Cemagref

Science, water & land management

Louise CROCHEMORE
2010 - 2011

Evaluation of hydrological models:

Expert judgement vs Numerical criteria

Final-year internship
From 1/2/2011 to 30/9/2011

POLYTECH
PARIS-UPMC
SCIENCES DE LA TERRE

Supervisor at Cemagref:
Charles Perrin
Hydrosystems and Bioprocesses Research Unit (UR HBAN)
1 rue Pierre-Gilles de Gennes
CS10030
92761 Antony Cedex
Phone: +33 (0)1 40 96 60 86
Mail: charles.perrin@cemagref.fr

Supervisor at Polytech'Paris-UPMC:
Roger Guérin
Tour 56 Couloir 56-46
4 place Jussieu
75252 Paris Cedex 05
Phone: +33 (0)1 44 27 45 91
Mail: roger.guerin@upmc.fr

Acknowledgements

Several persons have contributed to the success of the project and the writing of this thesis. Without these persons, working at Cemagref would not have been such a human, instructive and pleasant experience.

First of all, I would like to thank Charles Perrin who initiated the project and trusted me to work on it. He supervised me all along with simulating ideas, constant enthusiasm and great patience, which turned this project into a thrilling challenge.

Then I would like to thank Vazken Andréassian for his great pedagogic skills, his constant good humour and his pleasant stories and most of all for welcoming me in his research team.

Maria-Helena Ramos very kindly helped me in the process of creating the survey by giving me advice from her past experiences in designing surveys. She also encouraged me all along.

I want to thank the three of the above-mentioned persons for their optimism and insights on the scientific research field and their orientation advice. They really helped me in my reflections and I won't forget any of their advice.

Then, I want to thank John Ewen, Uwe Ehret and Simon Seibert for their collaboration in adding the visual numerical criteria to the study.

I also would like to thank the whole Hydro team: François, Florent, Julien, Damien, Pierre, Pierre-Yves, Ioanna, Annie, Raji and Gianluca, who were really welcoming and who patiently contributed to the survey by being the first guinea pigs.

Finally, a special thank to the other Master students of the HBAN research unit: Hajer, Sophie, Virginie, Cyril, Nadège, Marie, Abderrahmen and Trang, to the PhD students from the other teams: Cécile, Olivier and Violaine and I would like to add an extra thank to Valérie.

Table of contents

Introduction 1

I. Background of the study 3

I.A Introduction to hydrological modelling 3

I.A.1 From the catchment to hydrological modelling 3

I.A.2 Step 1: Model Study Plan 4

I.A.3 Step 2: Data and conceptualisation 5

I.A.4 Step 3: Model set-up 6

I.A.5 Step 4: Calibration and validation 6

I.A.6 Step 5: Simulation and evaluation 6

I.B Evaluation criteria 7

I.B.1 A need for evaluation and consensus 7

I.B.2 Visual criteria 8

I.B.3 Mathematical criteria 9

I.B.4 Comparing expert judgement and mathematical techniques 10

I.C Description of the study 12

I.C.1 The study 12

I.C.2 Organization of the work 12

II. Methodology 14

II.A Design of the survey on expert judgement 14

II.A.1 Objectives, process and interface 14

II.A.2 The graphs presented 14

II.A.3 The questions to answer 17

II.B Computation of mathematical criteria 17

II.B.1 Listing and computing of the mathematical criteria 17

II.B.2 Diversity of the criteria and gathering 18

III. Results and discussion 19

III.A Statistical results on the survey 19

III.A.1 General facts 19

III.A.2 Who answered the survey? 19

III.B Results and discussion on the collected expert judgements 21

III.B.1 How lenient were the experts? 21

III.B.2 Do experts evaluate alike? 23

III.B.3 How did the experts say they visually evaluated the hydrographs? 27

III.C Results and discussion on the comparison of evaluation methods 31

III.C.1 What does equivalent simulation mean in terms of performance? 31

III.C.2 Which criteria give answers most similar to expert judgement? 34

III.C.3 What does a Good / Average / Poor simulation mean for the expert? 40

Conclusion 44

Appendices 46

Appendix A 46

Appendix B 49

References 51

Internet Sources 54

Summary 1

Introduction

This thesis is a summary and an analysis of the final-year internship I did at Cemagref (Centre national du Machinisme Agricole, du Génie Rural, des Eaux et des Forêts), from February 1st to September 30th 2011.

Created in 1981, Cemagref is a Public Scientific and Technical Research Institute directed by Roger Genet and under the authority of the French Ministries of Agriculture and Research. In a context of growing awareness to environmental issues and as a response to current environmental challenges, its research missions focus on water, ecotechnologies and territories at the French territory scale. In order to do so, Cemagref works in close cooperation with public institutions, other research institutes and private companies. It hires about 1600 persons including 950 researchers spread in 25 research units on 10 regional sites.

The Hydro team I worked in is part of the Hydrosystems and Bioprocesses (HBAN) research unit based in Antony. Vazken Andréassian is the Hydro team leader. The team's research focuses on developing hydrological models. Research themes include flood forecasting, ensemble forecasting and applications on ungauged basins. The team collaborates with EDF, BRL and Canadian institutes among others. Their research results can be applied to water resources management, design and management of water control facilities (e.g. dams), and the anticipation of flood or low flow periods. During my internship, the team was composed of 7 permanent researchers and technicians, 3 temporary contracts, 7 PhD students and 4 MSc students.

The internship I did under the supervision of Charles Perrin came within the scope of the research on evaluation criteria for hydrological models. The topic was to study the relation between numerical evaluation criteria and visual evaluation by experts; an issue that received limited attention until now. My work was divided into two main steps. The first step consisted in designing a survey to collect expert judgements on a certain number of hydrographs. The second step consisted in calculating corresponding numerical criteria. Last, results from the two approaches were compared and questions on how the two types of evaluation criteria relate to one another were answered.

The work for the internship was organized in the scope of the XXVth IUGG Assembly held in Melbourne in July this year. Indeed, part of the results were presented by Charles Perrin on the 3rd and 4th of July 2011 during the workshop entitled “Expert judgement versus statistical goodness-of-fit for hydrological model evaluation”.

In this thesis, I will first introduce the context of the study and provide a literature review on the topic. The two technical steps i.e. the design of the survey and the computation of criteria will then be described. Finally, results of the study will be presented and questions on the two evaluation criteria will be analyzed and discussed.

I. Background of the study

In this section, the reasons for carrying out this study are presented. After a general introduction on hydrological modelling, the literature on model evaluation is reviewed. Then the project and its different steps are detailed.

I.A Introduction to hydrological modelling

I.A.1 From the catchment to hydrological modelling

The **catchment** or river basin is the geographical scale commonly adopted to manage water. A catchment is a geographical system in which precipitations will contribute to the flow in a same and single outlet. It is a suitable unit for studying rainfall-to-runoff processes (cf Figure I-1). At this scale, water can be considered as a resource (for water supply, irrigation, etc.), or as a risk (e.g. protection against floods). In both cases, one needs to **estimate the volumes and temporal distributions of water that flows in the stream**. Although streamflow can be measured, its estimation by mathematical means is often required for management and scenario testing. This is a major objective of quantitative hydrological modelling.

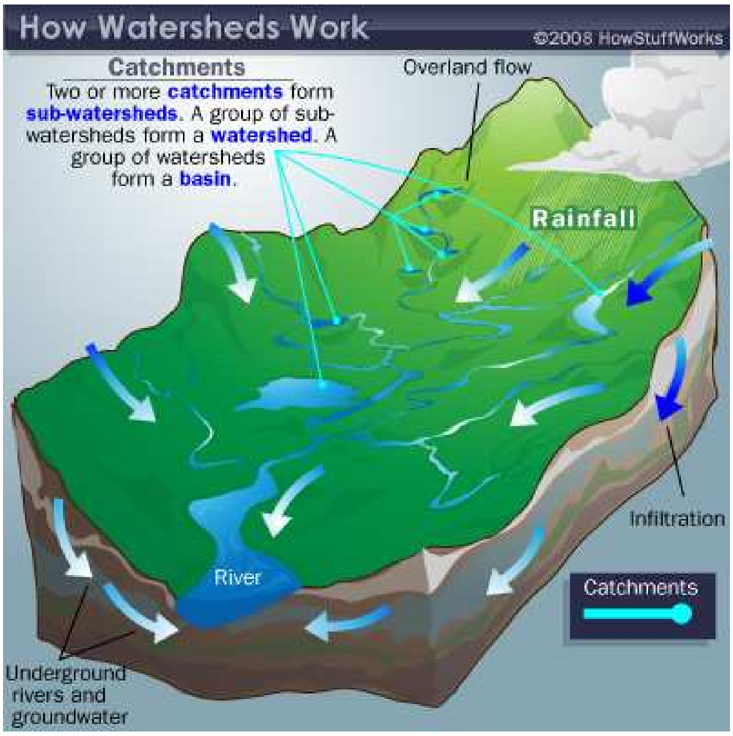

How Watersheds Work
©2008 HowStuffWorks
Catchments
Two or more catchments form sub-watersheds. A group of sub-watersheds form a watershed. A group of watersheds form a basin.
Overland flow
Rainfall
Infiltration
River
Catchments
Underground rivers and groundwater

Figure I-1: Schema of catchments in a watershed and internal water flux [d]

A **hydrological model** is a simplified mathematical representation of these processes involved in precipitation-runoff transformation at the catchment scale. Starting from precipitations that are the main drivers of flow, a hydrological model simulates the runoff at the catchment outlet. An example is given in Figure I-2.

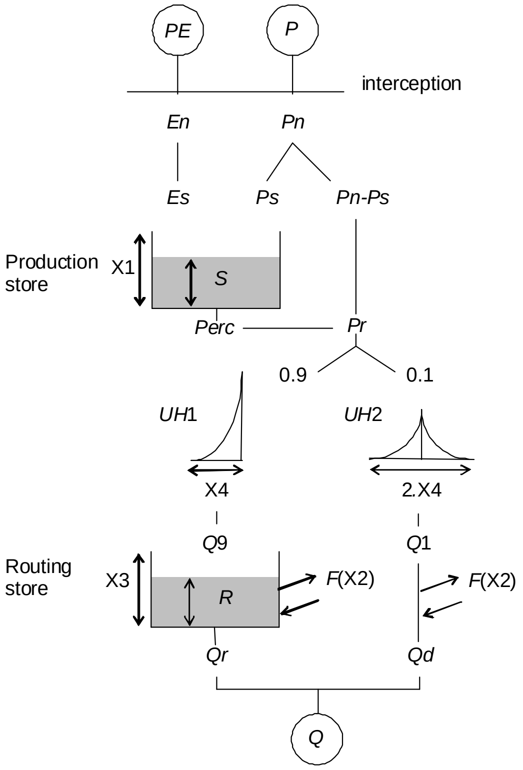


Figure I-2: Schema of the GR4J model (Perrin et al., 2003)

Building and implementing a hydrological model is a complex process. In the following, we present the hydrological modelling process using the five steps described by Refsgaard and Henriksen (2004) and Scholten et al (2007).

I.A.2 Step 1: Model Study Plan

A hydrological model is designed to **answer a specific water-related question**. For example, it can be built to predict in high-flow or low-flow periods. Therefore, modelling objectives may strongly influence the way the model is developed. The relevance of the model may also depend on climatic and physical conditions of applications. Last, the development of the hydrological models may be limited by external factors such as the availability of recorded data and the observation of the physical processes at stake.

These aspects must be accounted for at the very beginning of the modelling process, as they may play a key role on model efficiency.

I.A.3 Step 2: Data and conceptualisation

A hydrological model transforms input meteorological variables (mainly precipitations and potential evapotranspiration) into an output hydrological variable (mainly flow) over a time period (see Figure I-3). A hydrological model is made of **mathematical representations of the key processes** like evapotranspiration, infiltration and transfer in streams.

The model technically consists in a **set of hydrological parameters describing the catchment properties,** and **algorithms describing the physical processes**. There are many ways to build hydrological models. As mentioned above, the choice of the modelling approach (or of an existing model) may depend on the objectives, available data and user experience. In this modelling step, the way the natural system should be represented is defined.

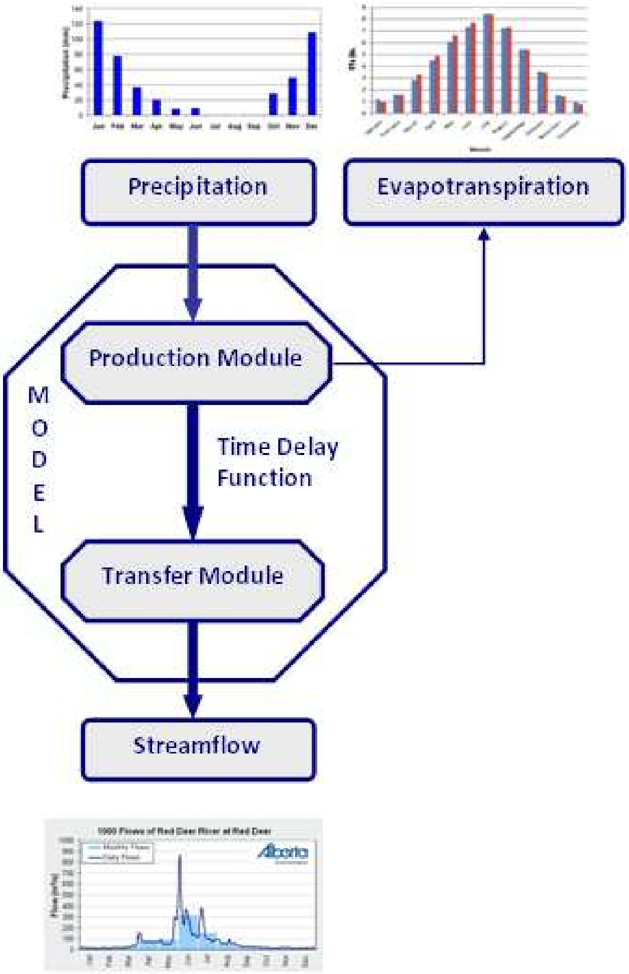


Figure I-3: Example structure of a hydrological model ([a], [b] and [f] modified)

I.A.4 Step 3: Model set-up

The structure of the model is then implemented. The typical structure of a hydrological model is composed of two parts (see Figure I-3):

- A **production module** assesses the portion of precipitation that feeds runoff at the catchment outlet (effective rainfall). The remaining part either is stored, infiltrates to deep aquifers or returns to the atmosphere through evapotranspiration,
- A **transfer module** routes effective rainfall to the catchment outlet, depending on the water pathways (surface, sub-surface or groundwater flow).

I.A.5 Step 4: Calibration and validation

The values for the parameters included in process equations cannot always be determined from field measurements. Therefore numerical optimization is often necessary to **obtain a set of parameters suitable for the studied catchment**. Optimization requires observed output data on a given time period.

A variety of optimization approaches were developed. The simplest ones start from an initial set of parameters that is iteratively changed to improve the quality of an objective function (a criterion defining the quality of the fit between observed and simulated flow values) until an optimum value is reached.

The calibration step should be systematically associated with a validation test, in which one assesses the results of the model on an independent data set.

It should be noted that uncertainty always remains due to errors in models structure, parameters, data, etc. This uncertainty should be properly assessed for a more informed model application.

I.A.6 Step 5: Simulation and evaluation

Once the model has been set up, calibrated and validated, it can be **run for the target application**. Examples of applications requiring hydrological models include the design and management of water control facilities such as dams, and the anticipation of flood or low flow periods.

The uncertainty in results has to be analyzed in order to assess the reliability of the simulated variable and support the decision making process.

Finally, the **results of the model have to be evaluated**, either by visually comparing observed and simulated plots (cf. Figure I-4, Figure I-5 and Figure I-6) or by applying a mathematical criterion to calculate the distance between observed and simulated data. Visual and mathematical evaluations may differ and generate different diagnosis on model quality and efficiency. The present study focuses on comparing these two evaluation approaches.

I.B Evaluation criteria

I.B.1 A need for evaluation and consensus

Since the early stages of hydrological modelling, there was a **need to evaluate the results of models and to quantify their efficiency for flow prediction**. In their early proposals of conceptual models, Linsley and Crawford (1960) and Dawdy and O'Donnell (1965) already quantified the residuals of their models, simply by plotting observed and simulated hydrographs or by calculating the percent difference between observed and simulated flows. At that time, computation times were an actual constraint and probably limited the calculation of various evaluation criteria. However, the question of how to evaluate models was rapidly identified as a key issue and Nash and Sutcliffe (1970) were among the first to propose an efficiency index for evaluating hydrological simulations. Their aim was to provide an **objective mean for giving a mark to a simulation**. Retrospectively, this proved to be a very good try as their index remains the most widely used in hydrological modelling despite its identified weaknesses (Gupta et al., 2009).

Since then, a **large variety of evaluation criteria and tools** were introduced in the literature, corresponding to various modelling objectives or target variables. This led researchers from various environmental disciplines to recommend the introduction of conventions and common references in the evaluation of models (Rykiel, 1996; Seibert, 2001; Perrin et al., 2006; Moriasi et al., 2007). As an answer to this need for a consensus in model evaluation, several articles have tried to **establish guidelines**:

- by proposing systematic testing schemes, (Klemes, 1986; Refsgaard and Henriksen, 2004; Bennett et al., 2010)
- by recommending specific or sets of criteria (Willmott, 1984; ASCE, 1993; Moriasi et al., 2007)
- and by proposing ranges of criteria values that correspond to simulations considered acceptable (Chiew and McMahon, 1993; Moriasi et al., 2007; Refsgaard et al., 2010).

Despite these efforts, model evaluation remains a quite ad hoc process and is strongly related to the modelling objectives, as pointed out by several authors (Mayer and Butler, 1993; Perrin et al., 2006). This makes the **results of various existing studies most often very difficult to compare**, due to the large panel of existing criteria even when the modelling objectives are similar.

I.B.2 Visual criteria

The most straightforward possibility to evaluate models is to use **graphical means** and **compare observed and simulated values**. This evaluation method is often considered approximate or **qualitative** since model fit is evaluated by eye.

Typical graphical representations of results from hydrological models include plotting:

- observed and simulated flow hydrographs over time (see example in Figure I-4),
- simulated flows against the observed flow (Q-Q plots, Figure I-5),
- the cumulative distribution function of observed and simulated flows (known as flow duration curves, see Figure I-6).

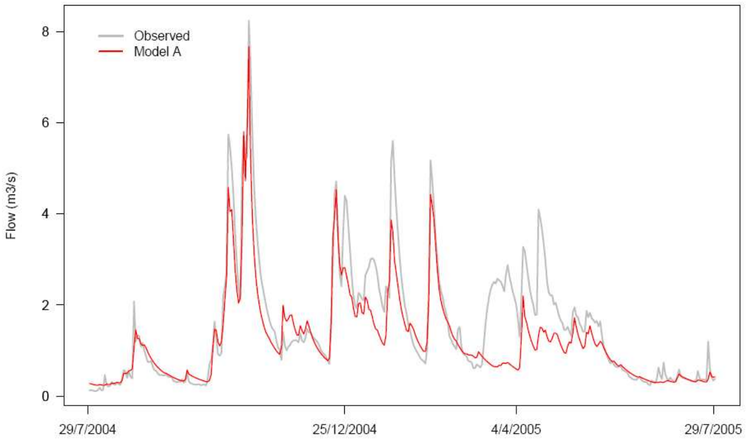


Figure I-4: Observed and simulated daily flow hydrograph over a 1-year period

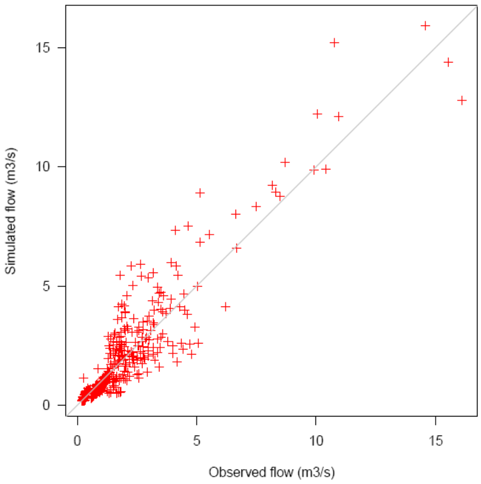


Figure I-5: Q-Q Plot

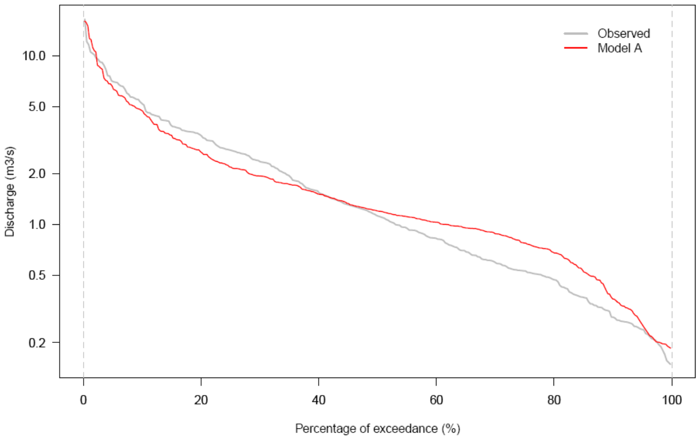


Figure I-6: Observed and simulated flow duration curve

Visual inspection benefits from the **expert comprehension, interpretation and experience**, which can be of real help in finely judging of model accuracy. As an example, Bennett et al. (2010) listed questions that experts should raise when visually evaluating hydrographs and that fully rely on the expert comprehension.

Nevertheless, this advantage can also be seen as a drawback as the dependence on the expert's own experience and references may make the evaluation subjective and quite different from one person to another (Houghton-Carr, 1999). Moreover, **depending on the kind of graphs that is used for the evaluation, the characteristics of model fit that are assessed will not be the same** (Chiew and McMahon, 1993).

I.B.3 Mathematical criteria

Another possibility for evaluating and comparing models is to compute a mathematical criterion which **evaluates a distance between the measured and simulated flow values** over a chosen time period. It usually takes the form of a norm that quantifies the distance between the observed and simulated series. These criteria are often considered **more objective** and are described as **quantitative**.

Depending on its form and calculation characteristics, a mathematical criterion will evaluate a certain aspect of the set of analyzed values (Jachner et al., 2007). It can be, among others:

- **absolute** (i.e. non-relative) when it consists in a dimensional difference between the observed and simulated values, (e.g. $MAE = \frac{1}{n}\sum_{i=1}^{n}\left|O_i - S_i\right|$)

- **relative to a benchmark or dimensionless** when it normalizes the difference between observed and simulated values by a standard quantity (Perrin et al., 2006), (e.g. $RAE = \frac{\sum_{i=1}^{n} |O_i - S_i|}{\sum_{i=1}^{n} |O_i - \overline{O}|}$, from Dawson et al. (2007))
- focusing on **timing errors** or **amplitude errors**, or combining both,
- calculated on a **continuous series** of flows or on **specific events** or periods (low flows, high flows, peaks, recessions, etc.),
- accounting for the **complexity of the model** by a dependence on the number of parameters or the number of values used for calibration (Dawson et al., 2007; Clarke, 2008b).

Because of these different characteristics, mathematical criteria have to be applied very cautiously as they **do not evaluate the same types nor ranges of values** (Krause et al., 2005). For example, a criterion effective to evaluate models on low flow periods will probably not be as suitable for evaluating floods simulated by the same model. Besides, in spite of their apparent simplicity, the behaviour of numerical criteria is still difficult to fully understand and can hide unexpected properties (Gupta et al., 2009; Berthet et al., 2010a; Berthet et al., 2010b).

As an answer to the variety of existing criteria and to obtain more comprehensive model evaluations, many authors (Chiew and McMahon, 1993; Krause et al., 2005; Dawson et al., 2007; Moriasi et al., 2007; Reusser et al., 2009) advised choosing a combination of criteria depending on the type of model application.

I.B.4 Comparing expert judgement and mathematical techniques

The existence of qualitative and quantitative criteria may put the hydrologist in a tricky situation, with potentially disagreeing diagnostics on model efficiency depending on the type of criteria used.

Many studies focused on the relevance of qualitative and quantitative evaluation criteria for hydrological modelling, but **only a few focused on comparing the two**. More generally, expert judgement in an evaluation context is little studied, whether it be for calibration (Boyle et al., 2000; Boyle et al., 2001) or evaluation of environmental models.

Nevertheless, visual inspection is acknowledged to be a full-fledged evaluation technique as essential as evaluation by mathematical criteria. Researchers have advised model users to complementarily use the two techniques (Chiew and McMahon, 1993; Mayer and Butler, 1993; Houghton-Carr, 1999; Moriasi et al., 2007; Bennett et al., 2010).

A few studies focused on relating visual and mathematical evaluation techniques:

- **Chiew and McMahon (1993)** who led a survey on 63 hydrologists asking them to select their preferences in visual and mathematical criteria when evaluating a model and asking them to assess the quality of 112 simulations on 28 catchments. From this survey, Chiew and McMahon extracted values for the Nash-Sutcliffe coefficient efficiency and for the coefficient of determination considered acceptable for the evaluation of hydrographs.
- Later, **Houghton-Carr (1999)** compared 2 expert judgements on 3 qualitative criteria with results from 10 quantitative criteria and evidenced the discrepancies of the two methods when ranking hydrological models. No universal criterion, quantitative or qualitative, could be selected from the studied set and precautions on choosing criteria adapted to the objective were suggested.
- **Olsson et al. (2011)** presented on a survey on 13 experts from SMHI who were asked to rank models on a scale ranging from 1 to 5. From this survey, Olsson plotted the visual evaluation score against the Nash-Sutcliffe criterion value for each graph, and could evidence values of the Nash-Sutcliffe criterion corresponding to models considered as acceptable and good.

More recently, a few studies have tried to combine the two approaches by **developing criteria reproducing the process of visual assessment**. Ehret and Zehe (2010) created a multi-step criterion composed of a threat score evaluating the correspondence between observed and simulated events, an error in amplitude and an error in timing. The process consists in only considering hydrological events over a chosen threshold, matching corresponding events from simulated and observed hydrographs, and comparing those corresponding events in time and amplitude by applying a regular mathematical criterion on the rising and recession limbs separately. Through this process, Ehret and Zehe tried to reproduce the trajectory and progression of the eye when comparing hydrographs. Ewen (2011) tried to traduce this same progression by means of elastic bends, the idea being to calculate the necessary work to fit the simulated hydrograph to the observed hydrograph.

I.C Description of the study

I.C.1 The study

The goal of this study is to **further investigate the relation between quantitative and qualitative criteria**.

In order to do so, a number of hydrographs was chosen and both quantitative and qualitative criteria were obtained for this set of hydrographs:

- A **web-based survey** was designed to collect as many expert judgements as possible. The survey displayed hydrographs and asked for the experts to evaluate the hydrographs both relatively (comparison of hydrographs) and absolutely (use of adjectives to describe the quality of the hydrographs).
- A large set of **mathematical criteria** based on a large literature review were computed on the sets of values.

In this respect, we wished for the conclusions of this study to be as general as possible.

I.C.2 Organization of the work

The organization of the tasks is illustrated in Figure I-7.

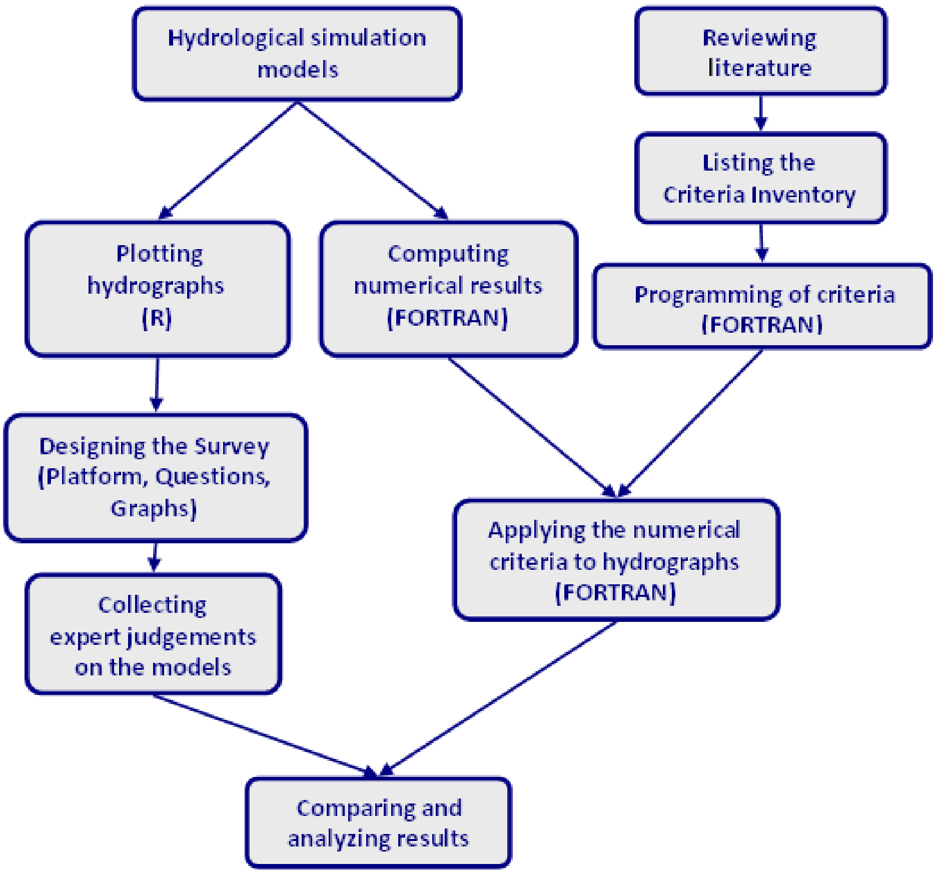

Hydrological simulation models
Reviewing literature
Listing the Criteria Inventory
Plotting hydrographs (R)
Computing numerical results (FORTRAN)
Programming of criteria (FORTRAN)
Designing the Survey (Platform, Questions, Graphs)
Applying the numerical criteria to hydrographs (FORTRAN)
Collecting expert judgements on the models
Comparing and analyzing results

Figure I-7: Graph of the tasks for the study

The programming tools used in order to complete the work include:

- FORTRAN: for the computation of the numerical validation criteria,
- R: for drawing hydrographs easy to read for the survey,
- SurveyGizmo: for designing and hosting the survey website.

In our attempt to further characterize the relation between quantitative and qualitative evaluation criteria of hydrological models, we have seen that two steps were necessary: the creation of a survey to collect expert judgements and the computation of numerical criteria. These two steps constituted most of my work at Cemagref and will be further detailed in next chapter.

II. Methodology

II.A Design of the survey on expert judgement

II.A.1 Objectives, process and interface

The whole survey was designed with the objective to make it **easy and quick to answer**.

Several trial versions of the survey were internally evaluated and tested. At each level, modifications and improvements on both the ergonomics and the hydrological questions raised by the survey were brought.

Technically speaking, the survey was designed using an online survey software named Survey Gizmo [e]. It was accessible **from June 15 2011 until August 15 2011** at the following URL: http://edu.surveygizmo.com/s3/561372/survey. The survey consisted in:

- A welcome page
- An explanation on the context of the study
- Instructions on how to answer the survey
- 20 low-flow hydrographs to judge
- 20 high-flow hydrographs to judge
- An information page including:
    - questions on how people visually evaluated the hydrographs
    - personal information on survey takers (work experience, background in hydrology and model using, name, country)
- A thank you page

The survey was submitted to a large panel of hydrologists **prior and during the IAHS conference in Melbourne** that took place on the 3rd and 4th of July 2011.

II.A.2 The graphs presented

The survey consisted in **comparing and ranking models by the visual evaluation of hydrographs**. The result graphs were chosen in order to make this study relevant and were drawn using R software.

We decided to focus on **hydrographs**, which are often used when visually evaluating models. An advantage of hydrographs is that they provide a global view of the observed and simulated flows over a chosen time period as well as a local view of peaks. Therefore, the user is able to dissociate local or peak errors from general or volume errors when cumulative graphs do not allow such error dissociation. Additionally, time differences and amplitude differences can be assessed separately more easily on hydrographs than on any other type of graphical representation (Chiew and McMahon, 1993).

Characteristics used to choose the displayed hydrographs

The questions that were raised during the **choice of hydrographs** included:

- the type of result graphs that should be presented for evaluation (hydrograph, FDC),
- time periods on which models should be assessed,
- the number and type of models evaluated during the survey,
- catchment behaviours that should or should not be included.

Here we used hydrographs produced by **existing models**. These models were chosen with the following characteristics:

- **deterministic**: the input variables and model states will entirely determine the output variable. The evaluation of probabilistic modelling approaches (in which not only a single output value but a distribution of values is provided) is out of the scope of this study;
- applied in **simulation mode**: the model is applied using precipitation and potential evapotranspiration data over a period of time to assess the resulting flow over the same period of time. We do not consider forecasting application mode in which one tries to anticipate future conditions;
- with simple lumped **conceptual structures**: the models used represent the rainfall-runoff transformation in a simplified way by a series of storages. The catchment is considered as a lumped unit.

Description of the 40 displayed hydrographs

The graphs displayed in the survey are **daily flow** hydrographs. (cf. Figure II-2)

The time periods were chosen to be **1- to 6-month long** so that peaks and dry spells would appear clearly therefore making the evaluation easier.

The hydrographs were computed for **29 catchments** (see Figure II-1) spread all over France. Thus, different catchment dynamics due to various climates and physical conditions are presented in the survey.

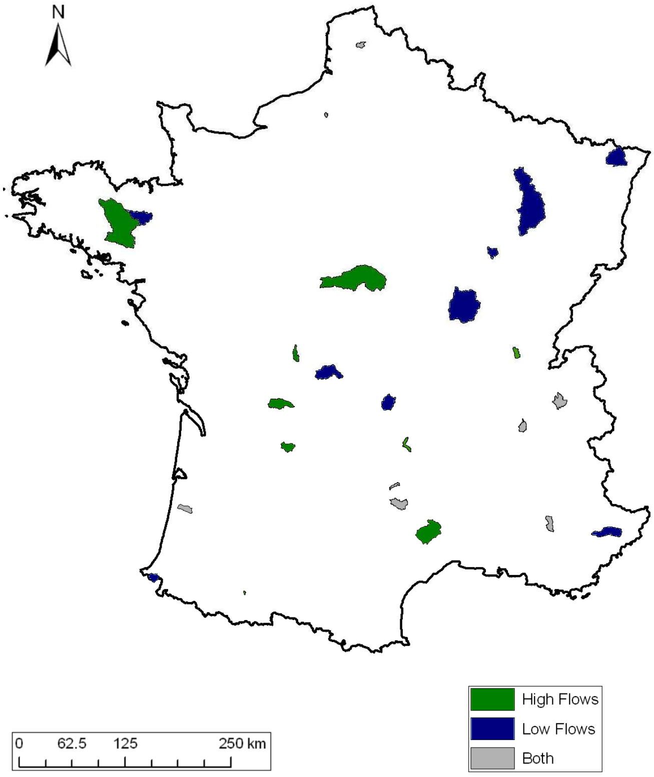


Figure II-1: Repartition map of the 29 catchments chosen for hydrograph results

Flows simulated by **five different models** were used in order to generate different hydrographs behaviours and shapes for the experts to compare (see Table 1). In addition, five versions of the GR4J model were generated to obtain simulations better on low flows and simulations with various timing errors.

Table 1: Description of the models used to generate the hydrographs (all models are modified versions of original models)

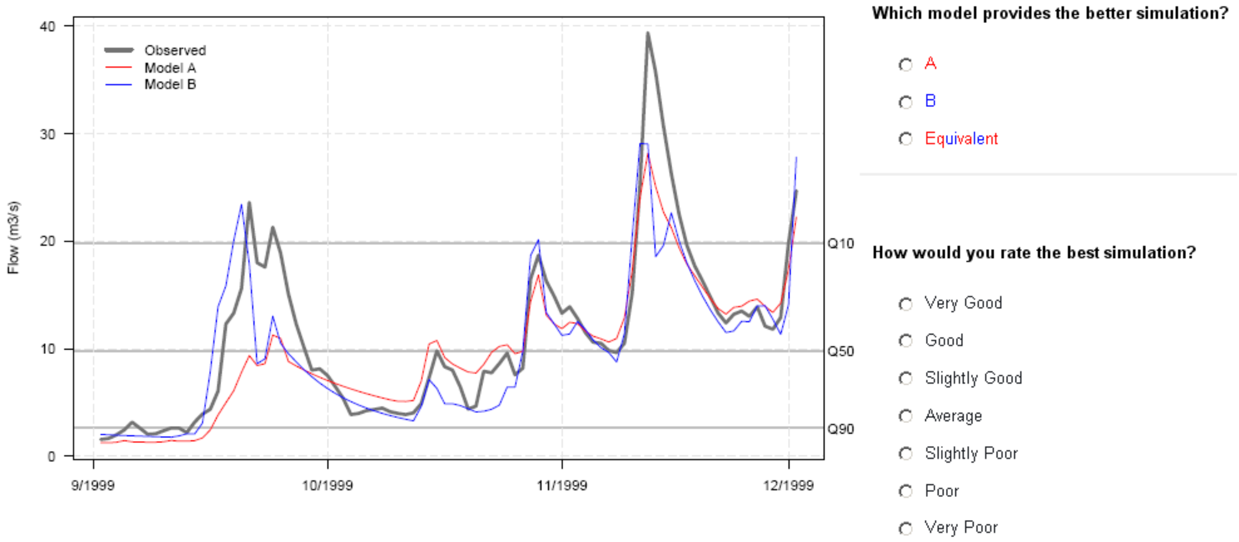
Models	Type	Time step	Number of free parameters	Reference of original version
GR4J	Empirical	Daily	4	Perrin et al. (2003)
Mordor	Conceptual	Daily	6	Garçon (1996)
Sacramento	Conceptual	Daily	13	Burnash (1995)
SMAR*	Conceptual	Daily	9	O'Connell et al. (1970)
Topmodel	Conceptual	Daily	8	Beven and Kirkby (1979)

*SMAR: Soil Moisture Accounting and Routing

II.A.3 The questions to answer

Two questions are asked on each hydrograph (cf. Figure II-2):

- **Question 1** asks the expert to pick the better of two models, therefore asking for a **relative evaluation** of hydrographs.
- **Question 2** is more **absolute** and asks for a rating evaluation of the better model on a 7-level scale.



Which model provides the better simulation?

○ A

○ B

○ Equivalent

How would you rate the best simulation?

○ Very Good

○ Good

○ Slightly Good

○ Average

○ Slightly Poor

○ Poor

○ Very Poor

Figure II-2: Example of a displayed hydrograph and related questions

II.B Computation of mathematical criteria

II.B.1 Listing and computing of the mathematical criteria

A list of the 60 computed mathematical criteria and the abbreviations used in this thesis can be found in Appendix A.

A **literature review** on numerical criteria used in hydrology and other domains such as ecology, environment and physical geography was done.

Two existing lists of criteria were used as a starting point:

- the list established by **Smith et al. (2004)** to compare hydrological models in the Distributed Model Intercomparison Project;
- the list proposed by **Dawson et al. (2007; 2010)**to create the HydroTest web site which automatically calculates numerical criteria values on flow simulations.

Hence a list of 60 criteria was established. The computation of these criteria consisted in programming them in a **FORTRAN routine designed to calculate numerical criteria on streamflow data**.

Finally, criteria values obtained for the selected hydrographs from the FORTRAN routine were calculated in parallel through Excel for verification.

II.B.2 Diversity of the criteria and gathering

The criteria thus computed present a **wide range of characteristics** described in I.B.3.

Several attempts at creating **classes of criteria** were made:

- a **manual classification** depending on the criteria mathematical properties and types of errors used: absolute, mean, squared, squared relative, etc.
- **hierarchical clusters** using R computed for High Flows and Low Flows (see Appendix B),

Later, from the results we were able to extract the criteria that corresponded to the same models ranking (see Table 2). From this observation, our previous classifications were validated.

Table 2: Composition of the classes of similar criteria

Name	Common Formula	Criteria
Squared Errors	$\sum (O_i - S_i)^2$	RMSE, CVR, EI, MEI, SEI, MSE, AIC, BIC
Cumulative Errors	$\sum (O_i - S_i)$	MC, B, RVE
Squared Relative Errors	$\sum_{i=1}^{n} \left(\frac{O_i - S_i}{O_i} \right)^2$	MSRE, EIrel
Absolute Errors	$\sum \lvert O_i - S_i \rvert$	MAE, RAE

The results obtained from the two evaluation methods: the expert judgement collected through the survey and the numerical values computed thanks to a FORTRAN routine were then analyzed and linked. From all the possible results, we had to choose the most relevant ones, which are presented in section III.

III. Results and discussion

III.A Statistical results on the survey

III.A.1 General facts

The survey was open from June 15th to August 15th 2011. **150 answers** were collected, including 39 during the IUGG workshop in Melbourne. Hydrologists from **20 countries** (France, Australia, Germany, Sweden, South Africa, etc.), and from over **79 institutions** (Cemagref, EDF, SMHI, Federal Institute of Hydrology, CSIRO, etc.) provided answers.

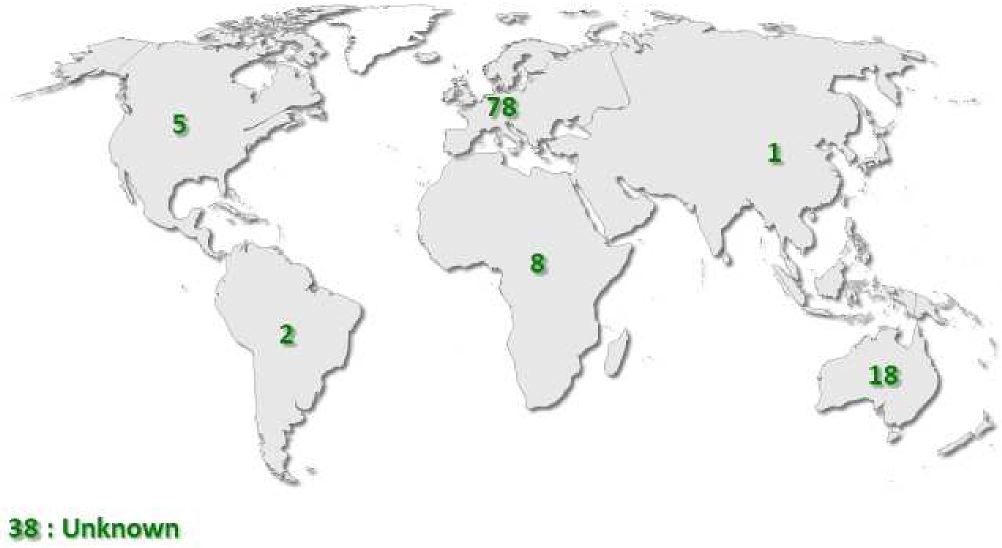


Figure III-1: Map of repartition of the experts who answered the survey [c]

III.A.2 Who answered the survey?

Pie and bar charts shown in Figure III-2 to Figure III-5 illustrate data on the set of survey takers. 68% of the experts come from the **research field** (see Figure III-2), 58% are **Senior** in their work field (see Figure III-3) and 35% have **1 to 5 years of experience in hydrological modelling** (see Figure III-4). Most of them are working as **developer and user of hydrological models** (see Figure III-5).

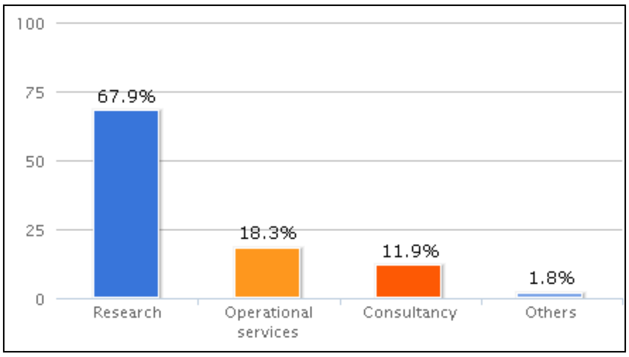


Figure III-2: Bar graph of the main sectors the experts work in [e]

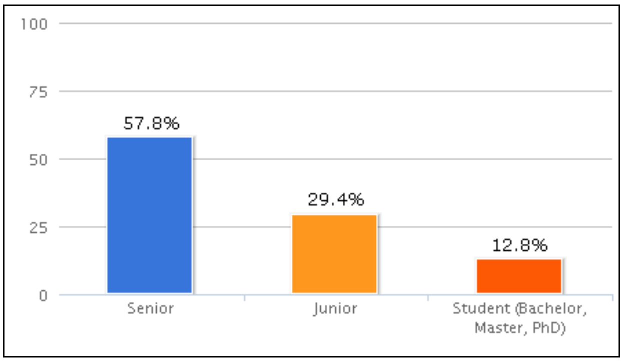


Figure III-3: Bar graph of the work status of the experts [e]

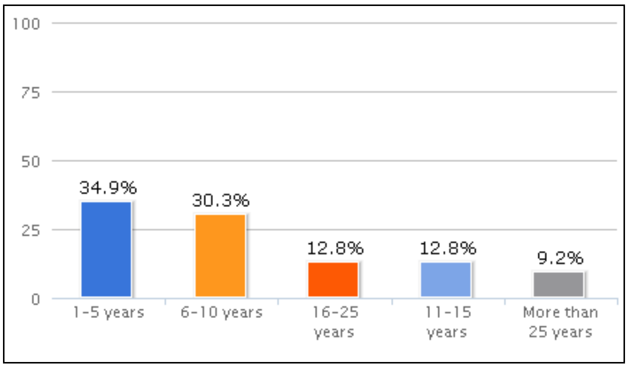


Figure III-4: Bar graph of the years of experience the experts have in hydrological modelling [e]

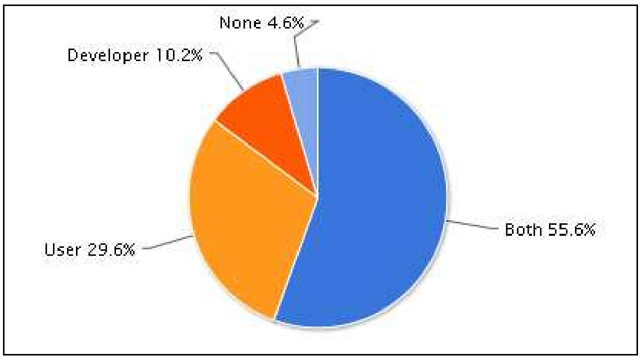


Figure III-5: Pie chart of how the experts' work relate to hydrological modelling [e]

III.B Results and discussion on the collected expert judgements

In this section, Question 1 will refer to the relative comparison question asked for each hydrograph and Question 2 will refer to the absolute rating question.

III.B.1 How lenient were the experts?

Definition of the leniency score

The leniency score is meant to represent how experts distributed their answers on the 7-level scale. For question 2, the answers are weighted as follows:

Very Poor	Poor	Slightly Poor	Average	Slightly Good	Good	Very Good
0	1	2	3	4	5	6

Given the sets of answers to question 2 for judges G, $\left(b_i^G\right)_{i\in[1,40]}$, N the number of questions considered (N=20 in high flows or low flows, N=40 in all flows) the leniency score of judge G can be expressed in percentage using:

$$D_G = \frac{\sum_{i=1}^{N}\left(b_i^G\right)}{N}$$

A leniency score of 0 means that the expert picked Very Poor for all the hydrographs. Similarly, a leniency score of N means that the expert picked Very Good for all the hydrographs. A survey taker who has equally divided his answers between the 7 choices will have a leniency score of 3.

Objective

We will analyze the leniency scores for the experts in order to see **if the answers from different experts are comparable and if experts actually used similar ranking scales**. The repartition of the experts according to their leniency score will be plotted. Also, a bar graph of the repartition of 7 answers for all the answers will be presented.

The analysis was made using answers of experts for Question 2.

Results were computed on both high and low flows.

Results

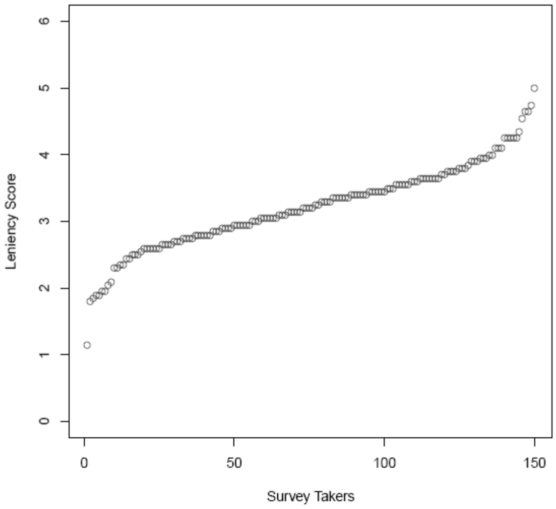


Figure III-6: Ordered leniency scores on high flows for the set of survey takers

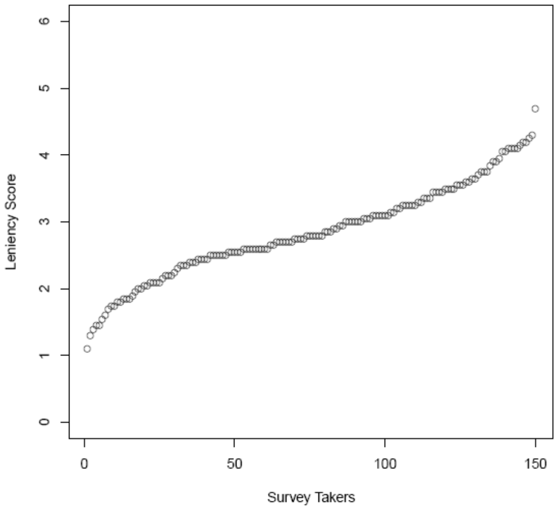


Figure III-7: Ordered leniency scores on low flows for the set of survey takers

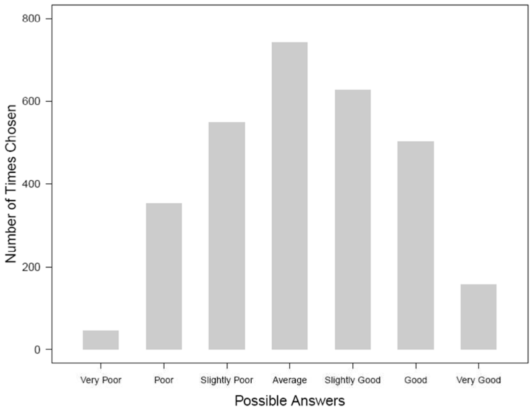


Figure III-8: Distribution of the high flow experts answers over the absolute evaluation answers

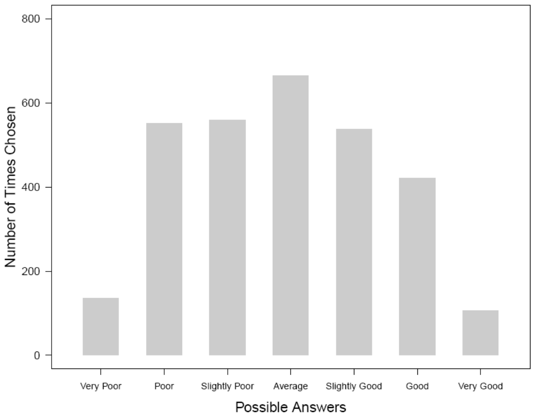


Figure III-9: Distribution of the low flow experts answers over the absolute evaluation answers

Discussion

From Figure III-6 and Figure III-7, we can conclude that the survey takers wisely used their rating scale as the mean leniency score is around 3. This means that on the overall, **survey takers centred their answers around "Average"**. We can notice the presence of a very demanding survey taker who scored much lower than any of the others, i.e. below 20%.

However, a score of 3 could also mean that survey takers picked Average for a majority of the hydrographs. From Figure III-8 and Figure III-9, we can see that this case is not to be excluded as "**Average" is the answer most often picked on the overall**.

III.B.2 Do experts evaluate alike?

Definition of the distances between two judges

The distance between two experts is meant to quantify the number of different answers between those two judges.

- For question 1, the answers are weighted as follows:

A	Equivalent	B
1	2	3

Given the sets of answers to question 1 for judges G and F, $\left(a_i^G\right)_{i\in[1,40]}$ and $\left(a_i^F\right)_{i\in[1,40]}$, the distance D_{GF}^1 between the two judges can be expressed in percentage using:

$$D_{GF}^1 = \frac{\sqrt{\sum_{i=1}^{40}\left(a_i^G - a_i^F\right)^2} *100}{4\sqrt{10}}$$

where $4\sqrt{10}$ is the maximum distance between two judges on question 1.

- For question 2, the answers are weighted as follows:

Very Poor	Poor	Slightly Poor	Average	Slightly Good	Good	Very Good
1	2	3	4	5	6	7

Given the sets of answers to question 2 for judges G and H, $\left(b_i^G\right)_{i\in[1,40]}$ and $\left(b_i^F\right)_{i\in[1,40]}$, the distance D_{GF}^2 between the two judges can be expressed in percentage using:

$$D_{GF}^2 = \frac{\sqrt{\sum_{i=1}^{40}\left(b_i^G - b_i^F\right)^2} *100}{12\sqrt{10}}$$

where $12\sqrt{10}$ is the maximum distance between two judges on question 2.

Objective

The idea is to **evidence experts who have the same way of thinking** whether it be on the absolute comparison question or on the relative rating question. We also **study if experts who are similar on the comparison question will also have similar answers for the rating question**.

The analysis is made using the answers of experts for Question 1 and Question 2.

Results are computed so that Questions 1 and 2 are considered separately.

Method

The couples of judges will be ranked from the closest (D close to 0%) to the furthest (D close to 100%), depending on D_{GH}^1 and to D_{GH}^2 separately.

For each couple of experts, a graph of the D_{GH}^{1} ranks over the D_{GH}^{2} ranks will be plotted in order to see if the two ranks are correlated.

Results

Table 3: Most similar survey takers for Question 1 and 2

Question 1 Relative Comparison Question		
ST1	**ST2**	**Similarity %**
137	130	100
73	41	80.6
130	4	
137	4	
141	46	79.1
87	41	
73	10	77.6
73	62	
77	**73**	
87	46	
87	77	
90	**73**	
110	43	
110	**73**	
123	**73**	
143	**73**	

Question 2 Absolute Rating Question		
ST1	**ST2**	**Similarity %**
137	130	94.7
16	**9**	89.8
99	36	87.6
36	16	
31	**9**	87.1
71	**9**	86.8
85	**9**	
66	43	86.6
74	14	
99	16	
112	109	
73	66	86.3
36	**9**	

Table 4: Least similar survey takers for Question 1 and 2

Question 1 Relative Comparison Question		
ST1	**ST2**	**Similarity %**
131	104	30.2
83	18	30.6
131	126	32
98	11	
131	118	32.4
131	98	33
131	45	
131	51	33.8
117	11	
101	11	

Question 2 Absolute Rating Question		
ST1	**ST2**	**Similarity %**
26	**19**	39
25	**19**	39.2
50	**19**	41.4
108	**19**	42.3
89	**19**	44.5
30	**19**	45.1
92	**19**	45.9
87	**19**	46.8
33	26	47

Table 5: Statistics on the range of similarity percentages for each question

	Maximum Value	Minimum Value	Median
Question 1	(100%) - 80.6%	30.2%	57.4%
Question 2	(94.7%) - 89.9%	39.0%	74.9%

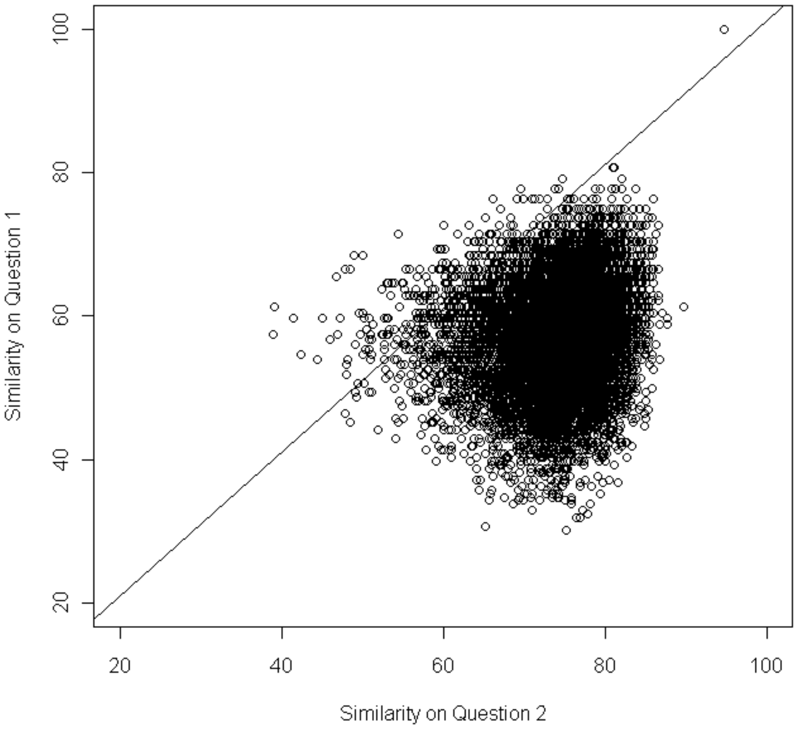


Figure III-10: Correlation graph of the similarity percentages for Question 1 against Question 2 for each couple of experts

Discussion

As indicated in green in Table 3 and Table 4, we identified **experts that can be compared to hubs as extreme results concentrate around them**, in couples including them:

- The best similarity percentages for Question 1 often are couples including Expert 73,
- The best similarity percentages for Question 2 often are couples including Expert 9,
- The worst similarity percentages for Question 1 often are couples including Expert 131,
- The worst similarity percentages for Question 2 almost always are couples including Expert 19.

From this, we can conclude that experts 9 and 73 are good go-betweens on Question 2 and 1 respectively for the set of experts. By comparing Figure III-12 to Figure III-11, we can conclude that the standing out of Expert 19 is due to his overuse of the adjective "Poor" (chosen for more than 30 hydrographs out of 40). No clear explanation on why Expert 9 stands out can be drawn from Figure III-13. We can just see that this expert often chose "Average", which is a characteristic common to many experts.

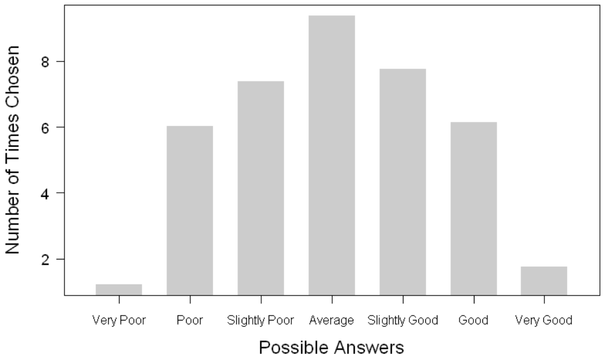


Figure III-11: Mean repartition of answers for all experts

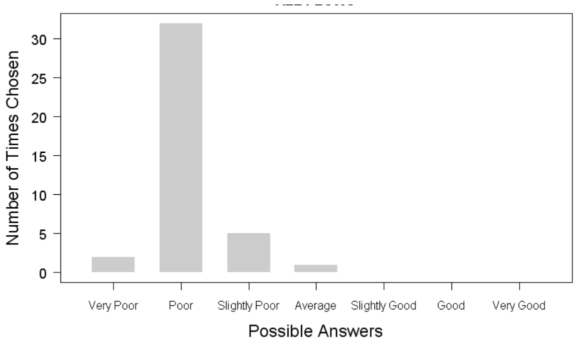


Figure III-12: Repartition of answers for Expert 19

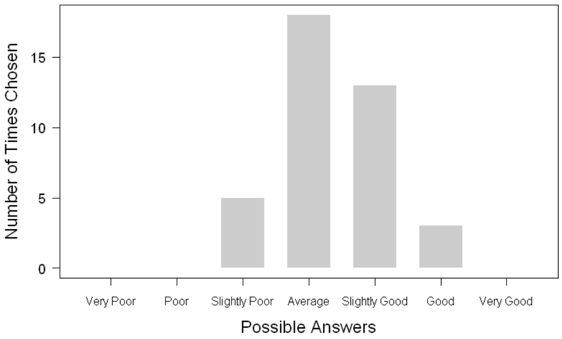


Figure III-13: Repartition of answers for Expert 9

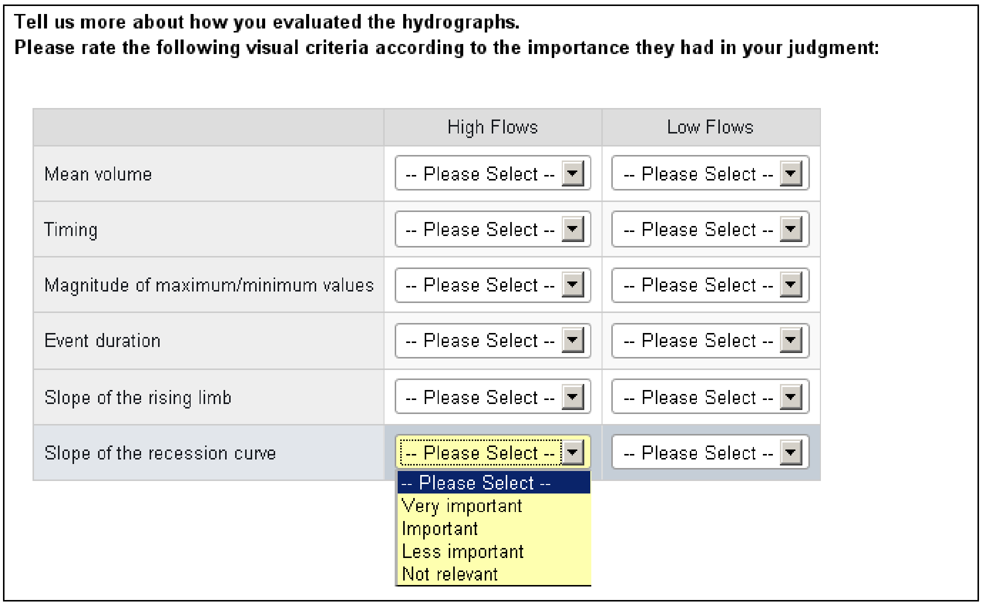
If we put aside the couple formed by experts 130 and 137 who only have a single different answer and both took the survey during the conference, the range difference observed in Table 5 indicates that **higher similarity percentages are obtained for Question 2**. An explanation is that Question 1 allows 3 answers when Question 2 allows 7 answers and that distances were normalized by the maximum possible distance. Indeed, the way distances are calculated minimizes distances between nearby answers (e.g. "Good" and "Average" for Question 2 and "A" and "Equivalent" for Question 1) in Question 2.However, from the median in Table 5, we can see that more couples achieve better similarities for Question 2 anyway.

Figure III-10 visually confirms that most of the times for a given couple of experts, similarity will be higher for Question 2 than for Question 1. However, **no general correlation between the similarity percentages obtained for the two questions can be established**.

III.B.3 How did the experts say they visually evaluated the hydrographs?

Objective

At the end of the survey, experts were asked to **assess the visual methods they used for evaluating hydrographs and the importance they gave to each of these criteria**. A screen caption of the question is presented in Figure III-14.



Tell us more about how you evaluated the hydrographs.
Please rate the following visual criteria according to the importance they had in your judgment:

	High Flows	Low Flows
Mean volume	-- Please Select --	-- Please Select --
Timing	-- Please Select --	-- Please Select --
Magnitude of maximum/minimum values	-- Please Select --	-- Please Select --
Event duration	-- Please Select --	-- Please Select --
Slope of the rising limb	-- Please Select --	-- Please Select --
Slope of the recession curve	-- Please Select --	-- Please Select --

Figure III-14: Screen caption of the visual evaluation criteria question from the survey

From these answers, we generated graphs to better understand the importance people gave to 6 different visual criteria:

- agreement in mean volume,
- agreement in timing,
- agreement in magnitude of maximum or minimum values,
- agreement in event duration,
- agreement in slope of the rising limb for events,
- agreement in slope of the recession curve for events.

Results were displayed for high and low flows separately so that importance could be set in the context of the evaluation and compared between the two cases.

Bar graphs and discussion

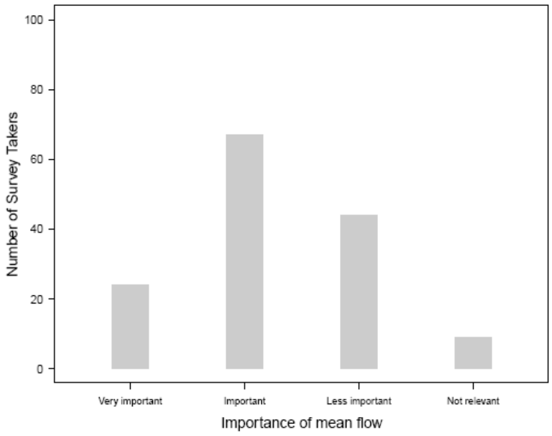


Figure III-15: Importance given by experts to agreement in mean flow for high flows

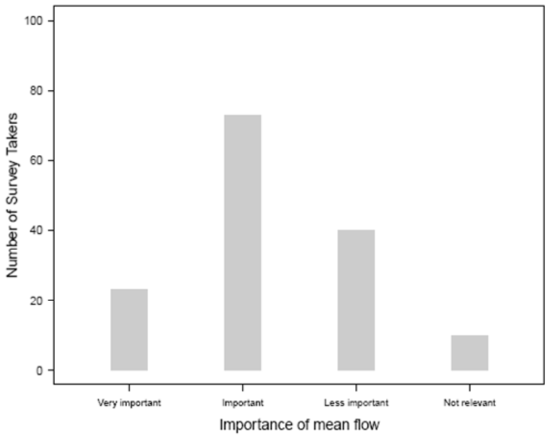


Figure III-16: Importance given by experts to agreement in mean flow for low flows

The agreement between the observed and simulated **mean flow is considered important or less important for both high and low flows** according to Figure III-15 and Figure III-16. The repartition of the votes is similar and quite independent from the flow period.

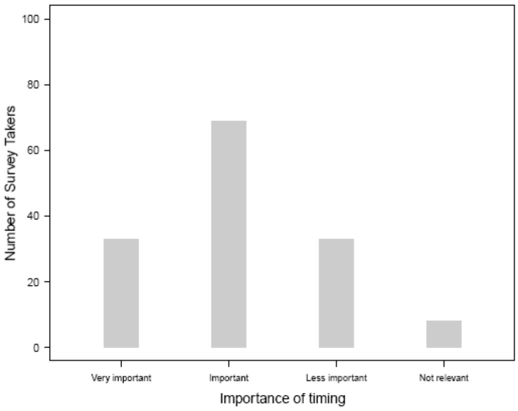


Figure III-17: Importance given by experts to agreement in timing for high flows

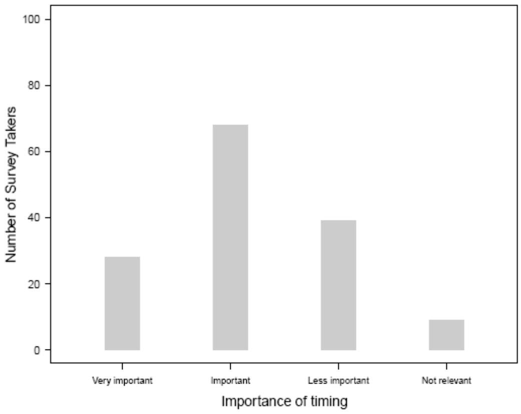


Figure III-18: Importance given by experts to agreement in timing for low flows

On Figure III-17 and Figure III-18, the agreement in **timing** between the observed and simulated hydrographs is **often considered important for both high and low flows**. The repartition of the votes is similar and quite independent from the flow period, but timing was considered **slightly less important for low flows than for high flows**, which is not so surprising.

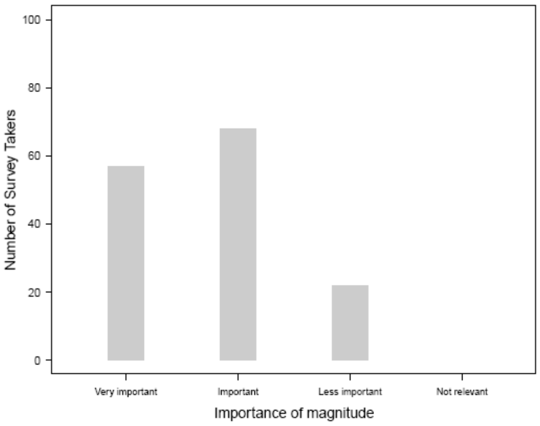


Figure III-19: Importance given by experts to agreement in magnitude for high flows

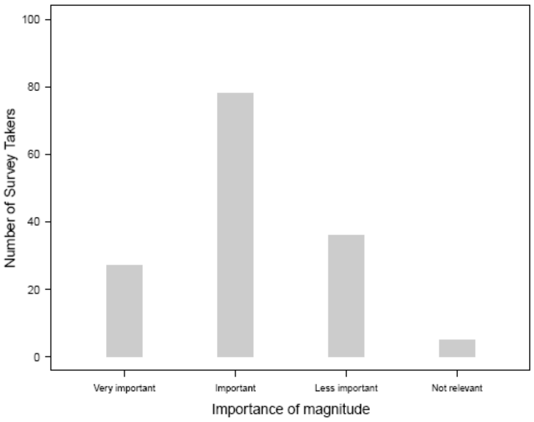


Figure III-20: Importance given by experts to agreement in magnitude for low flows

While the agreement in **magnitude** is considered **important or very important and always relevant for the evaluation of high flow hydrographs** (see Figure III-19), it was **not considered so important in low flows** (see Figure III-20). Indeed, when about 130 survey takers considered magnitude very important or important in high flows, about 105 considered it very important or important in low flows. Moreover, few survey takes even said magnitude was not relevant in low flows cases.

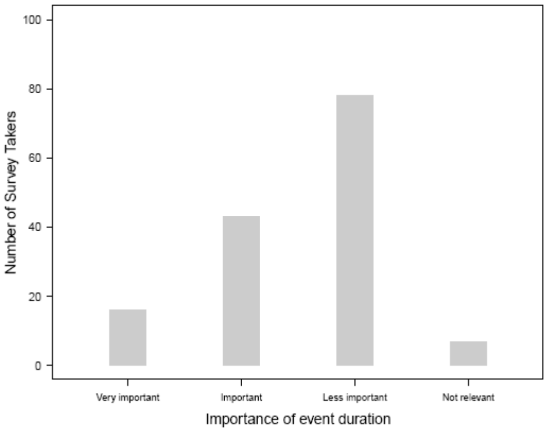


Figure III-21: Importance given by experts to agreement in event duration for high flows

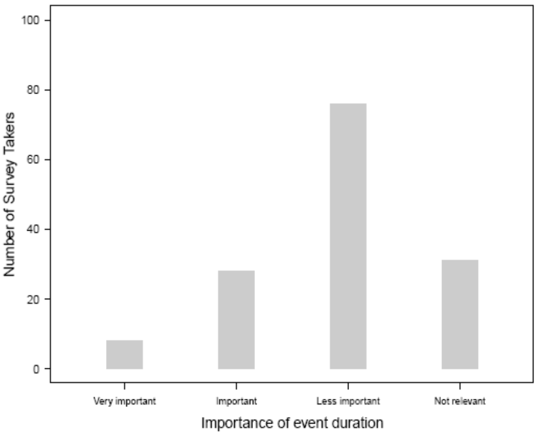


Figure III-22: Importance given by experts to agreement in event duration for low flows

From Figure III-21 and Figure III-22, we can notice that agreement in **event duration** is considered as a **slightly more relevant and important visual evaluation criterion in high flows than in low flows**. In both cases, a majority of people (about 80) qualified this criterion as **less important**.

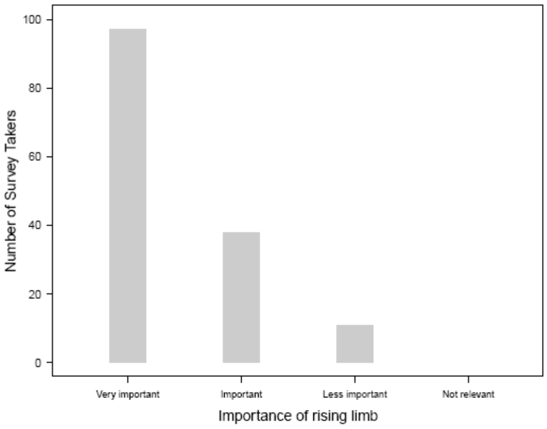


Figure III-23: Importance given by experts to agreement in rising limb for high flows

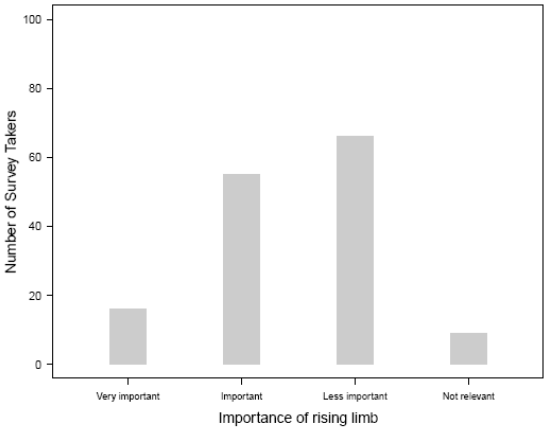


Figure III-24: Importance given by experts to agreement in rising limb for low flows

The agreement in **rising limb** is the visual criterion that has the **most radical change of importance between high and low flows**. **In high flows** (see Figure III-23), almost 100 survey takers said they considered the criterion as **very important**. Fewer people considered it important, even fewer less important. None said it was not relevant. **In low flows** (see Figure III-24), **"less important"** is the adjective mostly chosen (about 70 survey takers), and right after comes "important". Votes for "very important" and "not relevant" are not significant.

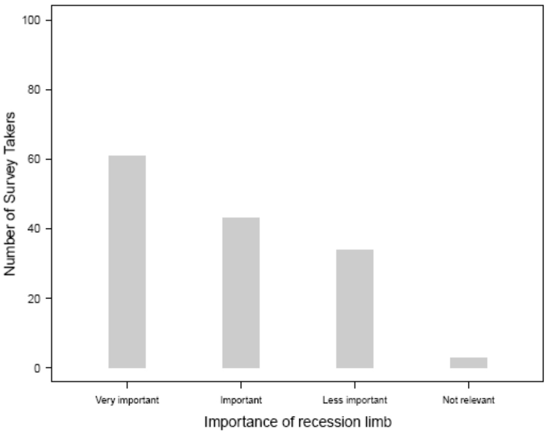


Figure III-25: Importance given by experts to agreement in recession limb for high flows

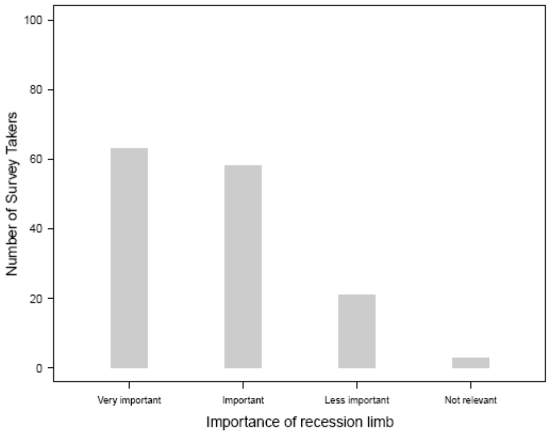


Figure III-26: Importance given by experts to agreement in recession limb for low flows

The agreement in **recession limb** is one of the few visual criteria that were considered **more important for low flows than for high flows** according to Figure III-25 and Figure III-26. Indeed, about 120 survey takers picked "very important" or "important" for low flows when 100 survey takers did so for high flows. However, in the two cases, the adjective most chosen is "very important".

To conclude, we can set **general patterns on how experts think they visually evaluate hydrographs**. In high flows, agreements in rising and recession limb will prevail, then agreements in magnitude of maximum or minimum values, mean flow and timing would be used and agreement in event duration would be considered last to rank models. In low flows, agreement in recession limb would be the most important, then agreements in magnitude of maximum or minimum values, mean flow and timing would be considered and in the end agreement sin rising limb and event duration might be used to decide between different models.

III.C Results and discussion on the comparison of evaluation methods

III.C.1 What does equivalent simulation mean in terms of performance?

Definition of equivalence

Let's consider criterion C and two simulations provided by models A and B over a time period. Corresponding criterion values are C_A and C_B respectively. A and B can be **considered equivalent if criteria values are close** enough, i.e. their ratio pertains to some interval:

$$\frac{\min(C_A;C_B)}{\max(C_A;C_B)} \in [r_n;1]$$

r_n **is a minimum ratio value under which the two criteria values can be considered significantly different**. For strictly positive criteria, a minimal ratio value of 0 (models A and B would be equivalent if the ratio pertains to $[0;1]$) would mean that models will be considered equivalent for any criteria values. Conversely, a minimal ratio value of 1 (models A and B would be equivalent if the ratio is equal to 1) would mean that only models that have strictly equal criteria values can be considered equivalent.

Objective

The objective is to find out **what the equivalence between simulations according to experts means in terms of differences in criteria values**. For each criterion, we will look for the r_n value that best matches the evaluation of experts.

The analysis was made using:

- the answers of all the experts for Question 1, i.e. their rankings of models A and B

- for each criterion, the values obtained by each model on each hydrograph, and the corresponding ranking of the criterion to Question 1.

Results are computed:

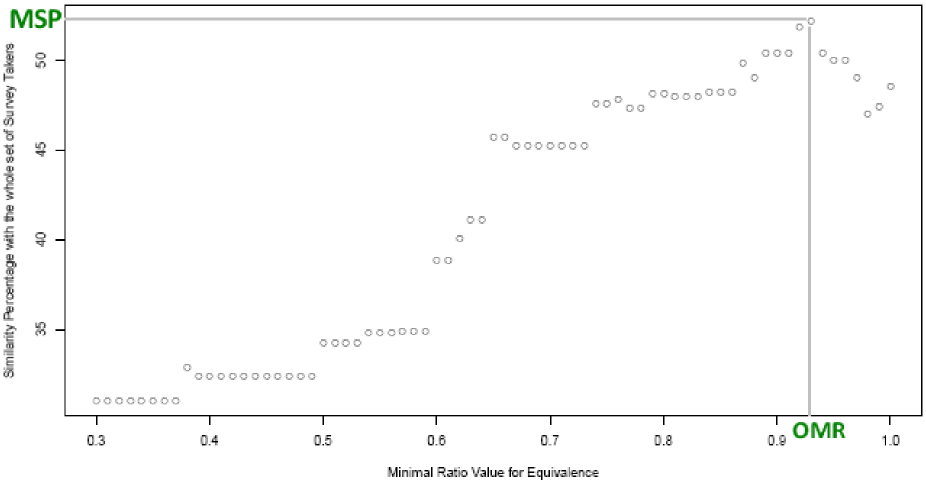
- for the group of experts, as the number of "equivalent" answers given by each expert is generally too small to obtain robust results (4 experts never ticked "equivalent");
- on the answers from high flows and low flows separately.

Method

To give an idea of the spread of criteria ratios that correspond to hydrographs judged equivalent by experts, one can plot the distribution of these ratios. However this is not very useful to determine the minimum ratio r_n we want to consider.

We determined r_n by an empirical approach, to **maximize the match between the occurrences of answers of simulation equivalence given by experts and numerical criteria**. To do this, we made the minimal ratio value vary from 0.3 to 1 by 0.01 increments. For each ratio value step, the answers to Question 1 (A better than B; B better than A; A and B equivalent) provided by numerical values of the criterion for each hydrograph might change. The similarity between this set of answers and the answers from the set of experts was then computed. This similarity was plotted against the value of the ratio to identify an optimum value.

Example of result graph



OMR: Optimal Minimal Ratio
MSP: Maximum Similarity Percentage

Figure III-27: Result graph for MAE and localization of the optimal minimal ratio value

Table 6: Optimal minimal ratio (OMR) for each criterion and the corresponding similarity percentage (MSP)

Criterion	High flows		Low flows	
	OMR	MSP	OMR	MSP
MAE	0.93	59.2	0.93	47
MC	0.96	45.1	0.93	36.1
RMSE	0.91	59.5	1	41.9
R4MS4E	0.88	56.9	0.96	39.4
Bias	0.44	47.3	0.58	40.9
RAE	0.93	59.2	0.93	47
R2	0.94	46.4	1	45.7
Rmod	0.98	50.4	0.75	32.5
CVR	0.91	59.5	1	41.9
EI	0.9	53.1	1	41.9
IA	0.97	55.2	0.99	34.7
EIFDC	0.98	55.4	1	40
EILN	0.98	48.2	1	54.9
EIR	0.95	53.7	1	46.6
EI2	0.85	53.5	0.97	39.3
MEI	0.9	53.1	1	42
SEI	0.95	54.6	1	42
PI	0.8	53.3	1	40.2
ESC	0.83	32.1	0.47	46.6
REMP	1	45.4	1	39
RBFI	0.96	44.1	1	33.8
FTEI	0.79	47.4	1	32.8
RMP	0.99	52.6	0.91	33.8
MRAF	1	42.5	0.94	41.1
RMV	0.97	54.1	1	36.3
EIHF	0.86	52.7	0.93	37.7
RLFD	0.37	33.6	0.97	45.9
RMLFV	0.84	39.1	0.98	58.2
EILF	0.72	43.3	1	53.5
MSE	0.83	59.5	1	41.9
MSLE	0.88	54.4	0.84	56.3

Criterion	High flows		Low flows	
	OMR	MSP	OMR	MSP
MSDE	0.85	47	1	36.3
AIC	0.93	53.8	1	41.9
BIC	0.92	53.8	1	41.9
MARE	0.84	55.7	0.96	60
MRE	0.33	40.3	0.6	49.7
MSRE	0.83	53.2	0.85	53.5
RVE	0.96	45.1	0.93	36.1
EIrel	1	51.4	1	50.1
IArel	0.99	49	1	47.3
EIparam	0.93	53.2	1	41.9
WR2	0.98	46.5	0.97	38.9
MSEA	0.77	46.8	0.58	41.3
MSEP	0.79	45.4	0.34	36
MSEL	0.42	43.9	1	33.9
MSES	0.85	54.9	0.34	37.4
MSEU	0.92	42.9	1	51.1
MdAPE	0.95	50.6	0.88	62
TEST	0.45	45	1	40.7
THREAT	0.93	33.2	0.96	29
ODMA	0.8	57.2	_	_
ODMT	0.6	40.2	_	_
RMVSX	0.84	50.4	1	39.6
MORPH-NSE (0,5,5)	0.98	57.1	1	39
MORPH-MAE (0,5,5)	0.94	58.9	0.99	51.4
AME	0.72	49.6	0.96	40.5
Pdiff	1	45.4	0.95	41
KGE	1	44.2	0.77	40.2
MORPH-NSE (1,5,5)	0.9	56.2	0.99	42.2
ODMA+ODMT/2	0.75	51.9	_	_

Discussion

For each criterion an optimal minimal ratio value can be identified on the computed graphs (see Figure III-27). The values are summed up in Table 6. This value corresponds to the **highest agreement between the criterion evaluation and the judges evaluations in terms of equivalence**.

When a criterion has an **OMR value equal to 1**, it means that models that have close values for this criterion and could therefore be considered "Equivalent" in regards to this criterion have not been considered "Equivalent" by experts. We could also say that the equivalence in regards to the criterion does not match the equivalence in regards to the experts. Therefore, a minimum of occurrences of "Equivalent" in the criterion answers on the 40 hydrographs is closer to the expert judgement. Indeed, OMR=1 means that only models with exactly the same criterion values will be considered "Equivalent"; this implies a set of answers with the least possible occurrences of "Equivalent".

If a criterion has an **OMR value distant from 1** (e.g. Bias, MRE, etc.), it means that models considered equivalent by this criterion correspond to models considered as equivalent by the experts. These particular criteria need their equivalence to be redefined.

We can notice more cases of OMR=1 in low flows than in high flows, with 5 cases in high flows against 27 cases in low flows. This can be due to a larger number of answers "Equivalent" in high flows than in low flows. Indeed, out of the 3000 answers in low flows (150 experts compared 20 hydrographs), 775 "Equivalent" answers appear in low flows against 953 in high flows. Therefore, **redefining the equivalence to fit the expert judgement is more necessary in high flows**.

Some OMR result graphs present plateau. This is due to the limited number of hydrographs and expert judgements. When the minimal ratio value varies, there might be ranges where no ratio needs to be redefined as equivalent. Also, even when ratios are redefined as equivalent, the hydrographs considered as equivalent might not have been identified as equivalent by the experts.

Criteria grouped in Section II.B.2 do not obtain similar OMR values. As a matter of fact, classes were formed by using the answers to Question 1 that only compares how models score and do not solicits the numeric criteria scores. Here, OMR is directly calculated from these scores.

III.C.2 Which criteria give answers most similar to expert judgement?

Definition of the similarity between a judge and a criterion

The distance between an expert and a criterion is meant to quantify the number of different answers between the criterion and the expert on the relative question.

Given J a chosen judge and C a criterion, W_J is the set of answers to Question 1 on all the high or low flow questions, and N_W the size of W_J. Similarity S_{JC} between the judge and the criterion will be defined in percentage and calculated using:

$$S_{JC} = \frac{\sum_{i=1}^{N_A} \delta_{JC,i} * 100}{N_W}$$

with $\delta_{JC,i} = \begin{cases} 1, \mathit{if} \;\; \mathit{judge} \; J \; \mathit{and} \; \mathit{criterion} \; C \; \mathit{have} \; \mathit{the} \; \mathit{same} \; \mathit{answer} \; \mathit{on} \; \mathit{question} \; i \\ 0, \mathit{if} \;\; \mathit{judge} \; J \; \mathit{and} \; \mathit{criterion} \; C \; \mathit{have} \; \mathit{different} \; \mathit{answers} \; \mathit{on} \; \mathit{question} \; i \end{cases}$

A similarity percentage of 100% means that the criterion and the expert answered identically, while a similarity percentage of 0% means that they have no common answers.

On the set of all the N experts, we define the similarity percentage of criterion S_C with the experts by:

$$S_C = \frac{\sum_{j=1}^{N} S_{C,j}}{N}$$

Objective

For each expert, the **similarity percentage with each criterion for high flows or low flows** is computed. The criteria best matching the expert judgements can be identified by the highest similarity percentages.

The **criteria best matching the whole set of experts** can also be identified by associating to each criterion the mean of its similarity percentage. The objective was to identify in various conditions the best ten out of the 60 criteria computed.

The analysis was made using the answers of experts and criteria for Question 1.

Results are computed so that:

- "Equivalent" answers of experts will not be considered,
- high and low flows are considered separately,
- each expert gets a ranking independent from the one for the group of experts.

Method

For each expert, a ranking of the criteria using S_{JC} values can be done for high flows and low flows separately. This can also be done for the set of experts by ranking the criteria using

S_c values. The same results were calculated using the Optimal Minimal Ratio calculated for each criterion in Section III.C.1 and obtained from Table 6.

Finally, for each criterion, the ordered list of similarity percentages with the survey takers was plotted.

Results

Table 7 : Global low flow ranking of criteria using OMR=1

Rank	Name	Mean Similarity Percentage
1	RMLFV - Ratio of mean low flow volumes	57. 7 %
2	MARE - Mean absolute relative error	55.7 %
3	MdAPE - Median absolute percentage error	55.6 %
4	EILN - Nash-Sutcliffe efficiency index on log-flow MSLE - Mean squared logarithmic error	54.9 %
5	EILF - Nash-Sutcliffe efficiency index on low flows	53.5 %
6	MSEU - Unsystematic error	51.1 %
7	**Squared relative errors**	50.1 %
8	MORPH-MAE (0,5,5) - Hydrograph Pattern Matching Algorithm	49.4 %
9	MRE - Mean relative error*100	48.7 %
10	IAREL - Relative deviation on the index of agreement	47.3 %

Table 8 : Global high flow ranking of criteria using OMR=1

Rank	Name	Mean Similarity Percentage
1	MORPH-NSE (0,5,5) - Hydrograph Pattern Matching Algorithm	56.8 %
2	ODMA - Overall distance of matching observed and simulated events with respect to amplitude, Part of Series Distance	56.2 %
3	MORPH-MAE – (0,5,5) - Hydrograph Pattern Matching Algorithm	55.1 %
4	MORPH-NSE (1,5,5) - Hydrograph Pattern Matching Algorithm	54.1 %
5	MSES - Overall systematic error	53.2 %
6	RMV - Ratio of mean flood volumes	53.1 %
7	EIFDC - Nash-Sutcliffe efficiency index based on flow duration curve	52. 5 %
8	**Absolute errors**	51.5 %
9	**Squared relative errors**	51.4 %
10	RMP - Ratio of mean flood peaks	50.6 %

Table 9 : Global low flow ranking of criteria using OMR from Section III.C.1

Rank	Name	Mean Similarity Percentage
1	MdAPE - Median absolute percentage error	62 %
2	MARE - Mean absolute relative error	60 %
3	RMLFV - Ratio of mean low flow volumes	58.2 %
4	MSLE - Mean squared logarithmic error	56.3 %
5	EILN - Nash-Sutcliffe efficiency index on log-flow	54.9 %
6	EILF - Nash-Sutcliffe efficiency index on low flows MSRE - Mean squared relative error	53.5 %
7	MORPH-MAE (0,5,5) - Hydrograph Pattern Matching Algorithm	51.4 %
8	MSEU - Unsystematic error	51.1 %
9	EIrel - Relative deviation on the Nash-Sutcliffe efficiency index	50.1 %
10	MRE - Mean Relative Error	49.7 %

Table 10 : Global high flow ranking of criteria using OMR from Section III.C.1

Rank	Name	Mean Similarity Percentage
1	RMSE - Root Mean Squared Error CVR - Coefficient of variation of residuals MSE - Mean Squared Error	59.5 %
2	MAE - Mean Absolute Error RAE – Relative Absolute Error	59.2 %
3	MORPH-MAE – (0,5,5) Elastic bends distances calculated with MAE, Ewen (2011)	58.9 %
4	ODMA - Overall distance of matching observed and simulated events with respect to amplitude, Part of Series Distance, Ehret and Zehe (2010)	57.2 %
5	MORPH-NSE – (0,5,5)Elastic bends distances with NSE, Ewen (2011)	57.1 %
6	R4MS4E - Fourth Root Mean Quadrupled Error	56.9 %
7	MORPH-NSE – (1,5,5)Elastic bends distances with NSE, Ewen (2011)	56.2 %
8	MARE – Mean Absolute Relative Error	55.7 %
9	EIFDC - Nash-Sutcliffe efficiency index based on flow duration curve	55.4 %
10	IA – Index of Agreement	55.2 %

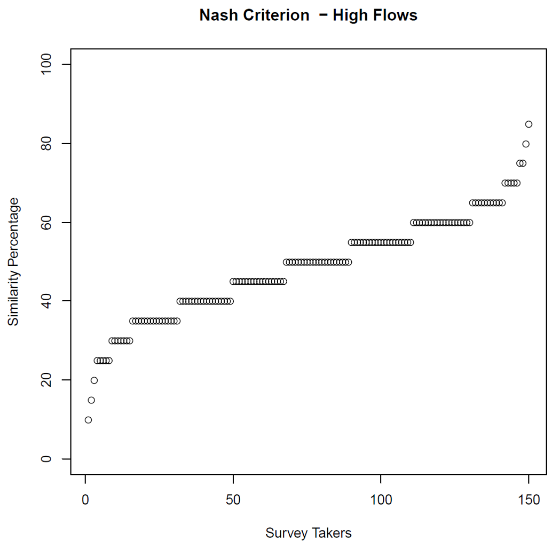


Figure III-28: Repartition of survey takers according to their similarity percentage to the Nash-Sutcliffe criterion in high flows

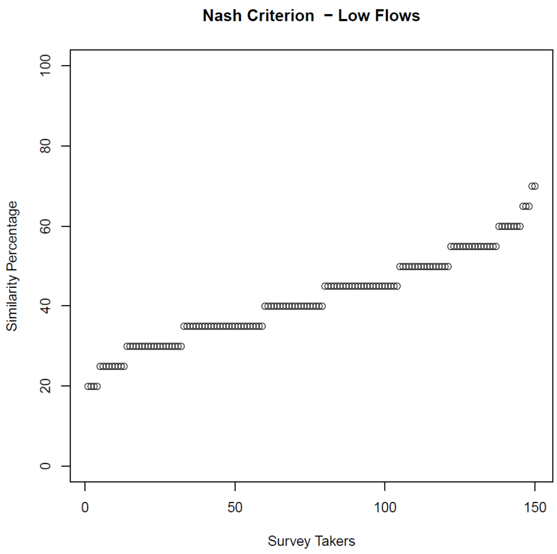


Figure III-29: Repartition of survey takers according to their similarity percentage to the Nash-Sutcliffe criterion in low flows

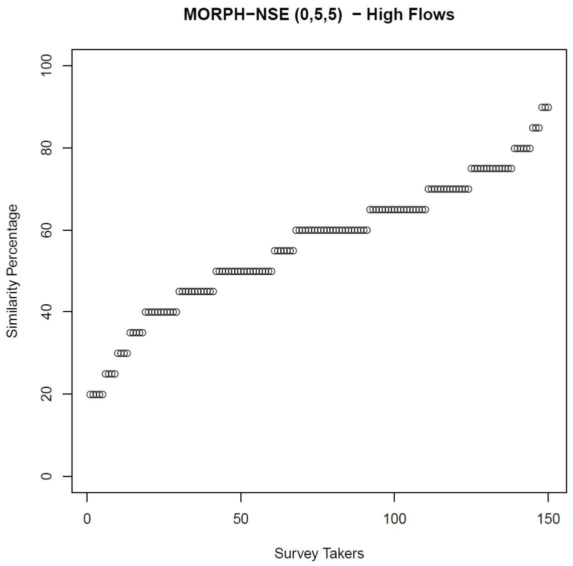


Figure III-30: Repartition of survey takers according to their similarity percentage to Morph-NSE (0,5,5) in high flows

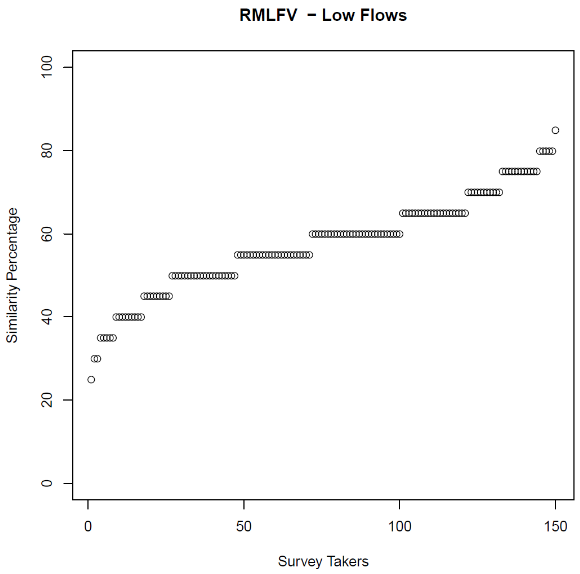


Figure III-31: Repartition of survey takers according to their similarity percentage to RMLFV in low flows

Discussion

Global Ranking without using OMR from Table 7 and Table 8

When evaluating High Flows, **numerical criteria designed to reproduce the expert judgement better match expert answers** than other numerical criteria. In low flows, only Morph-MAE remains in the ten best matching criteria which include criteria specific to low-flow evaluation such as RMLFV, EILN or EILF and more standard criteria such as MARE and MdAPE.

However, the **rankings must be carefully interpreted** as the similarity percentage never exceeds 60%. Moreover, the ranking is based on limited differences between similarity percentages; therefore the first top criteria are not so different.

The squared relative errors appear in both rankings. We could say that if a basic kind of error should be chosen to match the expert judgement, the mathematical structure of the error could be squared and relative.

OMR Global Ranking from Table 9 and Table 10

The OMR global rankings give **higher similarity percentages**. This is directly explained by the definition of the optimal minimal ration value for equivalence. The **criteria in the top 10 list are mainly the same** as the ranking made without taking into account the OMR values, but ranks change.

The **classes of criteria** present in the tables using OMR=1, “Absolute Errors” and “Squared Relative Errors” **are no longer coherent** due to different values of OMR for each criterion. This is why RMSE, MSE and CVR appear as representatives of the squared relative errors and MAE and RAE as representatives of the absolute errors.

Criterion – Experts Graphs

The graphs of similarity percentages between experts and criteria were plotted for the Nash-Sutcliffe criterion in high flows and low flows, and for the best criteria for high flows and low flows, respectively Morph-NSE (0,5,5) and RMLFV (see Figure III-28, Figure III-29, Figure III-30and Figure III-31). In the four cases, **experts are pretty well scattered on the range of scored similarity percentage** which depends on the criterion and the flow conditions. Also, we can note that **extreme values are never reached**.

III.C.3 What does a Good / Average / Poor simulation mean for the expert?

Definition of the absolute rating scales used for the visual and numerical evaluations

The absolute evaluation corresponds to the second question of the survey, in which the expert assesses models with qualitative adjectives. The qualitative scale used in the survey (from worst to best) is:

Very Poor / Poor / Slightly Poor / Average / Slightly Good / Good / Very Good

→ Increase in quality according to the expert→

This qualitative scale corresponds to a series of numerical ranges that should ideally not overlap:

Qualitative scale	Numerical ranges
Very Poor	[T_{VPN}; T_{VPX}]
Poor	]T_{PN}; T_{PX}]
Slightly Poor	]T_{SPN}; T_{SPX}]
Average	]T_{AN}; T_{AX}]
Slightly Good	]T_{SGN}; T_{SGX}]
Good	]T_{GN}; T_{GX}]
Very Good	]T_{VGN}; T_{VGX}]

Objective

For each criterion, the objective is to plot the **correspondence between the visual evaluation rating scale and the numerical criteria rating scale**. This can be done for each expert or on the whole set of experts.

The analysis was made using:

- the answers of experts for Question 1 and Question 2,
- for each criterion, the values obtained by each model on each hydrograph.

Results are computed so that:

- High and low flows are considered separately,
- Each expert gets plots independent from the ones for the group of experts.

Method

For each hydrograph, only the model selected by the expert as the best one in Question 1 is considered. When Equivalent has been chosen, any of the two values (Value on Model A) is kept, and when nothing has been answered, the hydrograph is skipped. The criterion value is then associated to the answer to Question 2 which corresponds to the absolute evaluation of the model (Very Poor – Poor / Slightly Poor / Average / Slightly Good / Good – Very Good). For each criterion, box plots of values corresponding to each qualitative range were drawn. This gives the numerical interval of variation that can be expected for each category.

Example of result graphs and ranges

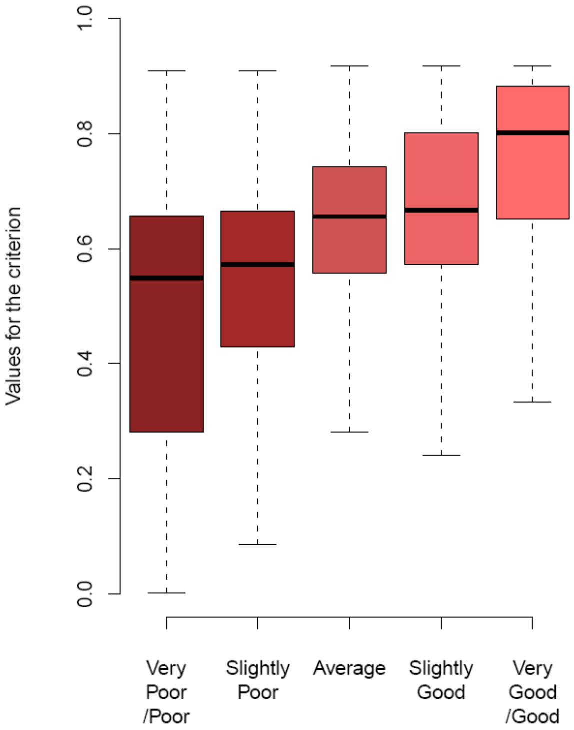


Figure III-32: Correspondence boxplot for the Nash-Sutcliffe efficiency criterion in High Flows

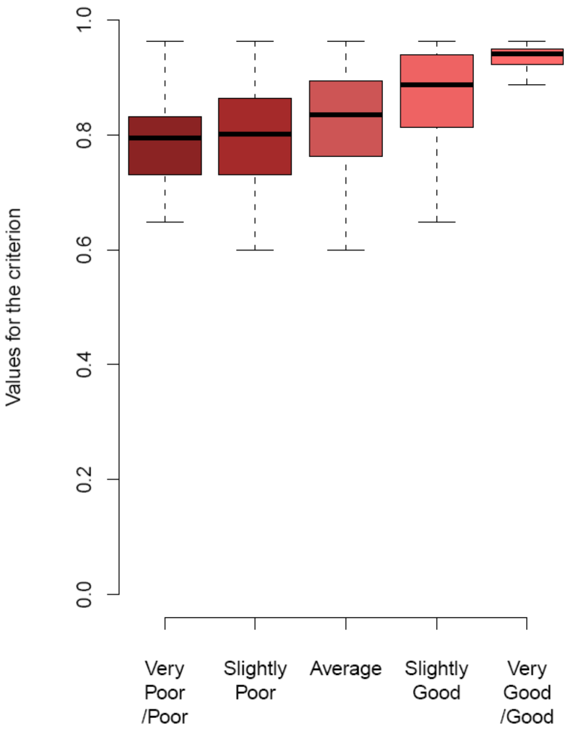


Figure III-33: Correspondence boxplot for the Morph-NSE (0,5,5) criterion in High Flows

Table 11: Numerical ranges for each evaluation class for the Nash-Sutcliffe efficiency criterion and Morph-NSE (0,5,5) in high flows

Qualitative scale	Numerical ranges for Nash	Numerical ranges for Morph-NSE
Very Poor - Poor	[0.3 ; 0.65]	[0.75 ; 0.82]
Slightly Poor	]0.45 ; 0.65]	]0.75 ; 0.85]
Average	]0.6 ; 0.75]	]0.79 ; 0.89]
Slightly Good	]0.6 ; 0.8]	]0.8 ; 0.95]
Good – Very Good	]0.65 ; 0.85]	0.95

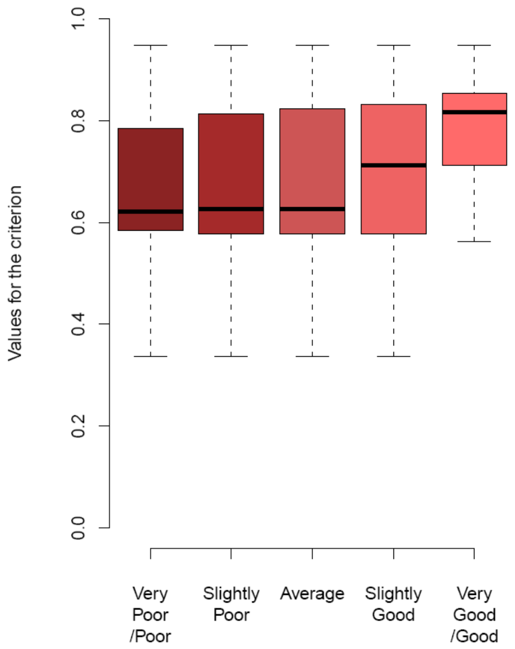


Figure III-34: Correspondence boxplot for the Nash-Sutcliffe efficiency criterion in Low Flows

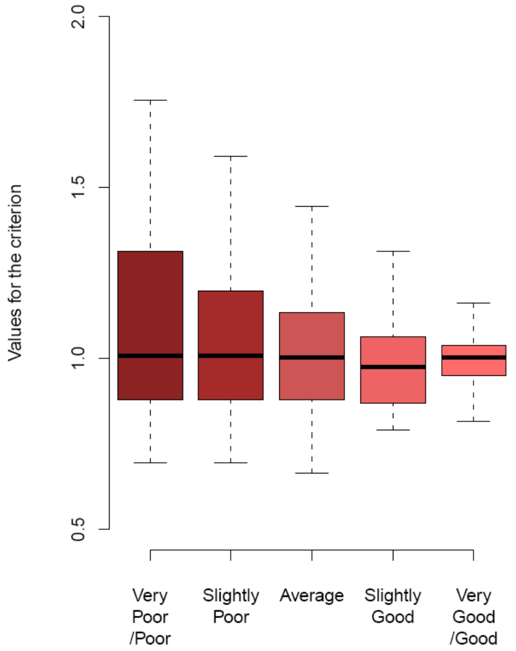


Figure III-35: Correspondence boxplot for the RMLFV criterion in Low Flows

Table 12: Numerical ranges for each evaluation class for the Nash-Sutcliffe efficiency criterion and the RMLFV criterion in low flows

Qualitative scale	Numerical ranges for Nash	Numerical ranges for RMLFV
Very Poor - Poor	[0.6; 0.79]	[0.9; 1.3]
Slightly Poor	]0.6; 0.81]	]0.9; 1.2]
Average	]0.6; 0.82]	]0.9; 1.15]
Slightly Good	]0.6; 0.83]	]0.9; 1.1]
Good – Very Good	]0.7; 0.85]	]1.0; 1.08]

Discussion

The coherence between the expert judgement and the criterion values depends on:

- the criterion, depending on its score range and consistence with the expert judgement (see section Best Match Criteria)
- the expert, depending on his indulgence and his use of the Average answer,

In high flows (see Figure III-32 and Figure III-33), **the two criteria** (that have a best match when they score 1) **have coherent evolutions with the expert judgements**. Even though the numerical ranges overlap (see Table 11), the mean values (black lines) increase as we get closer to Good rankings.

In low flows (see Figure III-34 and Figure III-35), we can notice the same kind of evolution of the **Nash-Sutcliffe criterion** as in high flows. However, numerical ranges overlap much more than in high flows, and mean values are constant for Very Poor / Poor, Slightly Poor and Average. Therefore, the criterion results are less consistent with the expert judgement in Low Flows than in High Flows. This conclusion had to be expected as the Nash-Sutcliffe criterion, due to its mathematical construction (squared error which emphasizes large errors than are encountered in high flows areas), is a better evaluator in high-flow conditions. The **RMLFV criterion** has a perfect match value of 1 and scores in [0; ∞[. Therefore, ranges are consistent with the expert judgement if we consider the top values of each range. Indeed, those values get closer to the perfect match value of 1.

Conclusion

From the comparison between the expert judgements and the numerical criteria, we were able to answer a set of questions on the relation between the quantitative and qualitative criteria and on the expert judgement itself.

Questions first regarded three aspects of the expert judgement itself. First, the experts' leniency is not supposed to bias the answers as experts uniformly spread on the leniency scale. Second, the similarities between expert judgements indicated that similarities were higher for the absolute ranking question than for the relative comparison question and outstanding experts were detected. Third, the visual criteria used by experts to evaluate hydrographs were ranked and patterns on how to decide between models could be set.

Then, conclusions on the direct relation between quantitative and qualitative criteria were drawn. The word "Equivalent" was redefined for each criterion to better match the expert answers and ratio values to consider models equivalent were extracted. Numerical criteria were ranked according to their likeness to expert answers and it was pointed out that numerical criteria designed to reproduce the expert judgment are the most similar to the expert judgement in high flows and that in low flows, criteria designed specifically for low flow conditions were better. Last, evaluation rating scales of quantitative and qualitative criteria were compared.

Thus, several questions on how the expert judgement relates to numerical criteria were answered. However, we should keep in mind that no absolute answers can exist when dealing with the human brain. Indeed, the Turing test is an illustration of that matter.

Analyses will be continued in September. Several aspects such as the influence of the experts experience on their judgement or the difference between how people say they evaluate hydrographs and how they actually do so, will be further analyzed. A scientific article to be submitted to an international journal will be written to present the main outcomes of this work. Individual assessments will be produced and sent to each survey taker to give them feedbacks o the survey.

Further studies in line with this research project could include the evaluation of probabilistic modelling approaches, deterministic or ensemble forecasting. Also, the comparison of evaluation criteria to expert judgement could be broaden to environmental and human sciences as evaluation criteria are in use in disciplines such as ecology, economics, sociology, etc.

As a conclusion, this internship taught me both technical and human skills. Indeed, I learnt to adapt to new tools as I knew little about hydrology and got to program in languages I had never used before. Also, I acquired skills in both engineering (through the design of the survey that needed to meet imposed conditions) and research (through the intensive literature review and the writing of a scientific article). I have grown more familiar to the research context with its constant evolutions and necessary collaborations. In my opinion, this permanently renewing knowledge environment is very precious and shall be an asset in any scientific position.

Appendices

Appendix A

List of criteria obtained from the literature review, corresponding abbreviations used in the thesis and bibliographical references

A. Calculated on whole period

Basic absolute criteria

Abbreviation	Criterion Name	References
MAE	Mean absolute error	Dawson et al. (2007) Jachner et al. (2007)
MCE	Mean cumulative error	Dawson et al. (2007)
MSE	Mean squared error	Houghton-Carr (1999)
MSEA	Additive systematic error	Willmott (1981)
MSEP	Proportional systematic error	Willmott (1981)
MSEL	Interdependance of MSEA and MSEP	Willmott (1981)
MSES	Overall systematic error	Willmott (1981)
MSEU	Unsystematic error	Willmott (1981)
RMSE	Root Mean Squared Error	Dawson et al. (2007) Jachner et al. (2007)
R4MS4E	Fourth Root Mean Quadrupled Error	Dawson et al. (2007)
MSDE	Mean Squared Derivative Error	de Vos and Rientjes (2007; 2008); Dawson et al. (2010)
AIC	Akaike Information Criterion	Akaike (1974)
BIC	Bayesian Information Criterion	Schwarz (1978)
AME	Absolute Maximum Error	Dawson et al. (2007)
PDIFF	Peak Difference	Dawson et al. (2007)

Relative criteria (dimensionless)

Abbreviation	Criterion Name	References
B	Bias	Smith et al. (2004)
MSLE	Mean squared logarithmic error	Houghton-Carr (1999) Dawson et al. (2010)
MARE	Mean absolute relative error	Dawson et al. (2007)
MdAPE	Median Absolute Percentage Error	Dawson et al. (2007)
MRE	Mean relative error	Dawson et al. (2007)
MSRE	Mean squared relative error	Dawson et al. (2007)

RVE	Relative volume error	Dawson et al. (2007)
RAE	Relative absolute error	Dawson et al. (2007)
R^2	Coefficient of determination	Smith et al. (2004) Krause et al. (2005)
wr^2	Weighted coefficient of determination	Krause et al. (2005)
r_{mod}	Modified correlation coefficient	Smith et al. (2004)
CVR	Coefficient of variation of residuals	World Meteorological Organization
EI	Nash-Sutcliffe efficiency index	ASCE (1993) Smith et al. (2004)
EImod	Modified Nash-Sutcliffe efficiency index	Krause et al. (2005)
EIrel	Relative deviation on the Nash-Sutcliffe efficiency index	Krause et al. (2005)
IA	Index of agreement	Willmott (1981)
IAmod	Modified index of agreement	Krause et al. (2005)
IArel	Relative deviation on the index of agreement	Krause et al. (2005)
EIFDC	Nash-Sutcliffe efficiency index based on flow duration curve	
EILN	Nash-Sutcliffe efficiency index on log-flow	
EIR	Nash-Sutcliffe efficiency index on root-squared flow	Chiew and McMahon (1994) Clarke (2008b)
EI2	Nash-Sutcliffe efficiency index on squared flow	
MEI	Modified Nash-Sutcliffe efficiency index	
SEI	Seasonal efficiency index	
PI	Persistence index	Dawson et al. (2007)
EIparam	Modified Nash-Sutcliffe efficiency index adjusted with parameters	Clarke (2008a)
KGE	Kling-Gupta Efficiency	Gupta et al. (2009)
ESC	Error sign count	
REMP	Relative error in maximum peak	ASCE (1993) Dawson et al. (2007)
RBFI	Ratio of base flow index	
TEST	t-test	Dawson et al. (2010)
IRMSE	Inertia Root Mean Squared Error	Dawson et al. (2010)

B. Calculated on single events

Calculated on flood events

A flood event is defined as a period of time with observed flow continuously exceeding the 0.8 percentile of observed flows over the test period.

Abbreviation	Criterion Name	References
FTEI	Flood threshold exceedance index	
RMP	Ratio of mean flood peaks	

MRAF	Mean ratio of annual flood	
RMV	Ratio of mean flood volumes	Andréassian et al., (2003)
RMVSX	Modified ratio of mean flood volumes	
EIHF	Nash-Sutcliffe efficiency index on high flows	

Calculated on low flow periods

A period of low flow is defined as a period where the observed flow remains under 0.2 percentile of observed flows over the test period.

Abbreviation	**Criterion Name**	**References**
RLFD	Ratio of low flow deficit	
RMLFV	Ratio of mean low flow volumes	
EILF	Nash-Sutcliffe efficiency index on low flows	

C. Criteria developed to combine qualitative and quantitative

Abbreviation	**Criterion Name**	**References**
THREAT	Series Distance - Threat score	Ehret and Zehe (2010)
ODMA	Series Distance - Error in amplitude	
ODMT	Series Distance - Error in timing	
Morph-NSE	Hydrograph Pattern Matching Algorithm Morph Version NSE	Ewen (2011)
Morph-MAE	Hydrograph Pattern Matching Algorithm Morph Version MAE	

Appendix B

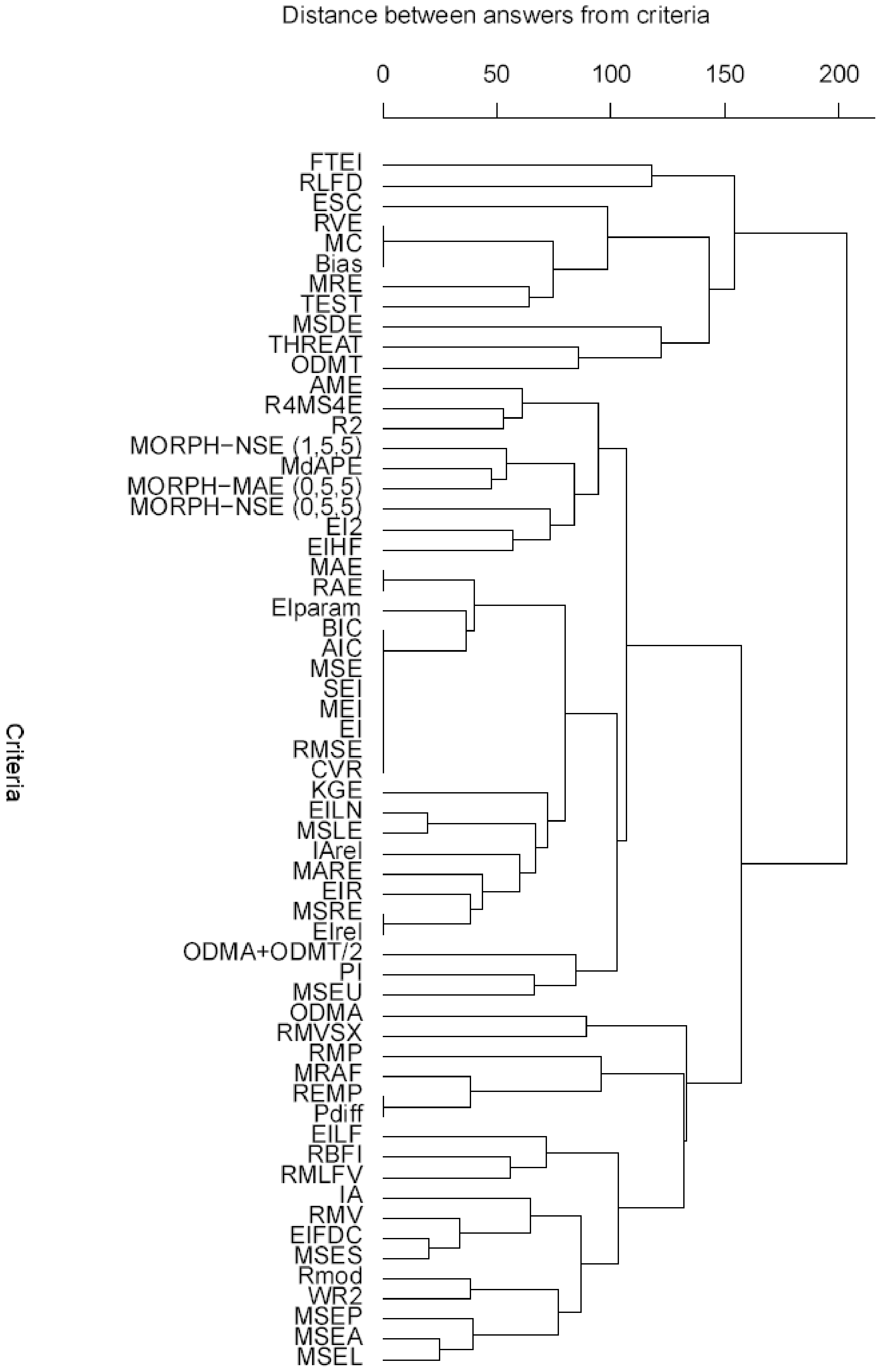


Figure 0-1: Hierarchical cluster of the criteria for high-flow periods

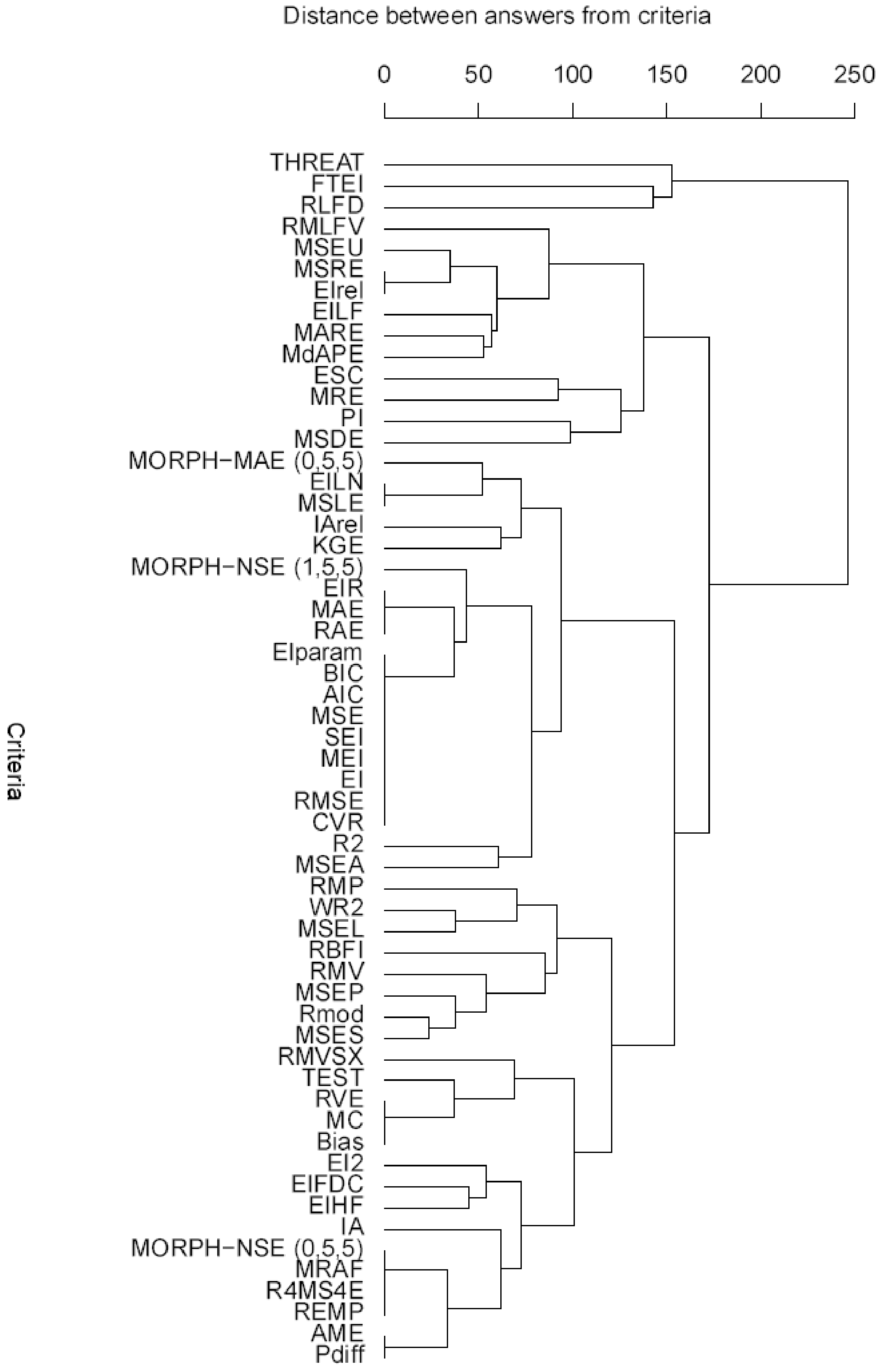


Figure 0-2: Hierarchical cluster of the criteria for low-flow periods

References

Akaike, H. (1974), A new look at the statistical model identification, *Automatic control, IEEE Transactions*, *19*, 716-723.

ASCE (1993), *Criteria for evaluation of watershed models*, American Society of Civil Engineers, Reston, VA, ETATS-UNIS.

Bennett, N. D., B. F. W. Croke, A. J. Jakeman, L. T. H. Newham and J. P. Norton (2010), Performance evaluation of environmental models, paper presented at Proceedings of International Environmental Modelling and Software Society (iEMSs) 2010 International Congress on Environmental Modelling and Software Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada.

Berthet, L., V. Andreassian, C. Perrin and C. Loumagne (2010a), How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion *Hydrological Sciences Journal*, *55*, 1063-1073.

Berthet, L., V. Andréassian, C. Perrin and C. Loumagne (2010b), How significant are quadratic criteria? Part 1. How many years are necessary to ensure the data-independence of a quadratic criterion value? , *Hydrological Sciences Journal*, *55*, 1051-1062.

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

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

Boyle, D. P., H. V. Gupta, S. Sorooshian, V. Koren, Z. Y. Zhang and M. Smith (2001), Toward improved streamflow forecasts: Value of semidistributed modeling, *Water Resources Research*, *37*, 2749-2759.

Burnash, R. J. C. (1995), The NWS River Forecast System - catchment modelling, in *Computer Models of Watershed Hydrology, Chapter 10*, edited by V. P. Singh, pp. pp. 311-366, Water Resources Publications.

Chiew, F. and T. McMahon (1994), Application of the daily rainfall-runoff model MODHYDROLOG to 28 Australian catchments, *Journal of Hydrology*, *153*, 383-416.

Chiew, F. H. S. and T. A. McMahon (1993), Assessing the Adequacy of Catchment Streamflow Yield Estimates, *Aust. J. Soil Res.*, *31*, 665-680.

Clarke, R. T. (2008a), A critique of present procedures used to compare performance of rainfall-runoff models, *Journal of Hydrology*, *352*, 379-387.

Clarke, R. T. (2008b), Issues of experimental design for comparing the performance of hydrologic models, *Water Resour. Res.*, *44*, 9pp.

Dawdy, D., R. and T. O'Donnell (1965), Mathematical models of catchment behavior, *American Society of Civil Engineers Proceedings*, *91*, 123-137.

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

Dawson, C. W., R. J. Abrahart and L. M. See (2010), HydroTest: Further development of a web resource for the standardised assessment of hydrological models, *Environmental Modelling & Software*, *25*, 1481-1482.

de Vos, N. J. and T. H. M. Rientjes (2007), Multi-objective performance comparison of an artificial neural network and a conceptual rainfall-runoff model, *Hydrol. Sci. J.-J. Sci. Hydrol.*, *52*, 397-413.

de Vos, N. J. and T. H. M. Rientjes (2008), Multiobjective training of artificial neural networks for rainfall-runoff modeling, *Water Resour. Res.*, *44*, 15pp.

Ehret, U. and E. Zehe (2010), Series distance - an intuitive metric for hydrograph comparison, *Hydrology and Earth System Sciences*, *7*, 8387-8425.

Ewen, J. (2011), Hydrograph matching method for measuring model performance, J. Hydrol., 408(1-2), 178-187, doi:10.1016/j.jhydrol.2011.07.038.

Garçon, R. (1996), Prévision opérationnelle des apports de la Durance à Serre-Ponçon à l'aide du modèle MORDOR, *La Houille Blanche*, *5*, 71-76.

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

Houghton-Carr, H. A. (1999), Assessment criteria for simple conceptual daily rainfall-runoff models, *Hydrol. Sci. J.-J. Sci. Hydrol.*, *44*, 237-261.

Jachner, S., K. G. van den Boogaart and T. Petzoldt (2007), Statistical methods for the qualitative assessment of dynamic models with time delay (R package qualV), *J. Stat. Softw.*, *22*, 1-30.

Klemes, V. (1986), Operational Testing of Hydrological Simulation-Models, *Hydrol. Sci. J.-J. Sci. Hydrol.*, *31*, 13-24.

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

Linsley, R., K. and N. Crawford, H. (1960), Computation of a synthetic streamflow record on a digital computer, *IAHS Publ.*, *51*, 526-538.

Mayer, D. G. and D. G. Butler (1993), Statistical validation, *Ecological Modelling*, *68*, 21-32.

Moriasi, D. N., J. G. Arnold, L. M. W. Van, R. L. Bingner, R. D. Harmel and T. L. Veith (2007), Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *Anglais*, *50*, 885-900.

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

O'Connell, P. E., J. E. Nash and J. P. Farrell (1970), River flow forecasting through conceptual models part II - The Brosna catchment at Ferbane, *Journal of Hydrology*, *10*, 317-329.

Olsson, J., J. Södling, J. Dahné, B. Arheimer, H. Amaguchi and A. Kawamura (2011), Man vs. Machine, a Swedish experiment on hydrological model performance assessment, edited, SMHI.

Perrin, C., V. Andreassian and C. Michel (2006), *Simple benchmark models as a basis for model efficiency criteria*, 24 pp., Schweizerbart, Stuttgart, ALLEMAGNE.

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

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

Refsgaard, J. C., L. Troldborg, H. J. Henriksen, A. L. Højberg, R. R. Møller and A. M. Nielsen (2010), God praksis i hydrologisk modellering, *Geo-vejledning*, *7*, 1-56.

Reusser, D. E., T. Blume, B. Schaefli and E. Zehe (2009), Analysing the temporal dynamics of model performance for hydrological models, *Hydrology and Earth System Sciences*, *13*, 999-1018.

Rykiel, E. J. J. (1996), Testing ecological models: the meaning of validation, *Ecological Modelling*, *90*, 229-244.

Scholten, H., A. Kassahun, J. C. Refsgaard, T. Kargas, C. Gavardinas and A. J. M. Beulens (2007), A methodology to support multidisciplinary model-based water management, *Environmental Modelling & Software*, *22*, 743-759.

Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, *6*, 461-464.

Seibert, J. (2001), On the need for benchmarks in hydrological modelling, *Hydrol. Process.*, *15*, 1063-1064.

Smith, M. B., D.-J. Seo, V. I. Koren, S. M. Reed, Z. Zhang, Q. Duan, F. Moreda and S. Cong (2004), The distributed model intercomparison project (DMIP): motivation and experiment design, *Journal of Hydrology*, *298*, 4-26.

Willmott, C. J. (1981), On the Validation of Models, *Physical Geography*, *2*, 184-194.

Willmott, C. J. (1984), On the evaluation of model performance in physical geography, *Spatial Statistics and Models*, 443-460.

Internet Sources

[a] Alberta Environment, http://environment.alberta.ca

[b] Cyprus Holiday Lettings, http://www.cyprusholidaylettings.co.uk

[c] Golf Fee Card International, http://www.golfcard-france.com

[d] How Stuff Works?, http://static.howstuffworks.com

[e] Survey Gizmo, http://www.surveygizmo.com

[f] UC Drought Management, http://ucmanagedrought.ucdavis.edu

[g] Webster's Online Dictionary, http://websters-online-dictionary.org

Summary

This thesis is the result of an 8-month internship in the Hydro team at Cemagref, a French public research institute. The topic of this thesis was the comparison of evaluation criteria for hydrological models.

Hydrological models are mathematical tools developed to simulate the rainfall-runoff relationship at the catchment scale. Simulated flows may be used for the design and management of water infrastructures or for flood and low flow predictions. The reliability of these applications requires accurately evaluating model results. This can be done in two ways: either using numerical criteria that quantify the distance between observed and simulated flows, or based on the expert judgement that visually evaluates result graphs such as hydrographs. In this thesis, the two evaluation methods were compared and relationships between them were investigated.

To this end, the study consisted in two main tasks. First, from a survey designed during the internship and displaying a set of 20 low-flow and 20 high-flow hydrographs, 150 expert judgements were collected. Two simulated and one observed time series were displayed for each hydrograph. Experts were asked to compare the models (relative evaluation) and then to qualify the best model (absolute evaluation) on a 7-level scale (Very Good / Good / Slightly Good / Average / Slightly Poor / Poor / Very Poor). In parallel, 60 numerical criteria obtained from a large literature review were applied to evaluate efficiencies using a large range of metrics.

The results of the study entirely rely on the experts who took the survey. Most of these experts were senior researchers. The analysis of expert judgements proved that leniency did not influence the results and that answers to the relative question were more universal than the ones to the absolute question. Moreover, visual criteria used by experts depend on whether high-flow or low-flow hydrographs are displayed. Then, from the comparison of the two approaches, we redefined the term "Equivalent" for each numerical criterion, ranked the numerical criteria according to their similarity with the expert judgement and compared the rating scales between the two types of criteria.

This study provides a better understanding of the relation between numerical and visual criteria, but further research is still needed on this issue. Such a study could now be broaden to other hydrological modelling applications or to other disciplines that use various types of criteria to evaluate model simulations.