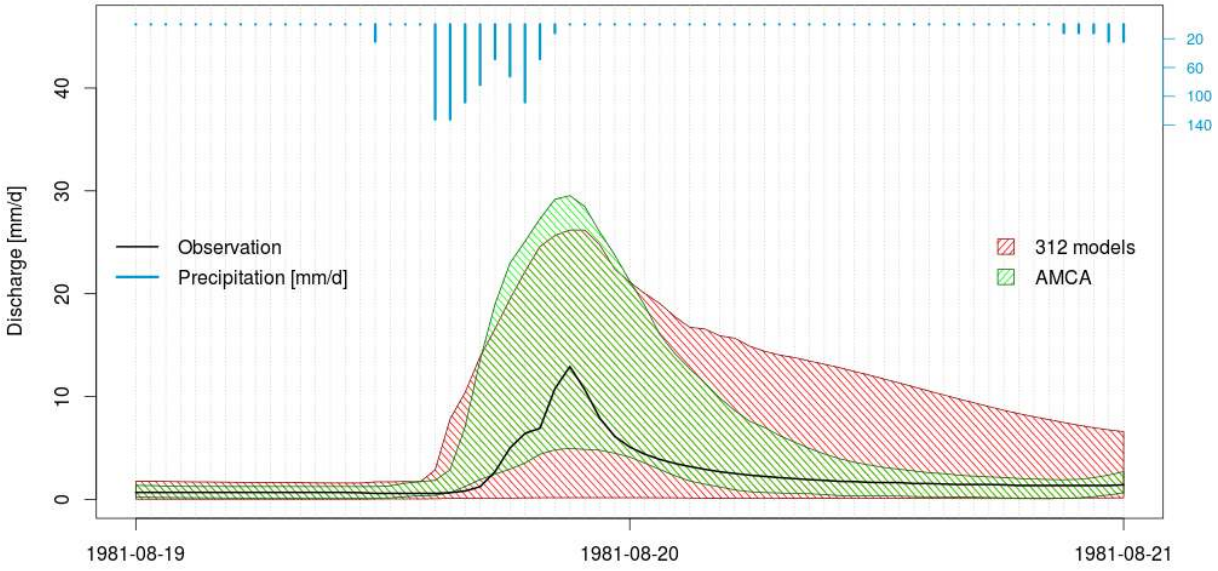

Figure 7.11: Posterior's 95% confidence intervals of streamflow predictions generated using 312 models (red polygon) and the AMCA configuration (green polygon) for event in August 1981 (low flows).

7.6 Concluding remarks

In this chapter a multi-model probabilistic framework is presented to allow predictions of land use changes impact. The proposed methodology builds upon a previous study by Bulygina et al. (2011) in which the authors show that the propensity of a catchment area to generate runoff and its baseflow contribution can be estimated in terms of two indices: the Curve Number and the Base Flow Index. Bulygina et al. (2011) suggested that in data-poor environments, these indices can be estimated from soil and vegetation data. In data rich environments, instead, the same information can be calculated from time series data and validated against the regionalised ones.

In this study, various steps forward were taken automating the existing procedure and extending it to: use topography, drainage, soil and vegetation maps to derive the regionalised indices, expand the model configuration options to make use of multiple model structures and to improve predictions by constraining the prior model configurations using the data mining workflow described in Chapters 5 and 6. The methodology was tested on the Plynlimon area in the United Kingdom, where subcatchments are characterised by similar soil and topography but different vegetation cover.

Results showed that prediction are highly accurate but bounds appear rather wide due to CN and BFI low constraining power. However these signatures clearly allow to discern amongst modelling options. Grouping the weights associated with each configuration showed that some of them are increasingly frequent as the probability increases. As the land use hydrological conditions worsen, predictions become more sensitive to changes, as the hydrological condition of the land use worsen. For instance, simulating the transformation of forest to pasture, both in good condition, led to 20% mean increase in the peak flow of a large event. The conversion of a forest in fair condition to pasture in poor condition, instead, led to 60% mean increase in the peak flow of the same event. This increase might be due to the greater water use of mature forests, compared to pasture lands (McCulloch and Robinson, 1993; Blackie and Robinson,

2007) and to the reduction in the infiltration capacities of the soil as result of soil compaction due to the use of vehicles and/or the presence of grazing animals. Bulygina et al. (2011) performed a similar analysis using the PDM model and assuming good hydrological conditions for both Severn and Wye. They found that for the same event in October 1980, the "median peak flow in the Severn increases by 9% when the afforested area becomes pasture" and added that, although "consistent with the difference in event-averaged unit hydrographs between the Wye and Severn catchments found by the Institute of Hydrology in their Plynlimon study (Figure 28, Kirby et al., 1991)", there is a significant uncertainty in this value.

Prediction performances improved significantly when conditioning the prior distributions using the AMCA approach. This suggests that a considerable part of the uncertainty in prediction is due to the definition of the prior and that more objective ways of constraining the prior using formal data-driven techniques are needed. AMCA is, however, a procedure that can only be applied to gauged catchment. Future experiments could test whether AMCA configurations could be regionalised or transferred to ungauged catchments on the basis of catchment characteristics.

Finally it was noticed that the improvement in prediction is due to constrains applied to both model and parameter space, but the parameter space seems to contribute the most.

Chapter 8

Conclusions

This thesis set out to explore the opportunities arising from the use of data mining techniques within the hydrology domain and focused on developing tools and methods to facilitate strategic problem solving for hydrologists, modellers and practitioners. This chapter provides a brief summary of the methodologies adopted, the main findings and scientific contributions. It iterates the main limitations of the approaches taken and points to possible future work.

8.1 Summary and contributions

The use of data mining, machine learning and cloud computing are no more exclusive to computer science. Many scientific disciplines have borrowed these technologies to improve the way domain specific datasets are handled, transferred over a network and manipulated. The hydrology domain is characterised by various assets and resources that are often accessible for educational/research purposes and there is a wider interest in information sharing, spanning from industry innovation to the governmental attempts to improve public involvement in decision and policy making. However, in the attempt to exploit data for driving decision making, scientists may end up "drowning in information, but thirsty for knowledge" (Königer and Janowitz, 1995). Having large volumes of data and countless models does not necessarily imply

better knowledge. Data can become "excessive and overwhelming, reaching a level of information overload", that hinders decision making rather than facilitating it (Bettis-Outland, 2012). From a modelling perspective, De Vos et al. (2011) highlight that environmental models[1] "lack essential quality characteristics in terms of transparency and reproducibility". In hydrological studies, this happens when too many choices are given to modellers. What model to use? What parameters to set? What data resolution to request? Being spoiled for choice is supposed to be an added value but often becomes a liability when there is no guidance. Ironically, when there is too much or too little information, modellers resort in expert elicitation to make decisions. But expertise is subjective and non-reproducible, therefore efforts are needed to provide transparent and reproducible guidelines to practitioners.

This work proposes a novel approach to guide the very first stage of any hydrological analysis: model selection and parameterisation. To achieve the thesis objectives, a case study site was identified: the Plynlimon catchments in the United Kingdom (described in Chapter 4). This area was selected because it is highly instrumented, geologically and climatically homogeneous, but divided in two major catchments with different land cover. These characteristics allowed to explore model structure variability as well as the effect of land use changes on the ideal model configuration. In order to explore ways of optimising model configuration to best replicate the hydrological response in the selected study area, an inventory of conceptual rainfall-runoff models was considered: the FUSE ensemble model (Clark et al., 2008). This framework has minimal data requirements, provides several hundreds model structures and allows users to switch between one model and another as well as build ensembles, maintaining a consistent definition of model structure components and parameters. FUSE was re-implemented as R-package as part of this PhD work and made openly available. It has rapidly become a widely used tool for building reproducible workflows for hydrological modelling and used by a number of NERC-funded projects such as the Environmental Virtual Observatory pilot for the UK (EVOp uses FUSE as web service), the Probability, Uncertainty and Risk in the Environment (PURE uses FUSE as workflow component) and the Environmental Research Workbench (ERW uses FUSE

[1] Hydrological models fall in the category of environmental models.

as cloud-enabled modelling tool on High Performance Computing facilities).

While the use of multi-model frameworks provides flexibility to modellers, often the selection of a set of suitable models to simulate scenarios and make predictions is dictated by pragmatic criteria rather than evidence from data. Chapter 5 illustrates the design of the Automatic Model Configuration Algorithm (AMCA). AMCA is a novel algorithm that consists of machine learning techniques to mine model performances, filtering out unsuitable model configurations, clustering similar behaviours to reduce redundancies and selecting the best configurations using time series matching algorithms. When tested, the algorithm returned a model configuration highly consistent with the synthetic inputs. This demonstrates that data-driven methods can provide guidance in making model selection and configuration more transparent (first research question). Results from various experiments suggested that model configurations could be considered interchangeable and that variability over the model structure space can often compensate for a sparsely sampled parameter space. However, the algorithm failed to identify optimal ranges for the majority of parameters.

In order to improve parameter identifiability, in Chapter 6, the interactions between model components and parameters were mined using association rules. This type of unsupervised machine learning technique was applied to the AMCA results in a recursive manner to identify the significance of simultaneous occurrences of parameter values and model components over multiple layers of complexities. Experiments showed that evidence of interaction are detectable, at first, among model components. This is probably because the model structure space is more limited compared to the parameter space. Constraining the model structure space first, allows to remove a layer of complexity and new interactions (also among parameters) become identifiable (second research question). Approaches such as the AMCA and its coupling to association rule mining are a novelty in hydrology and, although further testing is needed, they hold potential to become a valuable screening tool for hydrologists in a number of contexts.

Chapter 7 then demonstrates with a practical example the value of the newly designed data mining approaches applying them to the uncertainty analysis of simulated streamflow under land

use changes. In this modelling exercise, regionalised information related to soil and vegetation cover was used to assess the propensity of a catchment to generate runoff and its baseflow contribution through the use of two flow indices (Curve Number and Base Flow Index) that can be independently calculated for gauged as well as ungauged catchments. Experimental results showed good consistency between theoretical and empirical values which confirms that these indices can be used in regionalisation studies with good level of confidence. For all the catchments investigated, the predictions improved significantly initialising the prior using the AMCA approach. It was also proved that in gauged catchments, changes in land use can be interpreted in terms of shift in the model configuration and that ungauged catchments can also benefit from the AMCA approach as prior flow distributions can be constrained using donor-configurations.

8.2 Limitations and discussion

The modelling experiments, from Chapter 5 to 7, are limited to the use of a specific category of models, known as conceptual rainfall-runoff models. Within this category, the modelling framework adopted in this study is FUSE (Clark et al., 2008). It can be argued that the model selection problem is simply shifting from the subjective identification of a model structure to the equally subjective identification of a modelling framework. This is only relatively true, as in theory results from different frameworks should converge if the same processes are modelled. However at the current stage, hydrological models are far from being exhaustive and the adoption of one framework rather than another can still result in dissimilar predictions, as they represent only a limited number of processes and use specific forcing inputs (rather than all the available information).

The AMCA algorithm was tested using a synthetic dataset. The predictions reproduced the observed discharge with a conservative uncertainty estimation that bracketed the observations, even though neither model structure nor the parameter set used for generating the data were contained in the configuration space. The same procedure, coupled with association rule min-

ing techniques was also tested on few experimental catchments in the Plynlimon area in UK but more investigations are needed in order to find out whether a wide range of real hydrological responses can be accurately simulated. Problems may occur, for instance, in case this methodology is applied to catchment characterised by processes which are not described in the FUSE framework. This is, however, a limitation of the FUSE framework, not the proposed data mining procedure. As, previously mentioned, AMCA can work with any modelling framework, provided that parameters and model components are defined consistently across all the models contained in the framework. Distributed modelling systems could also be used with AMCA, as far as the processing time allows to generate, in a reasonable amount of time, an *Initial Ensemble* that is representative of the range of possible responses. Sensitivity tests in Chapter 5 have shown that the AMCA converges to stable performance values with as little as 2500 simulations for each model structure. However this can change if a different modelling framework is used.

The AMCA, on its own, can efficiently filter model structures but fails to identify optimal parameter ranges. This is compensated by coupling the algorithm with association rule mining techniques. The coupling, however, is a rather computational intensive task that demands the use of a cluster of computers to parallelise the workflow. This makes the methodology unsuitable for users that do not have access or budget to utilise multiple (virtual) machines. Same limitation applies to the workflow to perform uncertainty analysis under land use changes described in the last chapter. This workflow requires 10 computational days to perform the analysis on a single catchment. Part of the computational burden is due to a non-optimised computer code which could be optimized as part of future activities.

8.3 Further work

The AMCA data mining procedure was completely designed and developed as part of this PhD work. As such, the only applications of this novel methodology are those presented in this thesis. As the results are encouraging, there is certainly scope for testing it further and explore

its potentials.

There are many questions that could not be answered within the time frame of this research project and, if more time could be spent developing and testing the AMCA further, new research could move in many different directions:

- Is the AMCA able to automatically configure models for any catchment size and type of response? The AMCA was tested on a small area in the UK, characterised by high average precipitations, low permeability soils and fast hydrological responses. The encouraging results obtained in this study could be isolated to this particular setting. To prove whether this is the case, AMCA should be tested on a variety of catchments. A barrier could be to get hold of data, but public datasets could also be used, e.g. the MOPEX database which collects areal averaged precipitation, potential evapotraspiration and streamflow discharge time series (amongst other variables) for hundreds of catchments in the US (Duan et al., 2006).

- Is AMCA suitable for regionalisation studies? If the AMCA can be used in any location, it should be possible to examine how parameters and model components can be mapped and correlated to catchment characteristics so that model configurations can be extrapolated to data-poor areas. This type of experiment would add a number of additional variables to the problem such as the catchment characteristics but also another level of interactions due to spatial proximity. As such, it would make another rather interesting data mining problem that could be explored with spatially explicit techniques such as Bayesian Networks (Blangiardo et al., 2013), which are currently used in epidemiology to assess risk exposure to pollution.

- How is AMCA's performance affected by the variability of the routing scheme, the granularity of the observed data, the initial conditions and the size of SOMs to cluster redundant data? This could be explored by undertaking a more in-depth sensitivity analysis. The new analysis should be performed on a dataset much longer than the one used in Chapter

5.

- More in general, can algorithms be more useful than expert elicitation in the context of decision-making? Expert contribution is certainly highly valuable but extremely subjectivity, impossible to reproduce and assess in terms of uncertainty involved. Using algorithms, instead, results are always reproducible, consistent and it is possible to track the propagation of uncertainties through the entire data manipulation process. It would be interesting to set up a comparative study to quantify potential trade-offs.

Appendix A

Machine specifications

description:	Notebook
product:	U36SD ()
vendor:	ASUSTeK Computer Inc.
version:	1.0
width:	64 bits
CPU:	Intel(R) Core(TM) i5-2410M CPU @ 2.30GHz
configuration:	cores=2 enabledcores=1 threads=2
memory	System board or motherboard size=8GiB

Appendix B

Calculate CN and BFI from observed time series data

In case of gauged catchments, CN and BFI can be calculated from observed precipitation and streamflow discharge time series data, during steady state periods. The empirical BFI is calculate using the filter for baseflow separation described in Gustard et al. (1992). This filter assumes that observed precipitation and streamflow are recorded at least with daily frequency. Data is first aggregated from the original frequency to daily. The series is divided in blocks of five-day non-overlapping consecutive periods and the minima is calculated for each block. Turning points in the sequence of minima are first identified, then connected and linearly interpolated to daily time steps to return the base flow hydrograph. The BFI is the ratio between the volume below the baseflow hydrograph and the volume of the total recorded flow.

The observed CN, instead, is calculated using the asymptotic method suggested by Hawkins (1993). This method was slightly modified to allow automatic identification of the precipitation and flow events and illustrated in figure B.1. The plot shows on the x-axis the sequence of time steps during the event, on the primary y-axis (left) the discharge measured in mm/h and on the secondary y-axis (right) the precipitation, also measured in mm/h. Precipitation events are identified based on the assumption that an event starts when a non-null value is recorded

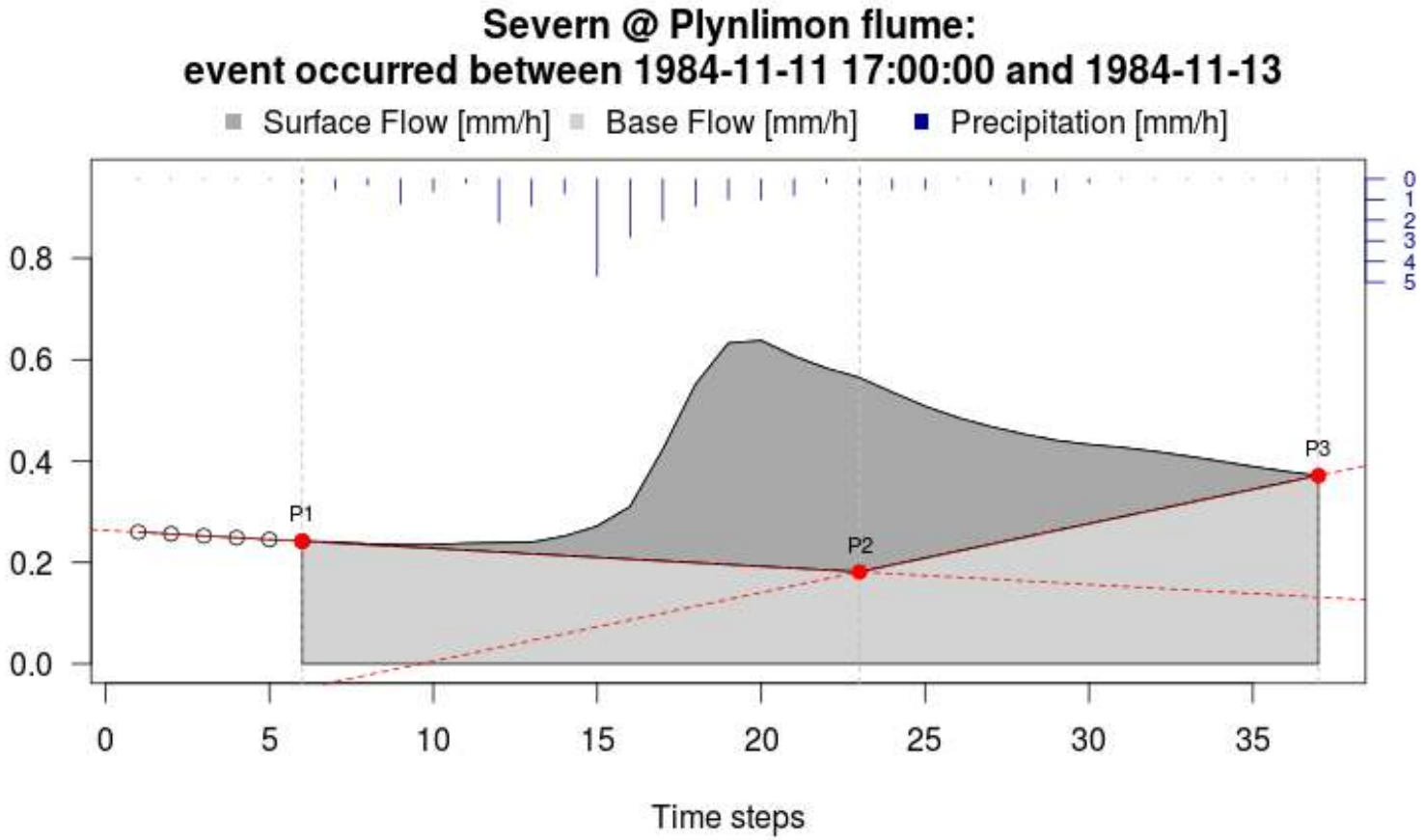

Figure B.1: Example event for which the slope of line P1-P2 is negative. The end of the flow event is set to 6 hours after the end of the precipitation event.

after at least 12 consecutive time steps of no rain (point P1). The event ends when no rain is recorded for two or more consecutive time steps. The time index that identifies the start of the precipitation event is taken as the start of the flow event as well.

At this point, Boorman et al. (1995) suggest to set the end of the flow event to the end of the rainfall event plus four times the lag (time difference between the rainfall and the runoff centroids). The procedure is recursive, as the flow-centroid (and therefore the lag) cannot be calculated without knowing the end of the flow event. Probably Boorman et al. (1995) identified the discharge centroid by visual inspection, or limited his analysis to single-peak events for which the lag can be calculated directly as the time between the rainfall and flow peaks. This approach, however, cannot be applied in case of multi-peak events. The procedure suggested in this work, instead, is entirely automated and the runoff centroid was calculated by simply extending the flow event for a certain number of time steps after the end of the rainfall depending on the average response time of the catchment under study (point P3).

Five time steps before the start of the rainfall event are used to linearly model the recession limb prior to the event. The linear model uses a fixed intercept (P1) while the slope is obtained by minimizing the sum of the squares of the distances between the observed points and the line. This line is then extrapolated through the event, up to the time of peak (point P2). A second linear interpolation is then carried out connecting points P2 and P3. The baseflow (light grey area) is the minimum between the two linear segments described above and the recorded flow, while the difference between the recorded flow and the baseflow is the surface runoff (dark grey area).

Figure B.2 shows an event in which the interpolated line through P1 has a positive slope. In this case, it is suggested to correct the slope to be zero. If the slope is not corrected, the baseflow is overestimated with the consequence that both surface runoff and CN become underestimated.

For each event the rainfall volumes (P) and surface runoff volumes (Q) are calculated. The events are sorted in descending order (independently) and the matching return periods calcu-

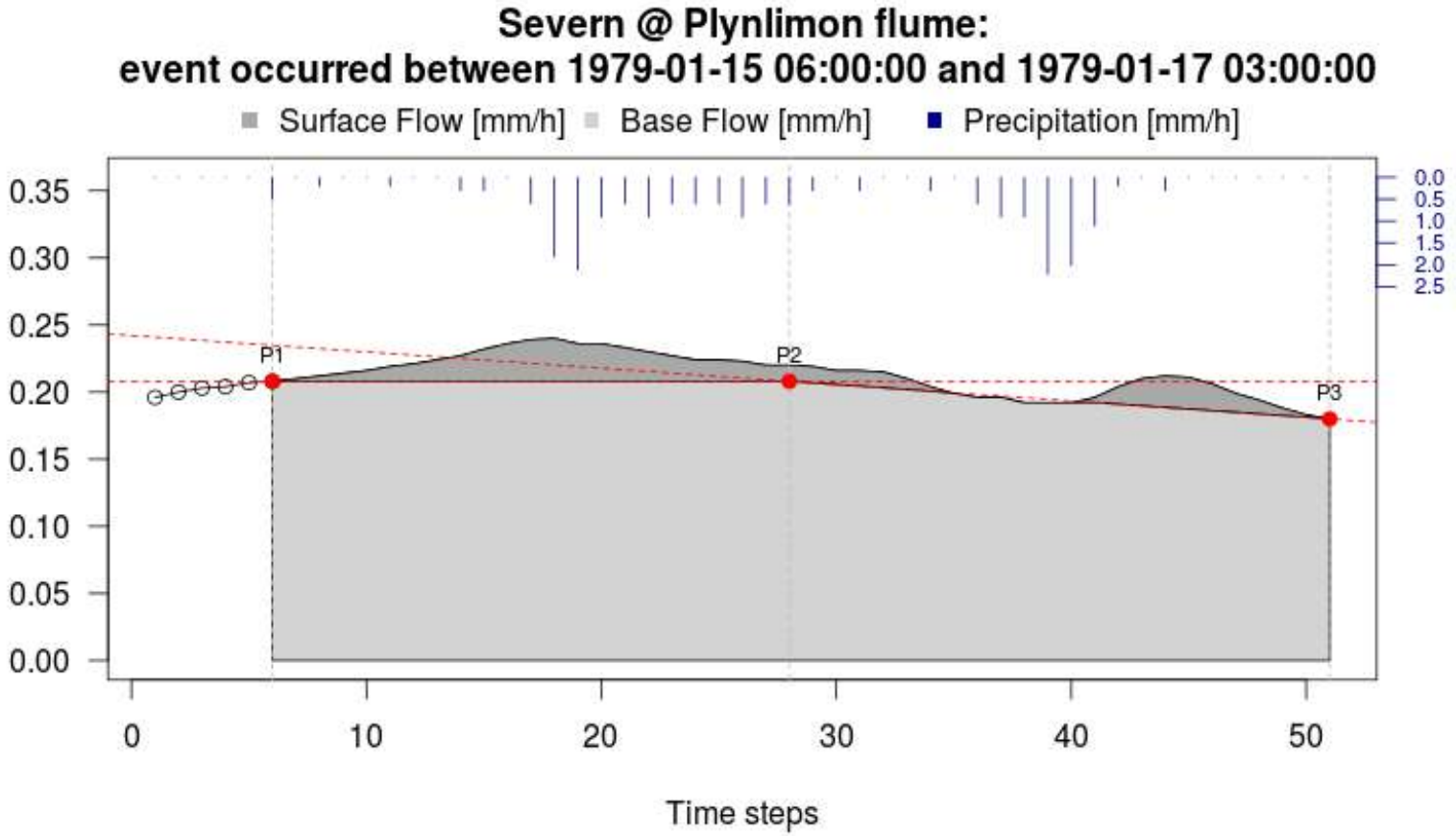

Figure B.2: Example event for which the slope of line P1-P2 was originally positive and the corrected to be zero.

lated. For each return period the CN is calculated as follows:

$$CN = \frac{25400}{254 + S} \tag{B.1}$$

where S is the potential maximum retention in mm and calculated using the following formula:

$$S = 5 \cdot (P + 2 \cdot Q - \sqrt{4 \cdot Q^2 + 5 \cdot P \cdot Q}) \tag{B.2}$$

Finally, the observed CN corresponding to the larger rainfall events is calculated fitting the CN-P points to a non-linear curve using the least squares method. Hawkins (1993) suggests to fit the CN-P relation using equation B.3, which was found to be appropriate for a wide array of catchments. In this equation CN_∞ is the CN that corresponds to larger storms and k_1 is a fitting constant.

$$CN(P) = CN_\infty + (100 - CN_\infty) \cdot \exp(-k_1 \cdot P) \tag{B.3}$$

The fitting returns a solution only in case the CN-P relationship follows what Hawkins (1993) defines a *standard response*, where the CN decreases asymptotically with increasingly larger storms, as in figure B.3. If no solution to the fitting is found, the CN-P relationship can follow other two possible responses. The first is called *complacent behaviour* and it is characterised by a progressive decrease of CN with every increase of P. The second is called *violent response*, it shows CN increasing asymptotically to a constant value with increasing P and was mainly observed in dry climates. Although it might be possible to adjust the fitting to different curves according to the type of response, the estimation of CN becomes particularly unreliable in case of complacent and violent behaviour. Therefore, in this work, only the case of standard behaviour is considered.

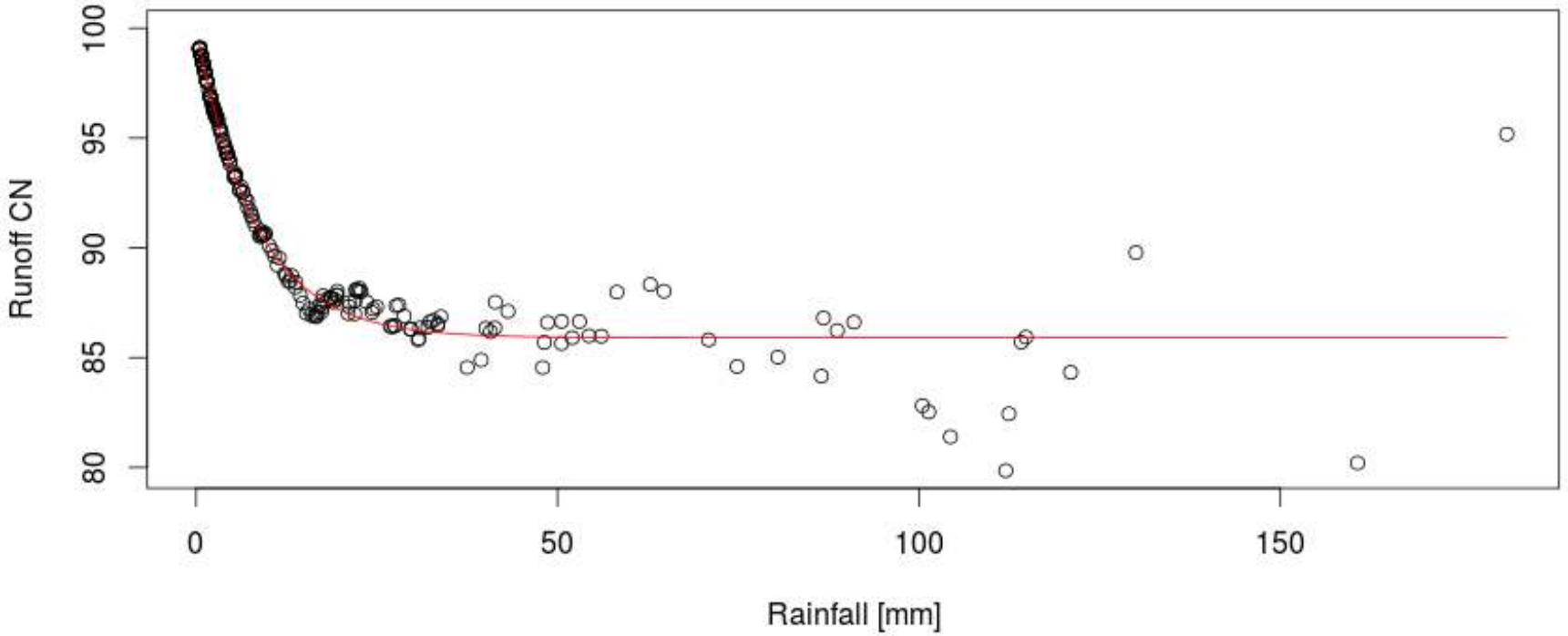

Figure B.3: Example of runoff CN-P relationship showing a standard behaviour. The fitted curve is shown as red line.

B.1 Empirical CN and BFI for Plynlimon

Tables B.1 and B.2 summarise the CN and BFI values for steady state periods in all the catchments in the Plynlimon area.

Table B.1: Empirical CN variability based on catchments and recording period. Felling period for the lower Hore: from May 1985 to Apr 1991. Felling period for the Tanllwyth: from February 1996 to Jan 1998 (but the period May 1994 - Feb 1996 was ignored because of the effect of the borehole drilling). No time series records where available for the Iago subcatchment.

Catchment	**Full record**	**Pre-fell**	1^{st} **felling** (May1985-Apr1991)	(May1991-Apr1994)	(May1994-Feb1996)	2^{nd} **felling** (Feb1996-Jan1998)	**Post-fell**
Severn	81	81 (-Apr1985)	83	81	79	81	81 (Feb1998-)
Tanllwyth	86	87 (-Apr1994)	-	-	84	86	86 (Feb1998-)
Hafren	79	-	-	-	-	-	-
Lower Hore	83	83 (-Apr1985)	85	-	-	-	83 (May1991-)
Upper Hore	83	-	-	-	-	-	-
Wye	86	-	-	-	-	-	-
Gwy	87	-	-	-	-	-	-
Cyff	87	-	-	-	-	-	-
Iago	-	-	-	-	-	-	-

Table B.2: Empirical BFI variability based on catchments and recording period. Felling period for the lower Hore: from May 1985 to Apr 1991. Felling period for the Tanllwyth: from February 1996 to Jan 1998 (but the period May 1994 - Feb 1996 was ignored because of the effect of the borehole drilling). No time series records where available for the Iago subcatchment.

Catchment	**Full record**	**Pre-fell**	1^{st} **felling** (May1985-Apr1991)	(May1991-Apr1994)	(May1994-Feb1996)	2^{nd} **felling** (Feb1996-Jan1998)	**Post-fell**
Severn	0.35	0.34 (-Apr1985)	0.37	0.39	0.40	0.34	0.34 (Feb1998-)
Tanllwyth	0.31	0.30 (-Apr1994)	-	-	0.34	0.29	0.31 (Feb1998-)
Hafren	0.38	-	-	-	-	-	-
Lower Hore	0.31	0.31 (-Apr1985)	0.32	-	-	-	0.30 (May1991-)
Upper Hore	0.33	-	-	-	-	-	-
Wye	0.30	-	-	-	-	-	-
Gwy	0.33	-	-	-	-	-	-
Cyff	0.30	-	-	-	-	-	-
Iago	-	-	-	-	-	-	-

Appendix C

Calculate CN and BFI from soil and vegetation data

CN and BFI can be calculated from on soil and vegetation and corrected based on topography and drainage information. The procedure consists of various steps illustrated in the workflow in Figure C.1. At first, the remotely sensed data are gathered from the relevant data providers. These are in the form of separate GIS layers representing soil, vegetation, topography (usually available as rasters) and drainage network (usually available as vectors). The slope raster is calculated from the topography raster using the algorithm suggested by Horn (1981) and implemented in the R library "raster". The GIS layers, represented by dark blue database objects in the workflow, are resampled to the same spatial resolution (e.g. 25m), projected in a common Coordinate Reference System (e.g. British National Grid, epsg: 27700) and overlaid. The new layers are represented by light blue database objects in the workflow.

A lookup table is compiled with all the possible combinations of HOST-USDA classes, land cover classes and corresponding theoretical CN for hydrologic soil groups A, AB, B, BC, C, CD and D (see example in table C.1, where grey shaded cells show the CN corresponding to a certain soil-vegetation combination). The intermediate groups are calculated as the average of the their extremes, e.g. if $A = 50$ and $B = 70$, $AB = 60$.

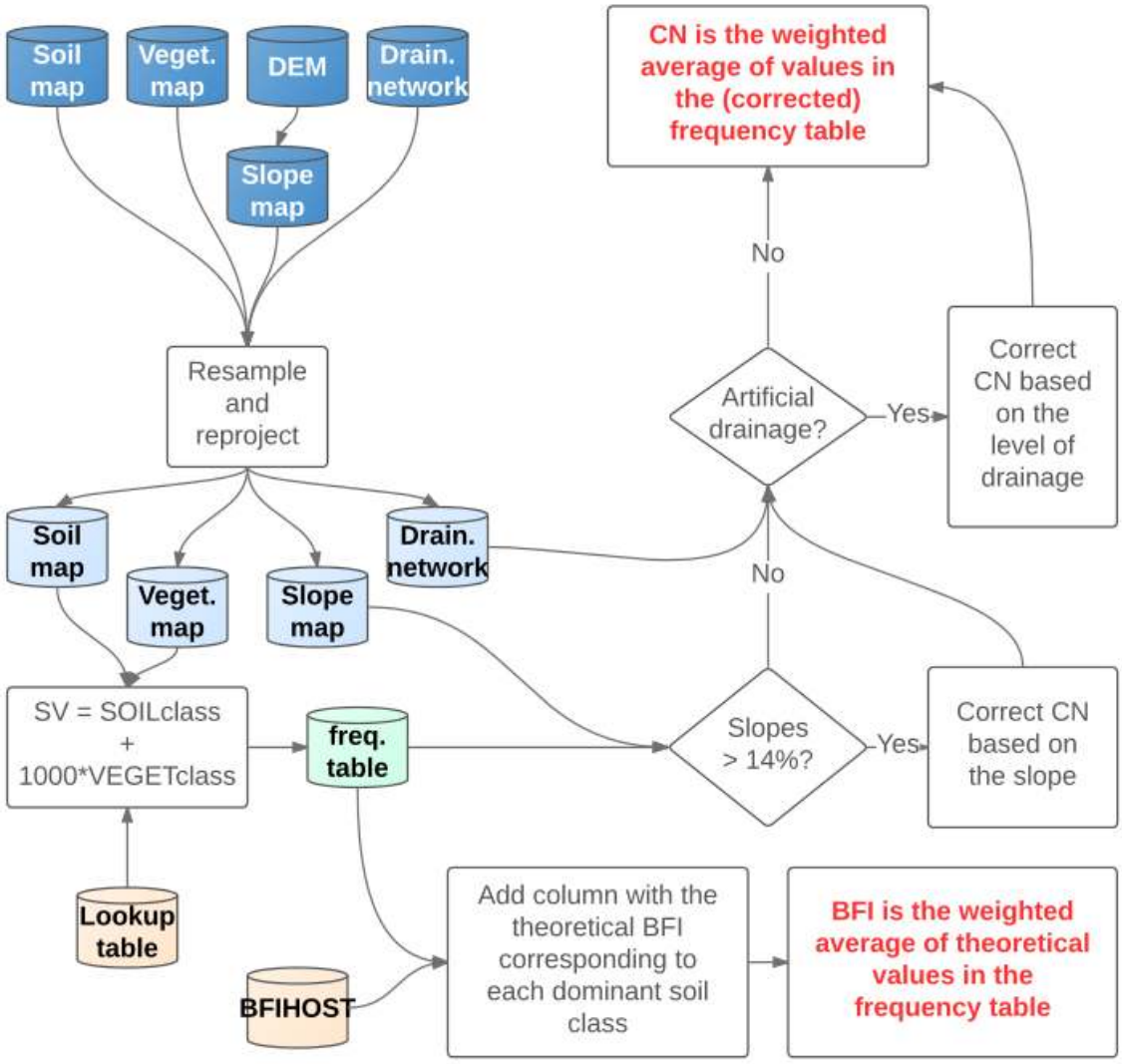

Figure C.1: Workflow for calculating CN from remotely sensed data.

Table C.1: Example of lookup table for Plynlimon's soil and vegetation classes. Grey-shaded cells illustrate the USDA/HOST mapping proposed by Bulygina et al. (2011).

HOST class	Vegetation class	Description	A	AB	B	BC	C	CD	D
15	1 (coniferous woodland)	Good, woods	30	42	55	62	70	74	77
17	7 (improved grassland)	Fair, pasture	49	59	69	74	79	82	84
29	6 (open dwarf shrub heath)	Fair, brush	35	46	56	63	70	74	77

The presence of artificial ditches increases the soil drainage worsening the natural hydrological conditions. In areas where the ratio between artificial drains and natural drains is below 50%, this work suggests to correct CN degrading areas in apparently good hydrological conditions to fair and areas in fair hydrological conditions to poor. No correction is applied if soils are already in poor conditions. In areas where the ratio is above 50%, the hydrological conditions are always considered poor.

Soil map units, as represented on the national soil maps, generate datasets with many HOST classes in each 1km grid cell. In the HOST data set there is a maximum of seven classes in each grid cell. HOST can be mapped by, for example, choosing only to display the dominant HOST class, which would ignore the other up to six classes that are present. In this case, a new raster (SV) is calculated by assigning to each cell the result of the following formula:

$$SV = SOIL_{class} + 1000 \cdot VEGETATION_{class} \tag{C.1}$$

Alternatively, the user can decide to utilise the information related to the seven classes in each grid cell. In this case the equation C.1 should be applied to each layer of the soil raster stack and then weighted based on the percentage coverage of each class.

For each SV value, the corresponding CN is extracted from the lookup table and the corresponding theoretical BFI is obtained from table 3.4 of Boorman et al. (1995).

The CN_0 frequency table summarises the number of cells for each calculated CN and the relative percentage coverage. The CN values calculated at this step, are acceptable values for slopes below 14% and no artificial drainage. Where the slope is above 14%, CN_0 is corrected using the equation proposed by Huang et al. (2006):

$$CN_{maps} = CN_0 \cdot \frac{322.79 + 15.63 \cdot S}{S + 323.52} \tag{C.2}$$

where S is the slope in percentage.

Signatures calculated using remotely sensed data are static and refer to the time/area for which the mapped geology and vegetation cover were observed. As land use changes over time and space, signatures are expected to change over these dimensions accordingly. A dynamic map of signatures could be implemented by repeating the methodology above for various location and update it as soon as changes are recorded. However, systematic and periodical updates of land cover maps are still uncommon. In the UK, for instance, land cover maps have been periodically published (in 1990, 2000, 2007 and 2013) but the classification is based on different methodologies which implies that a systematic measure of changes cannot be performed.

In order to investigate the observed changes in signatures over time, CN and BFI were also calculated from time series data of precipitation and streamflow discharge, as described in detail in the next section.

C.1 Regionalised CN and BFI for Plynlimon

The soil classes identifiable in the Plynlimon area are: H15, H17, H22, H26 and H29. The percentage coverages for each catchment and HOST class are summarised in table C.2. The first row shows the theoretical BFI, calculated by (Boorman et al., 1995, page 34) from multiple regression analysis. For each catchment, the BFI_{HOST} is calculated as the weighted average of these coefficients using the related percentage coverages as weights. The results are summarised in table C.3. The first column lists the catchments, the second column shows the BFI_{HOST} according to the percentages published in Boorman et al. (1995), the third column shows the values calculated using only the dominant soil classes, the fourth column shows the BFI_{HOST} calculated according to the percentages extracted from the latest soil map. These three columns are explicitly shown to inform the reader that subtle differences should be expected switching from one source of information to another. For clarity, the BFI_{HOST} in the third column will be considered as the theoretical/regionalised value. The fifth column of table C.3 lists the standard deviations expected for each catchment.

Table C.2: Percentages of HOST classes. When different, values calculated by Boorman et al. (1995) are added in parenthesis. Rounding error can cause the sum of the percentages to be different from 100. Percentages for the Upper Hore were not available before 1995.

HOST classes	**H15**	**H17**	**H22**	**H26**	**H29**
BFI coefficients from multiple regression analysis	0.387	0.613	0.294	0.247	0.232
Severn	0.58	0.01	0.00	0.14 (0.15)	0.27 (0.26)
Tanllwyth	0.64 (0.76)	0.00	0.00	0.14	0.21 (0.11)
Hafren	0.42 (0.43)	0.02 (0.03)	0.00	0.15	0.41 (0.40)
Lower Hore	0.66 (0.64)	0.00	0.00	0.14	0.20 (0.22)
Upper Hore	0.50	0.00	0.00	0.15	0.36
Wye	0.64	0.13	0.02	0.12	0.09
Gwy	0.67 (0.66)	0.00	0.00	0.14	0.19
Cyff	0.71 (0.70)	0.15 (0.16)	0.02	0.11	0.01
Iago	0.70 (0.69)	0.00	0.00	0.14	0.17

Table C.3: Theoretical BFI values calculated by Boorman et al. (1995) (second column), using only dominant classes (third column) and using the multi-layer HOST soil map (fourth column). The fifth column lists the weighted standard deviations.

Sub-catchment	**BFI_{HOST} (Boorman et al. 1995)**	**BFI_{HOST} (dominant classes)**	**BFI_{HOST}**	**$\sigma_{BFI_{HOST}}$**
Severn	0.328	0.339	0.328	0.020
Tanllwyth	0.354	0.346	0.331	0.020
Hafren	0.315	0.297	0.307	0.019
Lower Hore	0.333	0.367	0.336	0.021
Upper Hore	-	0.347	0.314	0.020
Wye	0.384	0.418	0.384	0.019
Gwy	0.334	0.387	0.338	0.021
Cyff	0.404	0.398	0.402	0.021
Iago	0.341	0.387	0.345	0.021

Table C.4 shows CN values calculated using dominant classes (second column) but also for a combination of good woods and fair pasture (third column) and fair woods and poor pasture (fourth column). This is to make explicit the uncertainty related to hydrological condition of the soil. However, the CN_{HOST} in the fourth column is considered the theoretical/regionalised value.

Table C.4: Theoretical CN values calculated using the HOST soil map and the land cover map updated in 2013.

Subcatchment	**CN_{HOST} (dominant classes)**	**CN_{HOST} (good woods - fair pasture)**	**CN_{HOST} (fair woods - poor pasture)**
Severn	76	76	78
Tanllwyth	74	75	77
Hafren	76	76	76
Lower Hore	75	77	79
Upper Hore	77	77	80
Wye	79	81	87
Gwy	82	83	88
Cyff	81	81	87
Iago	82	83	88

Bibliography

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'Connell, P. E. and Rasmussen, J. (1986), 'An introduction to the European Hydrological System. SHE, 1: History and philosophy of a physically 2: based distributed modeling system', *J. Hydrol.* **87**, 45–59.

Agrawal, R., Imieliski, T. and Swami, A. (1993), 'Mining association rules between sets of items in large databases', *ACM SIGMOD Record* **22**(2), 207–216.

Agrawal, R. and Srikant, R. (1994), Fast Algorithms for Mining Association Rules, *in* 'Proceedings of the 20th VLDB Conference', Santiago, Chile.

Allen, R. G., Pereira, L. S., Raes, D. and Smith, M. (1998), Crop evapotranspiration - Guidelines for computing crop water requirements - FAO Irrigation and drainage paper 56, Technical report, FAO - Food and Agriculture Organization of the United Nations, Rome.
URL: *http://www.fao.org/docrep/x0490e/x0490e08.htm*

Amatriain, X. (2015), 'In machine learning, is more data always better than better algorithms?'.
URL: *https://www.quora.com/In-machine-learning-is-more-data-always-better-than-better-algorithms*

Ames, D. P., Horsburgh, J. S., Cao, Y., Kadlec, J., Whiteaker, T. and Valentine, D. (2012), 'HydroDesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis', *Environmental Modelling & Software* **37**, 146–156.

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.-H. and Valéry, A. (2009), 'HESS Opinions "Crash tests for a standard-

ized evaluation of hydrological models"', *Hydrology and Earth System Sciences Discussions* **6**(3), 3669–3685.

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

Andrews, F. T., Croke, B. F. W. and Jakeman, A. J. (2011), 'An open software environment for hydrological model assessment and development', *Environmental Modelling & Software* **26**(10), 1171–1185.

Argent, R. M., Grayson, R. B. and Rahman, J. M. (2005), Model Selection and the Catchment Modelling Toolkit, *in* '29th Hydrology and Water Resources Symposium: Water Capital, 20-23 February 2005, p.[430]-[437]', Institution of Engineers Australia, Rydges Lakeside, Canberra.

Arnold, J. G., Srinivasan, R., Muttiah, R. S. and Williams, J. R. (1998), 'Large area hydrological modeling part 1: model development.', *Journal of the American Water Resources Association* **34**(1), 73–89.

Ballard, C. (2011), The role of physics based models for simulating runoff responses to rural land management scenarios, Thesis, Imperial College London.

Bayardo, R. J. and Agrawal, R. (1999), Mining the most interesting rules, *in* 'Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99', ACM Press, New York, New York, USA, pp. 145–154.

Beck, M. B. (1987), 'Parameter conditioning and prediction uncertainties of the LISFLOOD-WB distributed hydrological model', *Water Resources Research* **23**(8), 1393–1442.

Beran, B. and Piasecki, M. (2009), 'Engineering new paths to water data', *Computers & Geosciences* **35**(4), 753–760.

Berndt, D. J. and Clifford, J. (1994), Using Dynamic Time Warping to Find Patterns in Time Series, *in* 'KDD Workshop', pp. 359–370.

Bettis-Outland, H. (2012), 'Decision-making's impact on organizational learning and information overload', *Journal of Business Research* **65**(6), 814–820.

Beven, K. (2001*a*), 'How far can we go in distributed hydrological modelling?', *Hydrology and Earth System Sciences* **5**(1), 1–12.

Beven, K. (2006), 'A manifesto for the equifinality thesis', *Journal of Hydrology* **320**(1-2), 18–36.

Beven, K. J. (2000*a*), 'On the future of distributed modeling in hydrology.', *Hydrological Processes* **14**, 3183–3184.

Beven, K. J. (2000*b*), 'Uniqueness of place and process representations in hydrological modelling', *Hydrology and Earth System Sciences* **4**(2), 203–213.

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

Beven, K. J. (2007), 'Working towards integrated environmental models of everywhere: uncertainty, data, and modelling as a learning process.', *Hydrology and Earth System Science* pp. 460–467.

Beven, K. J. and Binley, A. M. (1992), 'The future of distributed models: model calibration and uncertainty prediction', *Hydrological Processes* **6**, 279–298.

Beven, K. J. and Freer, J. (2001), 'Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology', *Journal of Hydrology* **249**(1-4), 11–29.

Beven, K. J. and Kirkby, M. J. (1979), 'A physically based variable contributing area model of basin hydrology', *Hydrol. Sci. Bull.* **24**, 43–69.

Beven, K. J. and OConnell, P. E. (1982), On the Role of Physically-based Distributed Models in Hydrology, Technical Report 81, Technical Report Institute of Hydrology Report, Wallingford.

Beven, K. J., Warren, R. and Zaoui, J. (1980), 'SHE: towards a methodology for physically-based', *IAHS Publ.* **129**, 133–137.

Bird, S. B., Emmett, B. A., Sinclair, F. L., Stevens, P. A., Reynolds, B., Nicholson, S. and Jones, T. (2003), PONTBREN: Effects of tree planting on agricultural soils and their functions. CCW Contract Science Report No 550., Technical report, NERC Centre for Ecology and Hydrology.

Blackie, J. R. and Eeles, C. W. O. (1985), *Lumped Catchment Models in Hydrological Forecasting*, Vol. 3, John Wiley and Sons.

Blackie, J. R. and Robinson, M. (2007), 'Development of catchment research, with particular attention to Plynlimon and its forerunner, the East African catchments', *Hydrology and Earth System Sciences* **11**(1), 26–43.

Blangiardo, M., Cameletti, M., Baio, G. and Rue, H. (2013), 'Spatial and spatio-temporal models with R-INLA', *Spatial and Spatio-temporal Epidemiology* **7**, 39–55.

Blöschl, G. and Sivapalan, M. (1995), 'Scale issues in hydrological modelling a review.', *Hydrological Processes* **9**, 251–290.

Blöschl, G., Sivapalan, M. and Wagener, T., eds (2013), *Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales*, Cambridge University Press.

Boardman, J., Ligneau, L., de Roo, A. and Vandaele, K. (1994), 'Flooding of property by runoff from agricultural land in northwestern Europe', *Geomorphology* **10**(1-4), 183–196.

Boorman, D. B., Hollist, J. M. and Lilly, A. (1995), Report No. 126 Hydrology of soil types: a hydrologically-based classification of the soils of the United Kingdom, Technical report, Institute of Hydrology, Wallingford.

Bowling, L. C., Storck, P. and Lettenmaier, D. P. (2000), 'Hydrologic effects of logging in western Washington, United States', *Water Resources Research* **36**(11), 3223–3240.

Boyle, D. P., Gupta, H. V. and Sorooshian, S. (2000), 'Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods', *Water Resources Research* **36**(12), 3663–3674.

Breuer, L., Huisman, J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H.-G., Gräff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M. and Viney, N. R. (2009), 'Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use', *Advances in Water Resources* **32**(2), 129–146.

Brill, E. (2003), *Computational Linguistics and Intelligent Text Processing*, Vol. 2588 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg.

Brogi, A., Mancarella, P., Pedreschi, D. and Turini, F. (1994), 'Modular logic programming', *ACM Transactions on Programming Languages and Systems (TOPLAS)* **16**(4), 1361–1398.

Brown, A. E., Zhang, L., McMahon, T. A., Western, A. W. and Vertessy, R. A. (2005), 'A review of paired catchment studies for determining changes in water yield resulting from alterations in vegetation', *Journal of Hydrology* **310**(1-4), 28–61.

Bulygina, N., McIntyre, N. and Wheater, H. (2009), 'Conditioning rainfall-runoff model parameters for ungauged catchments and land management impacts analysis', *Hydrology and Earth System Sciences* **13**(6), 893–904.

Bulygina, N., McIntyre, N. and Wheater, H. (2011), 'Bayesian Conditioning of a rainfall-runoff model for predicting flows in unguaged catchments and under land use changes', *Water Resources Research* **47**.

Burt, T. P. (2001), 'Integrated management of sensitive catchment systems', *CATENA* **42**(2-4), 275–290.

Butts, M. B., Payne, J. T., Kristensen, M. and Madsen, H. (2004), 'An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation', *Journal of Hydrology* **298**(1-4), 242–266.

Buytaert, W. and Beven, K. J. (2009), 'Regionalization as a learning process', *Water Resour. Res.* **45**.

Buytaert, W., Reusser, D. E., Krause, S. and Renaud, J. (2008), 'Why can't we do better than Topmodel?', *Hydrological Processes* **22**(August), 4175–4179.

Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J. M. and Pascual-Montano, A. (2006), 'Integrated analysis of gene expression by Association Rules Discovery.', *BMC bioinformatics* **7**(1), 54.

Castronova, A. M., Goodall, J. L. and Elag, M. M. (2013), 'Models as web services using the Open Geospatial Consortium (OGC) Web Processing Service (WPS) standard', *Environmental Modelling & Software* **41**, 72–83.

Castronova, A. M., Goodall, J. L. and Ercan, M. B. (2013), 'Integrated modeling within a Hydrologic Information System: An OpenMI based approach', *Environmental Modelling & Software* **39**, 263–273.

Chandramouli, V. and Raman, H. (2001), 'Multireservoir Modeling with Dynamic Programming and Neural Networks', *Journal of Water Resources Planning and Management* **127**(2), 89–98.

Chang, F.-J. and Chen, Y.-C. (2003), 'Estuary water-stage forecasting by using radial basis function neural network', *Journal of Hydrology* **270**(1-2), 158–166.

Christoffersen, P. F. (1998), 'Evaluating Interval Forecasts', *International Economic Review* **39**(4), 841.

Clark, M. P., McMillan, H. K., Collins, D. B. G., Kavetski, D. and Woods, R. A. (2011),

'Hydrological field data from a modeller's perspective: Part 2: process-based evaluation of model hypotheses', *Hydrological Processes* **25**(4), 523–543.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T. and Hay, L. E. (2008), 'Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models', *Water Resour. Res.* **44**, 91–94.

Cormen, T. H. (2013), *Algorithms Unlocked*, The MIT Press.

Corne, D. W. and Knowles, J. D. (2007), Techniques for highly multiobjective optimisation, *in* 'Proceedings of the 9th annual conference on Genetic and evolutionary computation - GECCO '07', ACM Press, New York, New York, USA, p. 773.

Coxon, G., Freer, J., Wagener, T., Odoni, N. A. and Clark, M. (2014), 'Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments', *Hydrological Processes* **28**(25), 6135–6150.

CRCCH (2005), Series on model choice n.1: General approaches to modelling and practical issues of model choice, Technical report, Cooperative Research Centre for Catchment hydrology.

Cunderlik, J. (2003), Hydrologic Model Selection for the CFCAS Project: Assessment of Water Resources Risk and Vulnerability to Changing Climatic Conditions - Book 9, Technical report, Department of Civil and Environmental Engineering, The University of Western Ontario, London, Ontario, Canada.
URL: *http://ir.lib.uwo.ca/wrrr/9*

Cutrell, J. (2012), 'Understanding the Principles of Algorithm Design'.
URL: *http://code.tutsplus.com/tutorials/understanding-the-principles-of-algorithm-design–net-26561*

David, O., Markstrom, S. L., Rojas, K. W., Ahuja, L. R. and Schneider, I. W. (2002), *The Object Modeling System - Chapter 15 in Agricultural System Models in Field Research and Technology Transfer*, CRC Press.

Dawdy, D. R. and O'Donnell, T. (1965), 'Mathematical Models of Catchment Behavior', *Journal of the Hydraulics Division* **91**(4), 123–137.

Dawson, C. W., Abrahart, R. J. and See, L. M. (2007), 'HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts', *Environmental Modelling & Software* **22**(7), 1034–1052.

De Vos, M. G., Janssen, S. J. C., Van Bussel, L. G. J., Kromdijk, J., Van Vliet, J. and Tope, J. L. (2011), Are environmental models transparent and reproducible enough?, *in* '19th International Congress on Modelling and Simulation', Perth, Australia.
URL: *http://mssanz.org.au/modsim2011*

Demiriz, A., Bennett, K. P. and Shawe-Taylor, J. (2002), 'No Title', *Machine Learning* **46**(1/3), 225–254.

Dooge, J. C. I. (1957), 'Rational method for estimating flood peaks', *Engineering* **184**, 374–377.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H., Gusev, Y., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. and Wood, E. (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.

Duan, Q. Y., Sorooshian, S. and Gupta, V. (1992), 'Effective and efficient global optimization for conceptual rainfallrunoff models', *Water Resources Research* **28(4)**(4), 1015–1031.

Ebrahimian, M., Nuruddin, A. A. B., Soom, M. A. B. M., Sood, A. M. and Neng, L. J. (2012),

'Runoff Estimation in Steep Slope Watershed with Standard and Slope-Adjusted Curve Number Methods', *Polish Journal of Environmental Studies* **21**(5), 1191–1202.

Einax, J. W., Truckenbrodt, D. and Kampe, O. (1998), 'River Pollution Data Interpreted by Means of Chemometric Methods', *Microchemical Journal* **58**(3), 315–324.

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

Fenicia, F., McDonnell, J. J. and Savenije, H. H. G. (2008), 'Learning from model improvement: On the contribution of complementary data to process understanding', *Water Resources Research* **44**(6).

Garaud, D. and Mallet, V. (2011), 'Automatic calibration of an ensemble for uncertainty estimation and probabilistic forecast: Application to air quality', *Journal of Geophysical Research* **116**(D19), D19304.

Giorgino, T. (2009), 'Computing and Visualizing Dynamic Time Warping Alignments in R : The dtw Package', *Journal of Statistical Software* **31**(7), 1–24.

Goodrich, M. T. and Tamassia, R. (2002), *Algorithm Design: Foundations, Analysis, and Internet Examples*, John Wiley & Sons, Inc.

Grace, B. and Quick, B. (1988), 'A Comparison of Methods for the Calculation of Potential Evapotranspiration Under the Windy Semi-arid Conditions of Southern Alberta', *Canadian Water Resources Journal* **13**(1), 9–19.

Grayson, R. and Blöschl, G. (2001), *Spatial Patterns in Catchment Hydrology: Observations and Modelling.*, Cambridge University Press: Cambridge, UK.

Gupta, H. V., Sorooshian, S. and Yapo, P. O. (1998), 'Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information', *Water Resources Research* **34**(4), 751–763.

Gupta, V. K. and Sorooshian, S. (1985), 'The Automatic Calibration of Conceptual Catchment Models Using Derivative-Based Optimization Algorithms', *Water Resources Research* **21**(4), 473–485.

Gustard, A., Bullock, A. and Dixon, J. M. (1992), Report No. 108: Low flow estimation in the United Kingdom, Technical report, Institute of Hydrology.

Halcrow and University of Stirling (2011), Allan Water Natural Flood Management Techniques and Scoping Study (Scoping report - appendix A)., Technical report, Scottish Environment Protection Angency.

Halevy, A., Norvig, P. and Pereira, F. (2009), 'The Unreasonable Effectiveness of Data', *IEEE Intelligent Systems* **24**(2), 8–12.

Hawkins, R. H. (1993), 'Asymptotic Determination of Runoff Curve Numbers from Data', *Journal of Irrigation and Drainage Engineering* **119**(2), 334–345.

Herbst, M., Gupta, H. V. and Casper, M. C. (2008), 'Mapping model behaviour using Self-Organizing Maps', *Hydrology and Earth System Sciences Discussions* **5**(6), 3517–3555.

Hersperger, A. M., Gennaio, M.-P. P., Verburg, P. H. and Bürgi, M. (2010), 'Linking Land Change with Driving Forces and Actors: Four Conceptual Models', *Ecology and Society* **15**(4).

Horn, B. (1981), 'Hill shading and the reflectance map', *Proceedings of the IEEE* **69**(1), 14–47.

Hornberger, G. (1980), 'Eutrophication in peel inletI. The problem-defining behavior and a mathematical model for the phosphorus scenario', *Water Research* **14**(1), 29–42.

Hornberger, G. M. and Spear, R. C. (1981), 'An approach to the preliminary analysis of environmental systems', *Journal of Environmental Management* **12**(1), 7–18.

Huang, M., Gallichand, J., Wang, Z. and Goulet, M. (2006), 'A modification to the Soil Conservation Service curve number method for steep slopes in the Loess Plateau of China', *Hydrological Processes* **20**(3), 579–589.

Hudson, J. A., Gilman, K. and Calder, I. R. (1997), 'Land use and water issues in the uplands with reference to the Plynlimon study', *Hydrology and Earth System Sciences* **1**(3), 389–397.

Ireland, G., Volpi, M. and Petropoulos, G. (2015), 'Examining the Capability of Supervised Machine Learning Classifiers in Extracting Flooded Areas from Landsat TM Imagery: A Case Study from a Mediterranean Flood', *Remote Sensing* **7**(3), 3372–3399.

Jachner, S., van den Boogaart, K. G. and Petzoldt, T. (2007), 'Statistical Methods for the Qualitative Assessment of Dynamic Models with Time Delay (R Package qualV)', *Journal of Statistical Software* **22**(8).

Jiang, T., Chen, Y. D., Xu, C.-y., Chen, X., Chen, X. and Singh, V. P. (2007), 'Comparison of hydrological impacts of climate change simulated by six hydrological models in the Dongjiang Basin, South China', *Journal of Hydrology* **336**(3-4), 316–333.

Jones, J. (1997), 'Global Hydrology: Processes, Resources and Environmental Management', *Longman* .

Jordan, P., Argent, R. and Nathan, R. (2007), New generation models for catchment modelling: developments in the eWater CRC, *in* 'Proceedings of the 5th Australian Stream Management Conference. Australian rivers: making a difference. Charles Sturt University, Thurgoona, New South Wales.'.

Kampf, S. K. and Burges, S. J. (2007), 'A framework for classifying and comparing distributed hillslope and catchment hydrologic models', *Water Resources Research* **43**.

Kavetski, D., Fenicia, F. and Savenije, H. H. G. (2010), A flexible multi-model framework for catchment-specific calibration, and application to diverse European catchments, *in* 'Geophysical Research Abstracts, EGU General Assembly 2010'.

Kirby, C., Newson, M. D. and Gilman, K. (1991), Plynlimon research: the first two decades. Rep. 109, Technical report, Institute of Hydrology, Wallingford.

Klok, E. J., Jasper, K., Roelofsma, K. P., Gurtz, J. and Badoux, A. (2001), 'Distributed hydrological modelling of a heavily glaciated Alpine river basin', *Hydrological Sciences Journal* **46**(4), 553–570.

Kohonen, T., Hynninen, J., Kangas, J. and Laaksonen, J. (1996), SOM PAK: The Self-Organizing Map program package. Technical Report A31., Technical report, Helsinki University of Technology.

Königer, P. and Janowitz, K. (1995), 'Drowning in information, but thirsty for knowledge', *International Journal of Information Management* **15**(1), 5–16.

Kottek, M., Grieser, J., Beck, C., Rudolf, B. and Rubel, F. (2006), 'World Map of the Köppen-Geiger climate classification updated', *Meteorologische Zeitschrift* **15**(3), 259–263.

Krause, P., Boyle, D. P. and Bäse, F. (2005), 'Comparison of different efficiency criteria for hydrological model assessment', *Advances in Geosciences* **5**, 89–97.

Kuczera, G. (1997), 'Efficient subspace probabilistic parameter optimization for catchment models', *Water Resources Research* **33**(1), 177–185.

Kuczera, G. and Parent, E. (1998), 'Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm', *Journal of Hydrology* **211**(1-4), 69–85.

Laio, F. and Tamea, S. (2007), 'Verification tools for probabilistic forecasts of continuous hydrological variables', *Hydrology and Earth System Sciences* **11**(4), 1267–1277.

Leavesley, G. H., Markstrom, S. L., Restrepo, P. J. and Viger, R. J. (2002), 'A modular approach to addressing model design, scale, and parameter estimation issues in distributed hydrological modelling', *Hydrological Processes* **16**(2), 173–187.

Leavesley, G. H., Markstrom, S. L. and Viger, R. J. (2006), 'USGS Modular Modeling System (MMS) as e Precipitation-Runoff Modeling System (PRMS)', *CRC Press* p. 159e177.

Leavesley, G. H., Restrepo, P. J., Markstrom, S. L., Dixon, M. and Stannard, L. G. (1996), 'The modular modeling systemMMS: Users manual', *U.S. Geol. Surv. Open File Rep.* **142**, 96–151.

Lees, M. J. (2000), 'Data-based mechanistic modelling and forecasting of hydrological systems', *Journal of Hydroinformatics* **2**(1), 15–34.

Legates, D. R. and McCabe, G. J. (1999), 'Evaluating the use of goodness-of-fit Measures in hydrologic and hydroclimatic model validation', *Water Resources Research* **35**(1), 233–241.

Ley, R., Casper, M. C., Hellebrand, H. and Merz, R. (2011), 'Catchment classification by runoff behaviour with self-organizing maps (SOM)', *Hydrology and Earth System Sciences* **15**(9), 2947–2962.

Liang, X. (1994), *A two-layer variable infiltration capacity land surface representation for general circulation models*, Water Resour. Ser. TR140.

Lin, G.-F. and Chen, L.-H. (2006), 'Identification of homogeneous regions for regional frequency analysis using the self-organizing map', *Journal of Hydrology* **324**(1-4), 1–9.

Lin, G.-F. and Wu, M.-C. (2011), 'An RBF network with a two-step learning algorithm for developing a reservoir inflow forecasting model', *Journal of Hydrology* **405**(3-4), 439–450.

Lu, J. B., Sun, G., McNulty, S. G. and Amatya, D. M. (2005), 'A comparison of six potential evapotranspiration methods for regional use in the southeastern United States', *Journal of the American Water Resources Association* **41**(3), 621–633.

Ma, H. and Yang, D. (2010), 'Hydrological model comparison and combination for flood forecast in the Three Gorges Region, China', *Advances in Geosciences* **17**, 203–216.

Malhotra, K. and Venugopal, H. (2011), 'Association Rules of Data Mining for the Characteristic Analysis of Sub-basins of a River', *International Journal of Computer Applications* **27**(3), 5–8.

Manimaran, J. and Velmurugan, T. (2013), A survey of association rule mining in text applications, *in* '2013 IEEE International Conference on Computational Intelligence and Computing Research', IEEE, pp. 1–5.

Marshall, L., Nott, D. and Sharma, A. (2005), 'Hydrological model selection: A Bayesian alternative', *Water Resources Research* **41**(10).

Marshall, M. R., Francis, O. J., Frogbrook, Z. L., Jackson, B. M., McIntyre, N., Reynolds, B., Solloway, I., Wheater, H. S. and Chell, J. (2009), 'The impact of upland land management on flooding: results from an improved pasture hillslope', *Hydrological Processes* **23**(3), 464–475.

Mason, L., Bartlett, P. and Baxter, J. (1998), Direct Optimization of Margins Improves Generalization in Combined Classifiers, Technical report, Advances in Neural Information Processing Systems.

McCulloch, J. S. and Robinson, M. (1993), 'History of forest hydrology', *Journal of Hydrology* **150**(2-4), 189–216.

McGuire, M. P. and Gangopadhyay, A. (2006), Modeling, visualizing, and mining hydrological spatial hierarchies for water quality management, *in* 'ASPRS 2006 Annual Conference', Reno, Nevada.

McIntyre, N., Ballard, C., Bruen, M., Bulygina, N., Buytaert, W., Cluckie, I., Dunn, S., Ehret, U., Ewen, J., Gelfan, A., Hess, T., Hughes, D., Jackson, B., Kjeldsen, T., Merz, R., Park, J.-S., Connell, E. O., Donnell, G. O., Oudin, L., Todini, E., Wagener, T. and Wheater, H. (2013), 'Modelling the hydrological impacts of rural land use change: current state of the science and future challenges', *Hydrology Research* **45**(6), 737–754.

McIntyre, N., Lee, H., Wheater, H., Young, A. and Wagener, T. (2005), 'Ensemble predictions of runoff in ungauged catchments', *Water Resources Research* **41**(12).

McKay, M. D., Beckman, R. J. and Conover, W. J. (1979), 'A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code', *Technometrics* **21**(2), 239.

McMillan, H. K., Clark, M. P., Bowden, W. B., Duncan, M. and Woods, R. A. (2011), 'Hydrological field data from a modeller's perspective: Part 1. Diagnostic tests for model structure', *Hydrol. Process.* **25**(4), 511–522.

McMillan, H. K., Clark, M. P., Woods, R. A., Bowden, W. B. and Duncan, M. (2010), Improving Perceptual and Conceptual Hydrological Models using Data from Small Basins, *in* 'Status and Perspectives of Hydrology in Small Basins, Workshop held at Goslar-Hahnenklee, Germany, 30 March2 April 2009. IAHS Publ.', pp. 308–336.

Mersmann, O. (2015), 'CRAN- Package emoa'.
URL: *http://cran.r-project.org/web/packages/emoa*

Mo, X., Pappenberger, F., Beven, K. J., Liu, S., De Roo, A. and Lin, Z. (2006), 'Parameter conditioning and prediction uncertainties of the LISFLOOD-WB distributed hydrological model', *Hydrological Sciences Journal* **51**(1), 45–65.

Mobasher, B., Dai, H., Luo, T. and Nakagawa, M. (2001), Effective personalization based on association rule discovery from web usage data, *in* 'Proceeding of the third international workshop on Web information and data management - WIDM '01', ACM Press, New York, New York, USA, p. 9.

Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2012), *Foundations of Machine Learning*, The MIT Press.

Morris, E. M. (1980), 'Forecasting flood flows in grassy and forested basins using a deterministic distributed mathematical model', *IAHS Publ.* **129**, 247–265.

Mulvany, T. J. (1851), 'On the use of self-registering rain and flood gauges', *Trans. Inst. Civ. Eng.* **4**, 1–8.

Murdoch, L., Famiglietti, J., Lakshmi, V. and Hooper, R. (2008), CUAHSI Community Hydrologic Modeling Platform (CHyMP) Initiative, *in* 'EPA Integrated Modeling Workshop'.

Najafi, M. R., Moradkhani, H. and Jung, I. W. (2011), 'Assessing the uncertainties of hydrologic model selection in climate change impact studies', *Hydrological Processes* **25**(18), 2814–2826.

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

Nemeč, J. (1993), Comparison and selection of existing hydrological models for the simulation of the dynamic water balance processes in basins of different sizes and on different scales., Technical report, CHR/KHR-Report no. II-7 International Commission for the Hydrology of the Rhine Basin.

Oreskes, N., Shrader-Frechette, K. and Belitz, K. (1994), 'Verification, validation, and confirmation of numerical models in the Earth sciences.', *Science (New York, N.Y.)* **263**(5147), 641–6.

Paudel, M. (2010), An Examination of Distributed Hydrologic Modeling Methods as Compared with Traditional Lumped Parameter Approaches, PhD thesis, Brigham Young University.

Pechlivanidis, I. G., Jackson, B. M., Mcintyre, N. R. and Wheater, H. S. (2011), 'Catchment Scale Hydrological Modelling: a Review of Model Types, Calibration Approaches and Uncertainty Analysis Methods in the Context of Recent Developments in Technology and Applications', *Global Nest Journal* **13**(3), 193–214.

Peel, M. C., Finlayson, B. L. and McMahon, T. A. (2007), 'Updated world map of the Köppen-Geiger climate classification', *Hydrology and Earth System Sciences* **11**(5), 1633–1644.

Perry, C. (2013), 'Machine Learning and Conflict Prediction: A Use Case', *Stability: International Journal of Security & Development* **2**(3), 56.

Pianosi, F. and Wagener, T. (2015), 'A simple and efficient method for global sensitivity analysis based on cumulative distribution functions', *Environmental Modelling & Software* **67**, 1–11.

Pilászy, I. and Tikk, D. (2009), Recommending new movies, *in* 'Proceedings of the third ACM conference on Recommender systems - RecSys '09', ACM Press, New York, New York, USA, p. 93.

Refsgaard, J. C. and Knudsen, J. (1996), 'Operational Validation and Intercomparison of Different Types of Hydrological Models', *Water Resources Research* **32**(7), 2189–2202.

Reggiani, P., Sivapalan, M. and Majid Hassanizadeh, S. (1998), 'A unifying framework for watershed thermodynamics: balance equations for mass, momentum, energy and entropy, and the second law of thermodynamics', *Advances in Water Resources* **22**(4), 367–398.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M. and Franks, S. W. (2010), 'Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors', *Water Resources Research* **46**(5).

Reusser, D. E. (2014), 'CRAN - Package tiger'.
URL: *http://cran.r-project.org/web/packages/tiger/index.html*

Reusser, D. E., Blume, T., Schaefli, B. and Zehe, E. (2009), 'Analysing the temporal dynamics of model performance for hydrological models', *Hydrology and Earth System Sciences* **13**(7), 999–1018.

Reusser, D. E. and Zehe, E. (2011), 'Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity', *Water Resources Research* **47**(7).

Reynolds, B., Neal, C., Hornung, M., Hughes, S. and Stevens, P. A. (1988), 'Impact of afforestation on the soil solution chemistry of stagnopodzols in mid-Wales', *Water, Air, and Soil Pollution* **38**(1-2), 55–70.

Rogers, L. L. and Dowla, F. U. (1994), 'Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling', *Water Resources Research* **30**(2), 457–481.

Sakoe, H. and Chiba, S. (1978), 'Dynamic programming algorithm optimization for spoken word recognition', *IEEE Trans. on Acoust., Speech, and Signal Process* **26**, 43–49.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A. and Carrillo, G. (2011), 'Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA', *Hydrology and Earth System Sciences* **15**(9), 2895–2911.

Schaefli, B. and Gupta, H. V. (2007), 'Do Nash values have value?', *Hydrological Processes* **21**(15), 2075–2080.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A. and Blake, A. (2013), 'Real-time human pose recognition in parts from single depth images', *Studies in Computational Intelligence* **411**, 119–135.

Singh, V. P. and Chowdhury, P. K. (1986), 'Comparing some methods of estimating mean areal rainfall', *Journal of the American Water Resources Association* **22**(2), 275–282.

Singh, V. P., Frevert, D. K., Rieker, J. D., Leverson, V. and Meyer, S. (2006), 'Hydrologic Modeling Inventory: Cooperative Research Effort', *Journal of Irrigation and Drainage Engineering* **132**(2), 98–103.

Sivapalan, M., Zhang, L., Vertessy, R. and Blöschl, G. (2003), 'Downward approach to hydrological prediction', *Hydrological Processes* **17**(11), 2099–2099.

Spear, R. (1980), 'Eutrophication in peel inletII. Identification of critical uncertainties via generalized sensitivity analysis', *Water Research* **14**(1), 43–49.

Spear, R. C., Grieb, T. M. and Shang, N. (1994), 'Parameter uncertainty and interaction in complex environmental models', *Water Resources Research* **30**(11), 3159–3169.

Stephenson, G. R. and Freeze, R. A. (1974), 'Mathematical simulation of subsurface flow contributions to snowmelt runoff, Reynolds Creek Watershed, Idaho', *Water Resources Research* **10**(2), 284–294.

Storm, B. and Jensen, K. H. (1984), 'Experience with Field Testings of SHE on Research Catchments', *Nordic Hydrology* **15**, 283–294.

Strebel, O. (2013), 'A preprocessing method for parameter estimation in ordinary differential equations', *Chaos, Solitons & Fractals* **57**, 93–104.

Tarantola, A. (1987), *Inverse Problem Theory*, Elsevier-Sci., New York, New York, USA.

Thomas, T. M. (1965), Sheet erosion induced by sheep in the Pumlumon (Plynlimon) area, mid-Wales. Rates of erosion and weathering in the British Isles., Technical report, Institute of British Geographers, Bristol.

Todini, E. (1988), 'Rainfall-runoff modelingPast, present and future', *J. Hydrol.* **100**, 341–352.

Tormene, P., Giorgino, T., Quaglini, S. and Stefanelli, M. (2009), 'Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation.', *Artificial intelligence in medicine* **45**(1), 11–34.

Toth, E. (2013), 'Catchment classification based on characterisation of streamflow and precipitation time series', *Hydrology and Earth System Sciences* **17**(3), 1149–1159.

Tsoumakas, G., Angelis, L. and Vlahavas, I. (2005), 'Selective fusion of heterogeneous classifiers', *Intell. Data Anal.* **9**(6), 511–525.

Tsoumakas, G., Katakis, I. and Vlahavas, I. (2004), Effective Voting of Heterogeneous Classifiers, *in* 'Machine Learning: ECML 2004', Springer Berlin Heidelberg, pp. 465–476.

USDA (1986), Urban Hydrology for Small Watersheds TR-55, Technical report, USDA.
URL: *http://www.cpesc.org/reference/tr55.pdf*

Van Den Boogaart, K. G., Rost, S. and Petzoldt, T. (2014), 'CRAN - Package qualV'.
URL: *http://cran.r-project.org/web/packages/qualV/index.html*

Vaughan, M. and McIntyre, N. (2012), 'An assessment of DBM flood forecasting models', *Proceedings of the ICE - Water Management* **165**(2), 105–120.

Veldkamp, A. and Fresco, L. O. (1996), 'CLUE: a conceptual model to study the Conversion of Land Use and its Effects'.

Vitolo, C. and Le Vine, N. (2015), 'Soil Conservation Service Curve Number method (CurveNumber, R-package)'.
URL: *https://github.com/cvitolo/r_CurveNumber*

Viviroli, D., Zappa, M., Gurtz, J. and Weingartner, R. (2009), 'An introduction to the hydrological modelling system prevah and its pre- and post-processing-tools', *Environmental Modelling & Software* **24**, 1209–1222.

Voronoi, G. (1908), 'Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites.', *Journal für die reine und angewandte Mathematik (Crelle's Journal)* **1908**(133).

Vrugt, J. A., Gupta, H. V., Bouten, W. and Sorooshian, S. (2003), 'A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters', *Water Resources Research* **39**(8).

Vrugt, J. A. and Ter Braak, C. J. F. (2011), 'DREAM(D): An adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems', *Hydrology and Earth System Sciences* **15**(12), 3701–3713.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V. and Sorooshian, S. (2001), 'A framework for development and application of hydrological models', *Hydrology and Earth System Sciences* **5**(1), 13–26.

Wan, D., Zhang, Y. and Li, S. (2007), Discovery Association Rules in Time Series of Hydrology, *in* '2007 IEEE International Conference on Integration Technology', IEEE, pp. 653–657.

Wang, Q. J. (1991), 'The genetic algorithm and its application to calibrating conceptual rainfallrunoff models', *Water Resources Research* **27(9)**, 2467–2471.

Warden, P. (2011), *Big Data Glossary*, O'Reilly Media.

Weinmann, B., Schroers, J. O. and Sheridan, P. (2006), 'Simulating the effects of decoupled transfer payments using the land use model ProLand', *German Journal of Agricultural Economics* **55**(5/6), 248–256.

Western, A. W. and Grayson, R. B. (1998), 'The Tarrawarra Data Set: Soil moisture patterns, soil characteristics, and hydrological flux measurements', *Water Resources Research* **34**(10), 2765–2768.

Wohlfahrt, J., Colin, F., Assaghir, Z. and Bockstaller, C. (2010), 'Assessing the impact of the spatial arrangement of agricultural practices on pesticide runoff in small catchments: Combining hydrological modeling and supervised learning', *Ecological Indicators* **10**(4), 826–839.

Wong, P. C., Whitney, P. and Thomas, J. (1999), 'Visualizing Association Rules for Text Mining'.

World Meteorological Organization (1975), *Intercomparison of conceptual models used in operational hydrological forecasting*, Secretariat of the World Meteorological Organization.

Wu, G. (2013), 'Why More Data and Simple Algorithms Beat Complex Analytics Models'.
URL: *http://data-informed.com/why-more-data-and-simple-algorithms-beat-complex-analytics-models/*

Yadav, M., Wagener, T. and Gupta, H. (2007), 'Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins', *Advances in Water Resource* **30**, 1756–1774.

Yan, J. (2010), 'CRAN - Package som'.

URL: *http://cran.r-project.org/web/packages/som/index.html*

Yang, D., Herath, S. and Musiake, K. (1998), 'Development of a geomorphology-based hydrological model for large catchments', *Annual Journal of Hydraulic Engineering, JSCE* **42**, 169–174.

Young, P. (1998), 'Data-based mechanistic modelling of environmental, ecological, economic and engineering systems', *Environmental Modelling and Software* **13**(2), 105–122.

Zégre, N., Skaugset, A. E., Som, N. A., McDonnell, J. J. and Ganio, L. M. (2010), 'In lieu of the paired catchment approach: Hydrologic model change detection at the catchment scale', *Water Resources Research* **46**(11).

Zhang, Y., Burer, S. and Street, W. N. (2006), 'Ensemble Pruning Via Semi-definite Programming', *J. Mach. Learn. Res.* **7**, 1315–1338.