THRESHOLDS FOR FLOOD FORECASTING AND WARNING

EVALUATION OF *STREAMFLOW* AND *ENSEMBLE* THRESHOLDS

Werner H.A. Weeink

Enschede, June 2010

MSc thesis committee:
Dr. M.S. Krol
Dr.Ir. M.J. Booij
Dr. M.H. Ramos

Cemagref
Sciences, eaux & territoires

UNIVERSITY OF TWENTE.

THRESHOLDS FOR FLOOD FORECASTING AND WARNING

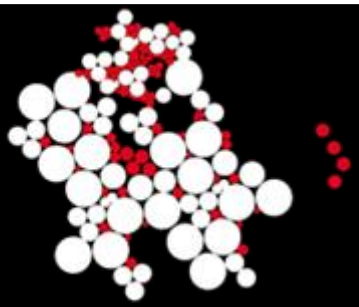
EVALUATION OF *STREAMFLOW* AND *ENSEMBLE* THRESHOLDS

Werner H.A. Weeink
Enschede, June 2010

MSc thesis committee:
Dr. M.S. Krol
Dr.Ir. M.J. Booij
Dr. M.H. Ramos

Version : MT_20100608_6.2
Status : Final

Cover photo : River Doubs at the village of Chalèze, France (10.03.2006)



Cemagref
Sciences, eaux & territoires

UNIVERSITY OF TWENTE.

COLOPHON

Author
Werner H.A. Weeink
Student Civil Engineering and Management | w.h.a.weeink@alumnus.utwente.nl

Streuweg 4
7663TC Mander
The Netherlands

Members MSc Thesis committee:
Dr. M.S. Krol [1]
Associate Professor | m.s.krol@ctw.utwente.nl
Dr.Ir. M.J. Booij [1]
Assistant Professor | m.j.booij@ctw.utwente.nl
Dr. M.H. Ramos [2]
Researcher | maria-helena.ramos@cemagref.fr

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

[2] Cemagref Antony
Unité de Recherche: Hydrosystèmes et Bioprocédés (HBAN)
Parc de Tourvoie, BP44
92163 Antony CEDEX
France

Institutions	: University of Twente Cemagref
Project	: MSc thesis
Extent document	: 89 pages
Author	: Werner H.A. Weeink
Date	: 8 June 2010

ABSTRACT

The use of ensemble weather predictions in flood forecasting is an acknowledged procedure to include the uncertainty of meteorological forecasts in streamflow predictions. Flood forecasters can thus get an overview of the probability of exceeding a critical discharge, and decide on whether a flood warning should be issued or not. This offers several challenges to forecasters, among which: 1) How to define critical thresholds along all the rivers under survey? 2) How to link locally defined thresholds to simulated discharges, which result from models with specific spatial and temporal resolutions? 3) How to define the number of ensemble forecasts predicting the exceedance of thresholds necessary to launch a warning?

In this study, *streamflow* thresholds are investigated for 75 catchments in France with defined operational thresholds. The emphasis lies on exceedances of this streamflow threshold -based on instantaneous observations- by daily discharges during a period of 10 years. The analysis shows that there is an overall optimal tradeoff among hits, misses and false alarms, expressed by the Critical Success Index (CSI), when the instantaneous streamflow thresholds are multiplied by an adjustment factor of 0.90 to give the daily streamflow thresholds.

The optimal *ensemble* threshold is also chosen to minimize the number of false alarms and misses, while optimizing the number of flood events correctly forecasted. Furthermore, in this study, an optimal ensemble threshold also considers flood preparedness: the gain in lead-time compared to a deterministic forecast. Data used to evaluate the ensemble thresholds come from a dataset of 208 catchments all over France, which covers a wide range of hydroclimatic conditions. The GRPE hydrological forecasting model, an ensemble version of the GRP model, a lumped soil-moisture-accounting type rainfall-runoff model, is used. The model is driven by the 10-day ECMWF deterministic and ensemble (51 members) precipitation forecasts for a period of 18 months.

From the results an overall ensemble threshold for the streamflow predictions based on the ECWMF forecast (i.e., a unique ensemble threshold to be applied to all catchments), which results in a higher CSI and a gain in lead-time compared to the deterministic forecast, could not be detected for the exceedance of the Q99 *streamflow* threshold (i.e. the 99^{th} percentile computed over the 18 month period). The search for optimal overall ensemble thresholds for lower *streamflow* thresholds also resulted in a negative preparedness score (i.e. a loss in lead-time). However, when the same analysis is conducted for a sub-selection consisting of 29 large catchments, ensemble thresholds resulting in higher CSI scores and gains in lead-time emerge for exceedances of the Q99 *streamflow* threshold: 10 ensemble members exceeding the threshold show up as an average optimal *ensemble* threshold. Furthermore, it was shown that both scores can be maximized when a catchment-specific *ensemble* threshold is applied. In this case, ensemble forecasts show an average gain in preparedness over deterministic forecasts of about 2-3 days for predictions of high flows (exceedances of the Q99% streamflow percentile).

RESUME

Seuils pour la prévision de crues et l'alerte

Evaluation de seuils de *débit* et seuils de *prévision d'ensemble*

L'utilisation de prévisions météorologiques d'ensemble pour la prévision de crue est une procédure reconnue pour prendre en compte l'incertitude des prévisions météorologiques dans les prévisions de débits. Les prévisionnistes peuvent conséquemment avoir une vision générale de la probabilité de dépasser un débit critique, et décider si une alerte aux crues devrait être émise. Cela présente plusieurs défis pour les prévisionnistes, parmi lesquels : 1) Comment définir les seuils critiques de débit le long de tous les cours d'eau surveillés ? 2) Comment relier les seuils définis localement aux débits simulés, lesquelles résultent de modèles avec des résolutions spatiales et temporelles spécifiques ? 3) Comment définir le nombre de prévisions d'ensemble prévoyant le dépassement des seuils nécessaire pour lancer une alerte ?

Dans cette étude, les seuils de débit sont évalués pour 75 bassins versants en France pour lesquels des seuils opérationnels sont définis. L'attention est portée sur les dépassements de ces seuils de débit – basés sur des observations instantanées – lors que l'on examine les débits journaliers pendant une période de 10 ans. L'analyse montre que il y a un compromis optimal entre *bonnes alertes, alertes manquantes et fausses alertes*, exprimé à l'aide du *Critical Success Index* (indice de succès critique – CSI), quand les seuils de débit instantanés sont multipliés par un facteur d'ajustement de 0,90 pour fournir les seuils de débits journaliers.

Le seuil optimal de prévision d'ensemble est également choisi pour minimaliser le nombre de fausses alertes et alertes manquantes, tandis que le nombre de crues correctement prévues est optimisé. En outre, dans cette étude, le seuil de prévision d'ensemble optimal considère aussi l'anticipation aux crues : le gain en délai de prévision comparé aux prévisions déterministes. Les données utilisées pour évaluer les seuils de la prévision d'ensemble proviennent d'une base de données de 208 bassins versants en France, qui couvre un large éventail de conditions hydro-climatiques. Le modèle GRPE de prévisions hydrologiques d'ensemble, version adaptée du modèle GRP, modèle pluie-débit global à réservoirs est utilisé. Le modèle est alimenté par les prévisions du Centre européen pour les prévisions météorologiques à moyen terme (CEPMMT - ECMWF en anglais). Il s'agit de prévisions déterministes et d'ensemble (51 membres) de précipitations, pour un horizon maximal de prévision de 10 jours et une période de 18 mois.

Les résultats n'ont pas permis de mettre en évidence un seuil de prévision d'ensemble global pour les prévisions de débit basées sur les prévisions ECWMF (i.e., un seuil unique qui pourrait être appliqué à tous les bassins versants), qui entraîne une plus grande valeur de CSI et un gain en anticipation comparé à une prévision déterministe pour le dépassement du seuil de débit Q99 (99ème percentile calculé sur la période de 18 mois). La recherche de seuils de prévision d'ensemble optimaux pour des seuils de débit plus bas a également conduit à un score d'anticipation négatif (i.e., une perte en délai d'anticipation). Néanmoins, la même analyse menée pour une sous-sélection consistant de 29 grands bassins versants a permis de détecter un seuil de prévision d'ensemble pour le dépassement du seuil de débit Q99 avec un score maximum de CSI et un gain en délai: 10 membres de la prévision d'ensemble dépassant ce seuil de débit apparait comme étant le seuil moyen optimal de la prévision d'ensemble. En autre, il a été montré que les deux scores peuvent être optimisés quand un seuil de prévision d'ensemble spécifique est appliqué à chaque bassin versant.

Dans ce cas, les prévisions d'ensemble montrent un gain moyen en délai d'anticipation d'environ 2-3 jours pour les prévisions de forts débits (dépassements du seuil de débit donné par le percentile 99%).

PREFACE

"He, who knows, does not predict. He, who predicts, does not know."
Lao Tzu (Chinese philosopher, 604-531 BC)

This report is the final product of my master Civil Engineering and Management, with a specialization in Water Engineering and Management, at the University of Twente. For this study, where I asses the thresholds involved in flood forecasting and warning, I joined the Hydrology group at Cemagref Antony in France for a period of four months. The months afterwards I spend my time at the Water Engineering and Management department at the University of Twente to finalize this MSc thesis.

I would like to thank all my colleagues at Cemagref. You all gave me a very warm welcome in France and in the world of flood forecasting and hydrological modeling. I appreciated the discussions about my work and other topics, your help developing my computer skills and the Frisbee games during the lunch breaks. Furthermore, I would like to thank F.Pappenberger from ECMWF for providing the forecast data, Météo-France for the observed precipitation data, the MEEDM (*Ministère de l'Ecologie, de l'Energie, du Développement durable et de la Mer*) for the discharge data and R. Lanblin and C. de Saint-Aubin from SCHAPI for the local thresholds data required for this research project.

Special thanks are for my supervisors Maria-Helena Ramos, Maarten Krol and Martijn Booij who helped me with their insights, comments and improvements to finalize this thesis.

Finally, I would like to thank my friends and my family, especially my parents, André and Marijke Weeink, for their love, support, and the fact that they gave me the opportunity to these chances in life.

Enschede, June 2010

Werner Weeink

TABLE OF CONTENTS

1 INTRODUCTION

Floods and inundations are a major natural hazard in several countries and pose a recurring risk. In Europe, according to the International Disaster Database (CRED, n.d.), 353 floods events have occurred during the last 20 years (1991-2009), killing about 2000 people and resulting in more than 83 billion US$ of damage. Only in France, three major floods have occurred between 1997 and 2007 (1999, 2002 and 2003), causing 60 casualties and a damage of 3.2 billion €. In 2008, areas vulnerable to inundation and floods in France covered 27000 km^2 (i.e. 16.134 communities with 5.1 million inhabitants) (Ministère de l'Écologie, de l'Énergie, du Développement durable et de la Mer, 2009).

The damage caused by flood events has become more important in the last fifty years due to the strong urban expansion and economic development at the floodplains. The report "Guidelines for Reducing Flood Losses" (UN, 2004) calls attention to the "alarming increasing trend in the number of people affected by natural disasters with an average of 147 million affected per year (1981-1990) rising to 211 million per year (1991-2000), with flooding alone accounting for over two-thirds of those affected". Effective measures to combat the risk associated with floods involve a number of activities and actions, including preventive measures, flood response and mitigative actions, post-disaster rehabilitation and economic recovery, as well as efforts to improve flood forecasting systems and increase preparedness for flood events. The severe impacts of flood events support the need for effective flood warning systems (FWS) to save lives and reduce economical damage.

This report focuses on the definition of thresholds for flood warning, e.g. which values should a forecasted event exceed in order to launch a warning? Paragraph 1.1 gives the main characteristics of a typical FWS and explains how thresholds are an integral part of a warning system. In paragraph 1.2, the focus lies on the specific context of this project: flood forecasting and warning in France. The organizational structure of the French flood forecasting authorities and their warning system is presented. The problem analysis and the motivation for this research are made explicit in paragraph 1.3. Furthermore, this problem analysis is converted into the objective and the research questions of this project in paragraph1.4. The final paragraph of this chapter (1.5) gives an overview of the outline of this report.

1.1 AN OVERVIEW OF FLOOD FORECASTING AND WARNING

Optimizing the thresholds of a flood warning system is the main goal of this project. In this report, we use the definition of Pingel *et al*. (2005) for a "flood warning system":

> "A flood warning system (FWS) is an integrated system of tools, data and plans that guides early detection of potential flood situations –flood forecasting– and coordinates response to flood emergencies."

Literature provides a wide overview of flood warning systems used around the world, e.g. EXCIFF (2005); Killingtveit and Sælthun (1997). In general, a flood warning system meets the main characteristics pictured in Figure 1. Individual FWS possibly deviate from this general structure and in- or exclude some (other) components or connections. In this schematic view, weather forecasts (a deterministic/single forecast or probabilistic/multiple scenarios forecasts), together with real-time data (precipitation, temperature, snow storage, discharge and/or water level), are the input for a rainfall-runoff model. The results of the rainfall-runoff model (hydrographs, maximum forecasted discharge/water level) are then compared to pre-defined "*streamflow* thresholds", which are often

based on historical observations. The threshold (non-) exceedance is evaluated and communicated to the decision-maker and/or the public. A warning is issued if a critical threshold, indicating the possibility of flooding, is exceeded. A FWS is usually based on a number of color-coded warning levels, which indicate the associated risk of the warning (e.g., moderate, high, severe). In the case of probabilistic predictions, an additional "*probabilistic forecast* threshold" is introduced: the forecaster has also to consider the percentage of forecasted scenarios exceeding a critical streamflow threshold (i.e., its probability to occur) when issuing warnings.

In this report, the focus lies on the component of a FWS corresponding to the definition of thresholds (*streamflow* thresholds and *probabilistic* thresholds) and on the evaluation of threshold exceedances, in order to find the best compromise between good and false alerts in flood forecasting.

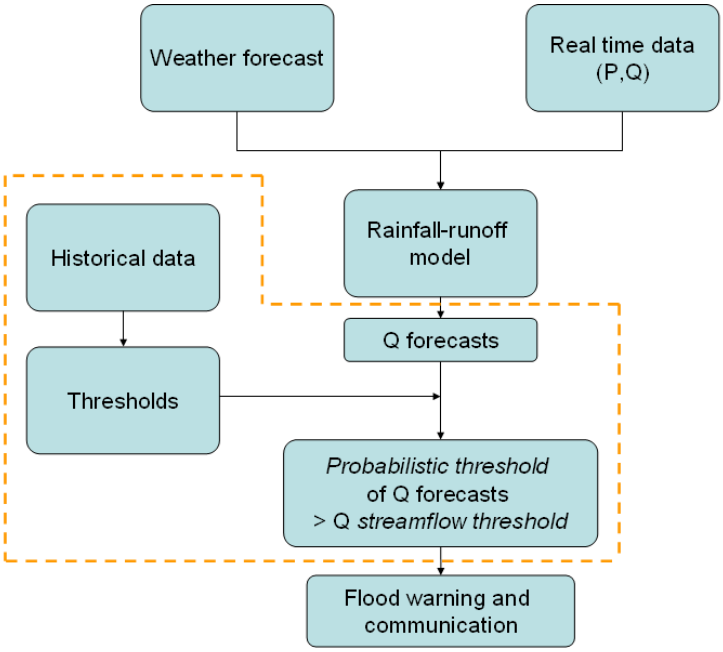


Figure 1. Schematic overview of a Flood Warning System.

1.1.1 *STREAMFLOW* THRESHOLDS

Streamflow thresholds are decision-making elements incorporated in a FWS to evaluate simulated hydrographs: is the simulated discharge higher than a predefined critical threshold? It is well-known that model results, used for simulation or discharge forecasting, are not "reality". The comparison between locally defined thresholds, based on historical observed data, and model results can therefore be a difficult task and eventually be at the origin of misleading conclusions.

A first source of this discrepancy is the uncertainty included in rainfall–runoff modeling, which goes along with results from the model and can introduce biases in its results. Beven (2001) identifies the following sources of uncertainty in rainfall-runoff modeling: errors in collecting rainfall data (measurements and forecasts), model uncertainty (structure and parameters), errors in streamflow data. Cloke and Pappenberger (2009) state that the meteorological input most often represents the largest source of uncertainty in flood forecasting. Moreover, they specify many sources of model uncertainty in the process of flood forecasting, for example: corrections and downscaling of the meteorological data, errors in the definition of the hydrological antecedent conditions, errors in the representation of the geometry of the system, possibility of infrastructure failure, and limitations of the models to fully represent physical processes.

Variations in the temporal and spatial scales among the data used to evaluate the thresholds and the hydrological and weather forecast models applied in the FWS need also to be considered. Thielen *et al.* (2008) discuss several additional reasons why critical *streamflow* thresholds for the European Flood Alert System (EFAS) could not directly be derived from historical discharge observations, but have to be evaluated from model-based simulations:

- information on management rules for lakes, reservoirs, polders or any other measures are not yet available;
- results have shown that the limited number of meteorological observations available for EFAS over Europe can lead to large discrepancies between model results and discharge observations;
- local critical values are generally derived from observations, and these are, however, only available at selected gauging stations and may not be valid for other river sections;
- EFAS is currently not able to reproduce hydrographs (especially peak discharges) well quantitatively in all river basins.

In summary, the choice of *streamflow* thresholds for guidance in flood warning is an essential step in a FWS. The strengths and limitations of the system, as well as its objectives, have to be considered. An optimum threshold should provide the best rate of detection of flood events, with a minimum acceptable of false alerts.

1.1.2 *ENSEMBLE* THRESHOLDS

One of the main differences that can be found among FWS, which strongly affect the communication of threshold exceedances, is the use of probabilistic or deterministic meteorological forecasts to drive the hydrological models.

Weather forecasts remain limited by the numerical representation of physical processes, the resolution of the simulated atmospheric dynamics and the sensitivity of the solutions to the pattern of initial conditions. A deterministic weather forecast in itself does not provide any information about the range of the resulting uncertainty. Ensemble prediction techniques attempt to take these uncertainties into account by changing the initial conditions slightly. This results in a number of weather forecasts (ensemble members) with the same probability of occurrence for the same location and time. Forecasts based on an Ensemble Prediction System (EPS) are an attractive product for flood forecasting systems since they can potentially extend forecasting lead-time; even though the range of uncertainty is often larger for meteorological forecasts with a longer lead-time (Cloke and Pappenberger, 2009).

In the case of implementing Ensemble Streamflow Predictions (ESP) in a probabilistic flood forecasting and warning system, one must also consider the *probabilistic forecast* threshold or *ensemble threshold.* The *ensemble* threshold is given by the number of ensemble members (i.e., the number of forecasts) exceeding each critical *streamflow* threshold. For example, in the Netherlands, 50% of the ensemble members is chosen to have to exceed a defined *streamflow* threshold to issue a pre-warning (Sprokkereef, 2009). In other terms, a pre-warning is issued if the *streamflow* threshold has 50% of probability to be exceeded. We note that for a warning of a higher category (flood event with lower probability or a more risky situation) two obvious options can be distinguished, since two thresholds are part of the system: 1) the forecaster can consider that a larger percentage of the ensemble members should exceed the same *streamflow* threshold or 2) the same amount of ensemble members should exceed a higher *streamflow* threshold. Alternative combinations can also be derived involving these two options.

1.1.3 THE ROLE OF THRESHOLDS IN FLOOD WARNING

As previously discussed, there are several cases resulting in a discrepancy between simulated and observed discharges/water levels and their corresponding thresholds (thresholds based on observations or on simulations), which can affect flood warning. Figure 2 illustrates why thresholds need to be defined with regard to these discrepancies.

In Figure 2 (a), the hydrological model [Qd(sim)] is not able to reproduce the exact quantities of discharge of the observed hydrograph [Qd(obs)], although it reproduces well the dynamics of the flow. In this case of underestimation of the discharges, the threshold based on observations will be exceeded a certain time after the observed discharge actually exceeds the same threshold. This has a significant impact on the warning of the flood event, since flood events would be "missed" by the system or warnings would be issued too late, when the flood is already occurring. It could also be the other way around: simulated discharges being systematically higher than the observed discharges. In this case, warnings based on the simulated discharges exceeding the observation-based threshold would result in frequent "false alerts". The ability of the model to forecast a hydrograph -including all sources of uncertainty- in the same way as the observed hydrograph affects the usefulness of thresholds based on observations. In the case illustrated in Figure 2 (a), the use of a threshold based on simulated discharges could be more appropriate to correctly detect the time of critical exceedances. This threshold would be lower than the threshold indicated in the figure, which is based on observed data.

In Figure 2(b) the effect of comparing daily mean discharges with "instantaneous" observed discharges is illustrated. Hourly and daily hydrographs are represented. Both graphs have the same daily mean discharge (Qd). It can be seen that the daily hydrograph (Qd) is not able to produce some of the peaks and threshold exceedances that are observed by the hourly hydrograph (Qh). Using thresholds based on hourly observations will probably lead to a higher number of misses (flood events/exceedances that are not forecasted), since most of the time the simulated (daily) peak discharges differ from the observed (hourly) peak discharges, especially during high flood events. This case highlights the need of defining a daily threshold in such a way that exceedances of simulated daily discharges correspond to the exceedances of hourly discharges to the observation-based threshold.

In Figure 2 (c), an observed daily discharged time series [Qd(obs)] is plotted as well as a simulated ensemble forecast [Qd(ens)]. The use of ensemble forecasting will influence the use of the critical threshold as well. Should a warning be issued if the threshold is exceeded by one ensemble member, a certain amount of members or all the members?

Furthermore, in ensemble forecasting, it could also be interesting to take into account the effect of possible amplified forecast uncertainty related to longuer lead-times. If there is a general trend in the accuracy of a flood forecast related to the lead time, then this trend will have a certain influence on the use of the *ensemble* threshold as well: should the number of members exceeding the critical *streamflow* threshold for a flood warning change according to the lead time? And should this *ensemble* threshold (number of ensemble members) vary according to the magnitude of the *streamflow* threshold?

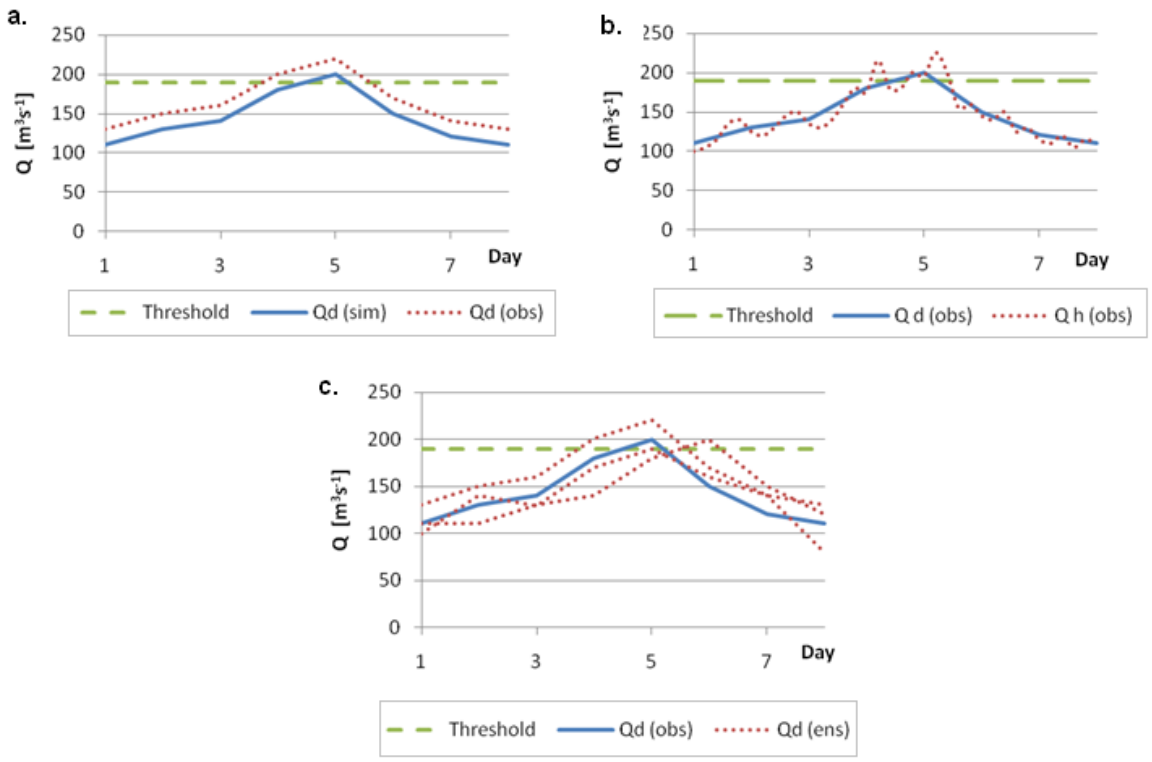


Figure 2 (a-c). The role of thresholds in flood warning when there is discrepancy between simulated and observed discharges.

1.2 FLOOD FORECASTING AND WARNING IN FRANCE

After the devastating floods of 1999 and 2002 in the Aude and Garde region (Delrieu *et al*., 2005; Gaume *et al.*, 2004), the French flood forecasting system and the involved organizations were totally reformed. A national hydrometeorological service SCHAPI ("Service Central d'Hydrométéorologie et d'Aide à la Prévision des Inondations", in French) was created to coordinate technical and financial programmes for 22 regional forecasting centres (SPC, "Service de Prévision de Crues") as well as to promote the development of flood forecasting tools and warning procedures, together with the national meteorological service (Météo-France). Currently, SCHAPI deals with information from several types of weather forecasts and hydrological models, including Météo-France deterministic and ensembles (Thirel *et al.*, 2008), the ensemble hydrological forecasts from the European Flood Alert System (EFAS) (Thielen *et al.*, 2009) and the ensemble streamflow prediction system developed by Météo-France, SIM-EPS (Rousset-Regimbeau *et al.*, 2007), based on 10-day ensemble predictions from the European Centre for Medium-range Weather Forecasts ECMWF (SCHAPI, 2008). Additionally, SCHAPI promotes the development of national wide flood forecasting platforms based on global and distributed hydrological models. Some local forecast centers use also locally calibrated systems; including the GRP forecast model developed at Cemagref, which is applied in this study (see Chapter 2.4).

Concerning flood warning in France, three *streamflow* thresholds are distinguished and visualized in the "Flood vigilance Map" (Figure 3), which defines the following colored levels:

- Red: risk of major flooding. Direct threat to the general safety of persons and property;
- Orange: risk of generating a significant level of inundation, which may have a significant impact on community life and on the safety of property and persons;

- Yellow: risk of flooding or rapid rise of water, which does not involve significant harm, but requires special vigilance in the case of seasonal and/or outdoor activities.

Each SPC defines these warning levels for their catchments under survey, i.e., where there is a need to forecast floods (human exposure, possibility of economic damages) and it is possible to forecast with enough lead-time to activate emergency procedures if necessary. These warning levels are based on historical, local observations and take the vulnerability of the area into account. This means that not all rivers in France are subject to operational flood forecasting and it might be the case that, for example, a high discharge is related to different warning levels in urban and rural areas.

According to SCHAPI (SCHAPI 2008), one of the current greatest challenges of their operational forecasters is to link the probabilistic model output to the operational (yellow/orange/red) alert levels used on the flood vigilance map (Figure 3).

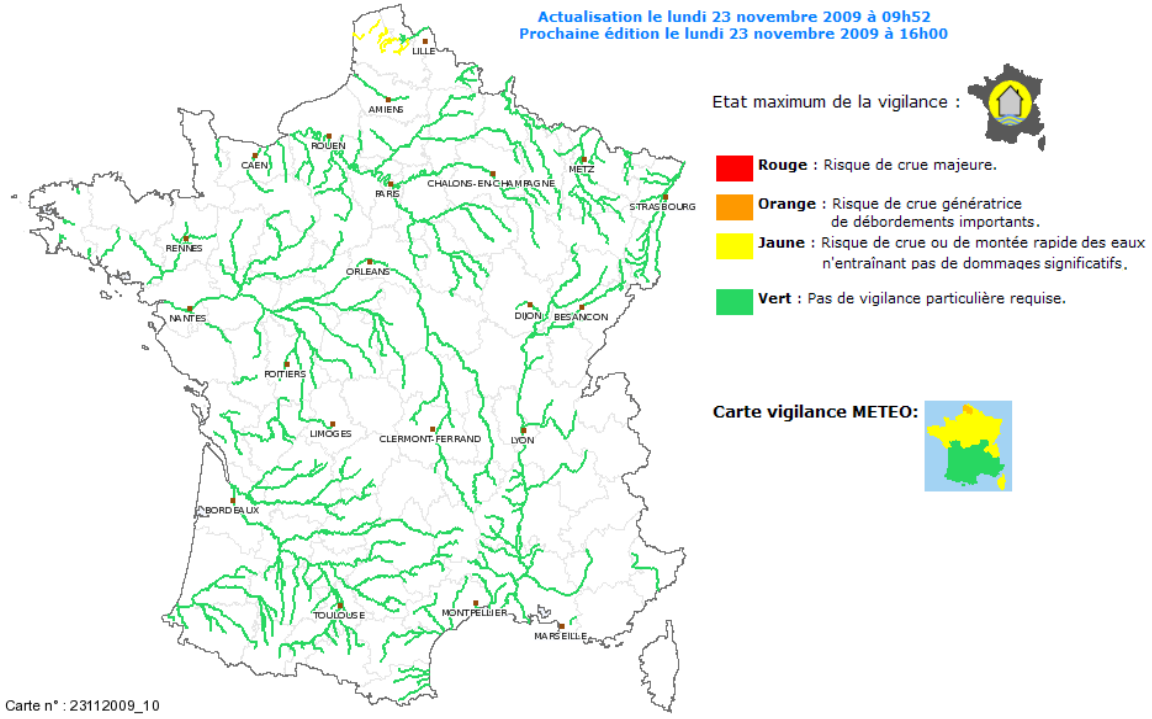


Figure 3. Example of a French "Flood vigilance" map (Carte de vigilance "crues", n.d.).

1.3 PROBLEM DEFINITION

From the previous paragraph (1.2), it becomes clear that the *streamflow* threshold and the *ensemble* threshold are two thresholds that are important for flood forecasting and warning in France. The definition of these thresholds raise some challenges described below.

The *streamflow* thresholds –triggering warnings if exceeded by the forecasted discharge- are the observed discharges linked to the colors (yellow-orange-red) in SCHAPI's flood warning system. However, these thresholds may not be appropriate to be applied directly to simulations from hydrological models that are setup to run at time steps different from the time step of the observed discharges at the origin of the threshold definition. In fact, the *streamflow* thresholds are based on "instantaneous" (hourly or shorter time steps) water level measurements, while several hydrological

forecast models run at larger time steps of several hours or day(s). This corresponds to the problem described in Figure 2(b).
The challenges for the observed streamflow threshold are to deal with:

- *agreement between the locally defined (instantaneous) threshold and a threshold adapted to the time step of the model;*
- *flood forecasting and warning in catchments without defined thresholds.*

As mentioned in paragraph 1.2, the *ensemble* threshold (i.e., the number of ensemble members exceeding the *streamflow* threshold to be considered for issuing a warning) is even a more complicated problem. The forecasted probability of exceedance, taken as the fraction of ensemble members exceeding the *streamflow* threshold, often does not represent the actual probability due to errors in the weather forecasts, the hydrological model, the estimation of initial conditions at the onset of the forecasts, etc. (Olsson and Lindström, 2008). The presence of two thresholds and the fact that the *ensemble* threshold does not represent the actual probability make it difficult to link the probabilistic model outcome to a warning procedure.

The challenge for the ensemble threshold is to find the average optimal number of ensemble members exceeding the streamflow threshold for a maximum of flood preparedness and a minimum of missing events or false alerts.

Finally, for both *ensemble* and *streamflow* thresholds, another *challenge to operational forecasters is to indentify links between catchment characteristics and thresholds values*.

1.4 OBJECTIVE AND RESEARCH QUESTIONS

Out of the problem definition and the challenges posed, the following objective is distilled:

> To determine optimal appropriate critical thresholds for operational flood forecasting and warning by analysing the performance of a flood forecasting system and the quality of its forecasts when different thresholds –*streamflow* thresholds and *ensemble* thresholds– are used, while taking into account the influence of catchment characteristics and the type of the weather forecast (ensemble/deterministic) used to drive the hydrological model.

This objective is converted into the following research questions:

- How should the *streamflow* thresholds based on instantaneous observations be adjusted for an optimal implementation in a (modeling) framework set up at daily time steps? What is the eventual relation between this "adjustment factor" and the catchment characteristics?
- What is the optimal *ensemble* threshold (i.e., the number of ensemble members exceeding the *streamflow* threshold) for a maximum preparedness in flood forecasting and warning? What is the eventual relation between this optimum and the catchment characteristics, the *streamflow* threshold levels and the forecasting lead-time?

1.5 REPORT OUTLINE

The next chapter (2) describes the data and the hydrological model used in the research project. Chapter 3 consists of a description of the methodological steps adopted, which fare the foundation for this research project. The results of the analysis of *streamflow* thresholds are presented and discussed in Chapter 4. The analysis of *ensemble* thresholds is the topic of Chapter 5. Conclusions are drawn and recommendations are given in Chapter 6.

2 DATA AND HYDROLOGICAL MODEL

This chapter gives an overview of the data and hydrological model used in this study. In paragraph 2.1, the catchment datasets used for the evaluation of the thresholds are presented. In paragraph 2.2 the focus lies on the observed precipitation and discharge archives. The probabilistic weather forecast (ECMWF) archives applied are introduced in paragraph 2.3. The structure and calibration of the GRP hydrological forecast model is highlighted in paragraph 2.4.

2.1 CATCHMENT DATASETS

In the problem analysis and the research objective (Chapter 1.3) a distinction is made between *streamflow* thresholds and *ensemble* thresholds. The criteria for the selection of a dataset of catchments to be used in the evaluation of these thresholds are not the same for both kinds of thresholds. For the *streamflow* thresholds, the main selection criterion is the availability of local operational thresholds (yellow-orange-red) defined by the local flood forecast centers and/or SCHAPI. The main selection criterion for the evaluation of the *ensemble* threshold is the availability of an archive of ensemble forecasts. The catchments selected for the evaluation of the observed *streamflow* thresholds are described in paragraph 2.1.1. The selection of catchments for the evaluation of the *ensemble* threshold is presented in paragraphs 2.1.2 and 2.1.3.

2.1.1 DATASET A: 75 CATCHMENTS

In order to evaluate the operational *streamflow* thresholds, the catchments in this dataset have to meet the following criteria: catchments should be of interest for operational services (real-time data available for forecasting and critical thresholds defined); catchments should have few missing data; catchments with a common period of data to compare the results between catchments; catchments should cover different hydroclimatic conditions. For this study, the first criterion was the most restrictive. A dataset of 75 catchments was finally selected.

The locations of these catchments are shown in Figure 4. This selection of catchments covers a wide range of the hydroclimatic conditions encountered in the country, including different geographical regions and catchment sizes. The catchment sizes range from 31 to 8900 km², with a median and mean size of respectively 747 and 1312 km². An overview of the catchment names, geographic coordinates and characteristics can be found in Appendix A. 1.

As explained in paragraph 1.2 the *streamflow* thresholds are based on historical observations and the local vulnerability and characteristics. This implies that they do not refer to the same statistical frequency or return period at all catchments. However, there appears to be some relation between the threshold levels and frequency periods, independent on the economical value and number of inhabitants of a catchment. For example, the yellow threshold is for most catchments often close to the two-year return period instantaneous flood (Figure 5). According to Carpenter *et al.* (1999), the discharge related to the two-year return period flood is a fraction larger than the bankfull discharges for natural rivers, causing potential damage in the inundated areas. For the smaller, natural catchments this description matches the definition of the yellow threshold as proposed by SCHAPI. The return period for floods exceeding the orange threshold is, for most of the catchments, between 2 and 5 years.

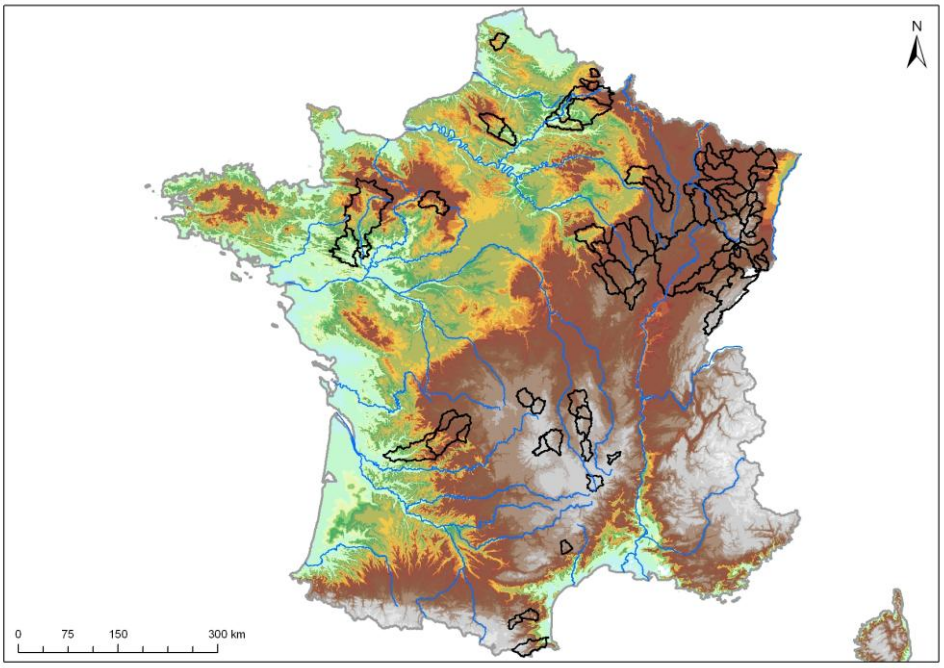


Figure 4. Location of 75 catchments in Dataset A.

For most catchments, the operational thresholds are available as water levels. Rating curves for these river sections are required to transform the water level thresholds into operational *streamflow* thresholds, which can then be compared to the output of the GRPE hydrological forecasting model, consisting of discharges only. However, rating curves were not available for this study and the final total number of catchments with thresholds defined was: 39 for the yellow threshold, 51 for the orange and 44 for the red *streamflow* threshold. For the catchments without locally defined thresholds, the analysis was carried with a threshold based on the 2-year return period, considering its similarity with the yellow threshold (Figure 5).

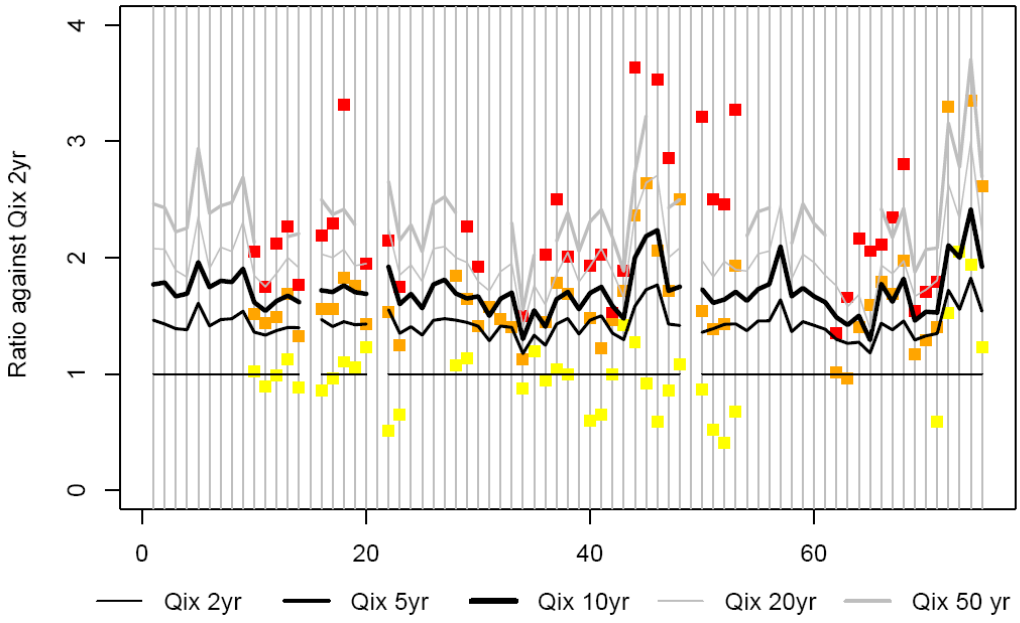


Figure 5. Relation between operational streamflow thresholds (yellow, orange and red squares) and instantaneous discharges of 2, 5, 10, 20 and 50 years of return period (lines) for 75 catchments. Both thresholds and discharges are represented by the ratio against the $Q_{ix\,2yr}$ discharge (y-axis). The catchments are ranked alphabetically on the x-axis.

2.1.2 DATASET B1: 208 CATCHMENTS

Another dataset of catchments is the starting point for the evaluation of the *ensemble* threshold. The general criteria that have to be met are: catchments with few missing data; catchments with a common period of data to compare the results between catchments; catchments covering different hydroclimatic conditions. Additionally, the most important condition is the availability of ensemble weather forecast archives for these catchments. In this study, the ECMWF ensemble forecast system (paragraph 2.3) is used for the evaluation of the *ensemble* threshold. Dataset B1 consists of 208 catchments ranging from 173 to 9390 km², with a median and mean size of respectively 879 and 1452 km². Their locations are shown in Figure 6. An overview of the catchment names, geographic coordinates and characteristics can be found in Appendix A. 1.

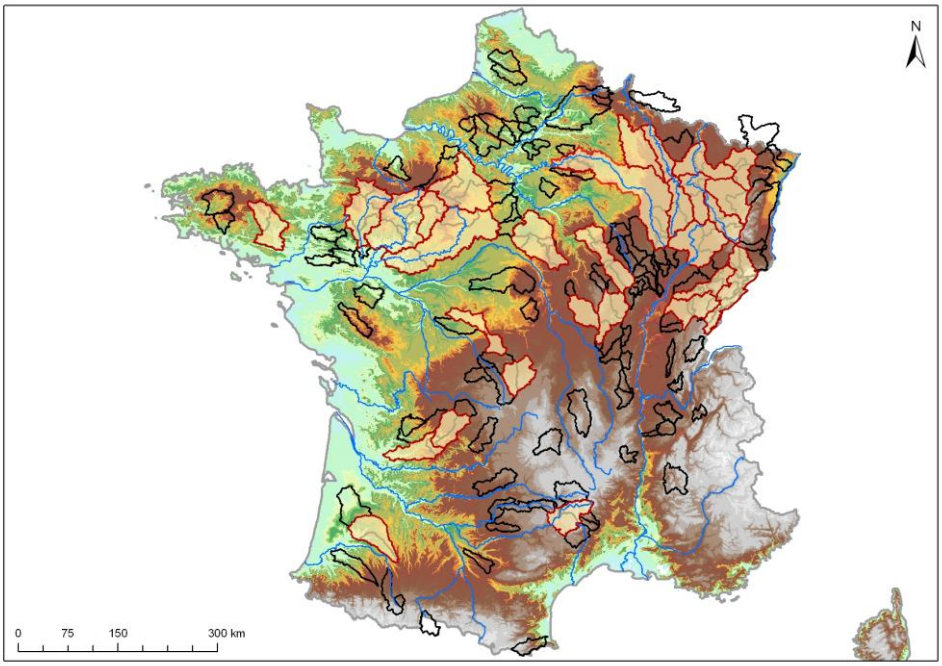


Figure 6. Location of 208 catchments in Dataset B1 (black contours) with the 29 catchments of Dataset B2 highlighted (red contours).

2.1.3 DATASET B2: 29 CATCHMENTS

The large grid size of the raw ECWMF data (0.5° x 0.5°, i.e., ~2000 km² of grid area over France) might influence the results of our analysis. Hence, a second dataset was created consisting of 29 large catchments selected out of dataset B1 (Highlighted catchments in Figure 6). This selection of catchments respects as well the criterion of covering a wide range of the hydroclimatic conditions encountered in the country. The catchment areas range from 1470 to 9390 km², with a median and mean size of respectively 3885 and 2990 km². An overview of the catchment names, geographic coordinates and characteristics can be found in Appendix A. 1.

2.2 OBSERVED DISCHARGE AND PRECIPITATION DATA

Observed precipitation and discharges are essential for the evaluation of the thresholds. Discharge observations are used for the comparison between the hourly and daily discharge values, the run of the hydrological model (calibration and forecasting) and the verification of the ensemble predictions, while precipitation data serve as input for the hydrological forecasting model. Observed precipitation

data come from the meteorological analysis system of Météo-France (SAFRAN) and observed streamflow data come from the French database Banque HYDRO.

DAILY AND HOURLY OBSERVATIONS

The archive of observed precipitation and discharge data consists of daily and hourly observations per catchment. The daily precipitation and discharge archive covers a period of 36 hydrological years (from 01.08.1970 to 31.07.2006); the hourly data covers a period of 10 years (from 01.08.1995 to 31.07.2005). Both, hourly as well daily discharges are required during the analysis of *streamflow* thresholds, which restrict the period of this analysis to 10 years. Figure 7 shows an example of the observed data for the year 2001 for catchment A1050310 the Ill River at Altkirch (Alsace). The red dot in the plot for the hourly discharge indicates missing data during April 2001. This means as well that during this period the observed daily discharge was most probably not constructed directly from the observed hourly discharges, but it has been reconstructed from other estimation procedures.

MISSING DATA

Missing data is the main problem concerning the data quality. The non-availability of hourly discharge data is for two reasons the most important problem:

- Daily discharges (mm) are equal to the sum of the hourly discharge (mm) during the day. So if hourly data is missing, the daily discharge is reconstructed and less accurate.
- Hourly data are often missing around the time of a threshold exceedance and, in this case, the magnitude and duration of exceedance are untraceable.

During the analysis (selection of time steps at which discharges exceed a given threshold), the possible influence of missing data is taken into account by taking these days out of the selection.

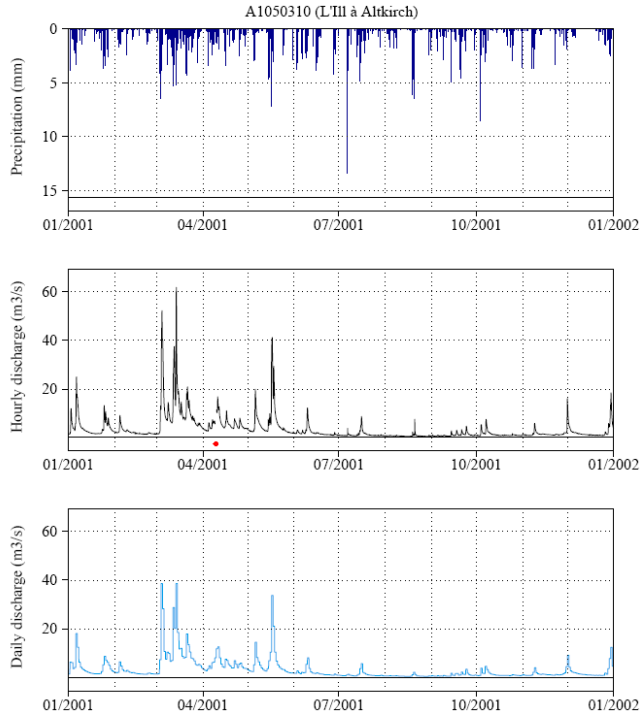


Figure 7. Example of time series of precipitation (top), hourly (centre) and daily (bottom) discharge data for catchment A1050310: The Ill River at Altkirch.

2.3 *ECMWF* ENSEMBLE PRECIPITATION FORECASTS

The atmosphere is a chaotic system, and small errors in the estimation of the current state can grow to have a major impact on the meteorological forecast. The errors in the meteorological forecast will have their impact on the forecasted discharge in the case of streamflow forecasting. Due to the limited number of observations and corresponding errors, there is always some uncertainty in the estimate of the current state of the atmosphere which limits the accuracy of weather forecasts. Taking into account the sensitivity of the prediction to uncertainties in the initial conditions, it is becoming common now to run in parallel a set, or ensemble, of predictions from different but similar initial conditions (Introduction to chaos, predictability and ensemble forecasts, n.d.; Palmer *et al.*, 2005).

The probabilistic weather forecast dataset -for precipitation only- available at Cemagref is issued by the European Centre for Medium-Range Weather Forecasts (ECMWF). The 52 rainfall forecasts consist of 50 ensembles, one high resolution deterministic forecast and one control forecast (same initial conditions as the high resolution deterministic forecast but at a coarser spatial resolution) (e.g. Goudeleeuw *et al.*, 2005). The ECMWF weather prediction model is run 51 times (control and 50 ensemble members) from slightly different initial conditions and each forecast is made using slightly different model equations. In this way, the effect of uncertainties in the model formulation and in the estimation of the initial conditions is taken into account.

Computer resources availability is one of the main factors that limits the resolution and complexity of numerical weather prediction models. In the case of meteorological ensemble forecasting, computer resources availability is the main reason that a tradeoff has to be made between resolution and the number of ensembles (Buizza, 2002). Hence, probabilistic forecasts often have a lower resolution than deterministic forecasts (i.e. it is not possible to forecast for *n* ensemble members on the same detailed resolution as the deterministic forecast within the time limits of operational flood forecasting) and (the uncertainties related) to small-scale atmospheric processes are not included in the ensemble weather forecast. The horizontal resolution of the ECMWF deterministic forecast is about 25x25km, and will be upgraded to 16x16 km in 2010 (Horizontal resolution increase, 2009).

The ECMWF ensemble prediction system (EPS) has 51 scenarios (or members) and a forecasting range of 10 days. It was provided within a grid size resolution of 0.5° latitude x 0.5° longitude (about 45x45 km of grid size over France). The 51 scenarios can be combined into an average forecast (the ensemble-mean) or they can be used to compute probabilities of possible future weather events. A precise estimation of the probabilities requires that the forecasts accurately describe the variability of the phenomenon being forecasted. However, the ECMWF forecast tends to underestimate the variability and spread; the relative large grid size of the ensemble forecast is debit to this performance (e.g. Buizza *et al.*, 2005). The advantage of the ECMWF ensemble forecast is its lead-time of 10 days and its large number of ensemble members. The disadvantage of this EPS is its relative large grid size (45x45 km), given that many catchments in dataset B1 have a substantial smaller surface area. The impact of this coarse grid size is addressed in Chapter 5.2, by taking into account dataset B2, a sub-set containing large catchments.

The ECMWF archive available at Cemagref (ensemble and deterministic forecasts) covers an 18-month period (11.03.2005 to 31.08.2006). ECMWF forecasts are issued at 12 UTC. In order to compare forecasts to observations available for the time lag from 0:00 to 23:59, the effective lead-time is reduced from 10 to 9 days, as indicated in Figure 8. Figure 9 shows an example of ensemble streamflow prediction based on ECMWF EPS and the GRPE hydrological model (forecast issued on 16.01.2006 and valid for the next 9 days – up to 25.01.2006).

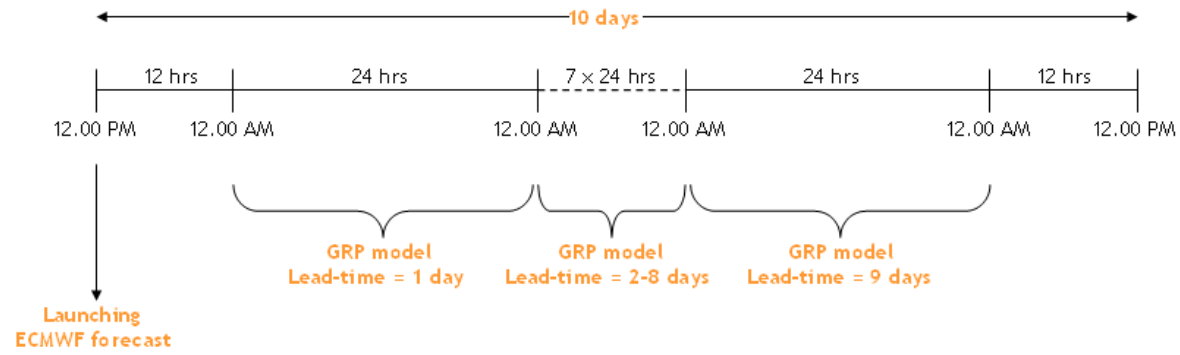


Figure 8. Lead times considered in this study for the ECMWF weather forecast and GRP hydrological model.

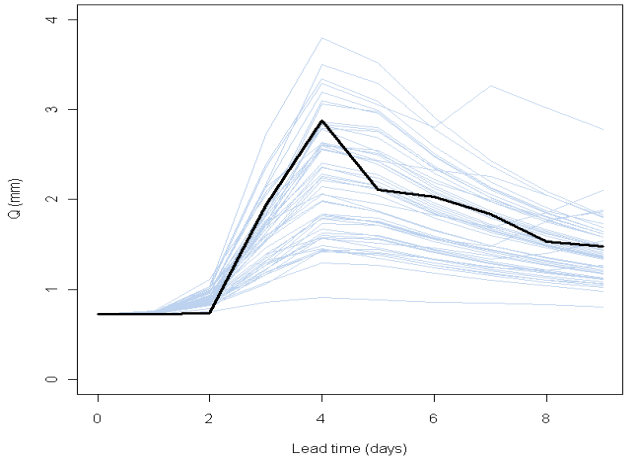


Figure 9. Ensemble Streamflow Prediction (Q in mm) for the Doubs River at Voujeaucourt (forecast issued on 16.01.2006 and for the next 9 days) with a lead-time of 1-9 days based on the ECMWF ensemble forecast and the GRPE hydrological model. The blue lines represent the ensemble members, the black bold line represent the observed streamflow.

2.4 *GRPE* HYDROLOGICAL MODEL

The hydrological ensemble forecasting model used is the GRPE model, based on the GRP model developed at Cemagref (Tangara, 2005) and recently adapted to run ensemble predictions (Ramos *et al.*, 2008). In paragraph 2.4.1, the model structure and parameters are presented. The calibration of the model is described in paragraph 2.4.2. The complete structure of the GRPE model -including its equations- is described in Appendix 0.

2.4.1 *GRPE* MODEL STRUCTURE

The GRPE model is a lumped soil-moisture-accounting type rainfall-runoff model, which is driven by daily precipitation forecasts (here ECMWF prediction sets) and mean evapotranspiration (daily averages computed from climatological data over the calibration period provided by Météo-France). The model structure (Figure 10) is derived from the GR4J hydrological simulation model (Perrin, 2002) and is specially designed for flood forecasting.

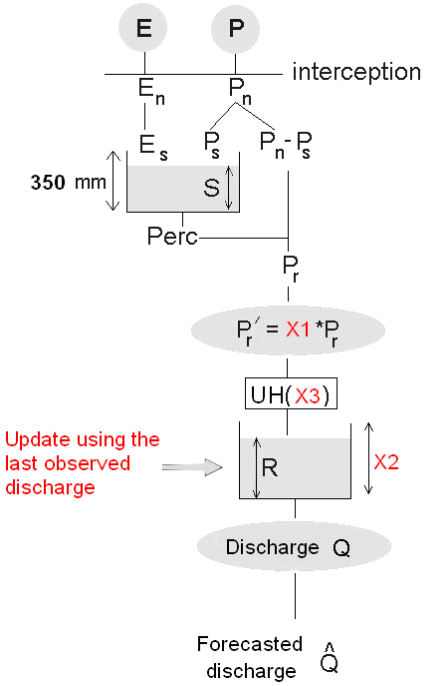


Figure 10. The GRP model structure and its 3 parameters (X1, X2, X3) (Tangara, 2005).

The model is composed of a production function, which computes the effective rainfall over the catchment, and a routing function, including a unit hydrograph and a non-linear routing store, which transforms effective rainfall into flow at the catchment outlet. The GRP model has 3 parameters that need to be calibrated against observed discharge: the first parameter (*X1*) corresponds to a volume-adjustment factor that controls the volume of effective rainfall; the second parameter is the capacity of the quadratic routing store (*X2*); the third parameter (*X3*) is the base time of the unit hydrograph. The maximum capacity of the production store is fixed. For flow forecasting, an updating procedure is applied based on the assimilation of the last observed discharge to update the state of the routing store and a model output correction according to the last model error (Berthet *et al.*, 2009). The Kalman filter -neither another filter- is not used in the model because it leads to performance losses during flood events when it assimilates streamflow alone (Berthet, 2010). The model used in this study runs at daily time steps and only the updating of the routing store is activated. Berthet (2010) shows that the impact of the model output correction is neglectable for time steps beyond 24 hours due to the stronger impact of the update using the last observed discharge.

2.4.2 CALIBRATION OF THE *GRPE* MODEL

The automatic calibration procedure minimizes the **root mean square error** (RMSE; Eq. 1) computed over sets of values of observed and forecasted daily discharges for the first lead-time of one day. Studies conducted at Cemagref showed that parameter values do not vary significantly with lead-time when the model is calibrated at daily time steps and with observed precipitation as "perfect rain forecasts".

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2}$$ Equation 1

Range: 0 to ∞. Optimal score: 0.

Where o_i are the observed values, f_i the forecasted values and n the number of forecasts

Figure 12 illustrates the procedure adopted in the calibration of the model. During the first step of the calibration process, the method uses the daily discharge data available for the catchment from 01.08.1970 up to 31.07.2000 to find the optimum set of parameters. The parameter values are then validated for the period 01.08.2000 to 10.03.2005. If the performance over the validation period is satisfying, the second step of the calibration process is launched. It uses daily discharge data available for the catchment from 01.08.1970 up to the start of the forecast period (11.03.2005) for calibration. These calibrated parameters are then used in the GRPE model to run the forecasting period. This means that the forecast period serves as well as validation period. From the results of model calibration, 3 catchments were taken out of the dataset B1 because the model calibration was not satisfying.

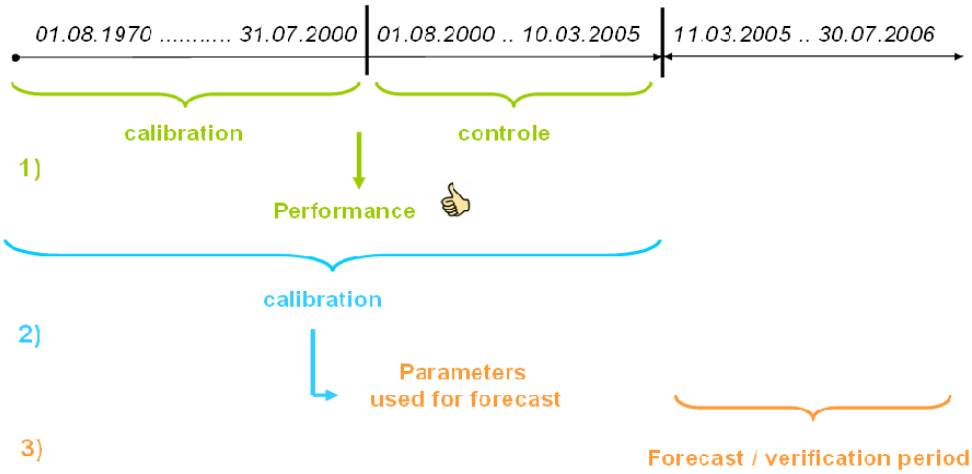


Figure 11. Calibration procedure of the GRPE model adopted in this study.

3 METHODOLOGY

In the problem definition (Chapter 1.4), two kinds of thresholds are distinguished: a *streamflow* threshold and an *ensemble* threshold. The *streamflo*w threshold represents a certain discharge and, if the forecasted discharge is higher than this threshold, a warning is issued. The *ensemble* threshold represents the number of ensemble members (probability) exceeding a certain *streamflow* threshold in order to issue a warning. In paragraph 3.1 the methodological research steps for the evaluation of the *streamflow* threshold are described. Paragraph 3.2 consists of a presentation of the methods used to evaluate the *ensemble* threshold.

3.1 STREAMFLOW THRESHOLDS

The greatest challenge for the *streamflow* threshold is to deal with the agreement between the locally defined (instantaneous) threshold and a threshold adapted to the time step of the model. In this study, hourly discharges are our "instantaneous" data. Therefore, we studied the moments (time steps) hourly discharges exceed the local streamflow threshold and searched for the daily discharges corresponding to each time of exceedance. These discharge values are then analysed to find an *optimal threshold* that optimizes flood warning.

In paragraph 3.1.1, we discuss how the contingency table and its statistical scores are used to study an optimal agreement between the instantaneous (hourly) threshold and a threshold adapted to daily time steps. The empirical frequency distribution (paragraph 3.1.2) allows finding an optimal threshold adjustment for the 75 catchments, by taking into account all exceedances for all catchments. In paragraph 3.1.3, the focus lies on the methodological steps addressing the question if a catchment-specific adjustment factor results in a better performance comparatively to an overall threshold adjustment factor that considers all catchments together. The procedure described is applied to the yellow and the orange thresholds, as well as to the 2-year return period flood for the instantaneous discharge. The red threshold is exceeded only 5 times in 44 catchments during the 10-year evaluation period of this study (1995-2005) and therefore is not part of the analysis.

3.1.1 THE CONTINGENCY TABLE, ITS SCORES AND THE OPTIMAL THRESHOLD

The search for an *optimal* threshold implies that there is no *perfect* threshold and that a tradeoff has to be made. In this report, the contingency table and the scores that can be computed from this table are used to make this tradeoff. In statistics, contingency tables are often used to record and analyse the relationship between two or more variables. Table 1 represents a contingency table suitable for analysing a flood warning system (FWS). To build such a contingency table, thresholds have to be defined for observed and forecasted events: e.g., a flood event is an "observed yes (no)" event if the observed discharge exceeds (does not exceed) a given threshold; a flood event is a "forecasted yes (no)" event if the forecasted discharge exceeds (does not exceed) the given threshold.

Table 1. The contingency table adapted to flood forecasting.

	# Floods observed			
# Floods forecast		Yes	No	Total
	Yes	*Hits*	*False Alarms*	*Forecast yes*
	No	*Misses*	*Correct negatives*	*Forecast no*
	Total	*Observed yes*	*Observed no*	*Total*

There are several statistical scores that can be computed from the contingency table and used to compare forecast methods mutually, e.g. the False Alarm Ratio (FAR), the Probability of detection (POD) and the critical success index (CSI) (WMO, 2007). These main statistical scores are defined as follows:

The **Probability of detection** indicates what fraction of the observed events was correctly forecasted. The POD is sensitive to hits, but ignores false alarms. The POD score is useful for rare events (like floods), but should always be combined with the FAR due to the ignorance of false alarms:

$$POD = \frac{hits}{hits + misses}$$ Equation 2

Range: 0 to 1. Optimal score: 1.

The **False alarm rate** indicates what fraction of the predicted "yes" events actually did not occur:

$$FAR = \frac{false\ \ alarms}{false\ \ alarms + hits}$$ Equation 3

Range: 0 to 1. Optimal score: 0.

The recommended joint use of POD and FAR scores indicate that a tradeoff has to be made among the number of hits, misses and false alarms. The **Critical success index** will take into account hits, false alarms and missed events, and is therefore a more balanced score. It indicates how well the forecast "yes" events did correspond to the observed "yes" events. It is sensitive to hits and penalizes both misses and false alarms.

$$CSI = \frac{hits}{hits + misses + \ false\ \ alarms}$$ Equation 4

Range: 0 to 1. Optimal score: 1.

By considering the slope of the CSI function with respect to POD and FAR, it was demonstrated by Gerapetritis and Pelissier (2004) that equal changes in FAR and POD produce an equal change in CSI when POD = 1 - FAR. When POD is greater than 1 - FAR, CSI is more sensitive to changes in FAR, and when POD is less than 1 - FAR, CSI is more sensitive to changes in POD.

A disadvantage of the CSI score is that it is a biased score that is dependent upon the frequency of the event that is forecasted (Schaeffer, 1990). On one hand, this plays only a role when events with different frequencies are compared, and not when threshold exceedances based on a certain frequency are evaluated. On the other hand this makes it difficult to identify which CSI score is acceptable and which CSI score is not acceptable anymore, since these limits are as well dependent on the frequency of the event.

The CSI does not distinguish the source of error, since both false alarms and misses will be counted together and lead to lowering the score. However, in the case of flood forecasting, since false alarms might have a higher level of acceptance than misses (for instance, in flood pre-warning), it can be useful to make a distinction between false alarms on one side and misses on the other one. To handle this difference in the level of acceptance of false alarms, we introduced a weighting coefficient α. The

false alarms are multiplied by α (ranging from 0-1). Eq. 4 shows the resulting weighted critical success index ($CSI_{(\alpha)}$) used in this study:

$$CSI_{(\alpha)} = \frac{hits}{hits + misses + \alpha * false\ \ alarms} . \qquad \text{Equation 5}$$

Range: 0 to 1. Optimal score: 1.

Both scores, the critical success index and the weighted critical success index, are applied to find an optimal *streamflow* threshold.

3.1.2 EVALUATING 75 CATCHMENTS: EMPIRICAL FREQUENCY DISTRIBUTION

The cumulative empirical frequency distribution (EFD) is used in order to consider threshold exceedances for all 75 catchments of dataset A and have a more statistically robust assessment of the optimal threshold. The empirical frequency distribution is used to compare the distribution of daily discharges to that of instantaneous discharges when a threshold is exceeded. The cumulative distribution curve describes the probability distribution of a variable *X*. For every *x*, the cumulative distribution for *X* is given by:

$$F_X(x) = P(X \leq x) \qquad \text{Equation 6}$$

The first step to construct an empirical frequency distribution is to identify which observed daily mean discharges correspond to observed hourly discharges exceeding the threshold. Hereby, we assume that the hourly discharge represents the instantaneous discharge, since hourly time steps are the smallest time steps available. Another choice is the use of observed daily discharges instead of the simulated daily discharges. In this case, we conduct the analysis on observed discharges only, without including errors from the hydrological model. The impact of the inaccuracy of the hydrological model will be addressed in the evaluation of the *ensemble* thresholds.

Not all the thresholds and corresponding discharges are of the same order of magnitude among catchments. Hence, in order to compare various catchments, we calculate the ***ratio R, which is defined as the ratio between the observed discharge and the value of the threshold considered***.

- Therefore, the ratio R indicates the magnitude of the threshold (non)exceedance: for instance, a R = 3.0 means that the observed discharge is 3 times greater than the threshold, while a R = 0.80 indicates that it corresponds to 80% of the threshold value.
- For the hourly discharges, R is always larger than 1.0, since the time steps considered are selected when hourly discharges exceed the threshold.
- For the daily discharges, R can however be smaller than 1.0. In this case, the hourly discharge is exceeding the threshold, while the mean daily discharge does not exceed the threshold.

The empirical cumulative frequency distribution is constructed by plotting the values for the Ratio R on the x-axis, against their corresponding frequencies on the y-axis. Several formulas have been proposed to compute these frequencies, or plotting positions. In this study, we use the Benard & Bos-Levenbach formula (also known as the Chegodayev's formula; Chow et al., 1988):

$$F_{(i)} = \frac{i - 0.3}{n + 0.4}$$ Equation 7

Range: 0 to 1. Where *i* equals the relative rank of the value and *n* equals the total number of values.

A typical frequency distribution for the ratio R, when considering discharges exceeding the yellow threshold, is shown in Figure 12. The yellow vertical line (threshold line) indicates the ratio R=1.0 (discharge equals the threshold). The graph *Qh_all* shows the frequency distribution of the ratio R for the observed hourly time steps, when considering all catchments. R is always larger than 1.0, since only time steps at which hourly discharges exceeds the yellow threshold are considered. From Figure 12, it can be seen that, in general (over the study period and the catchments), there is only a 5% chance that observed ratios are larger than 2, which means that in only 5% of the occurrences (time steps), the hourly observed discharge is 2 times or more greater than the threshold. The median ratio (the one that is exceeded 50% of times) is equal to about 1.2. The graph *Qd_all* shows the frequency distribution of the ratio R for the corresponding observed daily data. It can be seen that in about 10% of the occurrences, the daily discharge is below the threshold.

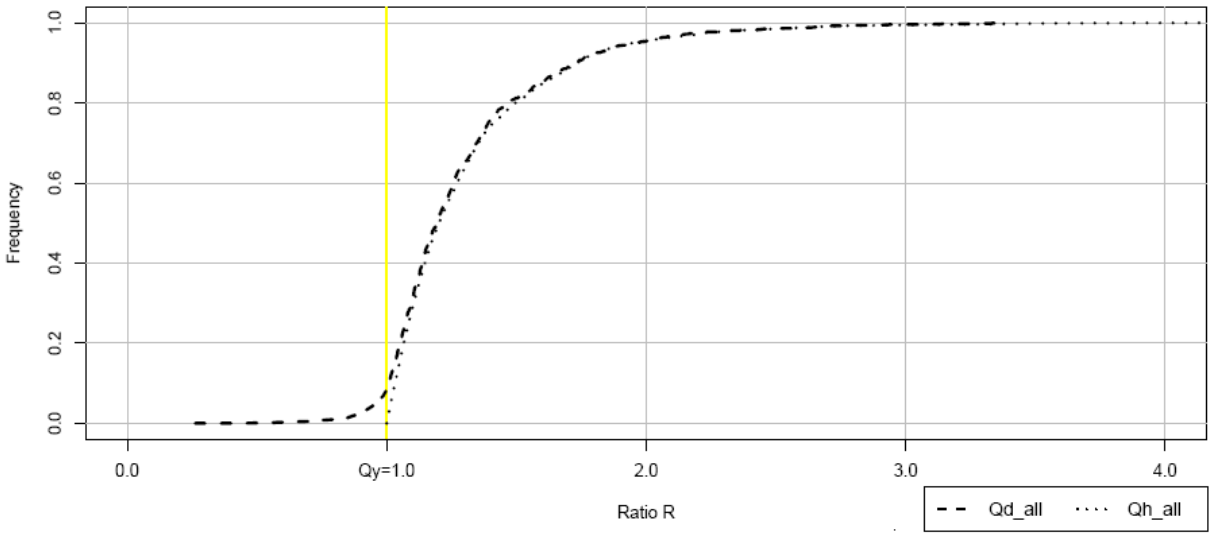


Figure 12 Empirical cumulative frequency of Ratio R values for dataset A: ratio between discharges and the yellow threshold for daily (Qd_all) and hourly (Qh_all) discharges observed when an hourly discharge exceeds the yellow threshold.

An empirical frequency distribution is constructed according to the steps explained above in order to find the optimum tradeoff among hits, misses and false alarms. In fact, the probability of detection (POD) score is directly related to the empirical frequency distribution. The intersection of the threshold (R=1) and the empirical frequency distribution defines the frequency of the misses -events observed but not forecasted- and its counterpart, the frequency of the hits (note that here all events are "observed yes" events, since the yellow threshold is exceeded). The y-value of this intersection is then the frequency of misses (e.g., for the hourly ratio the number of misses, by definition, is zero), and the frequency of hits (the POD score) is given by 1 - the y-value of the intersection.

In the case of the daily ratio R, the intersection between the curve Qd and the yellow line gives the frequency of misses that would be observed if the instantaneous yellow threshold was applied directly to the daily observed discharges. In Figure 12, this intersection is at approximately 0.1 and the POD = 1 – 0.1 = 0.9. In this case, the number of missed events increases and the POD is reduced.

In order to increase the POD and to decrease the number of misses when considering daily discharges, the threshold needs to be lowered before its application to the daily discharges, which means that it has to be multiplied by a factor x (0<x<1). This factor is here defined as the ***daily adjustment factor*** (Equation 8). For instance, from Figure12, a factor of 0.5 needs to be applied to increase the POD score to 1.0 for the daily time steps. In fact, if we shift the yellow threshold line (the vertical line) to 0.5, the y-value of the intersection between the distribution curve Qd and the threshold line becomes 0. This means that the frequency of misses equals 0 and the POD becomes $1-0=1$.

However, a shift to the left of the yellow threshold line, i.e., the use of a lower yellow threshold, can introduce "false alarms" regarding the "instantaneous" discharges. A false alarm here means that the daily discharge exceeds the adjusted (daily) threshold, but at any of the 24 hourly time steps of that day the hourly discharge exceeds the "instantaneous" (hourly) threshold. Therefore, when applying a daily adjustment factor, one has to pay attention that the increase in POD for the daily time step does not increase significantly the FAR when considering smaller time steps (time steps that better translate the local "instantaneous" situation).

$$Q_{\theta\ daily} = x \cdot Q_{\theta\ hourly} \qquad \text{Equation 8}$$

The POD score is useful for rare events (like floods), but should always be combined with the false alarm rate (FAR) due to the ignorance of false alarms in the computation of the POD (WMO, 2007). The FAR has first to be calculated independently from the empirical frequency curve of Ratios R and then it can be introduced in the graphical representation for a combined visualization of POD, FAR and frequency of misses in the same plot. The first step in the evaluation of FAR is to consider a range of (adjusted) daily thresholds: hourly thresholds are multiplied by the adjustment factors of 0.75, 0.80, 0.85, 0.90 and 0.95. The second step is to select the maximum hourly discharges for the days exceeding the (adjusted) daily threshold. Then the FAR is given by the frequency of hourly discharge which does not exceed the instantaneous (yellow) threshold. Applying this approach results in no false alarms for the original threshold, since the maximum hourly discharge of a day is always larger than the threshold (selection criterion).

The results of the POD analysis (number of hits and misses) and FAR analysis (number of hits and false alarms) are the input for the calculation of the Critical Success Index. The CSI is calculated for a number of adjustment factors of the original threshold – thresholds are multiplied by a factor of 0.75, 0.80, 0.85, 0.90 and 0.95 - in order to find the threshold with the maximum CSI score, which is the *optimal* threshold.

3.1.3 OVERALL vs. CATCHMENT-SPECIFIC OPTIMAL THRESHOLD

Figure 12 gives a general view, since the ratios are computed over all the 75 catchments of dataset A. It is however possible to apply the same procedure to every single catchment in order to have a catchment specific adjustment factor that could eventually result in a higher CSI score for that individual catchment. The disadvantage of applying a catchment-specific adjustment factor is certainly the limited number of events (especially for the higher thresholds) per catchment. The advantage is that it allows to evaluate the effect of a limited number of threshold exceedances on the search for the optimal daily threshold and, furthermore, to investigate a possible relationship between the level of adjustment (factor *x*) and the catchment characteristics.

3.2 *ENSEMBLE* THRESHOLD

The main objective in evaluating the *ensemble* threshold is to find the optimal number of ensemble members exceeding the *streamflow* threshold resulting in a maximum preparedness. We aim to address questions like: what is the optimal number of ensemble members exceeding the *streamflow* threshold resulting in the maximum Critical Success Index (CSI)? How many ensemble members exceeding the *streamflow* threshold is enough to launch a warning with a maximum in preparedness (anticipation for a flood event) and an optimal balance between hits, misses and false alarms?

The probabilistic weather forecast used in this study is the ECMWF ensemble forecast (Chapter 2.3), consisting of 51 ensemble members. In paragraph 3.2.1, we describe how the CSI (balance among hits, misses and false alarms) and the optimal *ensemble* threshold are calculated for an ensemble forecast. Paragraph 3.2.2 focuses on the evaluation of preparedness –gain/loss in lead-time compared to the deterministic forecast- related to the optimal *ensemble* threshold. The differences between an overall analysis (all catchments together) and a catchment-specific threshold are the topic of paragraph 3.2.3. The ECMWF ensemble forecasts have a lower spatial resolution (grid size) compared to the deterministic forecasts in the numerical weather model. In paragraph 3.2.4, we explain how the role of grid and catchment size is taken into account in the studies by shifting the focus from the whole set of 208 catchments (dataset B1) to the subset of 29 large catchments (dataset B2).

3.2.1 THE OPTIMAL THRESHOLD: MAXIMIZATION OF THE CRITICAL SUCCESS INDEX

The typical distribution of hits, false alarms and misses for a range of *n=1 to 51* ensemble members exceeding a given streamflow threshold is presented in Figure 13, together with the number of hits, false alarms and misses for a deterministic forecast. It can be seen that the number of false alarms and hits decreases when the *ensemble* threshold (required number of ensemble members exceeding the *streamflow* threshold to forecast an event) decreases, while the number of misses increases. A tradeoff between the number of hits, misses and false alarms can be made by calculating the Critical Success Index for each single *ensemble* threshold (1 to 51). Typically, the CSI score is low for a low *ensemble* threshold due to the large number of false alarms. Furthermore, the CSI is often low for a high *ensemble* threshold due to a relative high number of misses (non detection of observed events) and consequently a low number of hits.

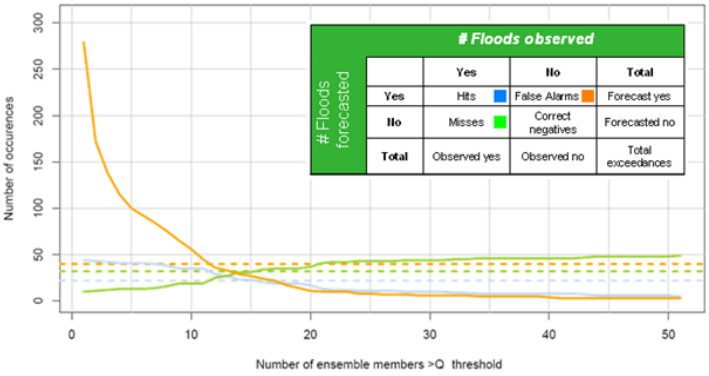


Figure 13. Typical distribution of hits (blue), misses (green) and false alarms (orange) for the range of 1 to 51 ensemble members exceeding a given *streamflow* threshold. The dotted lines represent the corresponding values for the deterministic forecast.

3.2.2 PREPAREDNESS: GAIN / LOSS IN LEAD-TIME

One of the advantages of using a probabilistic forecast method is the potential to extend the lead-time, which is given by the lag between the time an event is forecasted and the time it is observed (Cloke and Pappenberger, 2009; Clark and Hay, 2004; Buizza, 2002). The lead-time of a flood warning is an optimal balance among hits, false alarms and misses: a forecaster can decide to wait the forecasted event to approach (and thus decrease the lead-time for preparedness) to be more certain of his/her forecast, or decide to issue a warning earlier by accepting the possibility of a false alarm. Increased preparedness is also an indicator of success (and eventually usefulness) of a forecasting system: a successful flood warning (hit) which is issued 9 days in advance is more valuable than a successful warning with a lead-time of 1 day.

Despite the recognition of the potential usefulness of ensemble predictions to extend lead-time, comparatively to deterministic forecasts, literature does not provide a rigorous approach to objectively quantify this advantage. An evaluation was recently proposed by Ramos *et al.* 2009, but without considering impacts on forecast performance in terms of hits, misses and false alarms. In this study, a preparedness score is developed to compare lead-times between ensemble streamflow predictions and deterministic forecasts and a combined framework is used, where the CSI score is also taken into account.

The difference in preparedness (ΔP) is given by subtracting the deterministic lead-time (LT_{DET}) from the ensemble lead-time (LT_{ENS}) for all days with an observed threshold exceedance and is represented by the formula:

$$\Delta P = LT_{ENS} - LT_{DET} \qquad \text{Equation 9}$$

with LT_{DET} equals the number of days in advance where the deterministic forecast exceeds the *streamflow* threshold and LT_{ENS} equals the number of days in advance where the ensemble forecast exceeds the *streamflow* threshold with a specified number of members.

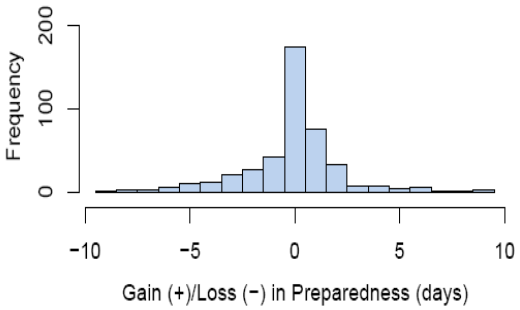


Figure 14. Preparedness histogram.

The results can be presented in a histogram with categories ranging from -9 to +9 days (Figure 14). Events on the negative side represent a loss in lead-time using the ensemble forecast compared to the deterministic forecast. The bars on the positive side represent a gain in lead-time by using the ensemble forecast. When ΔP equals zero, events have the same lead-time in both forecasts (including events with a lead time of zero, i.e. misses) for all observed exceedances.

It should be noted the preparedness score itself only focuses on the gain/loss in lead time. It has to be combined with the CSI in order to associate the number of false alarms, hits, and misses to its value. In this study, the preparedness is calculated for two probabilities/ensemble thresholds (number of ensemble members exceeding the *streamflow* threshold): 1) the probability corresponding to the

maximum CSI, and 2) the probability at which the CSI score of the ensemble forecast equals the CSI score of the ECMWF deterministic forecast. In order to compare various catchments and the results obtained by using different *streamflow* thresholds (resulting in different number of exceedances per catchment), we calculate the mean loss/gain per observed flood event.

3.2.3 OVERALL vs. CATCHMENT SPECIFIC OPTIMAL THRESHOLD

The same procedure is applied to every catchment individually in order to evaluate if there is a higher gain in lead-time for each individual catchment when catchment-specific ensemble thresholds are used. As already mentioned (paragraph 3.1.3), the disadvantage of carrying out a catchment-specific analysis is the limited number of events (especially for the higher thresholds) per catchment.

3.2.4 THE RELIABILITY DIAGRAM

The grid of ECMWF weather forecasts available for this study is approximately 45 x 45 km over France, while the smallest catchment in dataset B1 covers 39 km^2. The difference between the grid size of the weather forecast and the catchment size can have an impact on the reliability of the streamflow forecasts. The reliability diagram is often used in the verification of meteorological ensemble forecasts (e.g., van der Grijn, 2002) and is a useful tool for evaluating thresholds in ensemble forecasts. Olsson and Lindström (2008) and Renner *et al*. (2009) demonstrate the usefulness and applicability of the reliability diagram for the threshold-based evaluation of hydrological ensemble forecasts.

In the evaluation of a reliability diagram, the forecasted probability of exceeding a threshold is compared with the observed frequency of exceeding the same threshold over a set of probability classes (Figure 16(a)). The aim is to answer the question: How well do the predicted probabilities of an event correspond to their observed frequencies? The reliability diagram is applied in this study to indentify the sources of errors and the differences, if any, between the ensemble streamflow predictions' reliability based on the 29 large catchments (Dataset B2) and the ensemble prediction reliability based on the 208 catchments (Dataset B1, including smaller catchments).

In order to perform a more robust analysis (especially for the higher thresholds corresponding to rarer events), the ensemble predictions are transformed into a more comprehensive and legible representation of the forecasted probabilities by using the following percentiles (Figure 15): the minimum (2% forecast), the lower quartile (25%), the median (50%), the upper quartile (75%) and the maximum (98%).

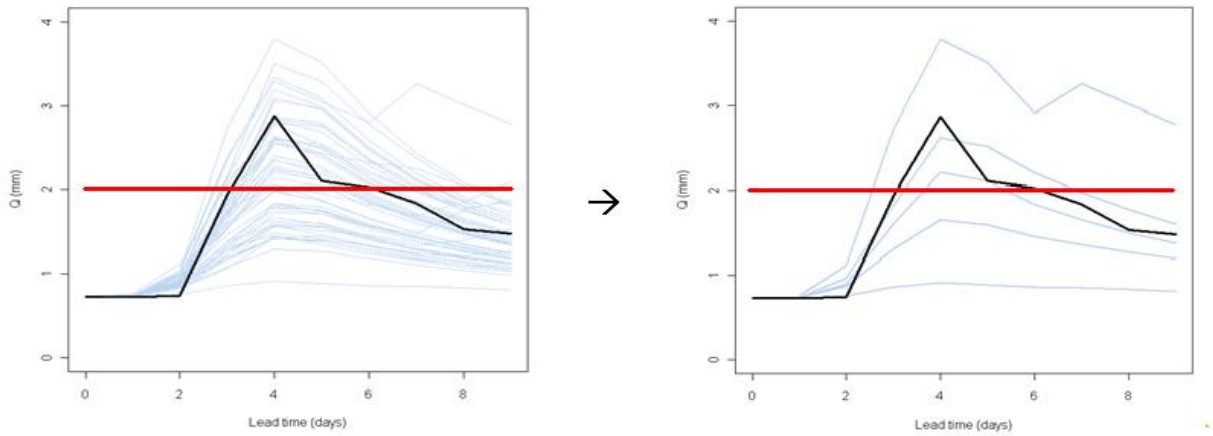


Figure 15. Transformation of the ensemble predictions into probability classes based on the 2%, 25%, 50%, 75% and 98% percentiles.

Therefore, for instance, in the case that, for a given forecast, the threshold (θ) is located between the lower quartile (x1=0.25) and the median (x2=0.5), as is the case illustrated in Figure 15 for the lead time of 4 days, the forecasted probability of exceeding this threshold is given by:

$$P_{(Q>\theta)} = 1 - \frac{1}{2} * (x1 + x2) \qquad \text{Equation 10}$$

where the assumption is made that 0.375 (the mean of x1=0.25 and x2=0.50), which corresponds to 19.5 ensemble members exceeding the threshold, equals the mean probability of events in this bin. The probability of a threshold exceedance equals in this case 1 – 0.375 = 0.625.

The percentiles defined above will then give the 6 probability classes that will be represented graphically in the reliability diagram: < 2%, between 2% and 25%, 25%-50%, 50%-75%, 75%-98%, and < 98%.

To carry out a reliability diagram analysis, it is also necessary to choose the event to consider (i.e., the *streamflow* threshold). In this study, the thresholds θ = Q70, Q90, Q95 and Q99 (percentiles 70%, 90%, 95% and 99% of the observed discharges, respectively) were considered. Since we are working with an archive of only 18 months and relatively high thresholds, we aggregated the results over all lead-times to have enough events per bin and to draw more robust conclusions.

In the reliability diagram, the forecasted probabilities are plotted on the x-axis and the observed frequencies on the y-axis. A perfect reliable forecast is the one where the forecasted probabilities equal the frequencies of observed exceedances, i.e., the one that falls in the line y=x (Figure 16(a)). If points fall below (above) the diagonal, the forecast system has a tendency to over(under)-forecast.

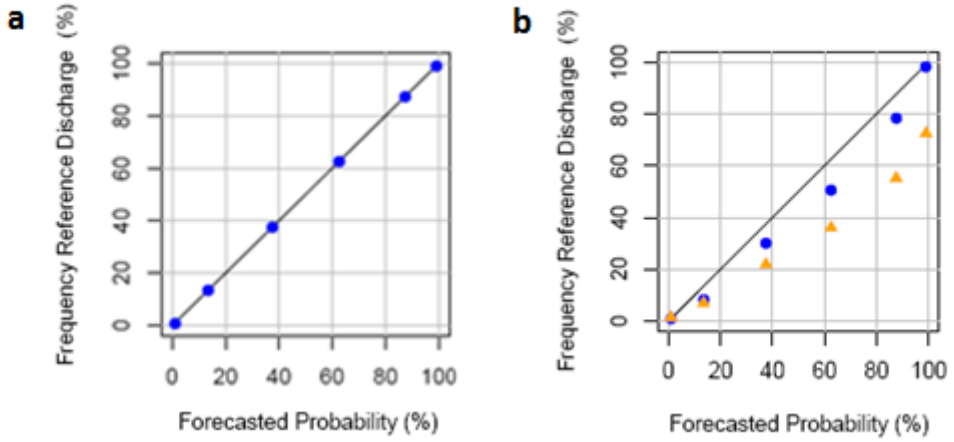


Figure 16(a). Reliability diagram showing the 6 probability classes (mean values: 1, 13.5, 37.5, 62,5 86.5, and 99 %) (b) Reliability diagram showing the 6 probability classes for the two reference discharges used in this study (observed frequencies in triangles and proxy-observed frequencies in dots).

When evaluating forecasts against observations, errors from both the meteorological forecast and the hydrological model are present in the evaluation. A useful procedure to separate these error components is to evaluate the forecasts also against a "proxy-observed" reference discharge, which consists of simulated discharges based on the observed amounts of precipitation (e.g., Olsson and Lindström, 2008). In this case, the proxy-observed discharges contain only the hydrological modeling error. According to Pappenberger *et al.* (2005), the hydrological model is the second major source of uncertainty involved in flood forecasting after the meteorological forecast data. Using both observed and simulated "perfect forecast" discharges allows indentifying and quantifying the magnitude of these uncertainty sources in the analysis.

Figure 16 (b) shows an example of a reliability diagram that compares forecast probabilities against relative frequencies from the two reference discharges: the proxy-observed (blue points) and the observed (orange triangles) reference discharge. We assume that observations (precipitation and discharge) are free of errors. The reliability diagram can be interpreted as follows:

- It is assumed here that bias come from two sources of errors: errors from the meteorological forecast and errors from the hydrological model (observations are assumed to be error-free).
- For the two references used:
  - Observed reference discharge (**orange triangles**): gives an indication of the bias of the streamflow probabilistic forecast system, when considering errors from the meteorological forecasts and the hydrological model.
  - Proxy-observed reference discharge (**blue dots**): gives an indication of the bias of the streamflow probabilistic forecast system, when considering errors from the meteorological forecasts only (since the reference discharge used for the observed frequencies already includes the errors of the hydrological model).
- When comparing plots, the distances between plots and between plots and the diagonal will give information on the relative importance, on average, of the errors of the hydrological model comparatively to the errors of the meteorological forecasts for the attribute of “reliability:
  - If the distance between **orange triangles** and **blue dots** is equal or close to zero, then the errors of the hydrological model can be considered negligible, and the errors of the meteorological forecasts can be considered to play a more important role in the reliability analysis of the forecasting system.

- If the distance between **orange triangles** and **blue dots** is greater than the distance between the blue dots and the diagonal, it can be an indication that the errors of the hydrological model play a more important role in the reliability analysis of the forecasting system.
- Finally, if the distance between **orange triangles** and **blue dots** is smaller than the distance between the blue dots and the diagonal, it can be an indication that the errors of the meteorological forecasts play a more important role in the reliability analysis of the forecasting system.

4 RESULTS I: *STREAMFLOW* THRESHOLDS

The challenge here is to address the frequency of exceedances of a locally-defined instantaneous threshold and find an optimal threshold adapted to the time step of the model used in forecasting. As explained in chapter 3.1, this is done by evaluating the exceedances of certain thresholds from hourly and daily discharge time series. The hourly discharges in this case represent the instantaneous discharges, as they are the discharges with the smallest time step available in our archive. The daily discharges refer to the time step applied in the flood forecasting model used here and in a pre-warning project piloted by the French national forecasting center. The yellow operational threshold, the 2-year-return-period flood threshold and the orange operational threshold –which all are based on the instantaneous observations- are adjusted by a daily adjustment factor *x* in order to be implemented in the flood warning system operating with a hydrological model running at daily time steps. Furthermore, from the analysis of a number of catchment descriptors, it is evaluated if the optimal adjustment factor should be catchment-specific or if an overall adjustment factor results in the same expected quality of flood warnings.

4.1 YELLOW OPERATIONAL *STREAMFLOW* THRESHOLD

4.1.1 OPTIMUM THRESHOLD: INTRODUCING THE START OF A FLOOD EVENT

Figure 17(left) shows the empirical frequency distributions of the ratio R between hourly and daily discharges for all time steps where the hourly discharge is exceeding the yellow threshold and for all catchments. From this figure, it can be seen, for instance, that in 20% of the time, hourly and daily discharges are of a magnitude of about 1.5 times greater that the threshold. Also, it can be seen that in only 9% of the occurrences the daily discharges are smaller than the yellow threshold: the relative frequency of daily ratios not exceeding the unity equals 0.09 (Freq (ratio ≤1)=0.09, red horizontal line in Figure 17). This means that 91% of the daily discharges, corresponding to days when "instantaneous" discharges exceed the yellow threshold, exceed the yellow threshold as well. In the case that the yellow threshold, based on instantaneous observations, is applied directly in the daily forecast model, 91% of the times a forecasted exceedance would be detected, supposing also a perfect -error free- forecasting model.

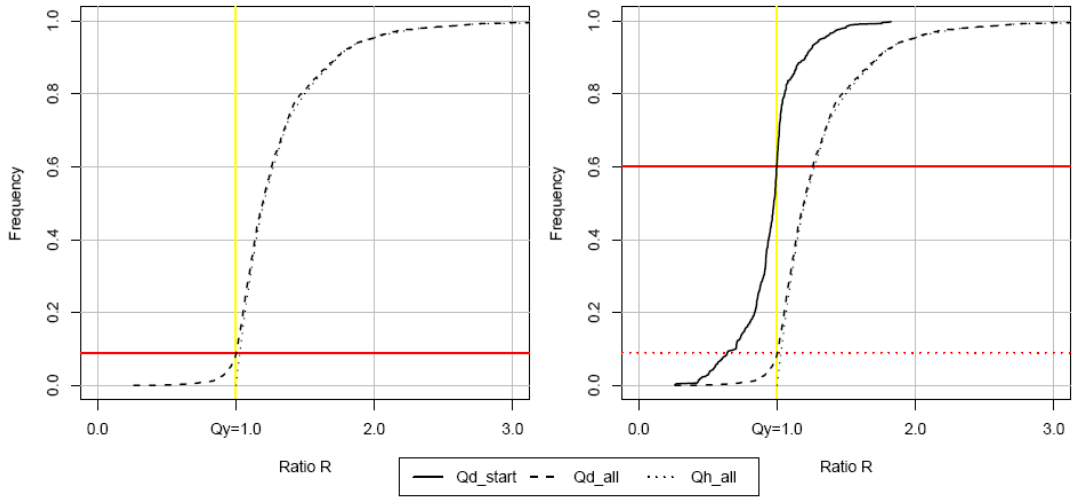


Figure 17. The empirical frequency distribution for all days (left) and including the start of a flood event (right) showing the cumulative frequency distribution of the ratios R between hourly (Qh) or daily (Qd) discharges and the yellow operational threshold for all catchments in Dataset A.

Having operational flood warning in mind, some questions remain: when is the 9% of non-detection occurring? What is the impact of these threshold non-exceedances (i.e., misses)? To answer these questions, we examined the hydrographs of the catchments. They showed two situations where the daily discharge is often below the threshold, while at least one of the instantaneous discharges is exceeding it, namely: 1) at the first day of a flood event and 2) at the last day of a flood event. A flood event is in this case defined as the period between the first and last time step with instantaneous discharges larger than the threshold (in this case the yellow threshold). For operational flood forecasting and warning, the start of a flood event is important information. A better knowledge of the first moment of a flood threshold exceedance can after all lead to better anticipation and a reduction of the impact of a flood event.

In order to quantify the impact of non-detection of the start of an event, we shift the focus of our analyses to the start of the flood event in the remaining part of this chapter. Figure 17 (right) reproduces Figure 17 (left), but adding the empirical frequency distribution for the Ratio R between daily discharges and the yellow threshold when only the days corresponding to the start of a flood event are considered in the analysis (i.e., a day is included in the analysis only if it is the day containing the first hourly time step when the hourly discharge exceeds the yellow threshold).

The frequency distribution shifts to the left for the start of a flood event, indicating smaller ratios R for equal frequencies: for instance, for the start of the event, in 20% of the time, daily discharges are of a magnitude of about only 1.1 times greater that the threshold (comparatively to 1.5 times greater observed in the Figure 17, left). The frequency of non-exceedance of the yellow threshold is now 0.60, compared to 0.09 when all time steps, and not only the start of the event, are considered. The POD thus decreases from 0.91 to 0.40 for the start of a flood event. This means that only 40% of the daily discharges exceed the instantaneous yellow threshold at the start of a flood event. It shows as well that on average up to 60% of the start-events are not detected when the yellow threshold is applied directly to the daily discharges. The need of lowering the value of the hourly-based *streamflow* threshold in order to find a better daily-based threshold to apply to daily discharge forecasts is highlighted. A better daily threshold means a threshold with a higher POD, without a significant increased of the number of false alarms.

Figure 18 shows the effects of adjusting the threshold by a factor $x < 1$ on the probability of detection (POD) and on the false alarm rate (FAR). The case of $x = 0.95$ is highlighted in red for the effects on the POD (Figure 18, left) and in blue for the effects on the FAR (Figure 18, right). It can be seen that lowering the threshold (i.e., moving the yellow threshold (vertical) line to the left, from 1.0 to 0.95) results in a higher probability of detection for the start of a flood event: the POD increases from 0.40 to 0.56 (Figure 18, left). The blue curve in figure 18 (right) represents the false alarm rate (values also on the frequency y-axis). It shows the fraction of the predicted "yes" events that actually did not occur (chapter 3.1.2.). Lowering the threshold for the daily discharges results in increased FAR, because events with an hourly discharge smaller than the original yellow threshold are included in the computation of the contingency table. For example, when $Q_{\theta\ yellow}$ is multiplied by 0.95, FAR becomes equal to 0.12 (12% of false alarms).

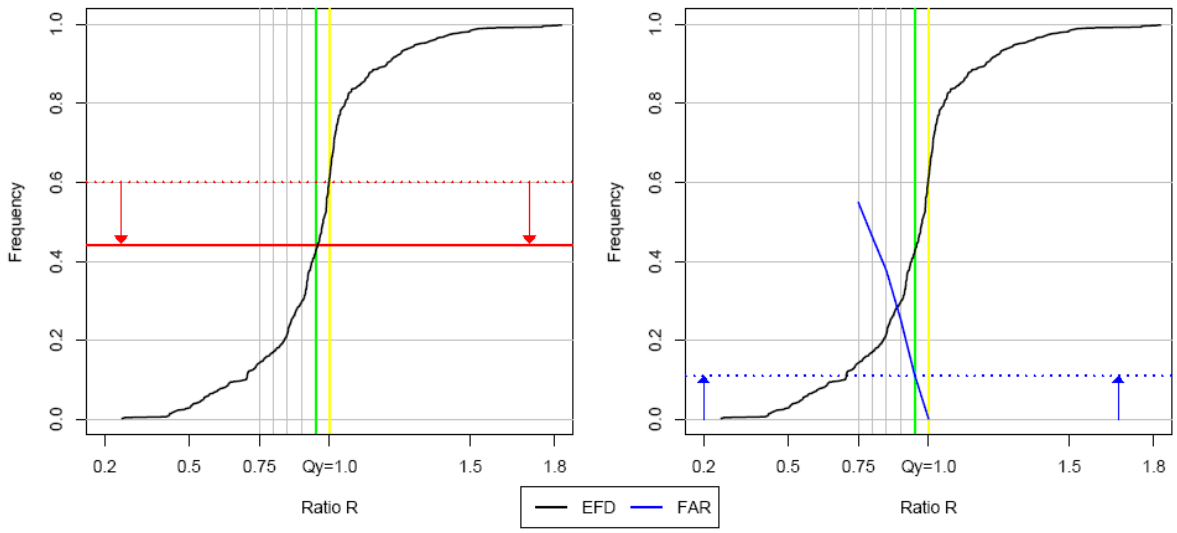


Figure 18. Effects of the increasing of the probability of detection (left, in red) and increasing of the false alarm rate (right, in blue) when an adjustment factor is applied to define a daily threshold: the case of an adjustment factor of 0.95 (green vertical line) is indicated by the arrows.

The tradeoff between POD and FAR, to find the best adjustment factor and hence the optimal daily *streamflow* threshold, is made with the help of the CSI score. Figure 19 shows again the empirical frequency distribution (EFD) of the ratios computed for the start of the flood event (curve in dark black) and four adjustment factors, 0.75, 0.80, 0.85, 0.90 and 0.95 (lines in light black). For each adjustment factor, the POD is given by the difference between the upper horizontal line 1.0 and the point at which the vertical line of the adjustment factor intercepts the EFD curve. The FAR is computed for all adjustment factors and represented in the graph (in blue). The Critical Success Index (CSI), a tradeoff between hits, misses and false alarms, is also represented (in red). The optimal adjustment factor is defined as the one associated with the maximum CSI score. From our results, its value is 0.90 and the optimal threshold for the daily time steps is equal to 0.90 times the instantaneous yellow threshold (Eq. 11):

$$Q_{\theta\ daily} = 0.90 \cdot Q_{\theta\ yellow} \ . \qquad \text{Equation 11}$$

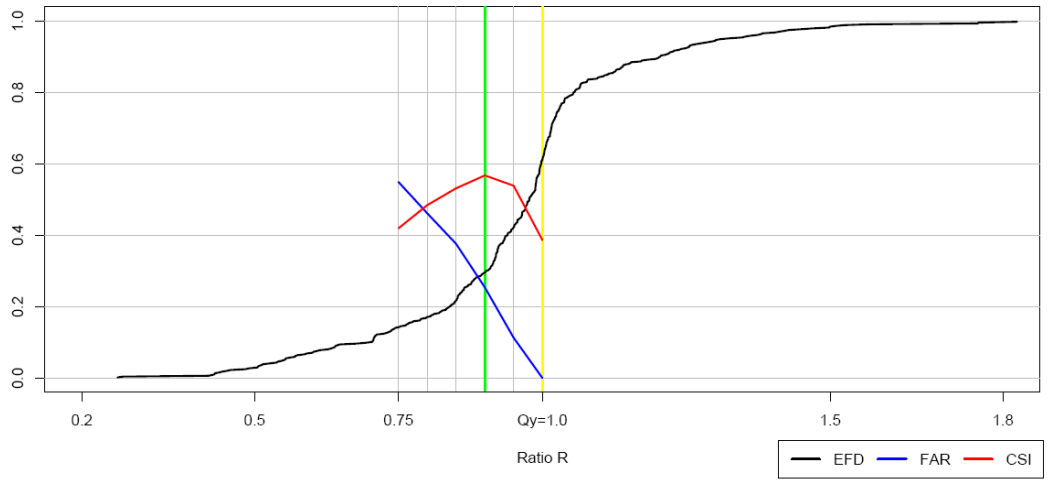


Figure 19. The critical success index indicating an optimal daily adjustment factor of 0.90 in the frequency distribution of the ratio R for discharges exceeding the yellow *streamflow* threshold.

The CSI represented in Figure 19 gives the same weight for false alarms and misses, both leading to lower scores. In practice, depending on the aims of the operational forecasting system and on the situation being forecasted, false alarms might have a higher level of acceptance than misses. Hence, the weighting coefficient α, which addresses this difference in the level of acceptance, was introduced in chapter 3.1.1. In the weighted CSI the false alarms are multiplied by α (ranging from 0-1). Using α<1 reduces the weight of the false alarms in the CSI and moves the optimal daily threshold to the left, as can be seen in Figure 20, where the CSI weighted by α = 0.5 is shown. $CSI_{(\alpha)}$ were computed for various values of α, resulting in different optimal values of the daily adjustment factor *x*, which are shown in Table 2. For example, α = 0.5 (green vertical line in Figure 20) results in the optimal daily threshold adjustment factor of 0.85. It is to the forecaster to decide on the level of acceptance to adopt: if it is decided to adopt α = 0.5, which means that two false alarms are accepted at the same level of one missed event, the forecaster can use the adjustment factor 0.85 and increase the probability of detection from 70% (case where α = 1.0 in Figure 19) to almost 80% (Figure 20). The upper limit is the case α = 0, where maximizing the CSI becomes in fact the same as maximizing the POD, without taking account of false alarms, which means moving the vertical threshold line up to the lowest ratio R of the empirical frequency distribution.

Table 2. Optimal daily adjustment factor for different values of α.

α in CSI(α)	Optimal x in: Qθ daily = x Qθ yellow
1.00 = CSI	0.90
0.75	0.90
0.50	0.85
0.25	0.80
0.00	0.30

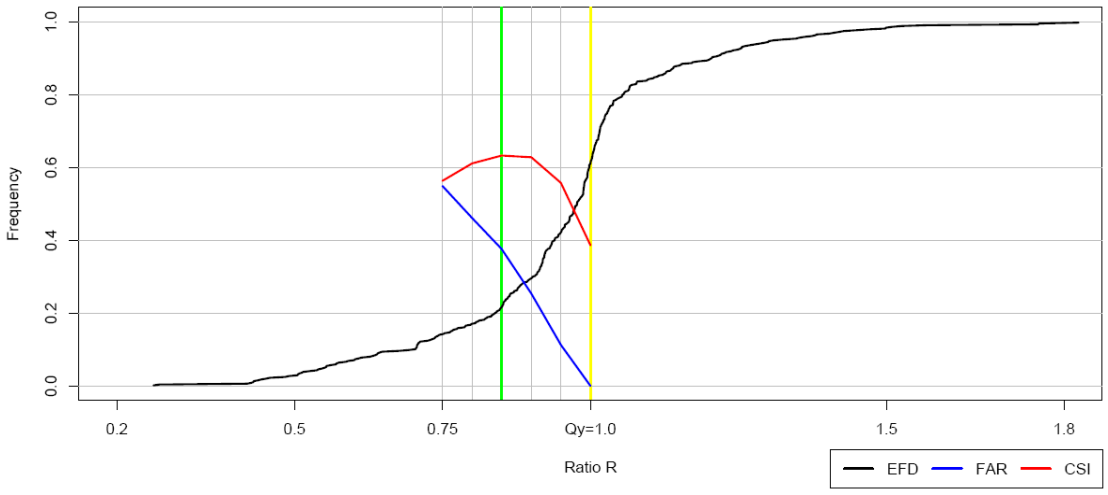


Figure 20. The weighted critical success index (α=0.5) indicating an optimal daily adjustment factor of 0.85 in the frequency distribution of the ratio R for discharges exceeding the yellow *streamflow* threshold.

4.1.2 CORRELATION WITH CATCHMENT CHARACTERISTICS

The analysis presented above considered together the exceedances of the yellow thresholds for all catchments in dataset A that have a yellow threshold defined. The same analysis was also conducted

for each single catchment. In this case, the number of flood events (consecutive period of exceedance of the yellow threshold) observed during the 10-year period studied is limited for most of the catchments (ranging from 0, the threshold is never exceeded, to 41; the catchments with no exceedances of the yellow threshold being, of course, excluded from the analysis). As a consequence, the empirical frequency distribution curves are based on a smaller number of exceedances and are not as smooth as the one for the dataset off all catchments together.

The optimal adjustment factors for the catchments with a yellow threshold and the number of exceedances are included in appendix 0. The map in figure 21 shows the geographic location of the optimal adjustment factors evaluated.

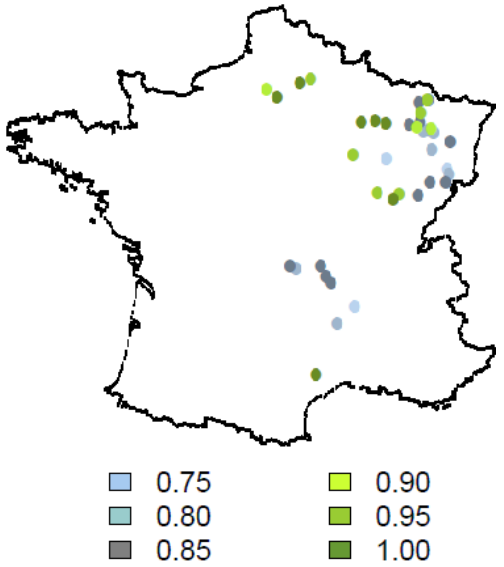


Figure 21. Location of the catchment-specific daily adjustment factors for the yellow threshold.

Despite the small number of events per catchment, a relationship can be detected between the location of the catchment and its optimal daily adjustment factor: smaller values are found for the catchments in the more mountainous areas. These catchments are typically characterised by a relative small area and large slope. The adjustment factor for these catchments is in most cases smaller than the *overall* adjustment factor of 0.90 as calculated in the previous paragraph. This indicates that a *catchment specific* adjustment factor leads to a better conversion of the instantaneous thresholds to the daily thresholds than applying an *overall* adjustment factor.

The box plots of Figure 22 show the correlation with catchment size (topographical upstream drainage area) and catchment's reactivity (based on the average time between the exceedance of the Q50 threshold and the corresponding peak discharge). The catchments are divided into four classes in order to construct these box plots. The first catchment area (reactivity) class consists of the 25 percent of catchments with the smallest area (highest reactivity), the second and third classes consist of the following groups of 25%, while the last class consists of the remaining 25% of catchments with the largest area (slowest reactivity). Correlation with other available catchment descriptors, as typical return period discharges, degree of vegetation, precipitation and evatransporation rates, productivity, etc., was also evaluated (not shown), but no evidence of a correlation link was identified.

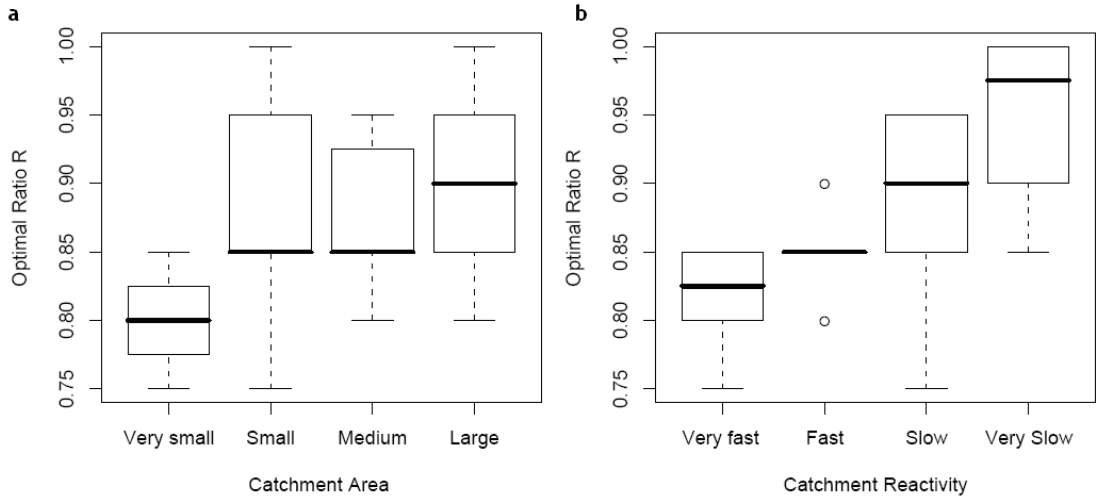


Figure 22. Correlation between the daily adjustment factor (optimal ratio R) and catchment size (a) and reactivity (b). Each class contains 25% of catchments. The top and the bottom of the box represent the 75th and the 25th percentile, respectively, while the top and the bottom of the tail indicate the 95th and the 5th percentile, respectively. The thick horizontal line is the median value.

The median value of the adjustment factor is smaller for smaller and faster catchments. It might be due to the fact that the rising limb of the hydrograph is higher and steeper for these catchments compared to larger and flatter catchments. Higher and steeper rising limbs might result in a larger difference between the daily and the maximum hourly discharge. Subramanya (2006) indentifies several ways how the shape of a hydrograph is controlled by the basin and the storm characteristics. The predominance of overland flow over channel flow in small, mountainous catchments influences the time base and magnitude of the peak discharges in smaller catchments. Besides this, higher catchment reactivity results in steeper rising limbs, especially in smaller and steeper catchments, where the overland flow is dominant during larger peak discharges. Slower catchments (low reactivity) might be the result of a higher drainage density, which is defined as the ratio of the total channel length to the total drainage area. One of the characteristic of catchments with smaller drainage densities is a slower rising limb. Furthermore, the intensity of the rainfall can also affect the shape (rising limb and peak flow) of hydrographs in (very) small catchments. Catchment and climatic characteristics influence the flow pattern of a catchment and seem to influence also the degree of threshold adjustment, as shown in Figure 22. The interplay of catchment and climatic (intense, orographic precipitation) characteristics can explain why the very small and fast catchments distinguish themselves more clearly in our analysis.

4.2 THE 2-YEAR RETURN PERIOD FLOOD

The yellow threshold is relatively close to the 2-year return period flood for most catchments (Chapter 2.1.1). In order to conduct an analysis including all the 75 catchments and to address the second challenge –to define a daily *streamflow* threshold for catchments without a yellow threshold - the analysis described in paragraph 4.1 was as well conducted for a threshold corresponding to the 2-year return period flood.

4.2.1 OPTIMAL STREAMFLOW THRESHOLD

The optimal daily adjustment factor is found by applying the critical success index and the weighted critical success index (as described in Chapter 3.1). The results are shown in Table 3 for the overall (all catchments together) analysis. As mentioned previously, only the time steps corresponding to the start of the flood event are considered.

Table 3. Optimal daily adjustment factor for different values of α Optimal daily adjustment factor for different values of α and for exceedances of the 2-year return period flood.

α in CSI(α)	Optimal x in: Qθ daily = x Qθ $Q_{ix\ 2\ yr}$
1.00 = CSI	0.90
0.75	0.90
0.50	0.90
0.25	0.80
0.00	0.30

4.2.2 CORRELATION WITH CATCHMENT CHARACTERISTICS

The catchment-specific optimal daily adjustment factor for the 2-year return period and its correlation with the catchment characteristics are shown in Figure 23 and Figure 24. They can be compared with the results presented in paragraph 4.1.2. Figure 23 shows that the daily adjustment factor (*x*) is in general relatively small for the more mountainous catchments and larger for the catchments with less relief. The optimal adjustment factors for the catchments with a 2-year return period threshold and the number of exceedances are included in appendix 0. It should be noted that the 2-year flood daily adjustment factor for most of the catchments with a defined yellow threshold is not exactly the same as the adjustment factor defined by the analysis of the yellow threshold: on average they differ of ±0.05, which, as the analyses are carried out for adjustment factors separated by 0.05 units, equals the accuracy of the method. The differences observed can also be explained by the fact that the 2-year return period is only an approximation of the yellow threshold. Local circumstances and vulnerability can be the reason for the differences between these thresholds.

Figure 24 shows that the median value of the adjustment factor is smaller for smaller and faster catchments, following the same tendency as observed in the analysis of the yellow streamflow threshold, although with less distinctive distributions.

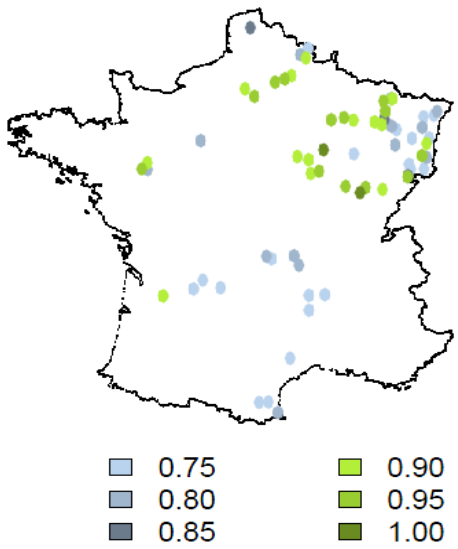


Figure 23. Location of the catchment-specific daily adjustment factors for the $Q_{ix\ 2\ yr}$ threshold.

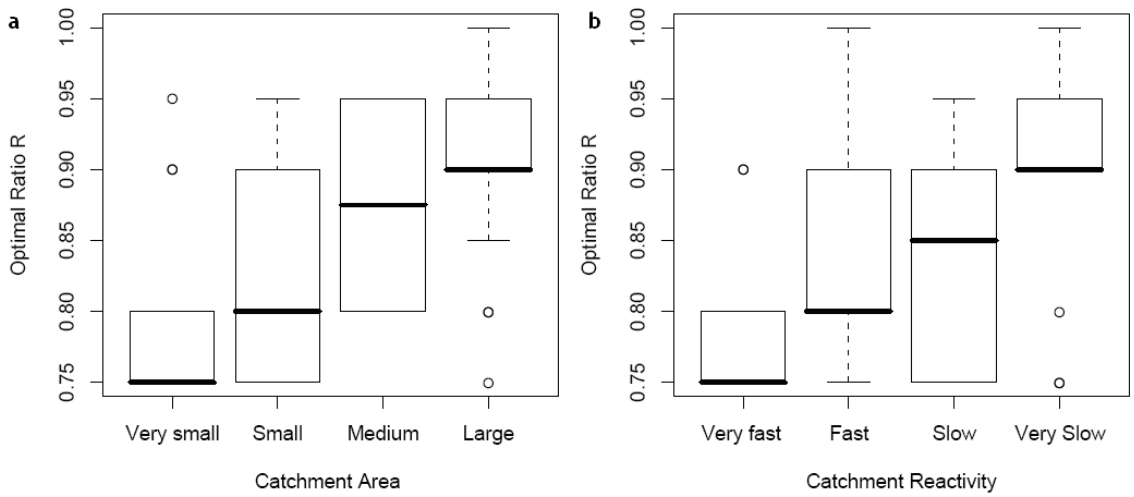


Figure 24. Correlation between the daily adjustment factor (optimal ratio R) and catchment size (a) and reactivity (b). Each class contains 25% of catchments. The top and the bottom of the box represent the 75th and the 25th percentile, respectively, while the top and the bottom of the tail indicate the 95th and the 5th percentile, respectively. The thick horizontal line is the median value..

4.3 HIGHER *STREAMFLOW* THRESHOLDS

The French flood warning system has three defined warning levels: yellow, orange and red. The orange and red thresholds were also evaluated according to the same procedure as described in the previous paragraphs. However, there are only a limited number of threshold exceedances for these warning levels due to the fact that they are related to rarer events with a higher return period. For most of the catchments, the orange threshold equals a discharge which is located between 2 and 5 years of return period, while the red threshold is often associated with return periods between 10 and 20 years. The number of exceedances for these thresholds is too small during the study-period: only 12 catchments have more than 2 orange threshold exceedances and a total of 6 events only is recorded for exceedances of the red threshold. Hence, a robust empirical frequency distribution could not be evaluated for the red threshold. However, in order to assess the impact of increased thresholds on the definition of the daily adjustment factor, we also considered, for each catchment, thresholds defined by the percentiles $Q9O_{ix}$, $Q95_{ix}$ and $Q99_{ix}$. In most cases these thresholds are lower than the corresponding yellow threshold. They represent instantaneous discharges of 90%, 95% and 99% of probability of non-exceedance over the 10 year period used in this analysis.

4.3.1 THE EMPIRICAL FREQUENCY DISTRIBUTION

The empirical frequency distribution for the orange threshold (Figure 25: graph 5) shows a probability of detection of only 0.33 for the start of a flood event (intersection of the empirical frequency distribution with the vertical line at y = 0.67). The probability of detection for the start of a flood event exceeding the yellow threshold equals 0.40 (Figure 25: graph 4; reproduced from paragraph 4.1.1). The empirical distribution curves for the percentiles $Q9O_{ix}$ (graph 1), $Q95_{ix}$ (graph 2) and $Q99_{ix}$ (graph 3) streamflow thresholds are also plotted in Figure 25. These three empirical frequency distributions support the hypothesis that POD decreases when the streamflow threshold increases. This result indicates that the role of the daily adjustment factor is even more important when considering higher discharges (detection of rarer events).

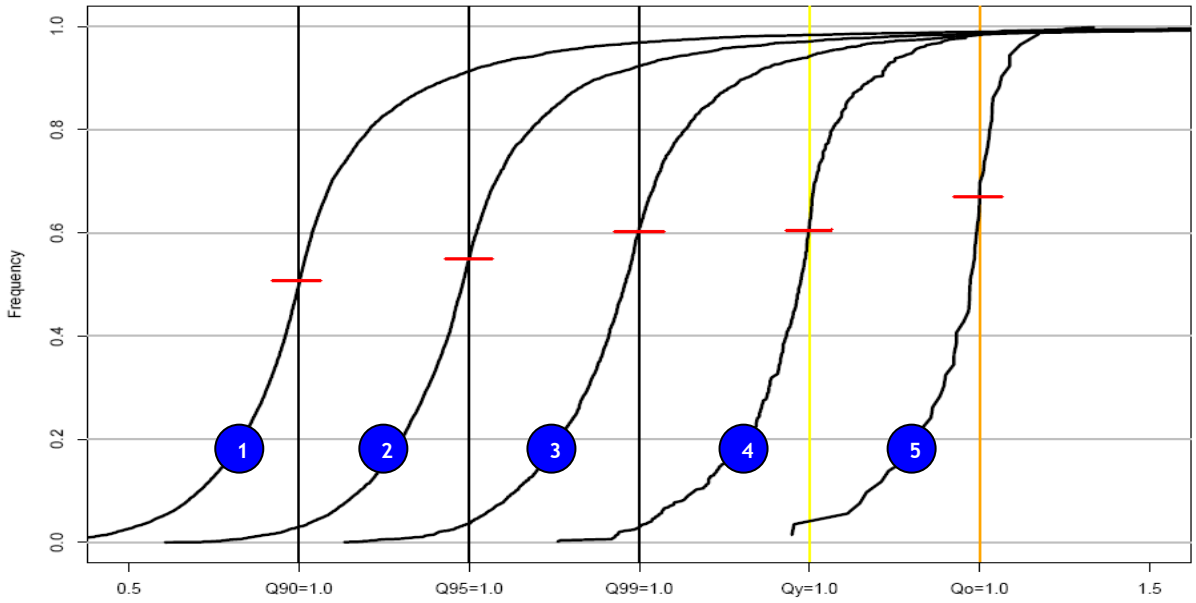


Figure 25. Empirical frequency distributions of the ratios R between daily discharges and the $Q9O_{ix}$ (1), $Q95_{x}$ (2), $Q99_{ix}$ (3), yellow (4) and orange (5) hourly streamflow thresholds for all catchments in dataset A.

4.3.2 OPTIMAL *STREAMFLOW* THRESHOLD

The CSI analysis is conducted for the orange threshold. We performed only an overall analysis and did not address the question if a catchment-specific adjustment would result in better performance due to the limited number of threshold exceedances and to the fact that the catchments concerned by this analysis do not represent the large range of hydro-climatic and geographic conditions of our dataset (most of the catchments with exceedances of the orange threshold are located in low altitude areas and are relatively large in size). Another remarkable point is the corresponding return period for the orange threshold in these catchments: it is relatively low compared to the other orange thresholds in the selection, which maybe explain the fact that threshold exceedances were recorded for these catchments. The vulnerability of these areas could be the underlying reason. The optimal daily adjustment factor of 0.95 (Table 4), found in our analysis for α=1.0, is therefore only valid for these catchments where threshold exceedances did occur and should be applied with caution.

Table 4. Optimal daily adjustment factors for different values of α and the orange streamflow threshold.

α in CSI(α)	Optimal x in: Qθ daily = x Qθ orange
1.00 = CSI	0.95
0.75	0.95
0.50	0.85
0.25	0.85
0.00	0.30

4.4 VALIDATION OF THE DAILY ADJUSTMENT FACTORS

The values found for the daily adjustment factors come from the analysis of a 10-year study period. Since we are investigating threshold exceedances, we needed to work with the whole period to obtain a significant number of occurrences, and thus more statistically robust results. Validation over an independent period could therefore not be performed. However, in order to carry out a "performance

check" of our methodology, we conducted an additional analysis by splitting the evaluation period in two periods. The first five years were used as 'calibration' period, while the next five years were used as 'validation' period. This performance check was only conducted for the Q_{ix} flood events with a return period of 2 years, since this is the threshold with more data (threshold exceedances for all catchments) available.

In the overall analysis, results from the calibration period of the first five years show an optimal overall adjustment factor of 0.90, which is the same adjustment factor obtained in paragraph 4.2 for the analysis of the entire 10-year period. The second period of 5 years (validation period) also shows the same result: an overall adjustment factor of 0.90. The catchment-specific adjustment factors are shown in Figure 26. They differ a little between the two datasets of the calibration and validation period, which might be the result of the limited number of events during one period (or both periods) in a specific catchment. The adjustment factor values per catchment are included in appendix 0.

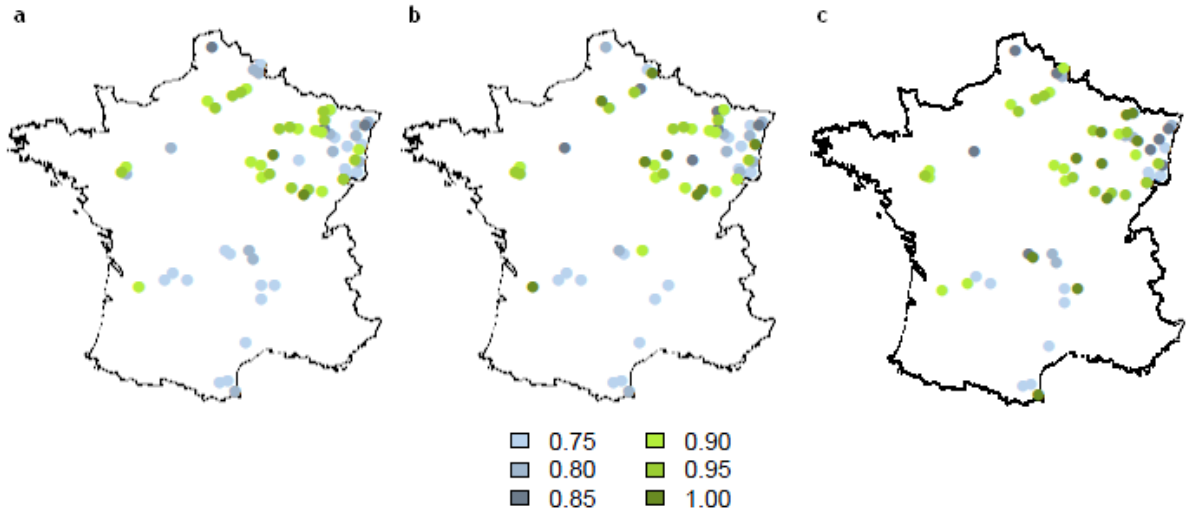


Figure 26. Catchment-specific daily adjustment factors for the $Q_{ix\ 2year}$ threshold for (a) the whole period 08.1995-07.2005, (b) the 5-year calibration period 08.1995-07.2000, and (c) the 5-year validation period 08.2000-07.2005.

4.5 DISCUSSION

The optimal overall adjustment factors for the thresholds defined by the yellow operational threshold, the $Q_{ix\ 2yr}$ and the orange operational threshold -based on the maximum value for the critical success index and its weighted formulation- are presented in Table 5. Table 6, in its turn, shows the number of exceedances and the optimal adjustment factors for the catchments where the yellow, orange and Q_{ix2yr} thresholds are exceeded at least one time during the evaluation period.

We remind that exceedances of the orange threshold only occur in a limited number of catchments, and therefore the results obtained do not represent the varied hydro-climatic condition present in the dataset. The optimal daily orange thresholds are therefore only valid for the catchments where this threshold is exceeded and should be used with caution.

Analyzing a catchment-specific daily adjustment factor gives, as expected, better results (in terms of CSI scores) than the overall daily adjustment factor for the yellow as well as the 2-year return period threshold.

The daily adjustment factors show a relation with the catchment characteristics for two of the analyzed catchment descriptors: size and reactivity. These catchment descriptors influence the characteristics of the hydrographs (rising limb, peak discharge). A multi-regression analysis would be the next step in order to identify the (inter)dependence of these catchment characteristics and the adjustment factor.

Furthermore, our analyses show that the 2-year return period discharge is a good alternative for catchments without a predefined warning level and for which a flood warning system is desirable. The catchment-specific adjustment factor for the 2-yr flood threshold varies ± 0.05 from the corresponding adjustment factor computed for the yellow threshold.

A calibration and validation check shows similar results, with only small differences at the level of accuracy in the catchment-specific adjustment factors.

Another remarkable point is the fact that the POD decreases when the magnitude of the *streamflow* threshold increases. However, a direct correlation between the level of the *streamflow* threshold and the daily adjustment factor was not found.

Table 5. Optimal daily adjustment factor for different values of α and for various thresholds.

α in CSI(α)	Optimal x in: Qθ daily = x Qθ yellow	Optimal x in: Qθ daily = x Qθ ix 2year	Optimal x in: Qθ daily = x Qθ orange*
1.00 = CSI	0.90	0.90	0.95
0.75	0.90	0.90	0.95
0.50	0.85	0.90	0.85
0.25	0.80	0.80	0.85

Table 6. Catchments with exceedances of the yellow, $Q_{ix\ 2\ year}$ and orange thresholds. Columns show the number of exceedances for every threshold and the corresponding optimal daily adjustment factor.

Catchment Code	**Yellow threshold**		**Qix2 threshold**		**Orange threshold**	
	Number of exc.	**Adjust. Factor**	**Number of exc.**	**Adjust. factor**	**Number of exc.**	**Adjust. factor**
A6761010	13	0.80	13	0.80	4	0.75
A7881010	4	0.85	10	0.95	4	0.95
A9752010	28	0.95	8	0.80	2	0.95
H5011020	8	0.75	5	0.75	3	0.75
H7401010	19	1.00	9	0.95	8	0.95
H7742020	13	1.00	40	0.95	4	1.00
K2070810	12	0.80	12	0.75	2	1.00
K2981910	38	0.85	7	0.80	2	0.95
U2354010	49	0.80	12	0.85	4	0.90

5 RESULTS II: *ENSEMBLE* THRESHOLD

The challenge here is to find an optimal number of ECMWF ensemble members exceeding the *streamflow* thresholds in order to launch a flood warning. Since the dataset of ECMWF weather forecasts covers only 18 months (March 2005-July 2006), we could not apply the operational thresholds evaluated in Chapter 4 because of the limited number of observed exceedances. Therefore, several quantile discharges computed over the same time period of the forecast archive are here applied as *streamflow* thresholds. Paragraph 5.1, consists of the results of the reliability diagram analysis. The reliability analysis investigates the difference between the reliability of the streamflow forecast computed for the dataset of 208 catchments and for the dataset of 29 large catchments. From this analysis, it becomes clear that the reliability of the streamflow forecasts is higher for the dataset of large catchments. In paragraph 5.2, the results of the analysis of the CSI score are presented for both datasets in order to find an optimal number of ECMWF ensemble members exceeding the *streamflow* thresholds required for launching a warning. Paragraph 5.3, consists of the results of the preparedness analysis, since an optimal *ensemble* threshold should not only emphasize the balance among hits, misses and false alarms, but should as well try to optimize the gain in lead-time compared to the deterministic forecast. In paragraph 5.4 two measures to increase the CSI score and preparedness are discussed, together with their impact on the ensemble threshold.

5.1 RELIABILITY ANALYSIS

A threshold-based evaluation focusing on the reliability of the ensemble streamflow predictions is conducted for both datasets B1 (29 catchments) and B2 (208 catchments). The reliability diagram analysis is conducted as described in paragraph 3.2 and compares the observed frequency to the forecasted probability of the threshold exceedances. The forecasted probability is deducted from the ECWMF ensemble forecast and subdivided in 6 probability classes (ranges: 0-2%, 2-25%, 25-50%, 50-75%, 75-98%, and 98-100%). E.g. 6 out of 51 ensemble members exceeding the threshold correspond to a probability of 11% and this fits in the probability class 2-25% with a mean value of 13.5% (Chapter 3.2.4 gives more details on the construction of a reliability diagram).

The aim of this evaluation is to find if there are statistical relationships among the reliability of the ensemble streamflow prediction (aspect of quality), the *streamflow* threshold level (Q70, Q90, Q95, Q99) and the catchment size (Dataset B1 and B2). Additionally, the reliability analysis can be used to identify the impact of the hydrological model and meteorological forecast errors. The analysis is conducted for the four *streamflow* thresholds mentioned above. Especially for the higher thresholds there are only a limited number of observed exceedances in our data series of 18 months (by definition, about 6 observed exceedances per catchment for the Q99 threshold during this period). The limited number of threshold exceedances is the reason why the exceedances are aggregated over all catchments and lead times to reduce the uncertainty in the observed frequencies and create more stable probabilities and frequencies especially for the higher *streamflow* thresholds (Q90 and Q99). An additional analysis is conducted in Appendix A. 4 to check the impact of lead-time on the reliability of the streamflow prediction. It shows that the reliability is lower for predictions with a short forecast range (1-2 days) and for the high probability classes of threshold exceedance prediction with a longer lead-time (7-9 days).

5.1.1 RELIABILITY DIAGRAM: DATASET B1

Figure 27 shows the results for the Q90 and Q99 *streamflow* thresholds when using both reference discharges: in blue dots, the proxy-observed reference discharge (i.e. the "perfect forecast"; run of the model with observed precipitation as input forecast), and, in orange triangles, the actual observed discharge.

It can be seen that there is a good agreement between the frequency of the reference discharges and the forecasted probabilities exceeding the Q90 threshold: for all probability categories, the blue points are quite close to the theoretical diagonal of perfect reliability. For example, an event with a forecasted probability of 86.5% is observed in 77% of times for the simulated discharge based on the actual observed amount of precipitation for the Q90 threshold. The exceedance probability is in this case slightly overestimated by the forecast model (i.e. the number of 'observed' events in the simulation with observed precipitation data is smaller than the number of forecasted events).

Another remarkable point is the agreement between the tendencies of the points in the diagrams comparing forecasted probabilities with observed frequencies (orange triangles) and frequencies of the simulated reference discharge (proxy-observed; blue dots) for the Q90 threshold. This is not only the case for the Q90 threshold, but as well for the Q70 and Q95 threshold (Reliability diagrams for these thresholds are included in appendix 0). The simple calibration procedure proposed by Olsson and Lindström (2008) could therefore be applied in this evaluation. For instance, if we multiply all the forecasted probabilities of exceeding the Q90 threshold by a factor of 0.90, the adjusted probabilities almost equal the corresponding frequencies of the simulated reference discharges (proxy-observed) exceeding the Q90 threshold. The other average estimated calibration factors are shown in Table 7 (column 2).

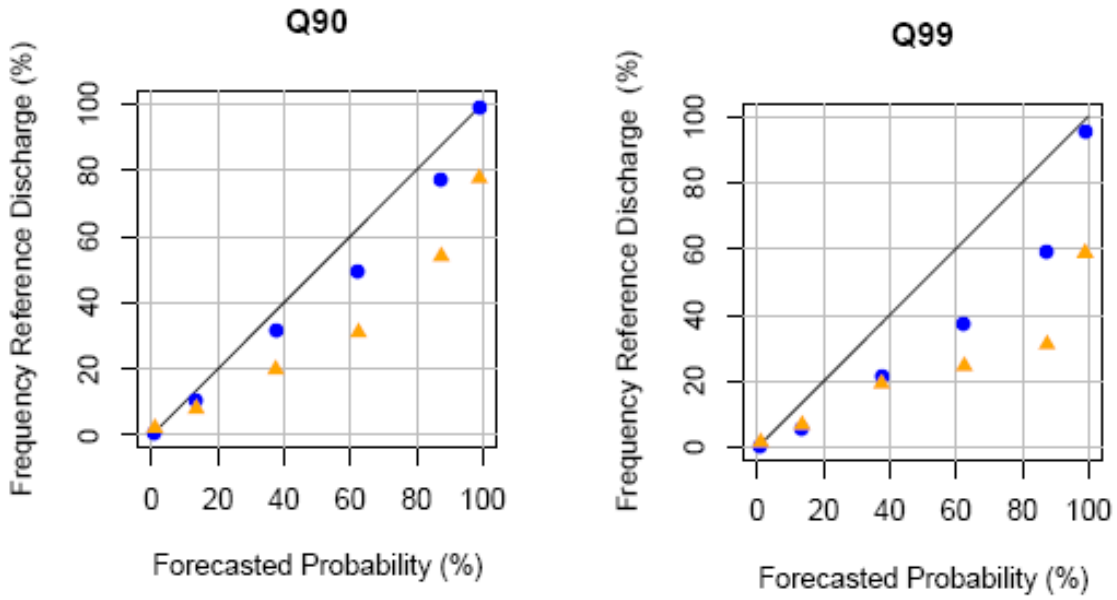


Figure 27. Reliability diagrams aggregated overall lead times and catchments in dataset B1 for the Q90 (left) and Q99 (right) *streamflow* thresholds over the 18 month evaluation period, and for the six probability classes (x-axis). The blue dots represent the plot of the frequency of threshold exceedances of the proxy-observed (simulated discharges with observed precipitation; "perfect forecast") against the ensemble forecasted probabilities; the orange triangles represent the frequency of exceedances of the observed discharge against the ensemble forecasted probabilities.

When considering the frequencies of the observed discharges (represented by the orange triangles), the analysis show a tendency of the forecast system to over forecast. For the Q90 diagram, for instance, if we multiply all the forecasted probabilities of exceeding the Q90 threshold by 0.70, the "calibrated" probabilities almost equal the corresponding frequencies of the actual observed discharges exceeding the Q90 threshold. The other average estimated calibration factors are also shown in Table 7 (column 3). Due to the fact that the events in the extreme highest probability categories (99% and 1%) are substantially closer to the y=x line, these probability calibration factors are only valid for the 13.5%, 37.5%, 62.5% and 86.5% probability categories.

Concerning the impact of the hydrological model error and the meteorological forecast error in the reliability of the streamflow predictions, the analysis of the reliability diagram for all thresholds shows that the impact of the error of the meteorological forecast is larger for Q99 than for the Q70 to the Q90 threshold exceedances, especially for the forecast categories between 13.5% and 62.5%. In these situations, the distance between the plots (blue dots and orange triangles) becomes smaller than the distance between the diagonal and the plot using the observed discharges as reference (blue dots), as the *streamflow* threshold increases. This impact is also reflected in the calibration factor: its value decreases as the threshold increases. This is an interesting result, because Olsson and Lindström (2008), studying 45 catchments in Sweden over an 18-month period, did not find a relation between the threshold, the observed frequency and the calibration factor. This can be explained by the fact that the Q99 discharge focuses on more extreme events or by the different characteristics of extreme precipitation events in Sweden and France (e.g. SMHI, n.d. and Météo France, n.d.)

Another remarkable point for the Q99 threshold exceedances is that the distance between the blue and orange points (hydrological model error) becomes larger for higher forecasted probabilities of exceedance. For the probability classes with a mean value of 86.5% and 99%, the hydrological model error is at least as important as the meteorological forecast error. The higher forecasted probabilities are often related to more extreme events and are of great importance for hydrologic forecasting and warning. In this case of forecasting high discharges (> Q99), the reliability diagram shows that the reliability of the streamflow forecasts is strongly affected by the hydrological model error and the meteorological forecast error in the highest classes of forecast probabilities.

Table 7. Probability calibration factors for both reference discharges and 4 selected thresholds, considering all 208 catchments in Dataset B1.

Threshold	Adjustment factor Proxy-Simulated RD	Adjustment factor Observed RD
Q99	0.70	-
Q95	0.90	0.70
Q90	0.90	0.70
Q70	0.95	0.75

In Appendix 0 a reliability diagram analysis for 4 different regional areas in France (22-48 catchments) for the Q99 threshold is included. The aim is to give an overview of the magnitude of the meteorological and hydrological errors for different geographical regions in France. The results show that the main average tendency to over forecast is observed at all regions. The reliability diagrams for the catchments in the Loire and Rhône river basins are the closest to the diagonal of perfect reliability. Errors from the meteorological forecasts seem to have less impact than errors from the hydrological

model in the Loire river basin, while errors from the hydrological model seem to have less impact in the Rhine river basin.

5.1.2 RELIABILITY DIAGRAM: DATASET B2

The first dataset consists of 208 catchments all over France with different catchment sizes. In order to evaluate the effect of a precipitation forecast with a coarse spatial grid on the reliability of ensemble streamflow predictions for catchments with a larger area, the same analysis is conducted for the 29 catchments in dataset B2 (size greater or approximately equal to the grid size of ECMWF forecasts).

Figure 28 shows the results of the reliability analysis over the 29 catchments for the Q90 and Q99 thresholds. Analyses for the Q70 and Q95 thresholds are included in appendix 0. For the Q90 threshold, the reliability diagram built with the proxy-observed reference discharges (dots) is almost equal to the one obtained from the catchments in dataset B1. This indicates that the average impact of the meteorological error has the same order of magnitude for the catchments in both datasets and is independent from the catchment size, i.e., the average impact of the meteorological error is not decreasing with an increasing catchment size. In order to adjust the forecasted probabilities, the same calibration factors for the proxy observed reference discharge (as presented in Table 7) can be applied for the 29 catchments.

However, for the Q99 threshold, the total reliability increases, specially for the three highest probability categories (62.5%, 86.5% and 99%). This is due to a decreasing impact of the hydrological model error, i.e. the hydrological model seems to have a higher capacity of forecasting (reproducing) high discharges (with high forecasted probabilities) for larger catchments.
Table 8 gives an overview of the average estimated calibration factors that could be applied for both reference discharges and all thresholds (Q70, Q90, Q95 and Q99).

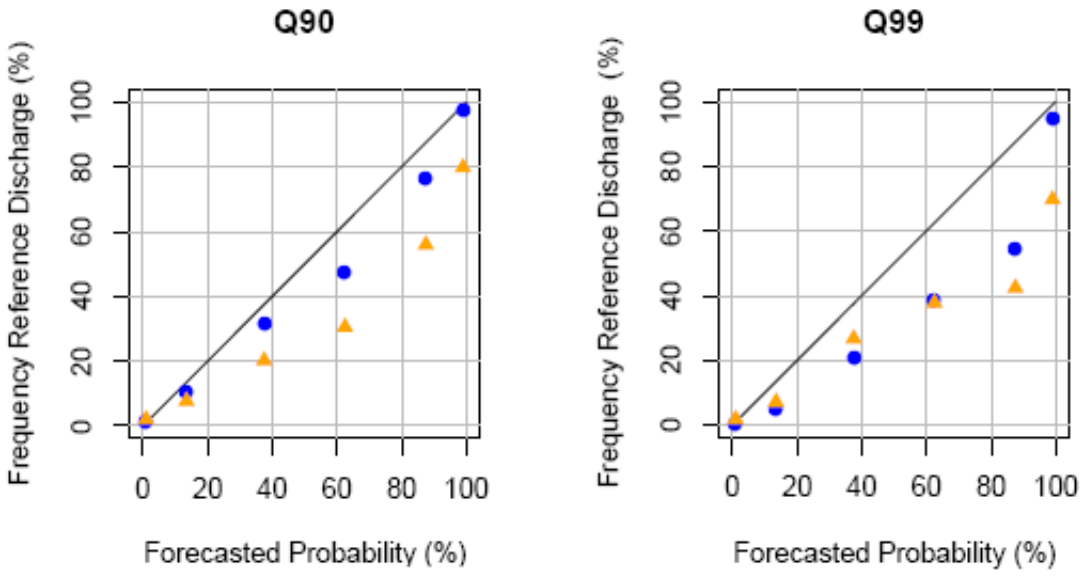


Figure 28. Reliability diagrams aggregated overall lead times and catchments in dataset B2 for the Q90 (left) and Q99 (right) *streamflow* thresholds over the 18 month evaluation period, and for the six probability categories (x-axis). The blue dots represent the plot of the frequency of threshold exceedances of the proxy-observed (simulated discharges with observed precipitation; "perfect forecast") against the ensemble forecasted probabilities; the orange triangles represent the frequency of exceedances of the observed discharge against the ensemble forecasted probabilities.

Table 8. Probability calibration factors for both reference discharges and 4 selected thresholds, considering all 29 catchments in Dataset B2.

Threshold	Adjustment factor Simulated RD	Adjustment factor Observed RD
Q99	0.70	-
Q95	0.90	0.75
Q90	0.90	0.75
Q70	0.95	0.85

5.1.3 DISCUSSION

The aim of the reliability analysis here conducted was to verify if the streamflow predictions were more reliable for a dataset containing larger catchments over different geographic areas in France. In paragraph 5.1.1 and 5.1.2, it is shown that the impact of the meteorological forecast error is of the same order of magnitude for both datasets of catchments. It seems that the large grid size of ECMWF ensemble weather forecast does not affect the average reliability of streamflow predictions for smaller catchments. However, since the overall reliability is lower for the first dataset (B1) than for dataset B2, especially for exceedances of the highest threshold Q99, there is an indication that it is the part of the error coming from the hydrological model that is affecting the total reliability: the bias coming from errors in the hydrological model appears to be smaller for large catchments (especially for high probability classes). Besides, for the higher probability classes, the distance between the orange and blue points (indicator of the hydrological model error) is often smaller for higher *streamflow* thresholds than for the lower ones.

The points listed below are the most important results from the reliability diagram analysis:

- The forecast prediction system has a general tendency to over forecast;
- Higher reliability of the streamflow predictions is obtained for Dataset B2 (large catchments);
- Lower reliability is obtained for higher *streamflow* thresholds;
- Extremely high (almost perfect) reliability is detected for forecasts with a high probability (category 99%);
- Reliability diagrams built with the proxy-observed forecasts as reference discharge are similar for both datasets (blue points). The effect of the error in precipitation data is the same for both datasets.
- The part of the hydrological model error is as important as the one from the meteorological forecast error for high forecast probabilities;
- The effect of the errors in the hydrological model is smaller for larger catchments, especially for high discharges.

The reliability diagram analysis demonstrates that there is a difference in the reliability of streamflow predictions for the datasets B1 and B2. In the next paragraphs, the analysis to determine the CSI and preparedness scores is also conducted for both datasets.

5.2 CRITICAL SUCCESS INDEX AND THE *ENSEMBLE* THRESHOLD

5.2.1 IMPACT OF CONSIDERING THE START OF A FLOOD EVENT

The CSI-optimal *ensemble* threshold is defined as the number of ensemble members exceeding the *streamflow* threshold required in order to launch a flood warning with the maximum Critical Success Index (CSI) over the study period (Chapter 3.2). From the evaluation of the *streamflow* thresholds (Chapter 4), it became clear that attention should be paid to the start of a flood event, which is one of the most difficult variables to forecast and at the same time one of the most important. Here, the impact of considering the start of flood events will be addressed for the evaluation of the ensemble threshold. CSI values are computed for all time steps as well as for only the time steps of the start of a flood event. For the start of a flood event, every time step with an observed discharge or a forecasted discharge (at least 1 EPS member) exceeding the *streamflow* threshold is considered. Score values are evaluated as a function of the number of ensemble members exceeding the given *streamflow* threshold in order to find the number of ensemble members resulting in the maximum CSI score.

Figure 29(a) shows the overall (all 208 catchments included) CSI curve for the exceedances of the Q99 percentile streamflow threshold. The blue curve represents the standard CSI (α=1) and the green curve represents the weighted CSI (α=0.5). The dotted horizontal lines represent the corresponding CSI scores for the high-resolution deterministic forecast. Figure 29(b) shows the overall CSI curve for only the days when the Q99 threshold was exceeded for the first time, i.e., the start of a flood event. The scores are aggregated over all catchments and all lead-times. It can be seen that the maximum CSI is lower for the start of flood event than for all exceedances. This can be explained by: 1) the fact that the frequency of events is lower for the start days only; and 2) the fact that it is easier to forecast a larger exceedance of the threshold (included more often in the analysis of all exceedances) than the exceedances at the first day of a flood event (often a smaller exceedance since the flood event has not yet fully developed). This supports the fact that forecasting correctly the start of a flood event is a difficult task in flood forecasting.

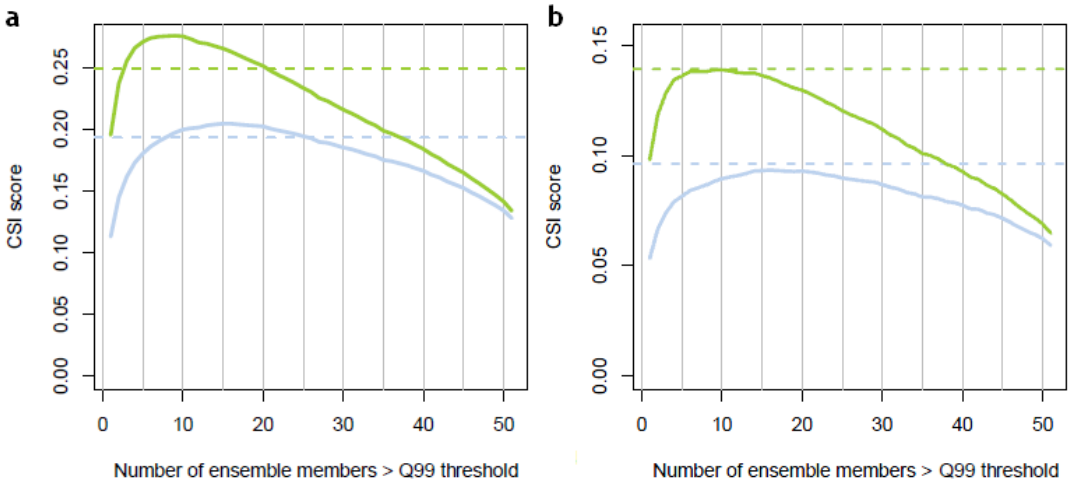


Figure 29. CSI curves for exceedances of the percentile Q99, for (a) all days and (b) only the start of a flood event. The blue (lower) curve represents CSI (α=1), the green curve represents CSI (α=0.5). The dotted lines are the CSI scores for the corresponding high resolution ECWMF deterministic forecast. The scores are aggregated over all 208 catchments of dataset B1 and all lead-times.

From Figure 29, it can also be seen that the CSI-optimal *ensemble* threshold (number of members exceeding the threshold required for a warning), which is the one with the highest CSI score, is of 16 ensemble members exceeding the Q99 *streamflow* threshold. If all time steps are considered, Figure 29(a), CSI values for the ensemble predictions are greater than the CSI score of the deterministic forecasts for a range of number of members exceeding the threshold. This was not observed when considering only the days of the start of the flood events. In this case the CSI of the deterministic forecast is slightly higher than the maximum CSI of the ensemble forecast. Table 9 gives the CSI-optimal *ensemble* thresholds for 5 selected quantiles (70%, 80%, 90%, 95%, 99%). The number between brackets is the lowest number, if any, of ensemble members resulting in the same CSI as the deterministic forecast. From Figure 29 and Table 9, it can be seen that the optimal ensemble threshold is of about 4 to 10 members lower when the weighted CSI (α=0.5) is considered (i.e., if forecasters accept a part of false alarms, warnings can be launched with a lower forecast probability, which might also mean with a greater anticipation).

Table 9. Number of ensemble members related to the maximum CSI score for five selected streamflow thresholds and two values for α (1.0 and 0.5). The number between brackets represents the minimum number of ensemble members required for an ensemble CSI score to be equal to the deterministic CSI score. The CSI scores are aggregated over all 208 catchments of dataset B1 and all lead-times.

Q_{th} LT 1-9	Start of a flood event	
	α=1.0	α=0.5
Q99	16 (-)	10 (-)
Q95	29(17)	21(15)
Q90	35 (21)	25(-)
Q80	27(15)	17 (9)
Q70	21 (8)	17 (6)

5.2.2 MEDIUM-RANGE FORECASTS: IMPACT OF LEAD TIME

The CSI analysis conducted in paragraph 5.2.1 is aggregated over all lead-times to allow the evaluation of high thresholds (i.e., analyses with a limited number of occurrences) over a data archive of only 18 months. The aggregation over all lead-times is however only valid if the CSI score and the *ensemble* threshold behave in a similar way for each specific lead-time in the range from 1 to 9 days.

Figure 30 (a) shows how the overall CSI-optimal number of ensemble members varies with lead time. Noticeable are the irregular highly variable values for the two shortest lead-times (1 and 2 days): for Q70, Q80, Q90 and Q95 thresholds, the optimal number of ensemble members for launching a warning one day ahead is around 50. The Q99 threshold behaves the other way around. For this threshold, the optimal number of ensemble members is close to 1. This can be explained by the fact that the ECWMF ensemble prediction system (see Chapter 2.3) is a medium-range weather forecast system, optimized for lead-times larger than 48 hours. In fact, the spread of ECMWF-EPS is smaller for the short range lead-times (as illustrated by the example of Figure 9 in Chapter 2.3). The limited spread for these short lead-times makes that the probabilistic forecast actually behaves like a deterministic forecast. During the analyses conducted in the remaining part of the chapter, only the lead-times from 3 to 9 days will be considered. During these lead-times the spread is larger and the probabilistic forecast distinguishes itself better from the deterministic one.

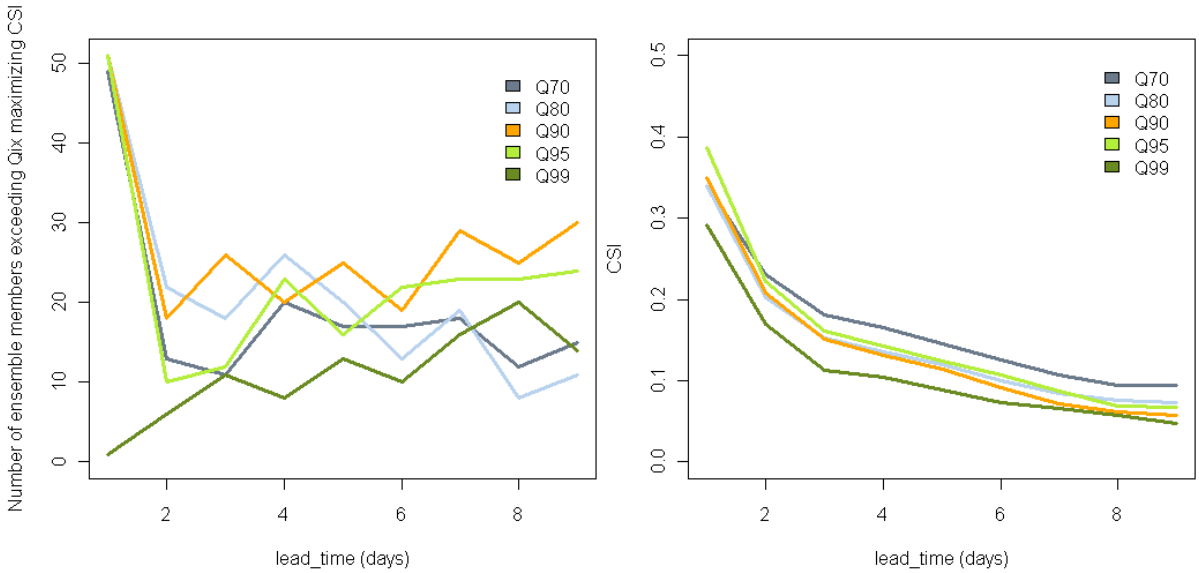


Figure 30 (a) Number of ensemble members resulting in the maximum CSI score for the five selected *streamflow* thresholds distinguished per lead-time: 1-9 days; (b) Corresponding maximum CSI scores score for the five selected *streamflow* thresholds distinguished per lead-time: 1-9 days.

5.2.3 DATASET B1: 208 CATCHMENTS

Excluding the lead-times of 1 and 2 days from the analysis will result in different values for the CSI score and the overall *ensemble* threshold. Figure 31 shows the overall CSI curve integrated over only lead-times 3 to 9 days for the Q99 threshold. It can be seen that the maximum CSI is lower than the maximum CSI for the whole lead-time range (Figure 29(b)). The lower CSI is not strange considering that it is more difficult to predict the weather and the resulting discharge 9 instead of 2 days in advance (Figure 30(b)).

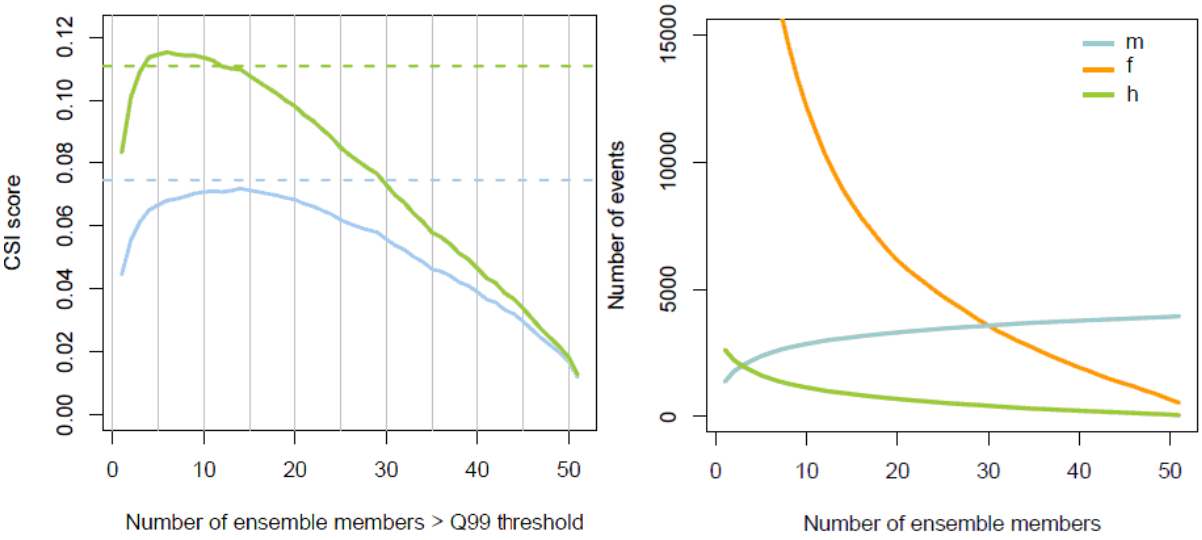


Figure 31 (a, left). CSI curve for exceedances of the percentile Q99 at the start of a flood event. The blue (lower) curve represents CSI (α=1) and the green curve, CSI (α=0.5). The dotted lines are the CSI scores for the corresponding high resolution ECWMF deterministic forecast. The scores are aggregated over all 208 catchments of dataset B1 and lead-times 3-9 days; (b, right) The number of hits (h-green), misses (m-blue) and false alarms (f-orange) corresponding to the CSI (α=1 and α=0.5) curves in figure a.

Furthermore, we see a shift in the *ensemble* threshold related to the maximum CSI, towards a smaller number of ensemble members.

Table 10 shows the CSI-optimal *ensemble* thresholds for the 5 selected *streamflow* thresholds when considering only lead times 3 to 9 days. The second row consists of the number of ensemble members exceeding the threshold resulting in the same CSI score as the deterministic streamflow forecast, i.e. the same balance among hits, false alarms and misses. The ensemble streamflow prediction does not result in a higher CSI score than the deterministic forecast for some thresholds (Q99 for α=1.0, blue curve in Figure 31(a), and Q90 for both α=1.0 and α=0.5). As a result, there is no number of ensemble members resulting in the same CSI for these *streamflow* thresholds ("-").

Every catchment behaves differently and therefore it is also interesting to calculate the CSI-optimal *ensemble* threshold for every individual catchment. Figure 32 shows an example of the catchment-specific optimal *ensemble* threshold for the Q95 *streamflow* threshold and the catchment P7001510 (River Isle at Bassilac; Dordogne). In this example, the catchment-specific *ensemble* thresholds (about 20 ensemble members for α=1.0 and α=0.5) equals the overall *ensemble* thresholds (20-22 ensemble members) for the Q95 streamflow threshold, although the CSI increases from 0.12 for the overall case to 0.26 for this catchment-specific case.

Table 10. The number of ensemble members related to the maximum CSI score for the five selected *streamflow* thresholds and the two values for α (1.0 and 0.5). The number between brackets represents the minimum number of ensemble members required for an ensemble CSI score to BE equal to the deterministic CSI score. The CSI scores are aggregated over all 208 catchments of dataset B1 and lead-times 3-9 days.

Q_{th} LT 3-9	Start of a flood event	
	α=1.0	α=0.5
Q99	14 (-)	6 (4)
Q95	22 (16)	20 (13)
Q90	27 (-)	19 (-)
Q80	18 (13)	11 (7)
Q70	18 (8)	14 (6)

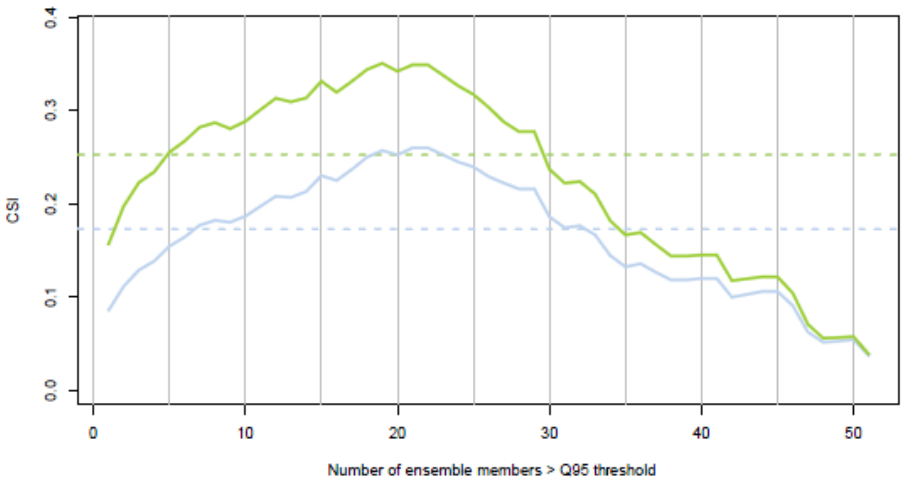


Figure 32**.** CSI as function of the number of ensemble members exceeding the Q95 streamflow threshold for the P7001510 catchment -River Isle at Bassilac (Dordogne)- at the start of a flood event. The blue (lower) curve represents CSI (α=1) and the green curve, CSI (α=0.5). The dotted lines are the CSI scores for the corresponding high resolution ECWMF deterministic forecast.

Figure 33 plots the differences between the optimal overall *ensemble* threshold (CSI α=1.0) and the catchment-specific *ensemble* thresholds for the Q70, Q80, Q90 and Q99 *streamflow* thresholds for all 208 catchments in Dataset B1. The overall *ensemble* threshold equals respectively 18, 18, 22 and 14 members (Table 10). In the plot, the blue points illustrate catchments with a threshold lower than the overall threshold, the orange points are those catchments with thresholds close to the overall threshold and the green points are catchments with a catchment-specific threshold higher than the overall threshold. Remarkable is the difference between the catchments in eastern and western part of France for the higher streamflow thresholds. Even more notable is the behaviour of the catchments in Eastern France. Most catchments have more or less a 'fixed' deviation from the overall *ensemble* threshold, but the deviation of this ensemble threshold changes with the level of the streamflow threshold for a large number of catchments located in this region.

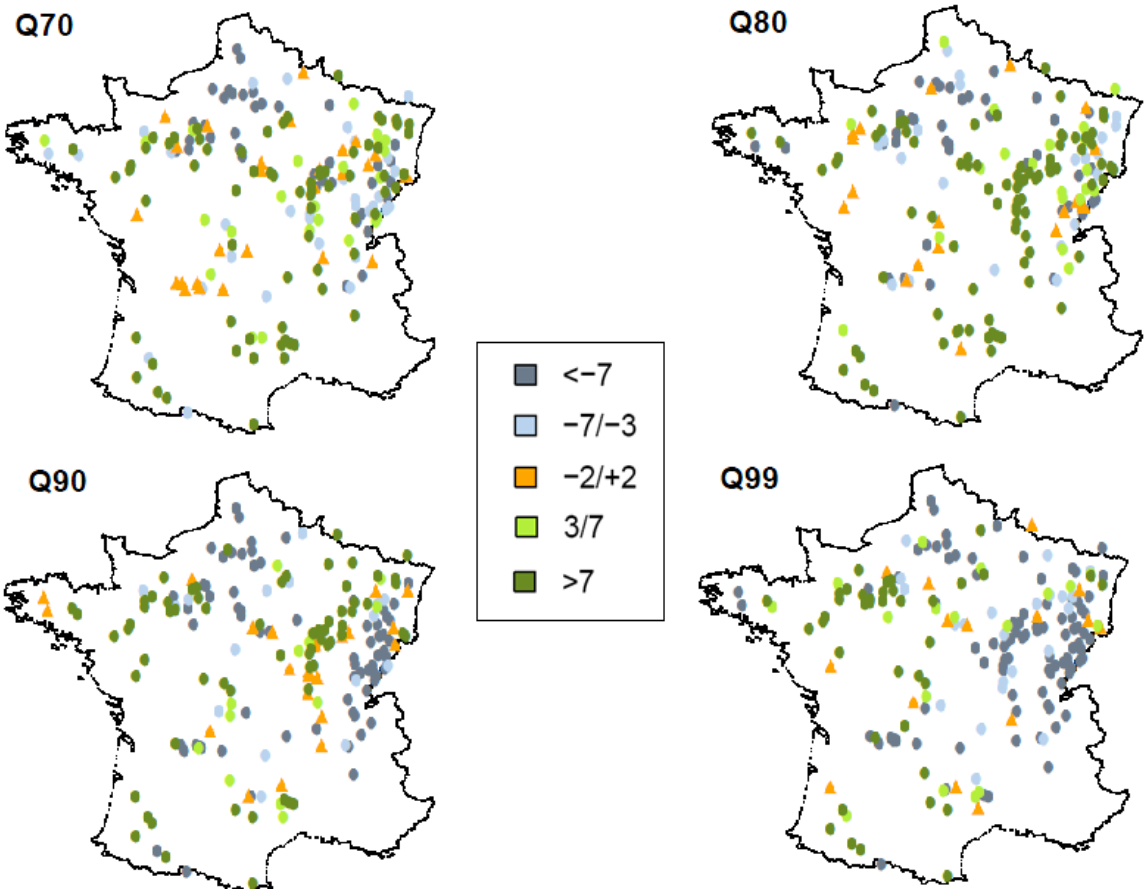


Figure 33. Difference between the overall *ensemble* threshold and the catchment-specific *ensemble* threshold for the 208 catchments in Dataset B1 (difference in number of ensemble members). The blue dots represent catchments with a lower optimal number of ensemble members, while the green dots represent catchments with a higher optimal number of ensemble members required to launch a warning than the overall optimum of respectively 18 (Q70), 18 (Q80), 22 (Q90) and 14 (Q99) members.

5.2.4 DATASET B2: 29 CATCHMENTS

CSI-optimal *ensemble* thresholds were also evaluated for the 29 catchments of dataset B2. The optimal number of ensemble members for the maximum CSI is calculated for streamflow predictions at the start of a flood event and for lead times ranging from 3 to 9 days. Figure 34 shows the CSI score value for the Q99 *streamflow* threshold. The optimal number of ensemble members exceeding this threshold is lower (N=10) comparatively to the results obtained for all 208 catchments in dataset B1

(N=14, Figure 31). Furthermore the maximum CSI score (0.09) is higher than the CSI score for the deterministic forecast (0.06).

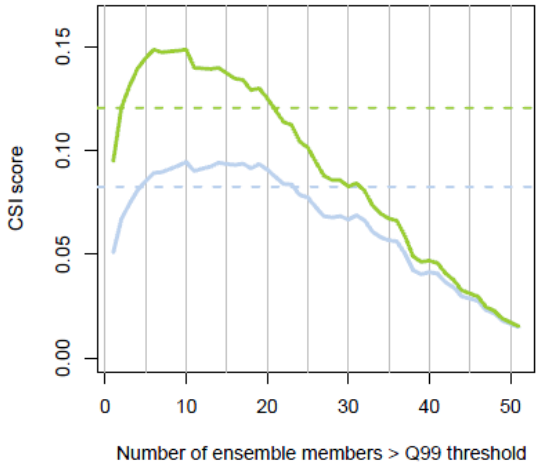


Figure 34. CSI curve for exceedances of the percentile Q99 and the start of a flood event only. The blue (lower) curve represents CSI (α=1) and the green curve, CSI (α=0.5). The dotted lines are the CSI scores for the corresponding high resolution ECWMF deterministic forecast. The scores are aggregated over all 29 catchments of dataset B2 and lead-times 3-9 days.

Table 11. Number of ensemble members related to the maximum CSI score for five selected *streamflow* thresholds and two values for α (1.0 and 0.5). The number between brackets represents the minimum number of ensemble members required for an ensemble CSI score to be equal to the deterministic CSI score. The CSI scores are aggregated over all 29 catchments of dataset B2 and lead-times 3-9 days..

$\theta_{streamflow}$ LT 3-9	Start of a flood event	
	α=1.0	α=0.5
Q99	10 (5)	6 (2)
Q95	22 (-)	20 (-)
Q90	27 (-)	16 (-)
Q80	21 (13)	15 (8)
Q70	12 (6)	11 (4)

5.2.5 DISCUSSION

Comparatively to the CSI values for the deterministic forecast, the overall *ensemble* threshold analysis for the 208 catchments leads to lower CSI values for the Q99 and Q90 *streamflow* threshold. For the other *streamflow* thresholds (Q95, Q80, Q70), the CSI scores related to the optimal ensemble threshold is a little higher than the deterministic ones for a range of approximately 5 to 25 members exceeding the *streamflow* threshold. The optimal ensemble threshold based on maximum CSI score ranges from 14 to 27 members out of 51 (i.e., empirical probabilities of about 25 to 50%) for the forecast of discharges exceeding the *streamflow* thresholds of Q99 to Q70 (smaller thresholds are obtained for the highest and lowest *streamflow* thresholds, i.e., Q99 and Q70). For the second dataset, consisting of 29 large catchments, the maximum CSI of the ensemble forecast does not exceed the deterministic one for Q95 and Q90. The maximum CSI of the overall analysis for the 29 larger catchments, i.e. the CSI-optimal *ensemble* threshold, is related to a lower (or equal) number of ensemble numbers exceeding the *streamflow* threshold comparatively to the overall *ensemble* threshold from the analysis of the 208 catchments. The overall CSI score itself is higher for both, forecasts, deterministic and ensemble, for the catchments in dataset B2. The CSI could be optimized

when a catchment-specific ensemble threshold is applied, which is specially the case for the catchments located in eastern France.

The behaviour of the CSI-optimal catchment-specific *ensemble* thresholds for the catchments in Eastern France deserves more attention. This *ensemble* threshold is lower than the overall threshold for most catchments for the lower *streamflow* thresholds, and it is gradually increasing for the higher *streamflow* thresholds. An explanation might be found in the reliability diagram for the contributories of the Seine River (H catchments in appendix A 5.4). The reliability for these catchments is low, partially due to bias from the hydrological model error.

5.3 PREPAREDNESS AND THE *ENSEMBLE* THRESHOLD

5.3.1 DATASET B1: 208 CATCHMENTS

The optimal balance among hits, false alarms and misses is one of the most important features of a forecasting system as well as its ability to anticipate a flood event with a as long as possible lead time. As shown in paragraph 5.2, an optimal number of ensemble members can be evaluated by maximizing the CSI score from ensemble predictions alone or by taking the minimum number of ensemble members necessary to equal the CSI score of the deterministic prediction. Here, the preparedness score is calculated for these two *ensemble* thresholds: (a) the ensemble threshold related to the maximum CSI score of ensemble predictions and (b) the ensemble threshold related to a CSI score of ensemble predictions that equals the CSI score of the deterministic forecast.

The gain/loss in lead-time from ensemble predictions, compared to deterministic forecasts, was computed following the methodological steps presented in Chapter 3. Table 12 shows the mean values of preparedness (gain/loss in lead-time) per observed flood event and for the 5 thresholds considered in this study (overall analysis of all catchments and α=1.0). The preparedness score shows the difference in lead-time between the probabilistic and deterministic forecast. E.g. if an observed flood event is foreseen 6 days in advance by the probabilistic forecast and 3 days in advance by the deterministic forecast, then the preparedness equals 3. Since we are calculating the mean preparedness, the misses are as well taken into account.

From Table 12 it can be seem that the preparedness a_{ov} -the preparedness for *ensemble* threshold related to the maximum of the overall CSI curve- is negative for all *streamflow* thresholds, which means that applying an optimal overall *ensemble* threshold does not lead to a gain in lead-time for the five selected *streamflow* thresholds.

The preparedness b_{ov} -the preparedness for the *ensemble* threshold related to the intersection with the deterministic CSI- is only calculated for the Q70, Q80 and Q95 *streamflow* thresholds, since the number of ensemble members when the ensemble CSI score is greater than the deterministic CSI score is not defined for the Q90 and Q99 *streamflow* thresholds. Results show that a small average gain in lead time per flood event is observed for predictions of discharge exceeding the thresholds Q80 (+0.39 days) and Q70 (+0.97 days).

Table 12. Mean preparedness (gain/loss in lead-time compared to the deterministic forecast: in days) per observed *streamflow* threshold exceedance at the start of a flood event for the 208 catchments in Dataset B1, when an overall *ensemble* threshold (number of ensemble members) is applied.

Streamflow threshold	Preparedness a_{ov}	Preparedness b_{ov}
Q99	-0.29	-
Q95	-0.94	-0.33
Q90	-1.30	-
Q80	-0.20	+0.39
Q70	-0.14	+0.97

The same analysis is conducted for the catchment specific *ensemble* thresholds. Table 13 shows the gain/loss in lead time for the *ensemble* threshold related to the maximum CSI (a_{cs}) and to the first intersection with the deterministic CSI (b_{cs}) per observed flood event averaged over all catchments.

Table 13. Mean preparedness (gain/loss in lead-time compared to the deterministic forecast: days) per observed *streamflow* threshold exceedance at the start of a flood event for the 208 catchments in Dataset B1, when a catchment-specific *ensemble* threshold (number of ensemble members) is applied.

Streamflow threshold	Preparedness a_{cs}	Preparedness b_{cs}
Q99	+1.89	+3.21
Q95	-0.07	+1.21
Q90	-0.27	+1.18
Q80	-0.24	+1.27
Q70	-0.10	+1.46

A number of issues can be highlighted:

- When using catchment-specific thresholds, the gain in preparedness generally increases (or the loss in preparedness decreases).
- The gain in preparedness for the Q99 *streamflow* threshold is relatively large compared to the other str*eamflow* thresholds. On average, it exceeds the anticipation given by a deterministic forecast by 2 to 3 days per flood event. The mean flood anticipation of an ensemble streamflow prediction is maximal for more extreme situations.
- The mean preparedness related to the maximum CSI is close to zero for the other *streamflow* thresholds.
- The *ensemble* threshold that equals the CSI score of the high resolution deterministic forecast results in a gain of preparedness for all *streamflow* thresholds. This gain is of approximately 1.5 days in lead-time compared to the *ensemble* threshold based on the maximum CSI score. This indicates that there is a range of values within which the *ensemble* threshold could be optimized (increasing CSI vs. extending lead-time).

Figure 35 shows the mean preparedness for the Q99 threshold for of the 208 catchments when a catchment-specific threshold is applied. For most of the catchments the use of a catchment-specific optimal *ensemble* threshold results in a gain in lead-time compared to the (high resolution) deterministic forecast.

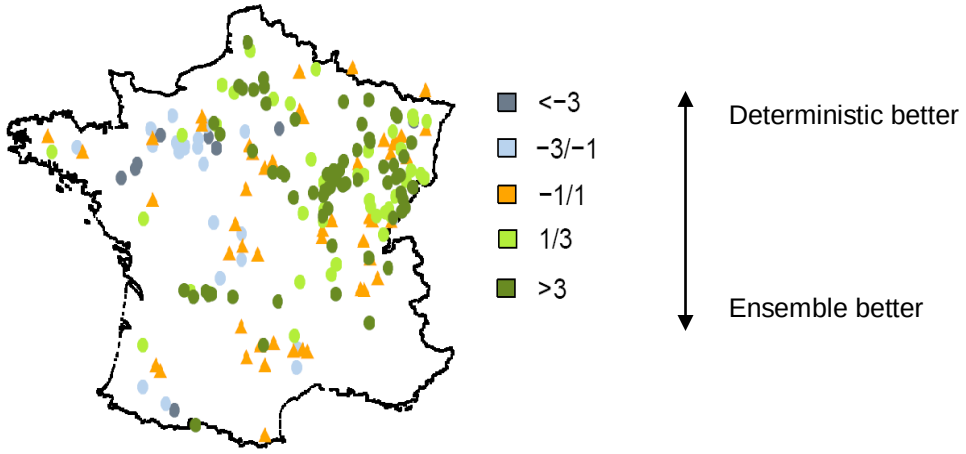


Figure 35. Mean preparedness (gain/loss in lead-time compared to the deterministic forecast: days) per observed Q99 threshold exceedance for the 208 catchments in Dataset B1 when a catchment-specific threshold is applied.

5.3.2 DATASET B2: 29 CATCHMENTS

Table 14 gives the mean preparedness (gain/loss in lead time compared to the deterministic forecasts) per observed exceedance for the 29 catchments using an *overall* ensemble threshold. The analysis is conducted in the same way as described in paragraph 5.3.1. The preparedness per flood event is calculated for streamflow predictions with a lead-time ranging from 3 to 9 days. Applying an overall *ensemble* threshold (based on the maximum CSI) has a small positive effect on the preparedness for the Q99 threshold (mean preparedness (ΔP) = +0.80). Using the optimal *ensemble* threshold, when the ensemble CSI is the same CSI as the deterministic forecast, results in a gain in lead-time of almost 2 days. The application of the overall ensemble Q99 thresholds on Dataset B2 results in a higher preparedness score compared to the average values given by the analysis of dataset B1.

Table 14. Mean preparedness (gain/loss in lead-time compared to the deterministic forecast in days) per observed *streamflow* threshold exceedance for the catchments included in Dataset B2, when an overall *ensemble* thresholds (number of ensemble members) is applied.

Streamflow threshold	Preparedness a_{ov}	Preparedness b_{ov}
Q99	+0.80	+1.94
Q95	-1.01	-
Q90	-1.22	-
Q80	-0.44	+0.52
Q70	+0.79	+1.57

Table 15 gives the mean gain/loss in lead-time compared to the deterministic forecasts per observed exceedance for the 29 catchments using a catchment-specific *ensemble* threshold. The use of a catchment-specific *ensemble* threshold leads to an increase in preparedness up to 1.5 day per event. In this dataset, the maximum gain in lead-time occurs for predictions of discharges exceeding the Q99 threshold. This again is an indication that the performance and the advantages of ensemble predictions are higher for more extreme *streamflow* thresholds.

Table 15. Mean preparedness (gain/loss in lead-time compared to the deterministic forecast in days) per observed *streamflow* threshold exceedance for the 29 catchments in Dataset B2, when a catchment specific *ensemble* threshold (number of ensemble members) is applied.

Streamflow threshold	Preparedness a_{cs}	Preparedness b_{cs}
Q99	+2.61	+3.49
Q95	+0.04	+0.55
Q90	-0.33	+0.86
Q80	-0.37	+1.01
Q70	+0.74	+1.69

Figure 36(a,b,c and d) show some alternative ways to visualize the preparedness score. Figure 36(e) shows the preparedness per flood event while applying the optimal overall *ensemble* threshold for the Q99 *streamflow* threshold (10 or more members exceeding the *streamflow* threshold). Figure 36(f) shows the preparedness after a catchment specific *ensemble* threshold is applied. Remarkable are the 3 'blue' catchments in western France: applying a catchment-specific *ensemble* threshold leads to a loss in lead-time compared to the deterministic forecast. More detailed analysis are needed to better understand the behaviour of this catchments, which is out of the scope of this study.

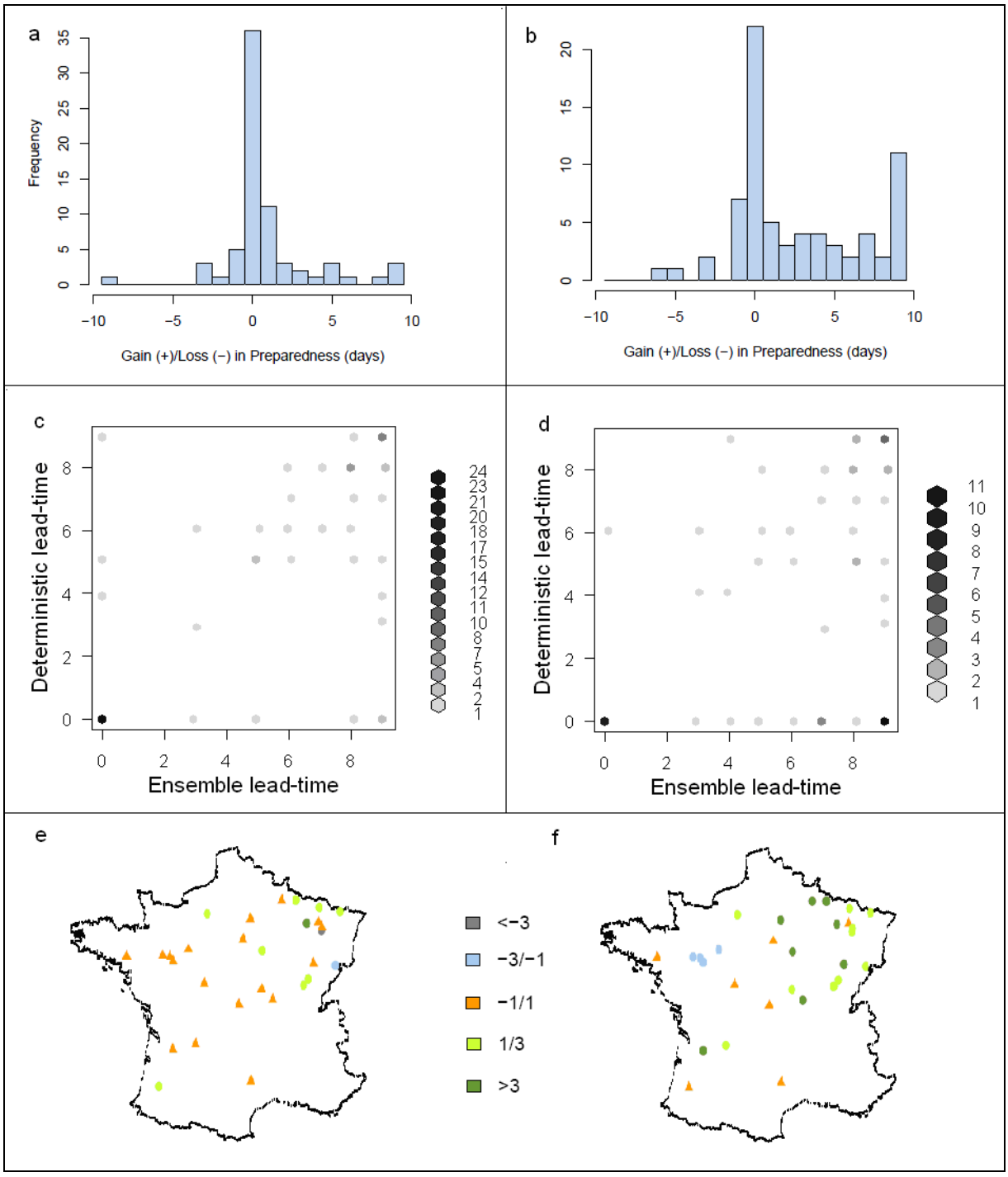


Figure 36. Mean preparedness (gain/loss in lead-time compared to the deterministic forecast: days) per observed Q99 threshold exceedance for the 29 catchments in Dataset B2. Figure (a) shows the mean preparedness related to the overall *ensemble* threshold, while (b) shows the mean preparedness related to the catchment-specific *ensemble* threshold. Figure (c and d) show the relation between the frequencies of the ensemble and deterministic lead-times corresponding to observed flood events for respectively the overall and catchment-specific *ensemble* threshold. Figure (e and f) show the mean preparedness per observed flood event per catchment for respectively the overall and catchment-specific *ensemble* threshold.

5.3.3 DISCUSSION

An overall (average over all catchments) *ensemble* threshold, based on the maximum CSI score obtained by the ensemble predictions, does not lead to an average gain in lead-time per flood event comparatively to a deterministic forecast. This result was obtained for all *streamflow* thresholds studies (Q70 to Q99) and for the 208 catchments included in dataset B1. For the 29 large catchments included in Dataset B2, a small average gain in lead-time is observed for the Q99 and Q70 thresholds (on average +0.80 days). However, when using the *ensemble* threshold, based on the ensemble CSI that equals the deterministic CSI, results show in almost all cases to a gain in preparedness ranging from 0.39 to 1.94 days (the highest value for the Q99 threshold).

Applying a catchment-specific *ensemble* threshold leads to a larger gain in lead time (or smaller loss in performance) for all *streamflow* thresholds, for both datasets and for both *ensemble* thresholds (maximum ensemble CSI and the one that equals CSI ensemble to the deterministic CSI). The largest gain in lead-time is observed for the Q99 *streamflow* threshold: 3.49 days in the 29 large catchments dataset and 3.21 in the dataset of all 208 catchments.

Only 26 catchments out of 208 (12.5%) show a loss in preparedness (catchments in blue in Figure 35) when using a catchment-specific ensemble threshold. The catchment-specific CSI curves show that the maximum (optimal) ensemble CSI score for these catchments is lower than the deterministic CSI score, indicating that the ensemble prediction has a lower performance, regarding this score, than the deterministic forecast. It might be the case that certain catchment characteristics (size, steepness) and climatic conditions (extreme heavy rainfall events) make the difference between the higher-resolution deterministic forecast and the probabilistic forecast.

5.4 MEASURES TO IMPROVE THE CSI AND PREPAREDNESS SCORES

The use of an ensemble streamflow prediction instead of a deterministic forecast results, on average, on better anticipation (longer lead-times) for high *streamflow* thresholds. The impact on the CSI score is also positive, but limited. A large number of threshold exceedances at the first day of a flood event remain undetected (missed) or are wrongly detected (false alarms). To tackle this problem in a preliminary analysis, we investigate two other possibilities to improve the CSI Score. The first one is the use of an adjustment factor of the *streamflow* threshold as is done in Chapter 4. The other approach focuses on the second day of the exceedance of the threshold in the flood event, given that the first day of a flood event is a critical one and difficult to forecast (paragraph 4.1.1). This part of the research is only conducted to explore other ways of improving the evaluation of the thresholds in flood forecasting and warning. That is why only dataset B2, Q99 threshold (closest to the actual thresholds used in France for operational forecast) and the overall results (average over al 29 catchments) are taken into account

ADJUSTING THE *STREAMFLOW* THRESHOLD

In Chapter 4 the *streamflow* threshold is multiplied by an adjustment factor (range: 0-1) in order to optimize the CSI: increasing the number of hits; decreasing the number of misses and increasing the number of false alarms.

Here, the adjustment factors (x) of 0.80, 0.85 and 0.90 are applied on the Q99 *streamflow* threshold. The maximum CSI is found when the Q99 *streamflow* threshold is multiplied by 0.85.

Figure 37 (a) shows the CSI curves for the conventional analysis (streamflow threshold = Q99; same as in Figure 34) and Figure 37(b) shows the CSI curves for the forecast where the streamflow threshold is adjusted by a factor 0.85. The CSI graphs and the results of this analysis show that:

- The maximum CSI is linked to a higher number of members (N) exceeding the adjusted threshold (20 instead of 10). As a consequence of lowering the streamflow threshold, more ensemble members (N) should exceed this threshold in order to issue an optimal warning.
- The adjustment factor has a low impact on the maximum ensemble CSI value (+0.003= ~3%).
- The adjustment factor has a larger impact on the improvement of the deterministic CSI (+0.05= ~20%).
- The preparedness for the adjusted Q99 threshold increases from 0.80 to 0.85 day per flood event, which is almost a negligible impact
- Even if the overall improvement seems to be small, a larger improvement, based on catchment-specific analyses, is to be expected.

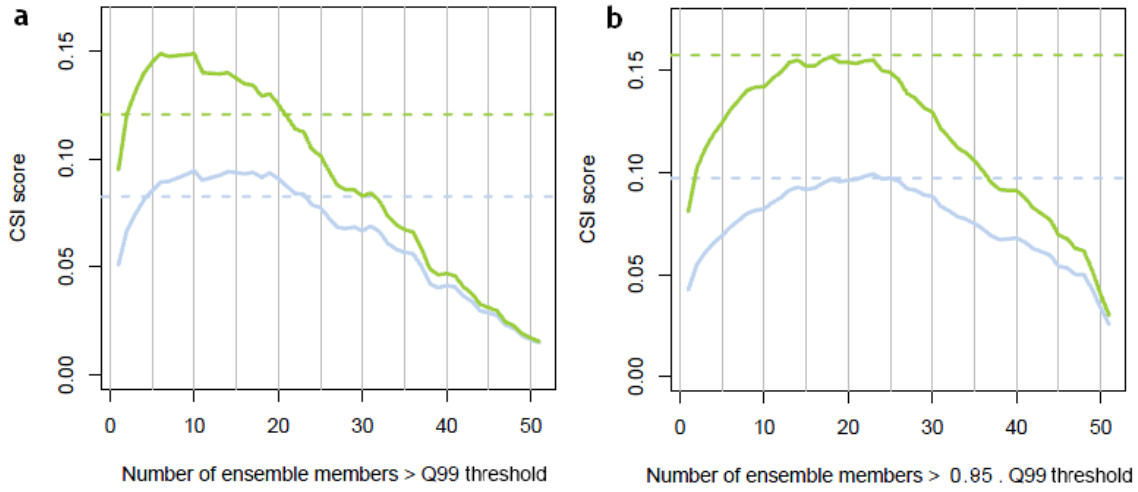


Figure 37 (a, left) CSI curve for the start of a Q99 threshold exceedance; (b, right) CSI curve for the start of an adjusted Q99 threshold exceedance (adjustment factor = 0.85). The blue (lower) curve represents CSI (α=1), the green curve represents CSI (α=0.5). The dotted lines are the CSI scores for the corresponding high resolution ECWMF deterministic forecast. The scores are aggregated over all catchments and the lead-times 3 to 9 days.

FORECASTING TOO LATE

The performance of the streamflow predictions is in general lower when considering only the ability of the system to forecast the start day of a flood event (paragraph 5.2.1), excluding therefore the subsequent days. Supposing that the forecast system is forecasting too late, due, for instance, to an error in the modeling chain, a possibility is to launch the warning a time step earlier than predicted. In this case, a warning is launched for *day X,* when a *streamflow* threshold exceedance is forecasted for *day X+1*.

Figure 38 (a, left) compares the CSI curves for the conventional analysis (same as in Figure 34) and Figure 38 (b, right) shows the CSI curves for the approach where the warning is launched a day before it is predicted (both curves concern predicted discharges exceeding the Q99 streamflow threshold). Remarkable is that the maximum CSI score for the second method is higher than the first one. For a target day X, a predicted threshold exceedance on day X+1 is a better indicator for an

actual exceedance of the Q99 threshold at day X than a predicted exceedance for the actual target day X. Figure 38 also shows that the optimal *ensemble* threshold increases when considering the second approach (focusing on the information of day X+1 to forecast for day X): 13 (instead of 10), i.e., more ensemble members have to exceed the Q99 threshold on day X+1 in order to launch a flood warning for day X. The preparedness (not shown) is however the same for both methods.

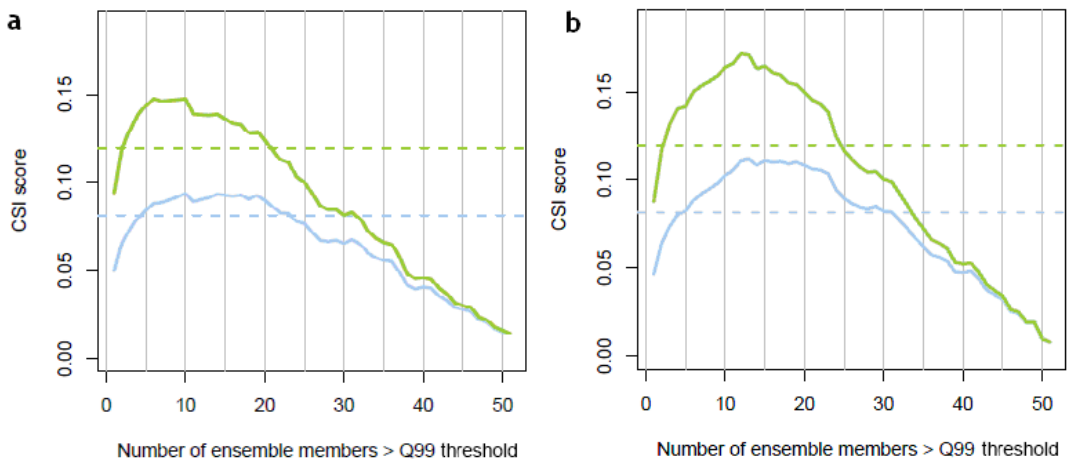


Figure 38 (a) CSI curve for the start of a flood event only. (b) CSI curve for the start of a flood event only focusing on the second day of a flood event. The blue (lower) cure represents CSI (α=1), the green curve represents CSI (α=0.5). The dotted lines are the CSI scores for the corresponding high resolution ECWMF deterministic forecast. The scores are aggregated over all catchments and the lead-times 3 to 9 days;

6 CONCLUSIONS AND RECOMMENDATIONS

Thresholds are a critical element in a flood forecasting and warning system. In the introduction and problem definition of this report (Chapter 1), two kinds of thresholds are distinguished: *streamflow* thresholds and *ensemble* thresholds.

The objective for this research project is defined as follows:

> To determine optimal appropriate critical thresholds for operational flood forecasting and warning by analysing the performance of a flood forecasting system and the quality of its forecasts when different thresholds – *streamflow* threshold and *ensemble* forecast probability thresholds – are used, while taking into account the influence of catchment characteristics and the type of the weather forecast (ensemble/deterministic) used to drive the hydrological model.

For the *streamflow* thresholds, a warning is issued if the discharge is higher than a predefined threshold. However, these thresholds cannot always be applied directly to most hydrological models. In fact, *streamflow* thresholds are usually based on instantaneous water levels that locally indicate flooding or river bankful conditions, while most hydrological forecast models work with an aggregated time step of one or several days. The challenge in the definition of the *streamflow* threshold is therefore to find an agreement between the locally defined (instantaneous) threshold and a threshold adapted to the time step of the model.

In this study, this is translated into the following research questions:

- How should the *streamflow* thresholds based on instantaneous observations be adjusted for the optimal implementation in daily time steps?
- What is the eventual relationship between this factor of adjustment and the catchment characteristics?

Together with *streamflow t*hresholds, there is a need for a second threshold when an ensemble streamflow prediction system is used: the *ensemble* threshold. This threshold is defined as the probability of exceeding a certain *streamflow* threshold required to launch a warning. In the case of a forecasting system based on ensemble predictions, as is the case of this study, where ECWMF ensemble forecasts are investigated, the question is: how many ensemble members exceeding a given streamflow threshold are necessary to launch a warning? The challenge for the *ensemble* threshold is therefore to find the optimal number of ensemble members exceeding the *streamflow* threshold, which is translated into the subsequent research questions:

- What is the optimal *ensemble* threshold (i.e. the number of ensemble members exceeding the *streamflow* threshold) for an optimal trade-off between hits, misses and false alarms and for a maximum preparedness in flood forecasting and warning?
- What are the eventual relationships between this optimum, the catchment characteristics, the *streamflow* threshold levels and the forecast lead-time?

The next paragraph (6.1) consists of the conclusions drawn from this study: the answers on the research questions addressed above. In paragraph 6.2 recommendations for further research are proposed and the question of how to implement the results and conclusions in operational flood forecasting and warning is addressed.

6.1 CONCLUSIONS

STREAMFLOW THRESHOLDS

An overall adjustment factor of 0.90 for the lowest operational threshold currently used by the French operational flood forecasting center SCHAPI (the yellow *streamflow* thresholds defined at each operational catchment) leads to the maximum Critical Success Index: i.e. the optimal trade-off between hits, misses and false alarms. This means that an optimum daily threshold would be the one given by multiplying the yellow threshold at each catchment by 0.90. A catchment-specific *streamflow* threshold adjustment factor leads even to higher CSI scores for every catchment. This catchment-specific threshold correlates with some of the catchment characteristics, namely: catchment size and reactivity. Both characteristics, which are themselves correlated as well, influence the steepness of the hydrograph's rising limb, which affects the discrepancy between the daily (model) and hourly (close to instantaneous) discharges. The adjustment factor for catchments with a smaller area and a higher reactivity is in general lower (range: 0.75-0.90) –i.e., the need for adjustment at these type of catchments is higher than the need for adjustment at the larger and slower catchments (adjustment factor range: 0.85-1.00).

Furthermore, we defined the *overall* adjustment factor for the $Q_{ix\ 2yr}$ (2-year return period flood) and for the second operational threshold from SCHAPI, the orange threshold, which are respectively: 0.90 and 0.95. A clear statistical link between the return periods of the thresholds and their corresponding adjustment factors could not be identified. However, the results indicate that the need for adjustment is higher for thresholds with a greater return period, since the POD-rate (Probability of Detection) decreases as the return period of the threshold increases.

The analysis was conducted to adjust the threshold for the discrepancy between instantaneous and daily observed discharge values. Therefore, only observed values were taken into account. Applying the adjusted threshold evaluated on the basis of observations in flood forecasting is very useful (it addresses the time-scales problem), but it might require an additional adjustment for the uncertainties related to the weather forecasts.

ENSEMBLE THRESHOLDS

The first step was to define an optimal *ensemble* threshold (i.e. number of ensemble members exceeding the *streamflow* threshold required for issuing a flood warning) for 208 catchments distributed all over France and covering a wide range of the hydroclimatic conditions and catchment characteristics (size and reactivity). The results show that there is no overall *ensemble* threshold for the streamflow predictions based on the ECWMF ensemble forecast which results in higher CSI (compared to a deterministic forecast) and also a gain in lead-time for the exceedance of the Q99 *streamflow* threshold. However, when a catchment-specific *ensemble threshold* is applied at the same Q99 threshold, the ensemble streamflow prediction results, for most catchments, in a higher CSI score and in an increase in preparedness (lead-time) of about 2-3 days. For the other thresholds (Q70, Q80, Q90 and Q95) a small loss in preparedness is also observed in the overall *ensemble* threshold analysis. However, in general, the CSI increases and the loss in preparedness reduces when a catchment-specific *ensemble* threshold is applied.

Under the hypothesis that the large grid size of the ECWMF forecast (about 45x45 km over France) might influence the performance of the streamflow predictions for the smaller catchments (<500 km^2), the analyses were also conducted over a dataset of 29 large catchments covering also a wide range of hydroclimatic conditions. The results of the reliability diagram analysis show that the reliability of

forecasted extreme events –forecasted exceedances of the Q99 *streamflow* threshold with a high probability- is higher for the catchments in this dataset of only large catchments. However, the contribution of the meteorological error is evaluated to be the same for both datasets. Under the hypotheses that observations are error-free, the difference in reliability seems therefore mainly due to the effect of a larger hydrological model error for the dataset containing smaller catchments. The results also show that the impact of the meteorological forecast error is higher for more extreme events. This is an interesting result since in a recent study by Olsson and Lindström (2008) no link was found between the level of the *streamflow* threshold and the magnitude of the meteorological forecast error.

The reliability diagram analysis shows a link between the size of the catchment and the reliability of the ensemble forecast: the streamflow prediction's reliability is higher for the dataset of large catchments. This probably explains the fact that a lower optimal *ensemble* threshold for the Q99 *streamflow* threshold was found for the dataset of 29 large catchments (10 members), compared to the optimal *ensemble* threshold found when all 208 catchments were analysed (14 members). In the case of large catchments an overall *ensemble* threshold of 10 ECWMF members leads to a higher CSI and a gain in lead-time of 0.80 day per flood event, compared to scores obtained when analyzing all 208 catchments (lower CSI and a no gain in lead-time).

For the large catchments, if we focus on the *ensemble* threshold which results in the same CSI score as the deterministic forecast (5 ECMWF members instead of 10 members), the gain in lead-time will even increase to 1.8 day per flood event. Additionally, when applying a catchment-specific *ensemble* threshold to these large catchments and for exceedances of the Q99 threshold, the results show a mean gain in preparedness of 2.6 days per flood event, related to maximum CSI score, and of 3.5 days, related to the CSI ensemble prediction score that equals the CSI of the deterministic forecast. Furthermore, it shows that ensemble flood forecasting is most valuable for the extreme events, where the increase of the CSI score and the gain in preparedness are the largest.

6.2 RECOMMENDATIONS

IMPLEMENTATION IN OPERATIONAL FLOOD FORECASTING AND WARNING

The adjustment factors for the yellow French operational threshold and for the *streamflow* thresholds corresponding to the 2-year return period flood found in this study can be implemented directly into an operational flood forecasting system running at daily time steps. The optimal balance among hits, false alarms and misses is reached when a catchment-specific adjustment factor is applied, and this is the recommended adjustment factor. However, if this is undesirable, a regional adjustment factor might be an alternative. This regional adjustment factor can be the outcome of a multiple regression analysis with basins characteristics as surface and reactivity as predictors. The adjustment factors for the orange French operational thresholds are based on a limited number of exceedances and therefore not robust enough to apply directly or without caution.

FURTHER RESEARCH

The analysis of the *streamflow* threshold was conducted with a limited range of possible adjustment factors. If the CSI score is determined within the same range (0.75-1.00), but with smaller steps (e.g. 0.01 instead of 0.05 used in this study), small shifts in the optimal adjustment factor might result. Besides, when considering other changes in the adjustment factors investigated, it might become easier to find a more clear relationship between the adjustment factors and the catchment characteristics.

Furthermore, it is desirable to conduct the same analysis with longer data series in order to evaluate the adjustment factors for higher thresholds (e.g. the French operational orange and red thresholds) for all catchments of interest. Longer data series are also desirable for the determination and the evaluation of the ensemble thresholds. The highest *streamflow* threshold (Q99) evaluated on the basis of the 18-month period here analyzed is, for most catchments, smaller than the operational yellow threshold.

Another point for further investigation is that the focus of a hydrological ensemble prediction analysis should not only be on improving the ensemble forecasts, but also on reducing the uncertainty related to the hydrological model. In the case of the GRP model used in this study, the impact of this error is in as large as the meteorological errors, especially when smaller catchments are considered in the analysis. Furthermore, a calibration of the model on peak discharges might lead to a reduction of the magnitude of the hydrological model error during flood events.

This report shows that ensemble flood forecasting is a valuable alternative to the high resolution deterministic flood forecasting. In most of the studied catchments, the optimal *ensemble* threshold leads to a gain in lead-time and a higher CSI score compared to the deterministic forecast. However, in some catchments, the deterministic forecast performs better than the ensemble forecast. Therefore, in an optimal flood warning system, it could be useful to implement some steering rules that could account for both types of forecasts: for example, a warning should be issued when ensemble predictions exceed the (adjusted) *streamflow* threshold with the catchment-specific (or regional) *ensemble* threshold **OR** when the deterministic forecast exceeds the (adjusted) *streamflow* threshold. These recommendations are the outcome of paragraph 5.4, which shows that the probabilistic and deterministic forecasts can be improved by various measures as well. Adjusting the *streamflow* threshold for variations between observed and forecasted values is one of the ways to improve the deterministic forecast. The other one is to take into account the temporal variation between the forecast and the observation by issuing a warning a day (time step) earlier. Both measures are topics for further research.

7 REFERENCES

Andréassian, V., Hall, A., Chahinian, N., and Schaake, J. (2006). *Introduction and Synthesis: Why should hydrologists work on a large number of basin data sets?* IAHS-AISH Publication, 307, 1–5.

Banque Hydro (n.d.). Catchment data available at: http://www.hydro.eaufrance.fr.

Bartholmes, J., Todini, E. (2005). Coupling meteorological and hydrological models for flood forecasting. *Hydrological and Earth System Sciences,* 9, 333-346.

Berthet L. (2010). *Prévision des crues au pas de temps horaire: pour une meilleure assimilation de l'information de débit dans un modèle hydrologique.* Thèse de Doctorat, Cemagref Antony, France and AgroParisTech, Paris, France.

Berthet L., Andréassian V., Perrin C., Javelle P. (2009). How crucial is it to account for the antecedent moisture conditions in flood forecasting? Comparison of event-based and continuous approaches on 178 catchments. *Hydrology and Earth System Sciences*, 13, 819–831.

Beven, J.K. (2001). *Rainfall-Runoff Modeling: the primer.* West Sussex, United Kingdom: John Wiley and Sons

Buizza, R. (2002*). Ensemble predictions*. Encyclopaedia of Atmospheric Sciences, Academic Press. Edited by Holton J.R., Pyle, J., Curry, J.A.

Buizza, R., Houtekamer, P.L., Toth, Z., Pellerin, G., Wei and Zhu,Y. (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, 133, 1076–1097.

Carpenter, T.M., Sperfslage, J.A., Georgakakos, K.P., Sweeney, T. and Fread, D.L. (1999). National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems. *Journal of Hydrology*, 224, 21-44.

Carte de vigilance "crues". (n.d.). Retrieved 23 November 2009 from: htpp://www.vigicrues.ecologie.gouv.fr.

Chow, V., Maidment, D.R., Mays, L.W. (1988) *Applied Hydrology*. McGraw-Hill, USA.

Clark, M. and Hay, L. (2004.) Use of medium-range numerical weather prediction model output to produce forecasts of streamflow, *Journal of Hydrometeorology,* 5, 15–32.

Cloke, H.L. and Papppenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology,* 375, 613-626.

CRED (n.d.).Disaster profiles. Retrieved at 10 april 2010 from : http://www.emdat.be/ result-disaster-profiles?disgroup=natural&dis_type=Flood&period=1991$2010

Delrieu, G., Ducrocq, V., Gaume,E., Nicol, J., Payrastre, O., Yates, E., Andrieu, H., Ayral, P., Bouvier, C., Creutin, J., Livet, M., Anquetin, S., Lang, M., Neppel, L., Obled, C., Parent-du-Châtelet, J., Saulnier, G., Walpersdorf, A. and Wobrock, W. (2005). The catastrophic flash-flood event of 8-9 September 2002 in the Gard region, France: a first case study for the Cévennes-Vivarais. *Journal of Hydrometeorology*.

Edlund, C. (2009). *Probabilistic flood forecasting at SMHI.* retrieved at 16 October 2009 from:http://www.ecmwf.int/newsevents/meetings/workshops/2009/EFAS/presentations/05_edlund_SMHI_probabilistic_flood_forecasting.pdf

EXCIFF - European exchange circle on flood forecasting (2005) *Current state-of-play and future needs of flood forecasting in Europe*. Proceedings 1st workshop, 14-15.04.05, Toulouse, France, MEDD & DG JRC, CD-ROM.

Gaume, E. Livet, M., Desborde, M. and Villeneuve, J.P. (2004). Hydrological analysis of the river Aude, France, flash flood on 12 and 13 November 1999. *Journal of Hydrology*, 286 1-4, 135-154.

Gerapetritis, H. and Pelissier, J.M. (2004). The Critical Success Index and Warning Strategy. Proceedings of the 17th Conference on Probability and Statistics in the Atmospheric Sciences, American Meteorological Society.

Van der Grijn, G. (2002). Tropical Cyclones forecasting at ECMWF: New Products and Validation. *ECMWF Technical Memorandum*, 386.

Goudeleeuw, B.T. Thielen, J., de Roo, A.P.J. and Buizza, R., (2005). Flood forecasting using probabilistic weather predictions. *Hydrological Earth System Sciences*, 9, 87-102.

Horizontal resolution increase (n.d.). Retrieved on 26 January 2010, from: http://www.ecmwf.int/publications/cms/get/ecmwfnews/251.

Horizontal resolution increase 2009 (n.d). Retrieved on 26 January 2010, from: http://www.ecmwf.int/products/changes/horizontal_resolution_2009/

Introduction to chaos, predictability and ensemble forecasts. (n.d.). Retrieved on 29 January 2010 from: http://www.ecmwf.int/research/predictability/background/index.html.

Killingtveit, Å. and Sælthun, N.R. (1997). *Hydropower Development (Vol. 7) : Hydrology.* Trondheim, Norway: Norwegian University of Science and Technology, Division of Hydraulic Engineering.

Matsueda, M. and Tanaka, H.L. (2008). Can MCGE Outperform the ECMWF Ensemble? *Scientific Online Letters on the Atmosphere*, 4, 77-80.

Météo France (n.d.). Climat en France. Retrieved on 24 April 2010 from: http://france.meteofrance.com/france/climat_france

Ministère de l'Ecologie, l'Energie, du Développement durable et de la Mer (2006). Circulaire relative à la production opérationnelle de la vigilance crues.

Ministère de l'Ecologie, l'Energie, du Développement durable et de la Mer (2009). *Prévention des inondations.* Retrieved on 15 January 2010 from: http://www.developpement-durable.gouv.fr/IMG/pdf/Prevention_des_inondations.pdf

Nobert, S., Demeritt, D. and Cloke H. (2009). Using Ensemble Predictions for Operational Flood Forecasting: Lessons from Sweden. *Journal of Flood Risk Management* 27 July 2009.

Palmer, T.N., Barkmeijer, J., Buizza, R. and Petroliagas, T. (1997). The ECWMF Ensemble Prediction System. *Meteorological Applications,* 4, 301-304.

Palmer, T., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., Smith, L. (2005) Ensemble prediction: a pedagogical perspective. *ECMWF Newsletter* 106, 10-17.

Pappenberger, F., Beven, K.J., Hunter, N.M., Bates, P.D., Gouweleeuw,B.T., Thielen, J., De Roo, A.P.J. (2005). Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions with the European Flood Forecasting System (EFFS).*Hydrology and Earth System Sciences*, 9, 381–393.

Pappenberger, F., Scipal, K., Buizza, R. (2008). Hydrological aspects of meteorological verification. *Atmospheric Science Letters*, 9 ,43-52.

Perrin, C. (2002). Vers une amélioration d'un modèle global pluie-débit au travers d'une approche comparative. *La Houille Blanche*, 6-7, 84-91.

Pingel, N., Jones, C. and Ford, D. (2005). Estimating Forecast Lead Time. *Natural Hazards Review,* 6, 60-66.

Olsson, J. and Lindström, G. (2008). Evaluation and calibration of operational hydrological ensemble forecast in Sweden. *Journal of Hydrology,* 350, 14-24.

Ramos, M.H., Thielen, J., Pappenberger, F. (2008) Using weather ensembles for operational flood forecasting and early warning. *Proc. Colloque SHF «Prévisions hydro-météorologiques»,* Lyon, 18-19 November 2008.

Ramos, M.H., Thielen, J., de Roo, A. (2009) Prévision hydrologique d'ensemble et alerte avec le système européen d'alerte aux crues (EFAS) : cas des crues du bassin du Danube en août 2005. In: Tanguy, J.-M. (Dir.), *Traité d'hydraulique environnementale - Volume 7 : applications des modèles numériques en ingénierie 1*, Chap. 5. Ed. Hermès Lavoisier, Oct. 2009, Paris, France, 190p.

Randrianasolo, A., Ramos, M.H., Thirel, G., Andreassian, V., Martin, E. (2009). *Impact of the use of two different hydrological models on scores of hydrological ensemble forecasts*. Poster presented at the 4th HEPEX Workshop, Toulouse, France, 15-18 June 2009.

Randrianasolo, A. (2009). *Evaluation de la qualité des prévisions pour l'alerte aux crues*. Mémoire de Master. AgroParisTech, ENGREF, Université Pierre et Marie Curie, Cemagref, Antony,

Renner, M., Werner, M.G.F., Rademacher, S. and Sprokkereef, E. (2009). Verification of ensemble flow forecasts for the River Rhine. *Journal of Hydrology*, 376, 463-475.

Rousset-Regimbeau, F., Habets, F., Martin, E., Noilhan, J. (2007). Ensemble streamflow forecasts over France. *ECMWF Newsletter* 111, 21-27.

SCHAPI (2008). Ensemble streamflow forecasting over France: use of the system SIM. Retrieved on 15 September 2009 from: http://www.ecmwf.int/newsevents/meetings/workshops/2009/EFAS/presentations/07_deSaintAubin_SIM_EPS_French.pdf

Schaeffer, J.T. (1990).The Critical Success Index as an Indcator of Warning Skill. *Weather and Forecasting,* 5, 570-575.

SMHI (n.d.). Nederbörd och klimatdata. Retrieved on 24 April 2010 from: http://www.smhi.se/klimatdata/meteorologi/nederbord

Subramanya, K. (2006). *Engineering Hydrology. Second edition.* New Dehli, India. Tata McGraw-Hill Publishing Company Limited.

Sprokkereef, E. (2009). Visualising and communicating probabilistic flow forecasts in The Netherlands. Retrieved at 12 October 2009 from: http://www.ecmwf.int/newsevents/meetings/workshops/2009/EFAS/presentations/06_EFAS_User_Meeting_29_30_January_2009_Sprokkereef.pdf

Tangara, M. (2005). *Nouvelle méthode de prévision de crue utilisant un modèle pluie-débit global*. Thèse de Doctorat, Cemagref Antony, France and Ecole pratique des hautes Etudes, Paris, France.

Thielen, J., Bartholmes, J., Ramos, M.H. & De Roo, A. (2009). The European Flood Alert System – Part 1: Concept and development. *Hydrology and Earth System Sciences*, 13, 125-140.

Thirel, G., Rousset-Regimbeau, F., Martin, E., Habets, F. (2008). On the Impact of Short-Range Meteorological Forecasts for Ensemble Streamflow Prediction. *Journal of Hydrometeorology*, 9, 1301-1317.

UN (2004). Guidelines for Reducing Flood Losses.

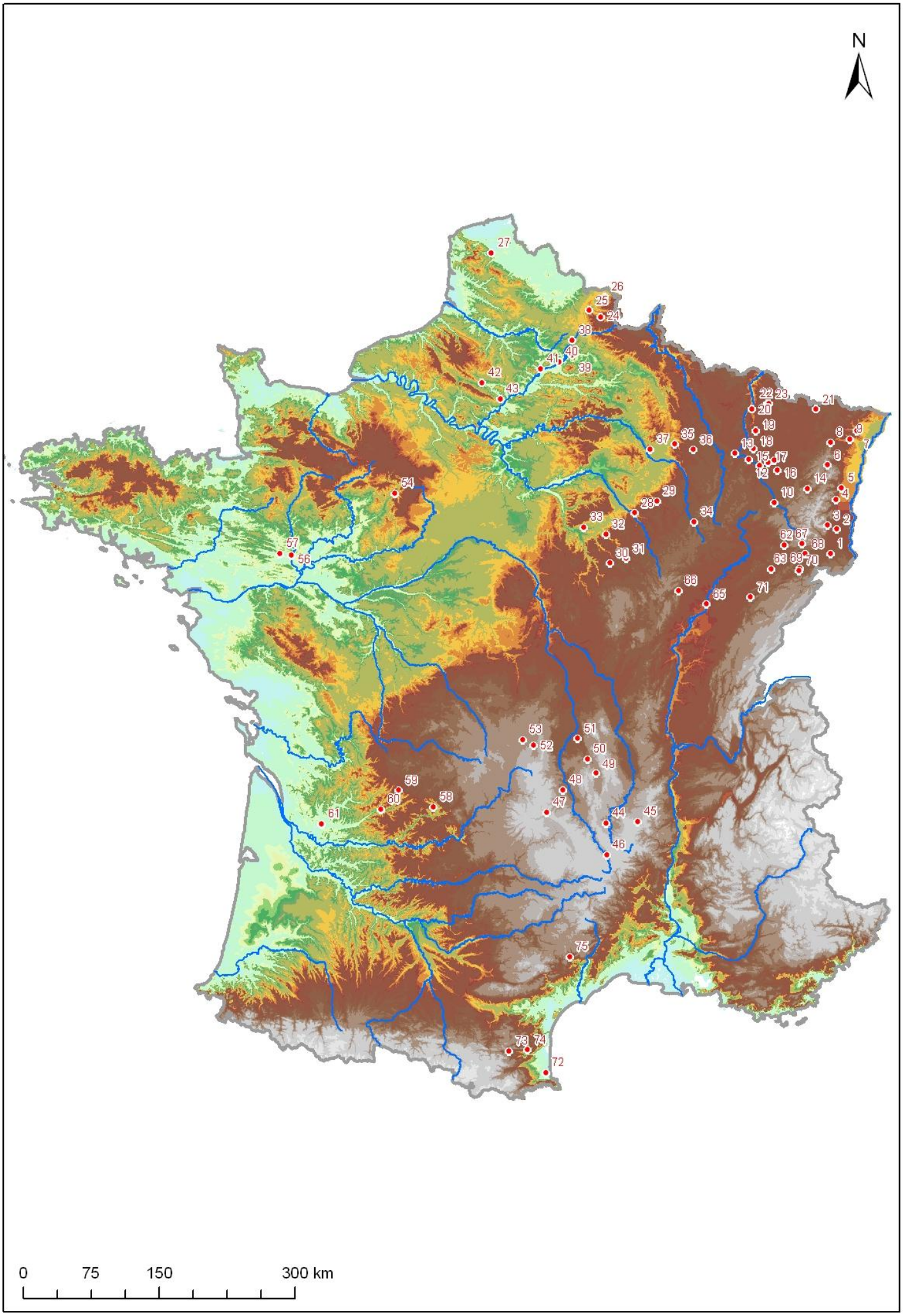
WMO (2007). Forecast verification – issues, methods and faq. WWRP/WGNE Joint Working Group on Verification.

APPENDICES

A. 1 LOCATION AND NAMES OF CATCHMENTS

DATASET A

Nb	Code	Name
1	A1050310	L'Ill à Altkirch
2	A1310310	L'Ill à Ensisheim
3	A1522020	La Lauch à Guebwiller
4	A2052020	La Fecht à Ostheim
5	A2352020	Le Giessen à Sélestat [Amont]
6	A2732010	La Bruche à Russ [Wisches]
7	A3301010	La Moder à Schweighouse-sur-Moder [Aval]
8	A3422010	La Zorn à Saverne [Schinderthal]
9	A3472010	La Zorn à Waltenheim-sur-Zorn
10	A4250640	La Moselle à Épinal
11	A5110610	La Moselle à Tonnoy
12	A5500610	La Moselle à Pont-Saint-Vincent
13	A5730610	La Moselle à Toul
14	A6051020	La Meurthe à Saint-Dié
15	A6341010	La Meurthe à Lunéville
16	A6731220	La Mortagne à Gerbéviller
17	A6761010	La Meurthe à Damelevières
18	A6941020	La Meurthe à Malzéville [2]
19	A7821010	La Seille à Nomeny
20	A7881010	La Seille à Metz
21	A9301010	La Sarre à Wittring
22	A9752010	La Nied Francaise à Condé-Northen [Pontigny]
23	A9832010	La Nied Allemande à Faulquemont
24	D0137010	L'Helpe Mineure à Étroeungt
25	D0137020	L'Helpe Mineure à Maroilles
26	D0206010	La Solre à Ferrière-la-Grande
27	E4035710	L'Aa à Wizernes
28	H0400010	La Seine à Bar-sur-Seine
29	H1201010	L'Aube à Bar-sur-Aube
30	H2332020	Le Serein à Dissangis
31	H2452020	L'Armançon à Aisy-sur-Armançon [Aval]
32	H2462020	L'Armançon à Tronchoy
33	H2482010	L'Armançon à Brienon-sur-Armançon
34	H5011020	La Marne à Marnay-sur-Marne
35	H5102030	La Saulx à Mognéville
36	H5122340	L'Ornain à Tronville-en-Barrois
37	H5172010	La Saulx à Vitry-en-Perthois
38	H7061010	L'Oise à Origny-Sainte-Benoite
39	H7162010	La Serre à Nouvion-et-Catillon [Pont à Bucy]
40	H7201010	L'Oise à Condren
41	H7401010	L'Oise à Sempigny
42	H7742010	Le Thérain à Beauvais
43	H7742020	Le Thérain à Maysel
44	K0253020	La Borne occidentale à Espaly-Saint-Marcel
45	K0403010	Le Lignon du Velay au Chambon-sur-Lignon
46	K2070810	L'Allier à Langogne
47	K2523010	L'Alagnon à Joursac [Joursac-le-Vialard]
48	K2593010	L'Alagnon à Lempdes
49	K2851910	La Dore à Ambert
50	K2871910	La Dore à Tours-sur-Meymont [Giroux]
51	K2981910	La Dore à Dorat
52	K3222010	La Sioule à Pontgibaud
53	K3273010	Le Sioulet à Pontaumur [La Prugne]
54	M0361510	L'Huisne à Nogent-le-Rotrou [Pont de bois]
55	M3600910	La Mayenne à Château-Gontier
56	M3630910	La Mayenne à Chambellay
57	M3851810	L'Oudon à Segré [Écluse de Maingué]
58	P3274010	La Loyre à Saint-Viance [Pont de Burg]
59	P6081510	L'Isle à Corgnac-sur-l'Isle
60	P7041510	L'Isle à Périgueux
61	P7261510	L'Isle à Abzac
62	U1014020	L'Ognon à Montessaux
63	U1044010	L'Ognon à Chassey-lès-Montbozon [Bonnal]
64	U1084010	L'Ognon à Pesmes
65	U1120010	La Saône à Auxonne
66	U1324010	L'Ouche à Plombières-lès-Dijon
67	U2345020	La Bourbeuse à Froidefontaine
68	U2345030	La Savoureuse à Belfort
69	U2354010	L'Allan à Courcelles-lès-Montbéliard
70	U2402010	Le Doubs à Voujeaucourt
71	U2512010	Le Doubs à Besançon
72	Y0284060	Le Tech à Argelès-sur-Mer [Pont d'Elne]
73	Y0624020	L'Agly à Saint-Paul-de-Fenouillet [Clue de la Fou]
74	Y0655010	Le Verdouble à Tautavel
75	Y2214010	La Lergue à Lodève

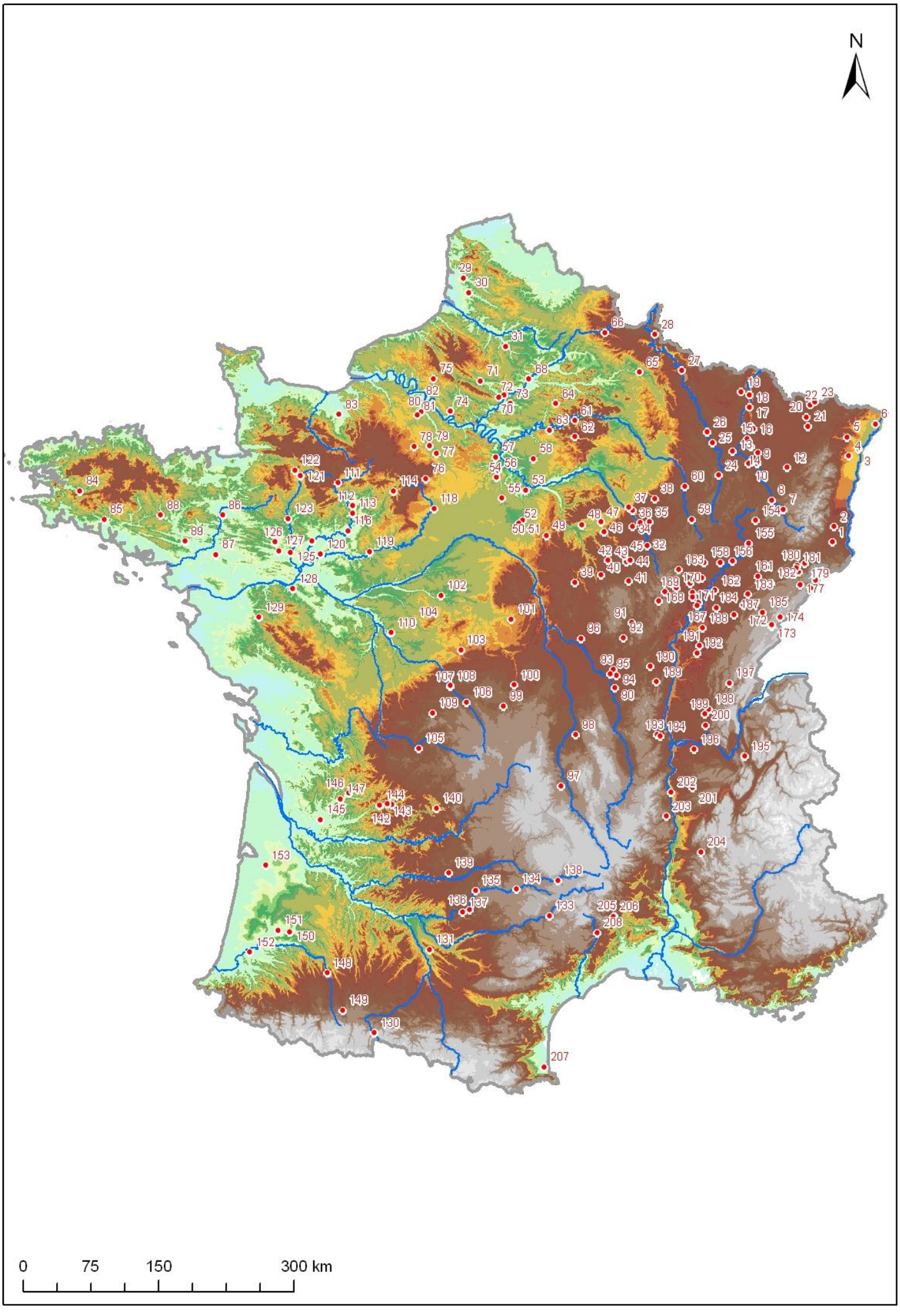

N
27
26
25
24
38
40
41
39
42
43
22
23
21
20
19
8
9
7
37
35
36
13
18
6
15
17
12
16
14
5
54
29
10
4
28
34
3
2
33
32
62
67
68
1
57
56
30
31
63
69
70
66
71
65
53
51
52
50
49
48
59
60
58
47
44
45
61
46
75
73
74
72
0
75
150
300 km

DATASET B1 AND B2

Catchments codes and names for the catchments included in Dataset B1. Dataset B2 consist of 29 large catchments, which are highlighted in this overview,

Nb	Code	Name
1	A1080330	L'Ill à Didenheim
2	A1310310	L'Ill à Ensisheim
3	A2280350	L'Ill à Strassbourg
4	A2860110	La Bruche à Holtheim
5	A3472010	La Zorn à Waltenheim-sur-Zorn
6	A3792010	La Sauer à Beinheim
7	A4200630	La Moselle à Saint-Nabord [Noir Gueux]
8	A4250640	La Moselle à Épinal
9	A5110610	La Moselle à Tonnoy
10	A5431010	Le Madon à Mirecourt [Station annonce de crue]
11	A5730610	La Moselle à Toul
12	A6221010	La Meurthe à Azerailles
13	A6921010	La Meurthe à Laneuveville
14	A6941020	La Meurthe à Malzéville [2]
15	A7010610	La Moselle à Custines
16	A7821010	La Seille à Nomeny
17	A7881010	La Seille à Metz
18	A7930610	La Moselle à Hauconcourt
19	A8431010	L'Orne à Rosselange
20	A9091050	La Sarre à Keskastel
21	A9091060	La Sarre à Diedendorf
22	A9221010	La Sarre à Sarreinsming
23	A9425050	La Bies à Bliesbruck
24	B1150010	La Meuse à Domremy la Pucelle
25	B2130010	La Meuse à Commercy
26	B2220010	La Meusse à St-Mihiel
27	B3150020	La Meusse à Stenay (2)
28	B6111010	La Semoy à Haulme
29	E5400310	La Canche à Brimeux
30	E5505720	L'Authie à Dompierre-sur-Authie
31	E6406010	L'Avre à Moreuil
32	H0100010	La Seine à Nod-sur-Seine
33	H0100020	La Seine à Plaine-St-Lange
34	H0203030	La laignes aux Riceys
35	H0321030	L'Ource à Autricourt
36	H0400010	La Seine à Bar-sur-Seine
37	H0400020	La Seine à Courtenot
38	H1201010	L'Aube à Bar-sur-Aube
39	H2062010	Le Beuvron à Ouagne (Champmore)
40	H2172320	Le Cousin à Availlon
41	H2322010	Le Serien à Biere-les-Sumur
42	H2332020	Le Serein à Dissangis
43	H2412010	L'Armancon à Quinchy-le-Vicomt
44	H2442340	La Brenne à Montbard
45	H2452020	L'armancon à Aisy -sur-Armancon
46	H2462020	L'Armancon Tronchoy
47	H2473010	L'Armance à Chessy-les-Pres
48	H2482010	L'Armancon à Brienon-sur-Armacon
49	H3122010	L'Ouanne à Chargny
50	H3122020	L'Ouanne à Gy-les-Nonains
51	H3201010	Le Loing à Chalette-sur-Loing
52	H3322010	La Bezonde à Pannes
53	H3621010	Le Loing à Pisy
54	H4022020	L'Esonne à Guigneville-sur-Esonne
55	H4022030	L'Esonne à Boulancourt
56	H4042010	L'Esonne à Boulancourt-sur-Esonne
57	H4252010	L'Orge à Morsang-sur-Orge
58	H4322030	L'Yerres à Courtomer
59	H5011020	La Marne à Marnay-sur-Marne
60	H5062010	Le Rognon à Doulan-Saucourt
61	H5302010	Le Surmelin à St- Eugene
62	H5412010	Le Petit Morin à Montmirail
63	H5412020	Le Petit Morin à Jouarre
64	H5522010	L'Ourcq à Chouy
65	H6221010	L'Aisne à Grivy
66	H7033010	Le Theon à Origny-en-Thierache
67	H7401010	L'Oise à Sempigny
68	H7423710	L'Aronde à Clairox
69	H7513010	L'Automne à Saintines
70	H7602010	La Breche à Nogent-sur-Oise
71	H7742010	Le Therain à Beauvais
72	H7742020	Le Therian à Maysel
73	H7813010	La Nonette à Courteuil
74	H8042010	L'Epte à Fourges
75	H8212010	L'Andelle à Vascoeuil
76	H9021010	L'Eure à St-Luperce
77	H9121010	L'Eure à Charpont
78	H9202010	L'Avre à Acon
79	H9222010	L'Avre à Muzy
80	H9331010	L'Eure à Cailly-sur-Eure
81	H9402030	L'Iton à Normanville
82	H9501010	L'Eure à Louviers
83	I2051040	La Dives au Mesnil-Mauger
84	J3811810	L'aulne à Chtaeuneuf-du-Faou
85	J4742010	L'Elle à Arzano
86	J7483010	La Seiche a Bruz (Carce)
87	J7963010	Le Don à Guemene-Penfao
88	J8202310	L'oust à Pleugriffet
89	J8502310	L'Oust à St-Grave
90	K1173210	L'Arconce à Montceaux-Etoil
91	K1251810	L'Arroux à Dracy-St-Loup
92	K1321810	L'Arroux à Etang-sur-Arroux
93	K1341810	L'Arroux à Rigny-sur-Arroux
94	K1383010	La Bourbince à Vitry-en-Charol
95	K1391810	L'Arroux à Digouin
96	K1773010	L'Aron à Verneuill
97	K2593010	L'Alagon à Lempdes
98	K2981910	La Dore à Dorat
99	K5183010	La Tardes à Ecaux-les-Bains
100	K5220910	Le Cher à St-Victor
101	K5552300	L'Yevre à Savigny-en-Septane
102	K6492510	La Sauldre à Selles-sur-Cher
103	K7202610	L'Indre à Ardentes
104	K7312610	L'Indre à St-Cyran-du-Jambot
105	L0563010	La Briance à Condat-sur-Vienne

106	L4210710	La Creuse à Glenic
107	L4220710	La Grande Cruesseà Fresselines
108	L4411710	La Petite Creusse à Fresselines
109	L5101810	La Gartempe à Besinnes-sur-Gartempe
110	L6202030	La Claisee au Granns-Pressigny (2)
111	M0050620	La Sarthe a St-Ceneri-le-Gerei
112	M0243010	L'Orne Saosnoise à MontBizot
113	M0250610	La Sarthe à Neuvilee-sur-Sarthe
114	M0361510	L'Huisne à Nogent-le-Rotrou
115	M0421510	L'Huisne à Montfort-le-Gesnoi
116	M0500610	La Sarthe à Spay
117	M0680610	La Sarthe à St-Denis-D'Anjourd
118	M1041610	Le Loir à St-Maur-sur-le-Loir
119	M1341610	Le Loir à Flee (Port Gautier)
120	M1531610	Le Loir à Durtal
121	M3060910	La Mayenne à Ambrieres-les-Valles
122	M3133010	La Varenne à St-Fraimbault
123	M3340910	La Mayenne à l'Huisserie
124	M3600910	La Mayenne à Chateau-Gontier
125	M3630910	La Mayenne à Chambellay
126	M3771810	L'Oudon àChatelais
127	M3851810	L'Oudon à Segre (Ecluse)
128	M5222010	Le Layon à St-Lambert-du-Latta
129	M7112410	La Sevre Nantaise à Tiffauges
130	O0010040	La Garonne à Saint-Beat
131	O2344010	Le Girou à Cepet
132	O3141010	Le Tarn à Mostuejouls
133	O3401010	Le Tarn à Millau
134	O5092520	L'Aveyron à Onet-le-Chateau
135	O5192520	L'Aveyron à Villefranche
136	O5292510	L'Aveyron à Lagueppie
137	O5664010	Le Cerou à Milhars
138	O7101510	Le Lot à Banassac
139	O8133520	Le Cele à Orniac
140	P3922510	La Correze à Brive-la-Gaillard
141	P6161510	L'Isle à Mayac
142	P6382510	L'Auvezere au Change
143	P7001510	L'Isle à Bassilac
144	P7041510	L'Isle à Perigeux
145	P7261510	L'sle à Abzac
146	P8284010	La Lizonne à St-Severin
147	P8312520	La Dronne à Bonnes
148	Q0280030	L'Adour à Estirac
149	Q0522520	L'Arros à Gourgue
150	Q2192510	Le Midou à Mont-de-Marsan
151	Q2593310	La Midouze à Campagne
152	Q3464010	Le Luy à St-Pandelon
153	S2242510	L'Eyre à Salles
154	U0124010	Le Coney à Fontenou-le-Chateau
155	U0474010	La Lanterne à Fleurey
156	U0610010	La Saone à Ray-sur-Saone
157	U0724010	Le Salon à Denevre
158	U0924010	La Vingeanne à St-Maurice-sur-Vingeanne
159	U0924020	La Vingeanne à Oisilly
160	U1044010	L'Ognon à Chassey-les- Montboz
161	U1054010	L'Ognon à Beaumotte Aubertrans
162	U1084010	L'Ognon à Pesmes
163	U1215030	L'Ignon à Villecomte
164	U1224010	La Tille à Arceau (Arcelot)
165	U1224020	La Tille à Cessey-sur-Tille
166	U1235020	La Norges à Genlis
167	U1244040	La Tille à Champdottre
168	U1314010	L'Ouche à la Bussiere-sur-Ouche
169	U1314020	L'Ouche à ste-Marie-sur-Ouche
170	U1324010	L'Ouche à Plombieres-les-Dijon
171	U1334010	L'Ouche à Trouhans
172	U2022010	Le Doubs à la Cluse-et-Mijoux
173	U2022020	Le Doubs à Doubs
174	U2102010	Le Doubs à Ville-du-Pont
175	U2122010	Le Doubs Goumois
176	U2142010	Le Doubs à Glere (Courclavon)
177	U2215020	Le Dessoubre à St-Hyppolyte
178	U2222010	Le Ddoubs à Mathay
179	U2334010	L'Allan à Feschest-le-Chatel
180	U2354010	L'Allan à Courcelles-les-Mont
181	U2402010	Le Ddoubs à Voujeaucourt
182	U2425260	Le Cusnacin à Baume-les-Dames
183	U2512010	Le Doubs à Besancon
184	U2542010	Le Doubs à Rochefort-sur-Nenon
185	U2604030	La Loue à Vuillefans
186	U2624010	La Loue à Chenecey-Buillon
187	U2634010	La Loue à Champagne-sur-Loue
188	U2722010	Le Doubs à Nneublans-Abergement
189	U3214010	La Grosne à Jalogny
190	U3225010	La Guye à Sigy-le Chatel
191	U3415030	La Brenne à Sen-sur-Seille
192	U3424010	La Seille à St-Usuge
193	U4624010	L'Azergues à Chatillion
194	U4644010	L'Azergues à Lozanne
195	V1315020	La Leysse à la Motte-Servolex
196	V1774010	La Bourre à Tignieu
197	V2444020	La Bienne à Jeurre
198	V2814020	Le Suran à Neuville-sur-Ain
199	V2814030	Le Suran à Pont-d'Ain
200	V2934010	L'Albarine à St-Denis-en-Buge
201	V3424310	Le Rival à Beaufort
202	V3434010	Les Collieres à St-Rambert
203	V3724010	Le Doux à Colmbier-le-Vieux
204	V4264010	La Drome à Saillans
205	V7124010	Le Gardon de Mialet à Generargues
206	V7135010	Le Gardon de St-jean
207	Y0284060	Le Tech à Argeles-sur-Mer
208	Y2102010	L'Herault à Laroque


N
0
75
150
300 km

A. 2 STRUCTURE OF THE *GRPE* MODEL

The GRPE model is an ensemble adaptation of the lumped hydrological forecasting model GRP, developed at Cemagref (http://www.cemagref.fr/webgr/). This appendix shows the structure of the GRP model and it is based on the description provided by Berthet (2010). Its entries are the accumulated rainfall P_t and the basin's potential evapotranspiration E_t. In practice, the climatological value for the potential rainfall is used. The GRP model consists in fact of a production function and a routing function. The model describes the sequence of operations taking place in one time step. This time step equals one day in the forecast model used in this study. Figure 39 gives the general structure of the GRP model:

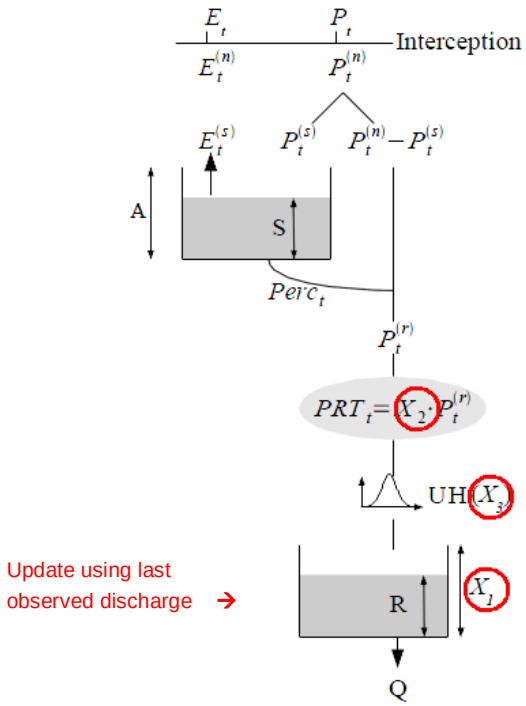


Figure 39. Model structure of the GPR hydrological forecasting model.

A 2.1 PRODUCTION FUNCTION

INTERCEPTION

The production function is used to calculate the effective precipitation. The model starts with the interception phase where it computes the net rainfall $P_t^{(n)}$ and the net evapotranspiration:

$$\begin{cases} P_t^{(n)} = P_t - E_t \text{ and } E_t^{(n)} = 0 \text{ if } P_t \geq E_t \\ E_t^{(n)} = E_t - P_t \text{ and } P_t^{(n)} = 0 \text{ if } E_t \geq P_t \end{cases}$$

PRODUCTION RESERVOIR

The net flow interacts with the production reservoir. If the net precipitation is positive, part of the precipitation $P_t^{(s)}$ is directed to the production reservoir while the rest $P_t^{(n)}$ - $P_t^{(s)}$ flows to the routing function (see below). The fraction of net rainfall stored in the reservoir depends on the level of filling of the reservoir (equivalent to the state of the soil moisture of the basin):

$$P_t^{(s)} = A \cdot \frac{\left(1 - \left(\frac{S_{t-1}}{A}\right)^2\right) \cdot tanh\left(\frac{P^{(n)}}{A}\right)}{1 + \frac{S_{t-1}}{A} \cdot tanh\left(\frac{P_t^{(n)}}{A}\right)}$$

where A is the capacity of production reservoir and S_t its level at time step t.

No water leaves the tank through evapotranspiration if $E_t^{(s)} = 0$. In the opposite case, when the net evapotranspiration escapes from the production reservoir (and is lost for the model) ,$E_t^{(s)}$ equals:

$$E_t^{(s)} = \frac{S_{t-1} \cdot \left(2 - \frac{S_{t-1}}{A}\right) \cdot tanh\left(\frac{E_t^{(n)}}{A}\right)}{1 + \left(1 - \frac{S_{t-1}}{A}\right) \cdot tanh\left(\frac{E_t^{(n)}}{A}\right)}$$

The level of the production reservoir is modified according the formula:

$$S_t^+ = S_{t-1} - E_t^{(s)} + P_t^{(s)}$$

The production reservoir should not be seen as a function which slows down the runoff, but acts as a "counter" memory of the soil moisture condition, which modulates the catchment runoff.

PERCOLATION

The reservoir loses water to the routing function, according the function: $dS = -k \cdot S^a dt$ with a=5. This function is called the percolation. Integrating this function by the time step results in:

$$Perc_t = S_t^+ \cdot \left(1 - \left(1 + \left(\frac{S_t^+}{K}\right)^4\right)^{-\frac{1}{4}}\right)$$

where K is related to the time step by the relation:

$$K = (k\,(a-1)\,\Delta t)^{\frac{1}{1-a}}$$

In the end the level of the production reservoir equals:

$$S_t = S_t^+ - Perc_t$$

The percolated water joins the direct runoff and undergoes an adjustment which gives the multiplicative effective precipitation entering the routing function:

$$PRT_t = X_2 \cdot \left(P_t^{(n)} - P_t^{(s)} + Perc_t\right)$$

where the coefficient of volume adjustment X_2 is a free model parameter.

A 2.2 ROUTING FUNCTION

The routing function delays the release of the effective precipitation PRT_t to the next time step. This function connects a linear (hydrograph unit) and a non-linear routing via a reservoir.

UNIT HYDROGRAPH

The effective precipitation, PRT_t, is the input for a symmetrical unit hydrograph whose ordinates γ_i are calculated by γ_i = UHC(i) - UHC (i - 1). Where UHC is the function of the cumulative unit hydrograph defined by:

$$\begin{cases} UHC(i)=0 & \text{if } i \leq 0 \\ UHC(i)=\dfrac{i^{\alpha}}{i^{\alpha}+(X_3-i)^{\alpha}} & \text{if } 0 < i < X_3 \\ UHC(i)=1 & \text{if } i \geq X_3 \end{cases}$$

The time base of the unit hydrograph X_3 is a free model parameter and exponent α is a fixed parameter. The output of the unit hydrograph is therefore written as:

$$quh_t = \sum_{i=1}^{[X_3]} \gamma_i \cdot PRT_{t-i+1}$$

The volume of water in the routing reservoir equals then the output of the unit hydrograph plus the volume of water which is already in the reservoir:

$$R_t^+ = R_{t-1} + quh_t$$

OUTPUT OF THE ROUTING RESERVOIR

The output of the routing reservoir follows the subsequent drain law:

$$\widehat{Q} = -dR = k' \cdot R^{\beta}\, dt$$

The reservoir is as in Tangara (2005) quadratic (β=2) and the drain law giving the model outflow is therefore simplified to:

$$\widehat{Q}_t = f_{X_1,\beta=2}\left(R_t^+\right) = \frac{R_t^{+2}}{R_t^+ + X_1}$$

Where X_1 is a free model parameter.

The reservoir level of production becomes:

$$R_t = R_t^+ - \widehat{Q}_t$$

UPDATE OF THE ROUTING RESERVOIR

Tangara (2005) proposes a combination of two updates processes for the GRP model. The first one uses the last observed discharges to adjust the level of the routing reservoir, while the second uses the last observed error to adjust the forecasted discharge at the end of the model. Only the first method is integrated in the GRPE hydrological ensemble forecasting model.

The update process calculates the new level of the routing reservoir using the last observed discharge Q_t. For a quadratic reservoir this results in:

$$R_{t|t}^+ = f_{X_1,\beta=2}^{-1}\left(Q_t\right) = \frac{\sqrt{Q_t^2 + 4X_1 \cdot Q_t} + Q_t}{2}$$

The updated routing reservoir level after the model outflow is then given by:

$$R_{t|t} = f_{X_1,\beta=2}^{-1}\left(Q_t\right) = \frac{\sqrt{Q_t^2 + 4X_1 \cdot Q_t} - Q_t}{2}$$

A 2.3 MODEL PARAMETERS

The GRP model consists of three free parameters (X_1, X_2 and X_3) which have to be calibrated for every catchment. Furthermore, a number of fixed parameter is defined for the forecasting model running at daily time steps. These parameters are given in Table 16:

Table 16. Fixed parameters of the GRP model used in this study.

Fixed parameter	**Symbol**	**Value**
Production reservoir capacity	A	350 mm
Percolation function coefficient	B	2.25
Unit hydrograph exponent	α	2.5
Outflow routing reservoir exponent	β	2.0

A. 3 CATCHMENT SPECIFIC ADJUSTMENT FACTORS

	Catchment	Surface [km^2]	# of exc yellow	# of exc orange	# of exc Qix2	Adj x yellow	Adj x orange	Adj x Qix2
1	A1050310	233	0	0	13	-	-	0.75
2	A1310310	1040	0	0	10	-	-	0.80
3	A1522020	68	0	0	4	-	-	0.95
4	A2052020	447	0	0	2	-	-	0.90
5	A2352020	260	0	0	6	-	-	0.75
6	A2732010	229	0	0	9	-	-	0.80
7	A3301010	622	0	0	13	-	-	0.80
8	A3422010	185	0	0	7	-	-	0.75
9	A3472010	688	0	0	14	-	-	0.85
10	A4250640	1220	6	1	7	0.80	-	0.80
11	A5110610	1990	2	0	1	0.80	-	-
12	A5500610	3080	7	0	9	0.90	-	0.90
13	A5730610	3350	9	1	11	0.85	-	0.90
14	A6051020	374	10	1	8	0.85	-	0.75
15	A6341010	1110	9	2	1	0.90	0.95	-
16	A6731220	493	9	0	8	0.80	-	0.75
17	A6761010	2280	13	4	13	0.80	0.75	0.80
18	A6941020	2960	10	0	12	0.85	-	0.85
19	A7821010	925	5	0	6	0.95	-	0.95
20	A7881010	1280	4	4	10	0.85	0.95	0.95
21	A9301010	1720	0	0	1	-	-	-
22	A9752010	499	28	2	8	0.95	0.95	0.80
23	A9832010	187	26	0	5	0.85	-	0.90
24	D0137010	175	0	0	5	-	-	0.80
25	D0137020	275	0	0	16	-	-	0.80
26	D0206010	115	0	0	10	-	-	0.75
27	E4035710	392	0	0	32	-	-	0.85
28	H0400010	2340	9	0	13	0.95	-	1.00
29	H1201010	1280	0	0	1	-	-	-
30	H2332020	643	0	0	6	-	-	0.90
31	H2452020	1350	0	0	5	-	-	0.95
32	H2462020	1970	0	2	9	-	0.80	0.90
33	H2482010	2990	0	2	7	-	0.90	0.90
34	H5011020	360	8	3	5	0.75	0.75	0.75
35	H5102030	477	5	0	11	1.00	-	0.95
36	H5122340	672	7	0	5	1.00	-	0.90
37	H5172010	2100	3	0	6	1.00	-	0.95
38	H7061010	1170	1	0	1	-	-	-
39	H7162010	1630	0	0	47	-	-	0.90
40	H7201010	3280	36	0	19	0.95	-	0.95
41	H7401010	4290	19	8	9	1.00	0.95	0.95
42	H7742010	747	13	0	25	0.90	-	0.90
43	H7742020	1200	13	4	40	1.00	1.00	0.95
44	K0253020	375	1	0	2	-	-	0.75
45	K0403010	139	7	0	6	0.75	-	0.75
46	K2070810	324	12	2	12	0.80	1.00	0.75
47	K2523010	310	0	0	1	-	-	-

	Catchment	Surface [km^2]	# of exc yellow	# of exc orange	# of exc Qix2	Adj x yellow	Adj x orange	Adj x Qix2
48	K2593010	984	1	0	1	-	-	-
49	K2851910	494	7	1	1	0.85	-	-
50	K2871910	800	6	1	4	0.85	-	0.80
51	K2981910	1520	38	2	7	0.85	0.95	0.80
52	K3222010	353	43	1	6	0.80	-	0.75
53	K3273010	472	24	0	9	0.85	-	0.75
54	M0361510	827	0	0	10	-	-	0.80
55	M3600910	3910	0	0	11	-	-	0.90
56	M3630910	4160	0	0	13	-	-	0.80
57	M3851810	1310	0	0	14	-	-	0.95
58	P3274010	274	0	0	11	-	-	0.75
59	P6081510	432	0	0	12	-	-	0.75
60	P7041510	2120	0	0	9	-	-	0.75
61	P7261510	3750	0	0	14	-	-	0.90
62	U1014020	168	0	0	1	-	-	-
63	U1044010	866	25	0	1	0.85	-	-
64	U1084010	2040	19	0	14	0.95	-	0.95
65	U1120010	8900	18	1	5	1.00	-	1.00
66	U1324010	655	9	0	9	0.95	-	0.95
67	U2345020	31	22	1	5	0.75	-	0.75
68	U2345030	141	27	0	10	0.80	-	0.75
69	U2354010	1120	49	4	12	0.80	0.90	0.85
70	U2402010	3420	40	0	12	0.85	-	0.95
71	U2512010	4400	41	1	10	0.85	-	0.90
72	Y0284060	729	0	0	11	-	-	0.80
73	Y0624020	216	0	0	14	-	-	0.75
74	Y0655010	305	0	0	10	-	-	0.75
75	Y2214010	228	3	0	11	1.00	-	0.75

A. 4 RELIABILITY DIAGRAMS

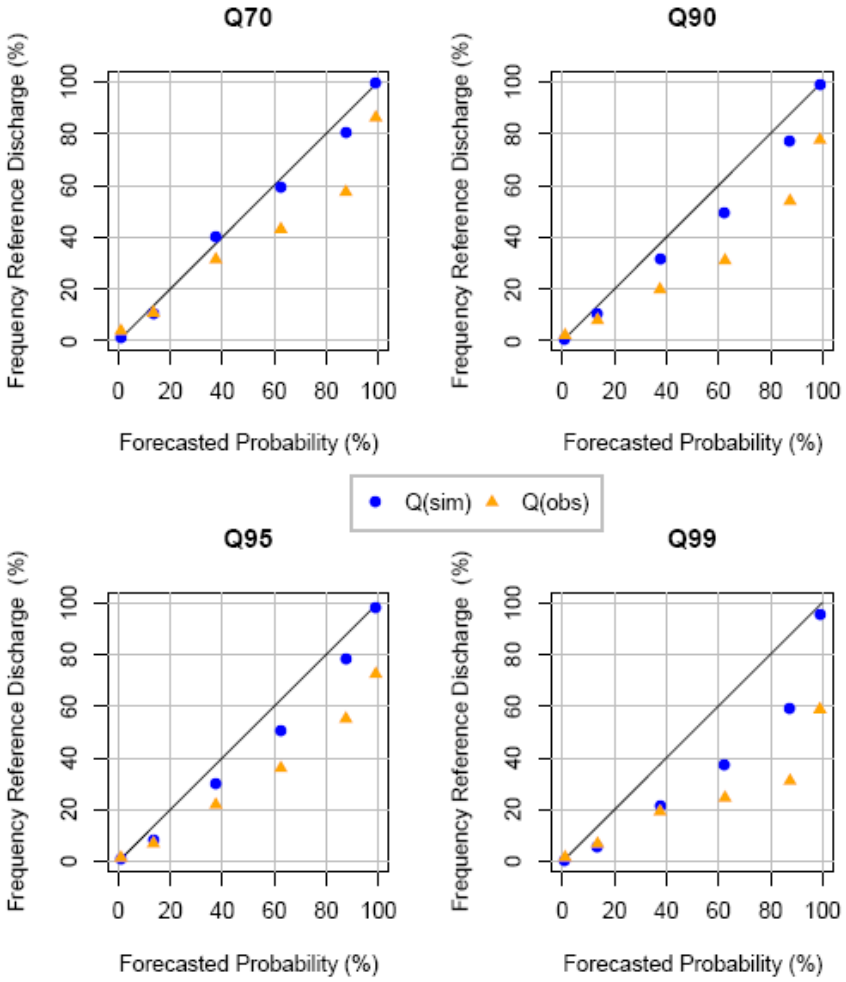
A 4.1 DATASET B1: 208 CATCHMENTS

The figure shows the reliability diagrams for the 208 catchments included in dataset B1. The diagrams give a quick overview of the quality of the hydrological ensemble forecast for the exceedance of four selected streamflow thresholds. The blue dots give the frequency of forecasted thresholds exceedances when the proxy-observed discharge (simulated discharges with observed meteorological data) is also exceeding the streamflow threshold for the six defined probability categories (x-axis: 1%, 13.5%, 37.5%, 62.5%, 86.5% and 99%). The orange triangles give the frequency of forecasted thresholds exceedances when the actual observed discharge is exceeding the streamflow threshold.

Example:

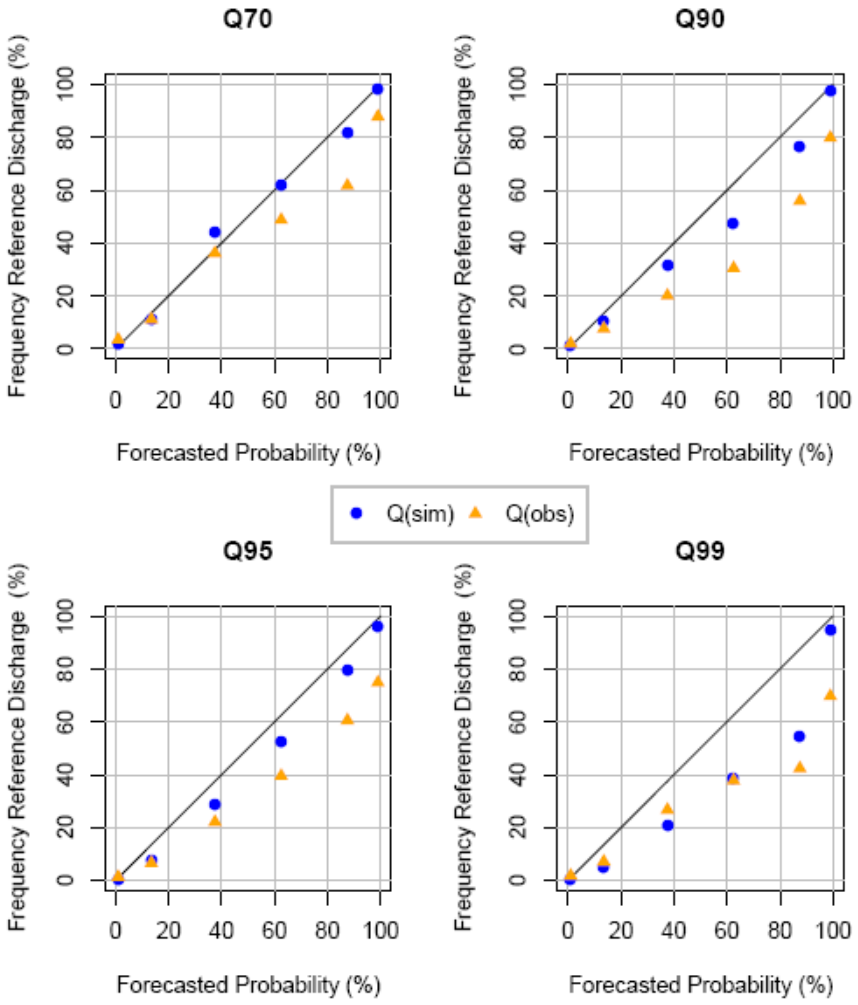
- 10 exceedances are forecasted with a probability between 0.75-0.98 (forecasted probability = mean value = 0.865 or 86.5%)
- 6 proxy-observed discharges are exceeding the threshold (frequency reference discharge = 6/10=0.60 or 60%)
- 5 observed discharges are exceeding the threshold (frequency reference discharge =5/10=0.5 or 50%)

Then the blue dot for this probability category is located at x=0.865, y=0.60, and the orange triangle will be located at x=0.865, y=0.50.



A 4.2 DATASET B2: 29 CATCHMENTS

The figure shows the reliability diagrams for the 29 catchments included in dataset B2. The diagrams give a quick overview of the quality of the hydrological ensemble forecast for the exceedance of four selected streamflow thresholds. The blue dots give the frequency of forecasted thresholds exceedances when the proxy-observed discharge (simulated discharge with observed meteorological data) is also exceeding the streamflow threshold for the six defined probability categories (x-axis: 1%, 13.5%, 37.5%, 62.5%, 86.5% and 99%). The orange triangles give the frequency of forecasted thresholds exceedances when the actual observed discharge is exceeding the streamflow threshold.



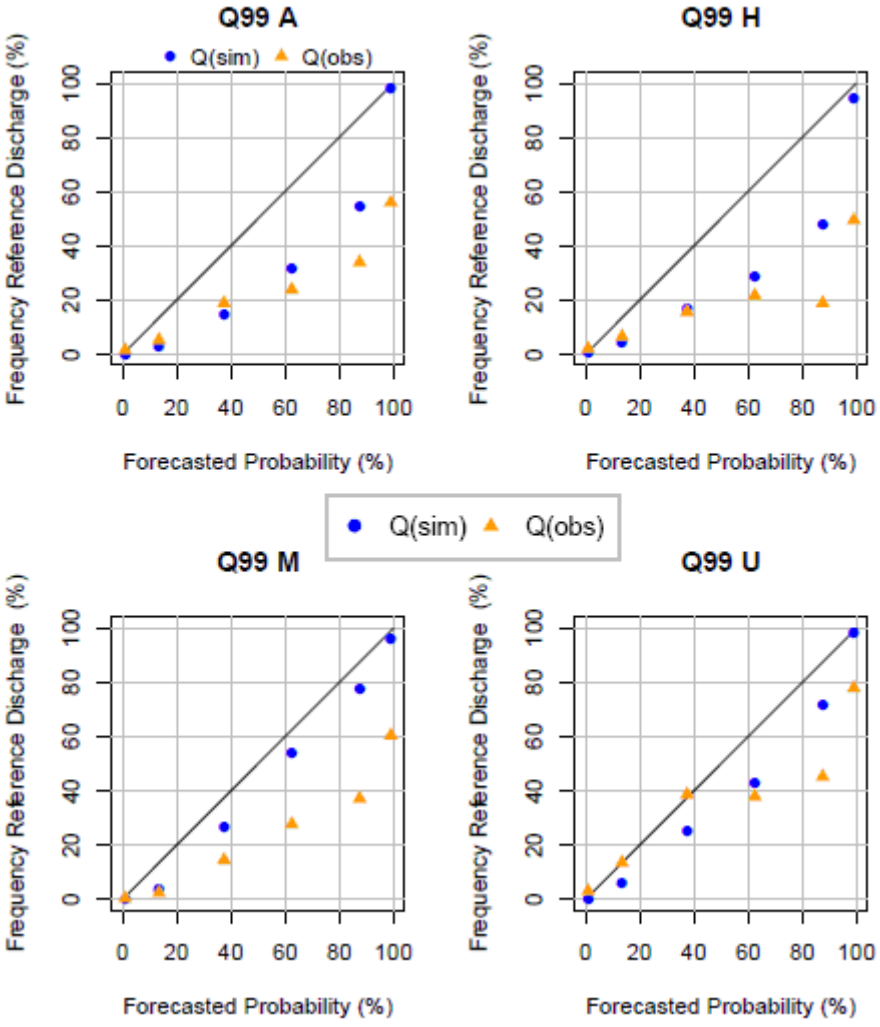
A 4.3 RELIABILITY DIAGRAMS PER CATCHMENT REGION

The figure shows the reliability diagrams for four main river basins in France, included in dataset B1, for the Q99 threshold.

The selection consists of:

- 23 A catchments (contributories of the River Rhine)
- 51 H catchments (contributories of the River Seine)
- 19 M catchments (contributories of the River Loire)
- 41 U catchments (contributories of the River Rhone)

The diagrams give a quick overview of the quality of the hydrological ensemble forecast for the exceedance of four selected streamflow thresholds. The blue dots give the frequency of forecasted thresholds exceedances when the proxy-observed discharge (simulated discharge with observed meteorological data) is also exceeding the streamflow threshold for the six defined probability categories (x-axis: 1%, 13.5%, 37.5%, 62.5%, 86.5% and 99%). The orange triangles give the frequency of forecasted thresholds exceedances when the actual observed discharge is exceeding the streamflow threshold.



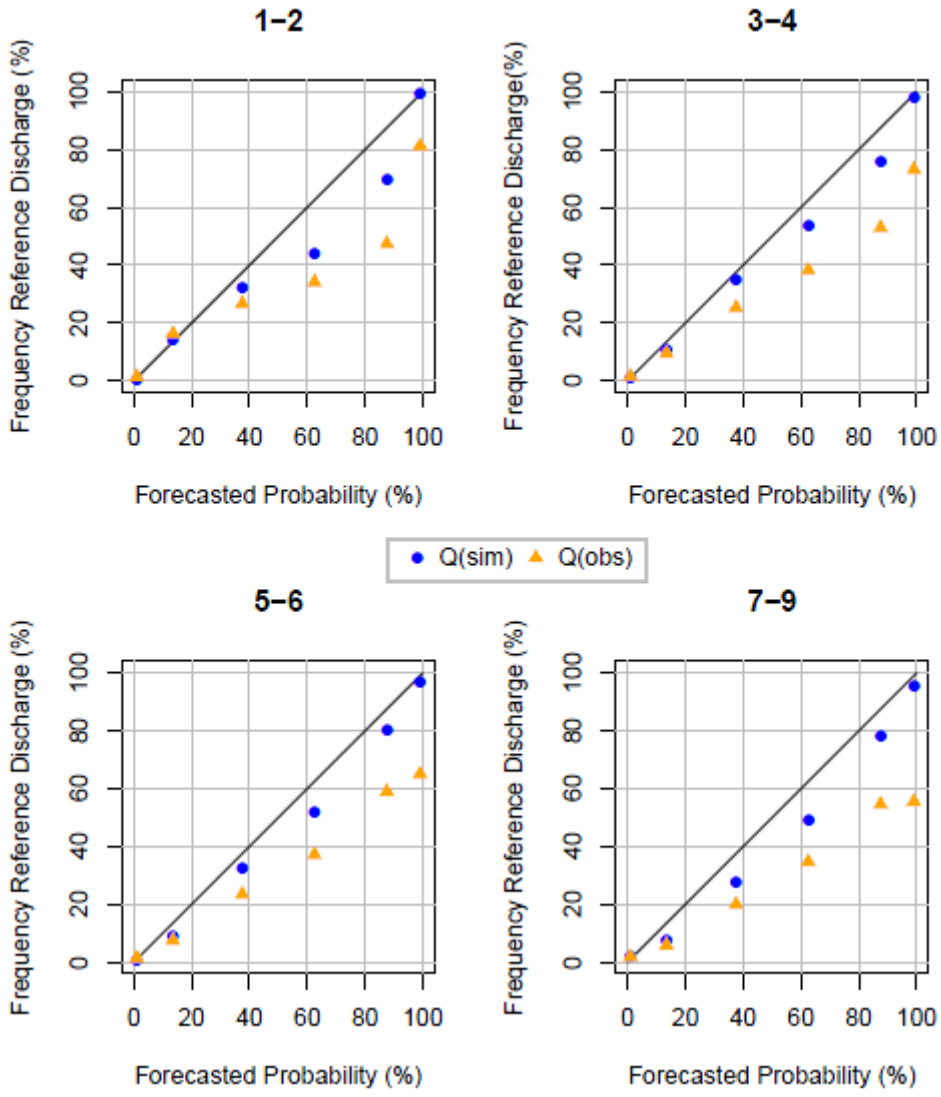
A 4.4 RELIABILITY DIAGRAMS AND LEAD-TIME (DATASET B1)

The figure shows the reliability diagrams split up per lead time class for the catchments included in dataset B1 and exceedances of the Q95 threshold..

The diagrams show the 4 lead time classes:

- Class 1: forecasts with a lead-time of 1 or 2 days
- Class 2: forecasts with a lead-time of 3 or 4 days
- Class 3: forecasts with a lead-time of 5 or 6 days
- Class 4: forecasts with a lead-time of 7, 8 or 9 days

The diagrams give a quick overview of the quality of the hydrological ensemble forecast for the exceedance of four selected streamflow thresholds. The blue dots give the frequency of forecasted thresholds exceedances when the proxy-observed discharge (simulated discharge with observed meteorological data) is also exceeding the streamflow threshold for the six defined probability categories (x-axis: 1%, 13.5%, 37.5%, 62.5%, 86.5% and 99%). The orange triangles give the frequency of forecasted thresholds exceedances when the actual observed discharge is exceeding the streamflow threshold.



A 4.5 RELIABILITY DIAGRAMS AND LEAD-TIME (DATASET B2)

The figure shows the reliability diagrams split up per lead time class for the catchments included in dataset B2 and exceedances of the Q95 threshold..

The diagrams show the 4 lead time classes:

- Class 1: forecasts with a lead-time of 1 or 2 days
- Class 2: forecasts with a lead-time of 3 or 4 days
- Class 3: forecasts with a lead-time of 5 or 6 days
- Class 4: forecasts with a lead-time of 7, 8 or 9 days

The diagrams give a quick overview of the quality of the hydrological ensemble forecast for the exceedance of four selected streamflow thresholds. The blue dots give the frequency of forecasted thresholds exceedances when the proxy-observed discharge (simulated discharge with observed meteorological data) is also exceeding the streamflow threshold for the six defined probability categories (x-axis: 1%, 13.5%, 37.5%, 62.5%, 86.5% and 99%). The orange triangles give the frequency of forecasted thresholds exceedances when the actual observed discharge is exceeding the streamflow threshold.