Using damage reports to assess different versions of a hydrological early warning system

Dimitri Defrance, P Javelle, D Organde, S Ecrepont, Vazken Andréassian, Patrick Arnaud

Hydrol. Earth Syst. Sci. Discuss., 11, 4365–4401, 2014
www.hydrol-earth-syst-sci-discuss.net/11/4365/2014/
doi:10.5194/hessd-11-4365-2014

Hydrology and Earth System Sciences
Open Access
Discussions

Using damage reports to assess different versions of a hydrological early warning system

D. Defrance[1], P. Javelle[1], D. Organde[2], S. Ecrepont[1], V. Andréassian[3], and P. Arnaud[1]

[1]IRSTEA, UR OHAX, Aix-en-Provence, France
[2]HYDRIS Hydrologie, Saint Mathieu de Tréviers, France
[3]IRSTEA, UR HBAN, Antony, France

Received: 25 March 2014 – Accepted: 3 April 2014 – Published: 22 April 2014

Correspondence to: P. Javelle (pierre.javelle@irstea.fr)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

In Europe, flash floods affect mainly the Mediterranean and mountainous regions, even if other regions also occasionally suffer from them. The catchments involved are usually small and ungauged, with short time of concentration. Forecasting this type of event remains difficult using hydrological models, and assessing the models is even more problematic. Typically, assessment is limited to gauged catchments that have relatively different geomorphological characteristics. The aim of this article is to present a method for assessing the models on real ungauged catchments through the use of damage reports and a multi-threshold approach, with assessment criteria that are based on a contingency table of the Critical Success Index type. The main conclusion, as demonstrated by Irstea's "Adaptation d'Information géographique pour l'Alerte en crue" for "Geographic information adaptation for flood warning" (AIGA) flood forecasting system and by the new version of AIGA for high-altitude catchments, is that while assessing hydrological models on gauged catchments is necessary, it is never sufficient and must be supplemented by assessments on ungauged catchments. This underlines the utility of building flood damage databases that are as exhaustive as possible. Such databases can be a valuable addition to more standard, often limited sources of data, especially for mountainous regions.

1 Introduction

1.1 Assessing the efficiency of early warning systems on ungauged catchments

The Aude River in Southern France in 1999 (Gaume et al., 2004), the Tagliamento River in the Italian Alps in 2003 (Borga et al., 2007), and the Nartuby River in Southern France in 2010 (Javelle et al., 2014) are three examples where flash floods caused major damage, although they occurred on small catchments located mostly outside

HESSD
11, 4365–4401, 2014

of the national monitoring networks. These floods were characterised by high specific discharge (Borga et al., 2011) and short response time. Unfortunately, it is impossible to monitor all of these streams, due to the high cost of building gauging stations and ensuring their upkeep (estimated at USD 10 000 year^{-1} (Gourley et al., 2013). And as Silberstein (2006) ironically pointed out, it is equally inconceivable to trust blindly in simulation models.

However, assessing an alert issued for an ungauged catchment poses specific problems, as you have to make up for the lack of continuous discharge data by relying on proxy data, which is obtained from field campaigns, newspaper clippings, etc. The information provided can be in the form of estimates from hydrological flow calculations or just damage data. Below, three examples of proxy data from damage reports adapted for ungauged catchments are presented: post-event reports, historical data, and damage data with quasi-real-time monitoring.

1.1.1 Post-event reports

Post-event reports consist of very comprehensive damage inventories that are made for major events, based on field surveys and hydrological calculations on small catchments that are not monitored by the national networks. These inventories serve for estimations of peak flows for comparison with the model's simulated flows. Borga et al. (2011) give many recommendations on these reports. The main drawback of this assessment technique is that it tests models only on major floods, with no possibility of assessing the model for smaller floods or of estimating false alerts emanating from the model, since only exceptional flood data are available.

1.1.2 Historical data

This technique consists of comparing simulated results from the hydrologic models with the damage identified in databases like the US National Weather Service (USNWS) Storm Events Database or the RTM Service (for Restauration de Terrains de Montagne,

HESSD
11, 4365–4401, 2014

or rehabilitation of land in mountain environments) in France. These two databases are fed from recent reports and newspaper clippings to provide information on past and recent floods (location, date, etc.) and on the damage observed. While there is an advantage to working with long time series, the information is not exhaustive enough to estimate false alerts from the hydrological models with any accuracy. No information is provided for cases where no flood occurred, and the reports usually focus on urban areas, where the risks are more concentrated (Gourley et al., 2010). But compared to the preceding technique, this one allows assessing the models based on a larger number and wider range of floods. Taking false alerts into account is still problematic, however, due to the lack of comprehensive data.

1.1.3 Damage data with quasi-real-time monitoring

This technique collects information on natural disaster damage from the historical database and adds information where reports are lacking. For example, a severe storm without damage will be added to the database. In the US, the Severe Hazards Analysis and Verification Experiment (SHAVE) project works along those lines, contacting businesses and individuals after each event to find out its geolocation and evolution over time, to try to rate the hazard as accurately as possible on a scale of ten categories from "No impact" to "Rescue/Fatality/Injury" (Gourley et al., 2010; Ortega et al., 2009).

The work of Calianno et al. (2013) highlights the advantage of gathering information on non-flood events, using a comparison of the NWS and SHAVE databases. The absence of a "no flood" category (in NWS) affects model assessment negatively by keeping false alerts from being taken into account. It is preferable, therefore, to add no-flood reports to the historical data to improve the assessment of hydrological models, for both missed and false alerts.

HESSD
11, 4365–4401, 2014

1.1.4 Summary

Each type of database has positive and negative aspects, then. Despite the advantages of quasi-real-time monitoring, however, this type of database is rarely used. The historical database from the RTM services in France is used for the present article. It was crucial to be able to work over a ten-year period to ensure there would be enough damage reports to assess the flash flood warning system presented in this article.

Damage data is important for assessing the hydrological models, but cannot be used directly. For the models to issue an alert by simulating flows on catchments, the flows must be quantified and ranked according to a catchment-specific alert threshold. This threshold must be in keeping with the observed damage. This is not easy to achieve and requires the use of the various techniques presented below.

1.2 Determining an alert threshold based on damage reports

When there is no available a priori information about floods, a theoretical alert threshold must be set. One recommendation is to set the threshold at a two-year return period that corresponds to the beginning of stream overflow (Carpenter et al., 1999), although this threshold does not appear to be well-adapted to small catchments (Reed et al., 2002). In reality, each site is flooded more or less often according to its degree of exposure. Ideally, the thresholds should be determined on a large scale, with integration of the exposure risk for human infrastructures (Naulin et al., 2013). Including damage reports can contribute significantly to the determination of flood alert thresholds.

Damage databases are a source of information about the flood risk exposure at the local level and at a large number of sites. However, the link between damage and flood is not clear-cut, for two reasons: (1) the damage is only reported when there are human stakes involved, and (2) floods of the same amplitude can cause different types of damage. (For example, streambank collapse can be caused by cumulative damage from a series of floods, but even if the last flood is the one which causes the

HESSD
11, 4365–4401, 2014

collapse, that flood is not necessarily the most severe one.) In spite of these limits, it is actually preferable to link damage reports with the model's alert threshold as much as possible, rather than to use the theoretical two-year threshold recommended by Carpenter et al. (1999).

In the literature, the research results from Versini et al. (2010) and Naulin et al. (2013) on road outages in France's Gard region (ANR Prediflood project) are interesting, in that ground surveys are used to calibrate thresholds for susceptibility, corresponding to flood probability for several return periods. These two studies show that taking the vulnerability of the location into consideration is crucial in defining accurate alert thresholds.

1.3 Scope of the paper

The scope of the paper is to develop a technique to assess hydrological models for actual ungauged catchments, using a damage database. This type of technique is rarely seen in the literature, where assessment is usually limited to flow data from gauged catchments that are much larger than those subject to flash floods.

After defining the assessment method, the article compares the AIGA method ("Adaptation d'Information géographique pour l'Alerte en crue" for "Geographic information adaptation for flood warning") (Javelle et al., 2010) developed for flash floods around the Mediterranean Basin with a new version developed specially for flash floods in mountainous regions.

Section 2 presents the catchment sample groups, the first consisting of gauged catchments and the second, of catchments where only damage reports are available, and then describes both the AIGA method and its new version. Section 3 describe the methodology developed to assess and compare the models on pseudo-ungauged catchments (flows) and real ungauged catchments (damage). The results obtained with both datasets are shown in Sect. 4 and discussed in Sect. 5. The conclusion outlines some avenues for improvement.

HESSD
11, 4365–4401, 2014

2 Material and methods

2.1 Reference dataset on gauged catchments: HYDRO

To develop the hydrological models, 118 gauged catchments (referred to as HYDRO in what follows and indicated in black in Fig. 1) were chosen from the HYDRO hydrometric database (http://www.hydro.eaufrance.fr/) in the Mediterranean, Northern Alps and Southern Pre-Alps regions, which are all subject to flash floods. To regionalise the flood warning system effectively, to work on a range of dissimilar catchments is necessary (Table 1): surface ranges between 8 and 897 km^2 (median: 110 km^2), with snow accounting for 0–58 % (median: 8 %) of the annual rainfall budget. The slopes, which are calculated by dividing the difference between the catchment's highest and lowest point by the stream length (Marchi et al., 2010), range between 0.003 and 0.17 (median: 0.04).

For each catchment, time series are available for total precipitation (P) (rainfall + snow), the snow fraction (%S) and temperature (T) at daily and hourly time steps, and also for daily evapotranspiration (E_0). P is the result of a ten-year radar reflectivity re-analysis (1997–2006), transformed into depth of rainfall with one-square-kilometre resolution using the current algorithms used by the Météo-France Weather Service and adding pluviograph measurements for the zones that are poorly covered (Tabary et al., 2012). The rainfall/snow ratio and the temperature are provided on 64 km^2 grid cells using Météo-France's Safran method, which combines soil measurements and atmospheric modelling (Vidal et al., 2010). E_0 is calculated on a formula proposed by Oudin et al. (2005), based on temperature and extraterrestrial radiation, which depends on latitude and the Julian day number.

The model used for the flash flood warning system is event-based, so it was necessary to define events for each catchment for performances assessment. Using the daily rainfall time series and a minimum rainfall threshold for each catchment, a total of nearly 26 000 events was obtained or an average 220 events per catchment. To avoid any ambiguity, each event corresponds to a specific start date, finish date and

HESSD
11, 4365–4401, 2014

catchment. Therefore, there can be more than one event on the same date, but the model manages each separately. The same process for the second catchment sample group (presented in the next section) was used, obtaining 26 000 other events.

2.2 The RTM damage reports dataset: a unique opportunity to explore real ungauged catchments

A second catchment sample group was assembled from catchments for which only damage reports are available. The Restauration des Terrains en Montagne (RTM) Service collects the reports in order to inventory past and current natural disasters. This database can be found at the website http://rtm-onf.ifn.fr. For clarity, these catchments will hereinafter be referred to as "RTM catchments".

Each event was recorded, located (a site can be a streambank, ravine, section of a river, etc.), identified with an individual number and assigned a specific natural risk. The information about the event was collected from various complementary sources: the export report from an RTM agent, press clippings and personal accounts.

Because the sites recorded in the database were not always comparable to catchments, a catchment should be associated to each site before being able to use the data for hydrological purposes. In the end, only 123 catchments were chosen (Ecrepont, 2012). Over a period of ten years, 179 damage reports were reported on all of these catchments, or 1.45 per catchment. These RTM catchments have different characteristics than the ones in the HYDRO sample group (Table 1). Their catchment area is smaller, with a comparable interval [5887 km^2] but a median of 22 km^2 (compared to 110 km^2), and a 27 % median for the snow fraction, more than three times higher than the one obtained on HYDRO catchments. Slopes on RTM catchments are also steeper: 0.02 minimum, 0.26 maximum, and a 0.09 median, or twice as high as for HYDRO catchments (0.17 maximum and 0.04 median). So RTM catchment characteristics are similar to those of the catchments subject to flash floods, and described in the Hydrate project (Marchi et al., 2010; Borga et al., 2011). The

HESSD
11, 4365–4401, 2014

precipitation (P) (rainfall + snow), the snow fraction (%S), the temperature (T) and also the daily evapotranspiration (E_0) time series were also available for all 123 catchments.

2.3 The AIGA early warning system and its new version

2.3.1 Overall operation

Operation for both the AIGA system and its new version is based on two hydrological models: the GR4J, a continuous daily soil moisture accounting model, and the GRD, an hourly event-based model used to simulate floods. Before presenting the coupling for the original AIGA method and for the new version, we will show how both models work.

Continuous GR4J model

The model GR4J is shown to the left in Fig. 2. It is a conceptual model that operates continuously on a daily time step and consists of a production function followed by a flood routing function. The production function consists of a nonlinear reservoir that is fed by precipitation (P) and emptied by evapotranspiration (E_0) and a drainage process. The runoff coefficient is obtained on a daily basis. To simulate flows, production is broken down into a rapid component (10 %) and a slow component (90 %). Both are preceded by a unit hydrograph. For the slow component, water enters the nonlinear routing reservoir. The flow is obtained by adding the two flow components together. For all the equations, refer to Perrin et al. (2003).

Hourly GRD model

As shown to the right in Fig. 2, the structure of the GRD hourly event-based model is similar to the GR4J's. The GRD serves to simulate flows during flood events. Hourly precipitation first passes through a unit hydrograph in two one-hour time steps, with 70 % of rainfall in the first step and 30 % in the second, before entering a production

HESSD
11, 4365–4401, 2014

reservoir that functions similarly to the one in GR4J. However, unlike with a daily time step, this reservoir does not empty out during an event: evapotranspiration and drainage are equal to 0 mm h^{-1}. The second reservoir (routing) transforms the portion of non-infiltrated precipitation into flow just as for the flow portion obtained by the GR4J routing reservoir. For more information, refer to the full description in Javelle et al. (2010).

2.3.2 The original AIGA method

The original AIGA method is based on the two conceptual hydrological models. Coupling the two models is simple, with only the hourly model's production reservoir being initialised by the fill rate for the GR4J's production reservoir. This fill rate, which takes daily precipitation (P) and evapotranspiration (E_0) into account, is corrected before being inserted into the GRD model, using a regional formula to remove bias for all HYDRO and RTM catchments (Eq. 1, Javelle et al., 2010). The fill value for the GRD routing reservoir is identical for all events and all catchments: 30 % of initial fill, leading to zero flow. Neither model takes snow into account.

$$(S/A)_{\mathrm{GRD(init)}} = a(S/A)_{\mathrm{GR4J}(j-1)} + b(S/A)_{\mathrm{stat}} + c \qquad \text{(Javelle et al., 2010)} \qquad (1)$$

with $a = 0.52$, $b = 0.38$, and $c = -0.17$: statistical parameter with a pixel value of 0–1

The parameters for both models are identical for all catchments in both sample groups. The values for production reservoir size in both the GRD and the GR4J are about 200 mm, and the routing reservoir size value in the GRD is 50 mm for the entire study area.

2.3.3 The new version of AIGA

Like the original method, the new version of AIGA relies on two global conceptual models: the continuous daily GR4J model, and the hourly event-based GRD model for flood simulation which operates exactly like the original method. Figure 2 shows the

HESSD
11, 4365–4401, 2014

coupling of the two models, with initialisation of the GRD production reservoir (next paragraph) and initialisation of the routing reservoir fill process as a function of the daily flow on the day before the event simulated by GR4J.

For the initialisation of the GRD production reservoir, a bias removal rule with a parameter "a" (Eq. 2) at the catchment scale (global model) is used to link the GR4J fill rate to the initial GRD rate:

$$(S/A)_{\mathrm{GRD(init)}} = a(S/A)_{\mathrm{GR4J}(j-1)} \quad (2)$$

with $(S/A)_{\mathrm{GRD(init)}}$: for the initial filling of the GRD hourly module's production reservoir, $(S/A)_{\mathrm{GR4J}(j-1)}$: for the filling of the GR4J daily module's production reservoir on the day before the event, and a: a parameter to be calibrated in order to link the fill rates for the GRD and GR4J modules.

So there are four parameters for GR4J ("A_{cont}", "B_{cont}", "C", "D") and two for the event-based GRD ("a" and "B_{hor}"), or a total of six parameters to be regionalised and which, unlike the original method, are not identical for all catchments. Based on the research of Wasson et al. (2001), the study area is broken down into four relatively homogeneous hydro-ecoregions (HER) to facilitate the regionalisation of both models. For the 1997–2006 period, the six parameters are calibrated on the 118 HYDRO catchments and then transferred to those HYDRO catchments that are considered to be ungauged, and to the 123 RTM catchments. The transfer method consists of taking the median values for the parameters of the three closest neighbour catchments in the same sub-region, obtained using the Euclidean distance between centroids. This technique makes it possible to obtain the best results for Irstea's GR models (Oudin et al., 2008).

In the new version of AIGA for studies on mountainous catchments, precipitation is fine-tuned to factor in snowfall. The snowfall portion (%S) of hourly and daily precipitation provides information about the quantity of snow to be subtracted from each rainfall depth estimated by the precipitation (P) radar, in order to keep only the liquid portion of total precipitation data (P). At a daily time step, snowmelt is added to this liquid portion (P) using a regionalised snowfall/snowmelt degree-day module

HESSD
11, 4365–4401, 2014

with a single parameter set for all catchments. The module estimates snowmelt based on the daily temperature provided by the Safran analysis system and on a threshold temperature of 0 °C (Folton and Arnaud, 2014).

3 Proposed assessment methodology to avoid censored-data issues

The damage-data-based assessment method started out with a similar approach to that of Naulin et al. (2013), linking damage report to a alert threshold. But because the database was not comprehensive enough and there were no "non-flood" reports, a multi-threshold approach was considered. This approach makes it easier to compare between hydrological models and to put the developments to best advantage, while reducing the impact of an absence of alerts in the database. A number of different alert thresholds are tested simultaneously for each catchment, and for each threshold, contingency criteria are worked out.

3.1 Contingency criteria

Before discussing the technique for determining the different thresholds, two contingency criteria are presented. They were obtained by comparing damage reports with the maximum flows that were modelled by the hourly hydrological model for each event. These criteria, which Schaefer (1990) described with precision, focus on three types of alert:

- correct alerts (CA): model exceeds threshold with damage observed,
- missed alerts (MA): damage observed without model exceeding threshold,
- false alerts (FA): model exceeds threshold with no damage observed.

Based on that, the correct alerts (Probability Of Detection) and false alerts (Success Rate) can be quantified, see Table 2. The POD informs on correct alerts (CA) compared

HESSD
11, 4365–4401, 2014

to the total number of alerts (CA + MA) and the SR, on the number of correct alerts (CA) out of the total number of modelled alerts (CA + FA).

3.2 A graph to compare models

The assessment criterium shares a common basis with the statistical concept of relative operating characteristic (ROC) curves introduced by Swets (1973). ROC analysis has been used in hydrology and meteorology since 1982, and was first applied to flash floods in 1992 (Krzysztofowicz, 1992; Manzato, 2005). Based on the same principle, the POD and SR that have been calculated for different alert thresholds are used to define a curve similar to a ROC curve, and the curve is plotted on a graph. Building the POD/SR curve is a two-step process:

1. Counting alerts (POD and SR) for different thresholds
   To test several alert thresholds simultaneously, the number of alerts simulated by the model must correspond to a certain number of damage reports. The detection threshold for each catchment is made to vary, which changes the number of alerts simulated, but maintains a set number of damage reports. The number of correct, missed and false alerts is then changed for each catchment. To define the different thresholds, the following relationship is used:

   $$\text{Nbr of modelled alerts} = N \cdot \text{nbr of reports} \quad (3)$$

   With N: an integral number

   To illustrate the threshold's impact on all three types of alerts (correct, missed, and false), Fig. 3 shows the change in the number of simulated alerts, with respect to RTM reports, on a virtual catchment comprising 11 events simulated by a model and two damage reports (green rectangle). The first threshold tested is determined with parameter N at a value of 1 (Fig. 3a). The number of simulated alerts and simulated flow exceeding the threshold corresponds to the number of damage reports. In this case, there are 2 reports (green rectangle) and 2

HESSD
11, 4365–4401, 2014

simulated alerts (one correct in green and one false in yellow). In Fig. 3b, the second threshold ($N = 2$) implies that the model has to simulate 4 alerts for the 2 reports. Consequently, the threshold is lowered to try to increase the number of correct alerts, but as a result, the number of false alerts also increases. So in this case, there is still one correct alert and one missed, and also 3 false alerts, or two more than with the first threshold. At the third threshold ($N = 3$ in Fig. 3c), no alerts are missed (2 correct alerts), but the number of false alerts rises to 4. At the $N = 4$ threshold (Fig. 3d), only the number of false alerts rises to 6.

To summarise, the threshold is gradually lowered, which increases the number of correct alerts and reduces the number of missed alerts; however, at the same time, the number of false alerts increases. So for each threshold tested (the N value), it is possible to associate a number of correct, missed and false alerts for all catchments as a whole, by adding up each category.

2. Plotting the POD/SR curve
For each threshold tested, the correct alerts for all catchments are added up to obtain a total number of correct alerts for the entire catchment sample group. The same procedure is performed for missed and false alerts, and the POD and SR contingency criteria are calculated for each threshold tested (N times). Using all the POD and SR values, the POD/SR graph (Fig. 4) is plotted and the model can be assessed for several thresholds at a time. Let's remember that to obtain an SR value equal to 1 (no false alerts), all reports would have had to correspond to the highest flows for all catchments. Inversely, to obtain a POD equal to 1, all alerts would have to be correct and the threshold would therefore be very low in some cases.

To compare two models, the POD/SR curves for each are plotted on the same graph to enable a comparative analysis of the results and to identify the better model, even when the performance is similar for both. The test is independent of the chosen alert threshold, because the POD and SR are simultaneously considered at a number N of

HESSD
11, 4365–4401, 2014

different thresholds. In the absolute, the best results would be represented by curves close to a value of (1.1), in the upper right corner. The ideal case would be a POD always equal to 1 for all thresholds, with an SR varying between 0 and 1. To distinguish between two close results, the area underneath the curve can be calculated, knowing that its maximum value would be close to 1.

4 Results

4.1 Results on pseudo-ungauged catchments

First, to demonstrate that adapting the new AIGA method to mountainous regions was necessary and beneficial, the performance of the hourly event-based models (for both the original and the new AIGA method) was analysed for all 118 HYDRO catchments using a Nash–Sutcliffe criterion limited to a range from −100 to 100 (Mathevet et al., 2006). Working with the flood events to obtain a criterion per model and per catchment, the events were pieced together to get a continuous time series (Nash pseudo-continuous, NPC). The ten years of data were used for validation, since all of the models were regionalised. The catchments were considered as ungauged, and the observed flows served to assess and compare the two GRD models (original and new version).

As shown in Fig. 5a, the overall results increased by an average 35 points in terms of Nash–Sutcliffe. The median for the results (thick black line) rose from 0 to nearly 30 points. Performance improved (Fig. 5b) on 80 % of the catchments (red dots). There was no change in performance between the two versions on only 10 % of the catchments (blue dots), and the remaining 10 % showed worse performance (green dots). The catchments in these last two cases were scattered all over the study area, which attests to the problems that can crop up with regionalisation.

On the relatively large catchments from the HYDRO database, the new version performs better, thanks to better initialisation (production and routing reservoirs from

HESSD
11, 4365–4401, 2014

the GRD model) and a regionalisation process that factors in the study area's regional differences by retrieving the parameters from neighbouring catchments within a homogeneous hydro-ecoregion (HER).

4.2 Results on real ungauged catchments

In the second step, the performance of the two AIGA versions was directly analysed on catchments in the RTM sample group. For these, assessment of the hourly models was based on the damage reports in the database, which were compared to the models' maximum flows using a multi-threshold approach.

For each one of the 123 RTM catchments, the choice was made to lower the threshold ten times using N values from 1 to 10. So for each one, there were 10 different thresholds and 10 values associated with correct, missed and false alerts. As the threshold drops, the number of correct and false alerts increases. As a result, for each threshold tested on each catchment, the number of correct, missed and false alerts varies. All of these are used to calculate the ten PODs and ten SRs required for the POD/SR curves.

As was suggested when the methodology was set up, priority should be given to comparative performance analysis via a common test, which has the advantage of penalising two competing systems equitably, as highlighted in the works of Andréassian et al. (2009). The results are presented by sub-region, corresponding to the hydro-ecoregions (HER). The RTM catchments are distributed over the Inner Alps, Mediterranean and Southern Pre-Alps HERs. On Fig. 6 the performances of both AIGA versions are variable as a function of region. In the Inner Alps (Fig. 6a), the initial flows and snowfall generate noticeable advantages, and the results of the event-based model in the new version are superior to those in the original version. And yet, it is in this sub-region that performance is the weakest.

On the other hand, for the other two HERs (Mediterranean and Southern Pre-Alps), the results are closer between the two hourly hydrological models, and superior to those for the Inner Alps. The improvements to the new version of AIGA have a very

HESSD
11, 4365–4401, 2014

limited effect, since the curves are similar for the Southern Pre-Alps. The new version's performance is worse in the Mediterranean HER, unlike the results for the HYDRO catchment study.

The improvements introduced by the new version are noticeable on the RTM catchments, but only on the high-altitude catchments (Inner Alps). However, performance for both versions (original and new) in this region is the weakest, highlighting various limits that are inherent to the assessment, to the input data, and also to the calibration and regionalisation method for the hydrological models.

5 Discussion

The above results showed improvements in the new version of the AIGA method compared to the original version. The performance remains fairly modest, however, with very low success rates (SR). In the following two paragraphs, the apparent reasons for these weak results are discussed:

- limits linked to the assessment method and to the damage data,
- limits linked to the models and to the rainfall data.

5.1 Limits linked to the use of the RTM database

The POD/SR assessment method is based on the damage reports in the RTM database. The method characterises the relative match between the alerts issued and the damage actually observed, and therefore relies on the following assumptions:

- the existence of a link between flow and the damage suffered: two similar flows cause the same damage, and so both are recorded in the database. This strong assumption implies that there are no effects of "memory" or accumulation and that all events that result in damage are independent. Also, the damage is really

HESSD
11, 4365–4401, 2014

caused by flood flows and not by debris flow, which causes greater damage by different processes;

- stationarity: over the ten-year period, we infer that there has been no change on the catchment, e.g. in terms of land use that would alter the type of damage caused during the floods. The flow that triggers the damage does not vary over time;

- the exhaustiveness of the database: over the ten-year period, we infer that the reports are obtained for all catchments and for all flood events in a like manner. For example, each field agent has the same understanding of what constitutes damage.

To verify these three assumptions, four catchments that were in both of the sample groups (HYDRO and RTM) were studied, giving access to both observed flow data and damage reports. These four catchments were the only ones common to both sample groups. Comparing observed flows and RTM reports enabled us to validate or invalidate the working assumptions. In the following examples, the fact of presenting the relative match between the damage and *the observed, unmodelled flows* enables us to temporarily exclude any model-related problems. If one of the assumptions is not validated on one of the four catchments, it could also be the case on the other RTM catchments. However, it would not be possible to determine so categorically.

In Fig. 7, the case of the Durance River (Fig. 7a) is ideal: the highest observed flow coincides with the only RTM report, so the highest threshold gives the best result and each time the threshold is lowered, performance declines. For the Mourachonne River (Fig. 7b), there are two reports, and one of the descriptions (green dot) does match the highest observed flow. However, two other observed flows of similar intensity (orange dots), gave rise to no reports, even though a lower observed flow coincides with the 22/11 report (red dot).

HESSD
11, 4365–4401, 2014

These two examples helped demonstrate that the assumptions on the assessment method were not always borne out. Nevertheless, it is not possible to determine which of the three assumptions is false.

Despite the problem raised, the method has its merits, which lie in its potential for comparing warning models even when flowmeter measurements are absent, which is often the case on catchments subject to flash floods. Performance should be assessed in a relative way from one model to another, and not in the absolute.

5.2 Limits linked to the models and to rainfall data

For limits that are linked to the models and rainfall data used, only the performance of the new AIGA version was analysed to characterise the shortcomings.

In Fig. 9, the results are represented as a function of (1) the two physical characteristics, surface and slope; and (2) the season during which the events occur. For the two physical descriptors, the sample group is divided in two, one comprising the 62 RTM catchments with the lowest values and the other comprising the other 61 catchments with the highest values. In this way, it is possible to determine on which type of catchment the model works the best. For surface (Fig. 9a), the largest catchments (more than 21.8 km^2) are shown in orange and the smaller ones in black. For slope (Fig. 9b), the catchments with gentler slopes (< 0.09) are shown in black, and those with the steepest slopes, in orange. For season (Fig. 9c), the results of the new AIGA version are shown in black for summer and in orange for autumn.

The results show that the smallest catchments register the lowest performance, with lower SR and POD values for all ten thresholds, with a roughly 0.1 point difference with respect to the large catchments. These small catchments are often characterised by steep slopes, because they are located mostly in the mountains (Inner Alps). In consequence, the very steep catchments are also the ones with the lowest SR values, at a maximum 0.2. The POD and SR values for catchments with gentler slopes are more acceptable, at a maximum 0.4 for each one of them. The performance is better

HESSD
11, 4365–4401, 2014

in autumn than in summer, when the SR values are particularly low (< 0.15). So the model is less adapted to small, steep catchments subject to summer flood events.

To explain these differences among catchments, three hypotheses can be proposed:

Hypothesis 1

The lower performances are observed on catchments for which there is the least rainfall data. In Fig. 9c, summer is when there appears to be a problem with accurate rainfall data. During this season, small catchments located (mostly) in the mountains of the Inner Alps are often hit with very local storms. In the database, these account for 57 RTM reports out of 179, or 32 %. Estimating the cumulated rainfall for these storms is very difficult. It is often underestimated, either because the storms occur in areas not covered by the raingauge network, which is sparser in mountainous areas (Gottardi, 2009), or because they are so intense that they cause radar signal black-out (par exemple: Berne and Krajewski, 2013; Diss et al., 2009). To illustrate this, the 27 July 2003 rainfall event is represented in Fig. 8.

The event is characterised by 50 mm of maximum cumulated rainfall over a two-day period. If the analysis is focused on the three catchments reported on by the RTM (red dots), the cumulated rainfall ranges between 15 and 25 mm, which is low for causing a flash flood. The RTM services did nevertheless find damage. By using a very low threshold, corresponding to 27 times the number of reports for each catchment (N = 27), i.e. lower than the thresholds used for the POD/SR curves (Eq. 3), no alert is issued in the location where damage is reported, constituting a missed alert (in red), and a false alert (modelled and not matching an RTM report) appears. This problem of modelled alerts is probably linked to an underestimate of cumulated rainfall.

The quality of rainfall depth data in mountainous regions, therefore, plays a crucial role in hydrological model performance, and can partially explain the poor results obtained on high-altitude catchments where little rainfall data is available (radar measurement problems or absence of raingauges to recalibrate radar-based rainfall).

HESSD
11, 4365–4401, 2014

Hypothesis 2

This hypothesis concerns certain catchment characteristics. In Fig. 9a and b, analysing performance as a function of slope reveals differences between large, flatter catchments and small, steeper ones. These results can be explained by other kinds of phenomena besides flash floods, because small mountain catchments are equally subject to floods and debris flows, where sediment can account for more than 50 % of the volume of the water runoff (Bertrand et al., 2013). In that case, it no longer constitutes a flash flood and the catchment processes are no longer the same. Because they are not taken into account in the models, such phenomena have a negative impact on the models, and to add a complementary module to factor in debris flows would be necessary, especially at an hourly time step.

Hypothesis 3

The size difference between the HYDRO catchments used to develop the new AIGA method and the RTM catchments, where the model was assessed, appears to influence the way the model works (Fig. 9a). The new AIGA method was regionalised based on gauged catchments. A "neighbourhood" technique was used within each HER region to transfer parameters from donor catchments to a target catchment. There was no transformation for catchment size, although the physical characteristics of the HYDRO catchments were averaged with respect to the small RTM catchments, which have more specific and more variable physical characteristics. Therefore, the hydrological processes are different (flood dynamics). In this situation, parameter transfer between catchments becomes problematic and the neighbourhood technique can have an adverse effect on the results.

HESSD
11, 4365–4401, 2014

6 Conclusion and future works

The method presented in this article enables the use of historical damage databases to compare hydrological models on ungauged catchments. The standard assessment method sticks to the use of catchments that are gauged but considered as ungauged, which poses a problem of scale. In some cases, the assessment concerns actual ungauged catchments, but it restricts itself to extreme events, in effect omitting an assessment of false alerts. To put damage reports to best use for comparing models, a POD/SR curve was developed to factor in the rate of correct and false alerts, and to eliminate the need to select an alert threshold.

As for the results, the improvements to the original AIGA method, e.g. taking snowfall into account, are visible on both gauged catchments and actual ungauged catchments (RTM). However there are still problems linked to insufficient knowledge on rainfall and to lowering the scale. For this reason, dual assessment of the hydrological models is important: while an assessment on gauged catchments is necessary, it is never sufficient for qualifying an early flash flood warning system on ungauged catchments.

To improve the hydrological models, studies are ongoing to try to calibrate the models on the ungauged catchments regionally or locally, using damage reports or limited streamflow data (Perrin et al., 2007; Rojas-Serna et al., 2006). Choosing to use the POD/SR curves directly as a criterion can also give better results. In this way, we can eliminate the need for calibration on large catchments and avoid problems related to scale lowering.

The lack of exhaustive databases remains a problem that can be solved only by improving field reports and setting strict standards for them. Also, to provide assessments that are more accurate than just a threshold that is exceeded, temporal data should be integrated into the databases to assess forecasting, as suggested by Calianno et al. (2013) and Ruin et al. (2008). For data gathering, Gourley et al. recommend more systematic reporting of both damage and the absence of damage (the SHAVE database) (2010) and assembling datasets as in

HESSD
11, 4365–4401, 2014

the FLASH project (http://www.nssl.noaa.gov/projects/flash/database.php) in the US, which groups the SHAVE database, NWS storm reports and USGS streamflow data (2013). Once this new generation of databases is created, it will be possible to use larger sample groups to improve early warning methods for small ungauged catchments.

Acknowledgements. The authors and their institute want to thank the European Union, the Provence–Alpes–Côte d'Azur Region and the French Ministry of Ecology for co-financing the Rhytmme (Risques HYdrométéorologiques en Territoires de Montagnes et Mediterranéens/Hydrometeorological Risks in the Mediterranean and Mountainous Regions) project. The authors also want to thank the many people involved in the project.

References

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.-H., and Valéry, A.: HESS Opinions "Crash tests for a standardized evaluation of hydrological models", Hydrol. Earth Syst. Sci., 13, 1757–1764, doi:10.5194/hess-13-1757-2009, 2009.

Berne, A. and Krajewski, W. F.: Radar for hydrology: unfulfilled promise or unrecognized potential?, Adv. Water Resour., 51, 357–366, 2013.

Bertrand, M., Liébault, F., and Piégay, H.: Debris-flow susceptibility of upland catchments, Nat. Hazards, 67, 1–15, doi:10.1007/s11069-013-0575-4, 2013.

Borga, M., Boscolo, P., Zanon, F., and Sangati, M.: Hydrometeorological analysis of the 29 August 2003 flash flood in the Eastern Italian Alps, J. Hydrometeorol., 8, 1049–1067, doi:10.1175/jhm593.1, 2007.

Borga, M., Anagnostou, E. N., Blöschl, G., and Creutin, J. D.: Flash flood forecasting, warning and risk management: the HYDRATE project, Environ. Sci. Policy, 14, 834–844, 2011.

Calianno, M., Ruin, I., and Gourley, J. J.: Supplementing flash flood reports with impact classifications, J. Hydrol., 477, 1–16, 2013.

Carpenter, T. M., Sperfslage, J. A., Georgakakos, K. P., Sweeney, T., and Fread, D. L.: National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems, J. Hydrol., 224, 21–44, 1999.

HESSD

11, 4365–4401, 2014

Diss, S., Testud, J., Lavabre, J., Ribstein, P., Moreau, E., and Parent du Chatelet, J.: Ability of a dual polarized X-band radar to estimate rainfall, Adv. Water Resour., 32, 975–985, 2009.
Ecrepont, S.: Aide au Développement d'un Système d'Alerte de Crues en Région Montagneuse, IRSTEA, Stage de fin d'étude, Université Jospeh Fourier (Grenoble 1), Aix-en-Provence, 60 pp., 2012.
Defrance, D.: Adaptation et évaluation d'un système d'anticipation de crues éclair sur des bassins de montagne non-jaugés. Valorisation de relevés de dégâts, Université Pierre et Marie Curie (Paris 6), Paris, 261 pp., 2014.
Gaume, E., Livet, M., Desbordes, M., and Villeneuve, J. P.: Hydrological analysis of the river Aude, France, flash flood on 12 and 13 November 1999, J. Hydrol., 286, 135–154, 2004.
Gottardi, F.: Estimation Statistique et Réanalyse des Précipitations en Montagne: Utilisation d'Ébauches par Types de Temps et Assimilation de Données d'Enneigement, Application aux Grands Massifs Montagneux Français, Institut National Polytechnique de Grenoble – INPG, Grenoble, 284 pp., 2009.
Gourley, J. J., Erlingis, J. M., Smith, T. M., Ortega, K. L., and Hong, Y.: Remote collection and analysis of witness reports on flash floods, J. Hydrol., 394, 53–62, 2010.
Gourley, J. J., Hong, Y., Flamig, Z. L., Arthur, A., Clark, R., Calianno, M., Ruin, I., Ortel, T., Wieczorek, M. E., Kirstetter, P.-E., Clark, E., and Krajewski, W. F.: A unified flash flood database across the US, B. Am. Meteorol. Soc., 94, 799–805, doi:10.1175/bams-d-12-00198.1, 2013.
Javelle, P., Fouchier, C., Arnaud, P., and Lavabre, J.: Flash flood warning at ungauged locations using radar rainfall and antecedent soil moisture estimations, J. Hydrol., 394, 267–274, 2010.
Javelle, P., Demargne, J., Defrance, D., Pansu, J., and Arnaud, P.: Evaluating flash flood warnings at ungauged locations using post-event survey: a case study with the AIGA warning system, Hydrolog. Sci. J., accepted, 2014.
Krzysztofowicz, R.: Performance tradeoff characteristic of a flood warning system, Water Resour. Bull., 28, 193–200, 1992.
Manzato, A.: An odds ratio parameterization for ROC diagram and skill score indices, Weather Forecast., 20, 918–930, 2005.
Marchi, L., Borga, M., Preciso, E., and Gaume, E.: Characterisation of selected extreme flash floods in Europe and implications for flood risk management, J. Hydrol., 394, 118–133, 2010.
Mathevet, T., Michel, C., Andréassian, V., and Perrin, C.: A Bounded Version of the Nash–Sutcliffe Criterion for Better Model Assessment on Large Sets of Basins, Large Sample

HESSD
11, 4365–4401, 2014

Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter ExperimentMOPEX, IAHS Publ. 307, AISH, Wallingford, 211–219, 2006.

Naulin, J. P., Payrastre, O., and Gaume, E.: Spatially distributed flood forecasting in flash flood prone areas: application to road network supervision in Southern France, J. Hydrol., 486, 88–99, 2013.

Ortega, K. L., Smith, T. M., Manross, K. L., Kolodziej, A. G., Scharfenberg, K. A., Witt, A., and Gourley, J. J.: The severe hazards analysis and verification experiment, B. Am. Meteorol. Soc., 90, 1519–1530, doi:10.1175/2009bams2815.1, 2009.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2 – Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling, J. Hydrol., 303, 290–306, doi:10.1016/j.jhydrol.2004.08.026, 2005.

Oudin, L., Andreassian, V., Perrin, C., Michel, C., and Le Moine, N.: Spatial proximity, physical similarity, regression and ungaged catchments: a comparison of regionalization approaches based on 913 French catchments, Water Resour. Res., 44, W03413, doi:10.1029/2007WR006240, 2008.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275–289, doi:10.1016/S0022-1694(03)00225-7, 2003.

Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C., and Mathevet, T.: Impact of limited streamflow data on the efficiency and the parameters of rainfall–runoff models, Hydrolog. Sci. J., 52, 131–151, 2007.

Reed, S., Johnson, D., and Sweeney, T.: Application and national georgraphic information system database to support two-year flood and threshold runoff estimates, J. Hydrol. Eng., 7, 209–219, 2002.

Rojas-Serna, C., Michel, C., Perrin, C., and Andréassian, V.: Ungauged Catchments: How Make the Most of a Few Streamflow Measurements?, Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter ExperimentMOPEX, IAHS Publ. 307, AISH, Wallingford, 230–236, 2006.

Ruin, I., Creutin, J.-D., Anquetin, S., and Lutoff, C.: Human exposure to flash floods – relation between flood parameters and human vulnerability during a storm of September 2002 in Southern France, J. Hydrol., 361, 199–213, doi:10.1016/j.jhydrol.2008.07.044, 2008.

HESSD
11, 4365–4401, 2014

Schaefer, J. T.: The critical success index as an indicator of warning skill, Weather Forecast., 5, 570–575, 1990.
Silberstein, R. P.: Hydrological models are so good, do we still need data?, Environ. Modell. Softw., 21, 1340–1352, doi:10.1016/j.envsoft.2005.04.019, 2006.
Swets, J. A.: The relative operating characteristic in psychology, Science, 182, 990–1000, 1973.
Tabary, P., Dupuy, P., L'Henaff, G., Gueguen, C., Moulin, L., Laurantin, O., and Merlier, C.: A 10 year (1997–2006) reanalysis of quantitative precipitation estimation over France: methodology and first results, in: International Symposium on "Weather Radar and Hydrology", Exeter, 18–21 April 2011, 672, 2012.
Versini, P.-A., Gaume, E., and Andrieu, H.: Application of a distributed hydrological model to the design of a road inundation warning system for flash flood prone areas, Nat. Hazards Earth Syst. Sci., 10, 805–817, doi:10.5194/nhess-10-805-2010, 2010.
Vidal, J. P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J. M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, Int. J. Climatol., 30, 1627–1644, 2010.
Wasson, J. G., Chandesris, A., Pella, H., and Souchon, Y.: Définition des hydroécorégions françaises, Méthodologie de détermination des conditions de référence au sens de la Directive Cadre pour la gestion des eaux. rapport de phase 1, Lyon, IRSTEA, 68 p., 2001.

HESSD
11, 4365–4401, 2014

Table 1. Characteristics of two catchment sample groups.

Sample group	Surface (km^2)	Snow (% annual precipitation)	Slope
HYDRO	[8.897], median: 110	[0.59], median: 8	[0.003, 0.17], median: 0.04
RTM	[5.887], median: 22	[0.77], median: 27	[0.02, 0.26], median: 0.09

HESSD
11, 4365–4401, 2014

Table 2. Contingency statistics and POD/SR/CSI criteria.

	Alert observed	No alert observed	Contingency statistics
Alert calculated	CA	FA	$POD = \frac{\sum CA}{\sum CA + \sum MA}$
No alert calculated	MA		$SR = \frac{\sum CA}{\sum CA + \sum FA}$

$\sum$CA: the number of correct alerts for all catchments (both modelled and observed);
$\sum$MA: the number of missed alerts for all catchments (both observed and unmodelled);
$\sum$FA: the number of false alerts for all catchments (both modelled and unobserved).

HESSD
11, 4365–4401, 2014

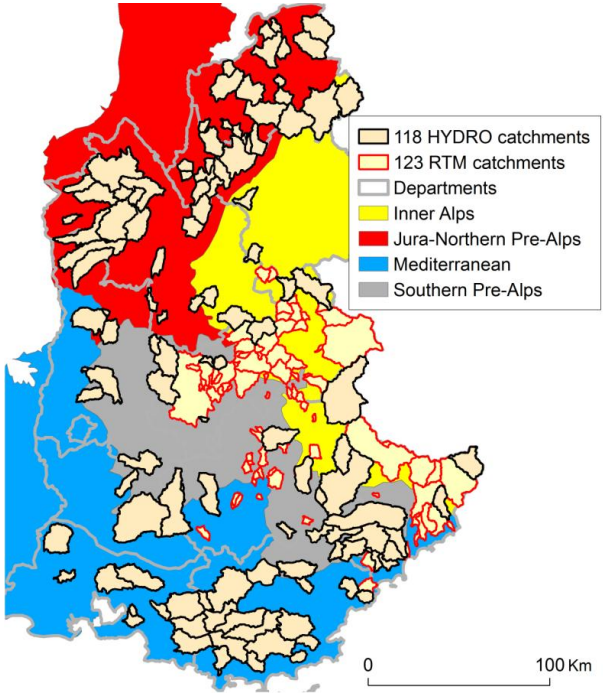

Fig. 1. Locations of the 118 HYDRO catchments (black) and the 123 RTM catchments (red).

HESSD
11, 4365–4401, 2014

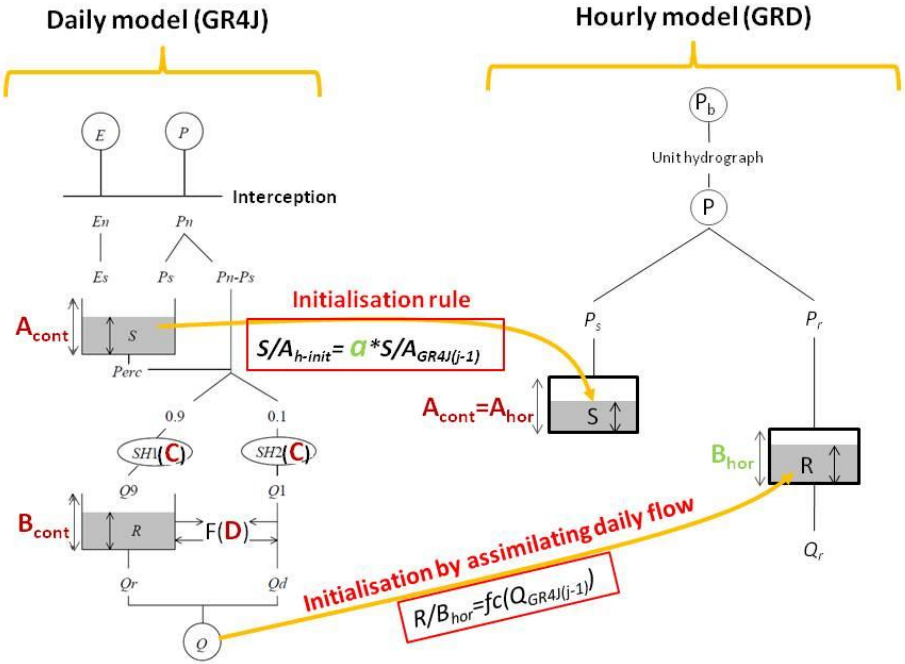

Fig. 2. The coupling between the daily model (GR4J) and the hourly model (GRD), with the 2 initialisation rules (production and routing, in red) and the parameters to be calibrated (4 GR4J parameters in dark red, and 2 GRD parameters in green).

HESSD
11, 4365–4401, 2014

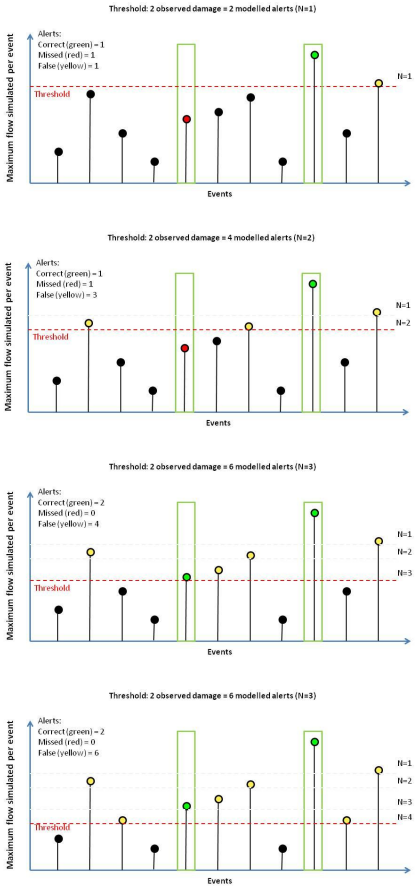

Fig. 3. The change in number of alerts (correct missed, and false) as a function of the given threshold. The values for *N* vary from 1 to 4.

HESSD
11, 4365–4401, 2014

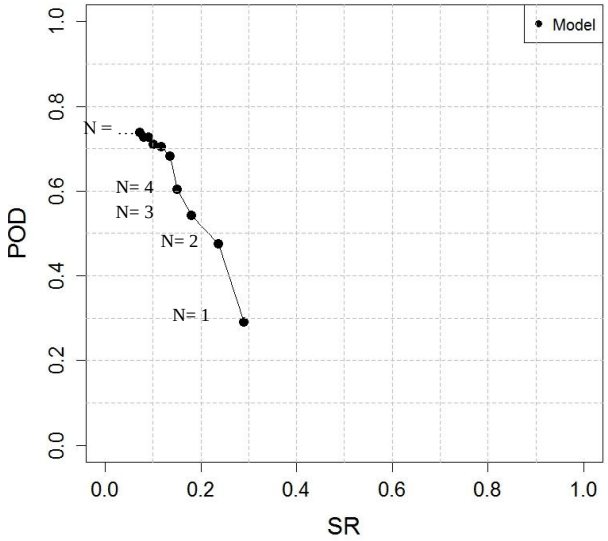

Fig. 4. Performance of a model estimated by the POD as a function of SR. The dots correspond to the contingency criteria for *N* thresholds.

HESSD
11, 4365–4401, 2014

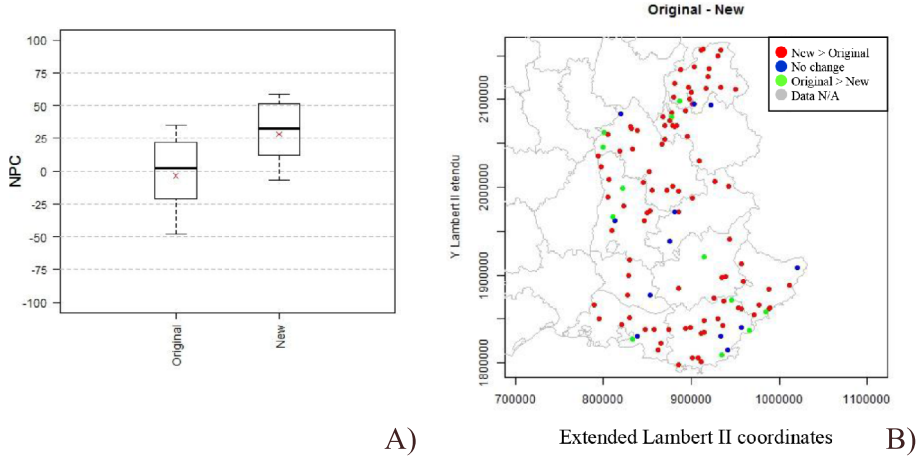

Fig. 5. Comparison of the original and new versions of the AIGA method in terms of Nash pseudo-continuous (NPC). **(A)** Box-and-whisker plot with the original version on the left and the new version on the right. The 1st and 3rd quartiles are inside the box and the 1st and 9th decile at whisker ends. **(B)** The range between the two versions: red dots show an improvement with the new version, green dots, a deterioration with the new version, and blue dots, no difference (under 5 points of NPC).

HESSD
11, 4365–4401, 2014

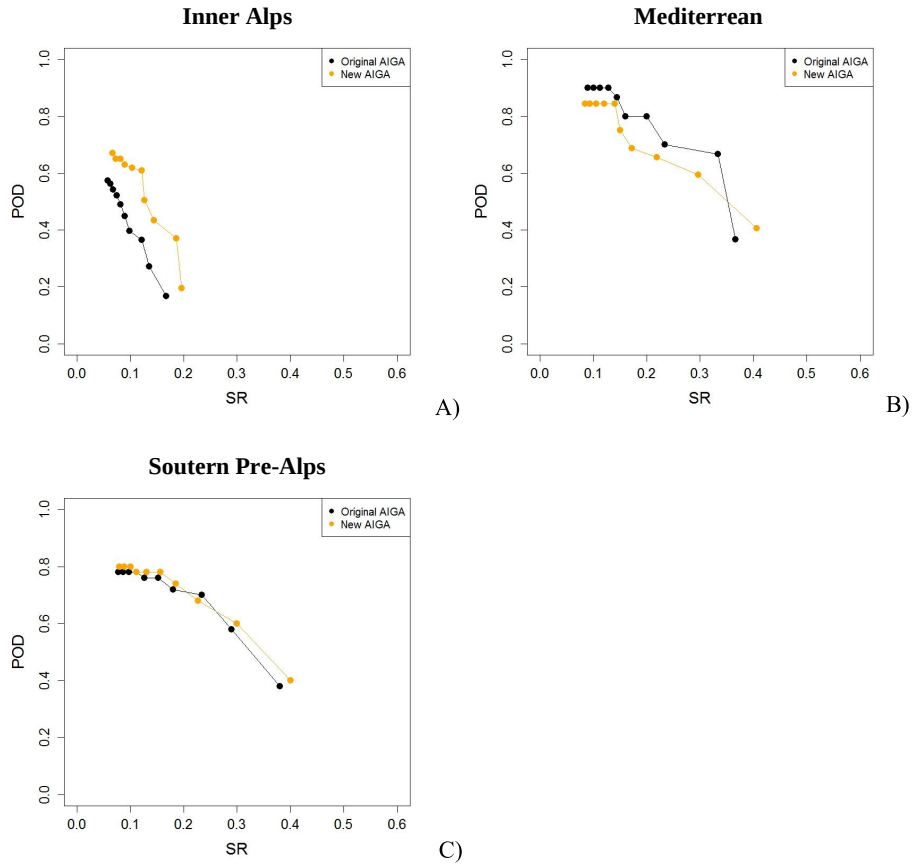

Fig. 6. Comparison of the performance of both AIGA methods using the POD/SR graph, with the original version in black and the new version in orange, for each hydro-ecoregion: **(A)** Inner Alps; **(B)** Mediterranean; **(C)** Southern Pre-Alps.

HESSD
11, 4365–4401, 2014

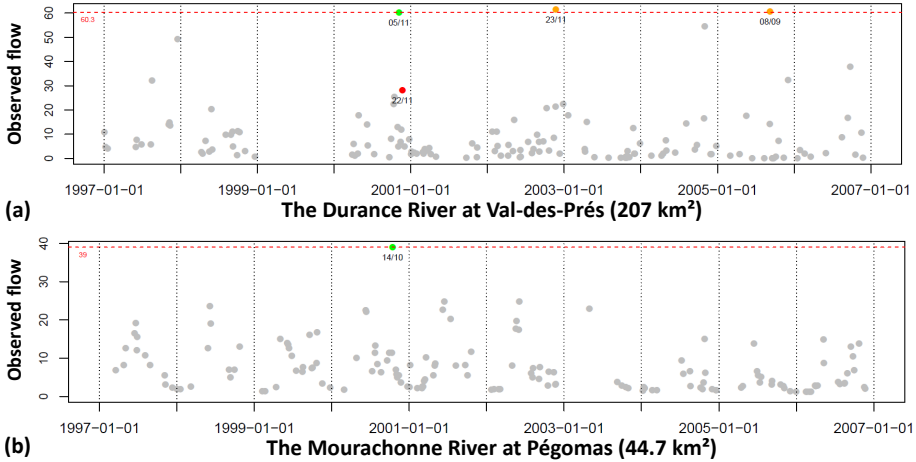

Fig. 7. Maximum flows observed during each event over a 10-year period for the four catchments that are in both sample groups (RTM and Hydro); the dotted red line indicates the highest threshold (number of alerts = number of times threshold is exceeded); grey dots indicate all the observed maximum flows; the green dots, an RTM damage report that coincides with a flow observed to be higher than the alert threshold; the orange dots, the flows observed to be higher than the alert threshold, but not corresponding to a damage report; the red dot, a damage report where the flow is under the threshold.

HESSD
11, 4365–4401, 2014

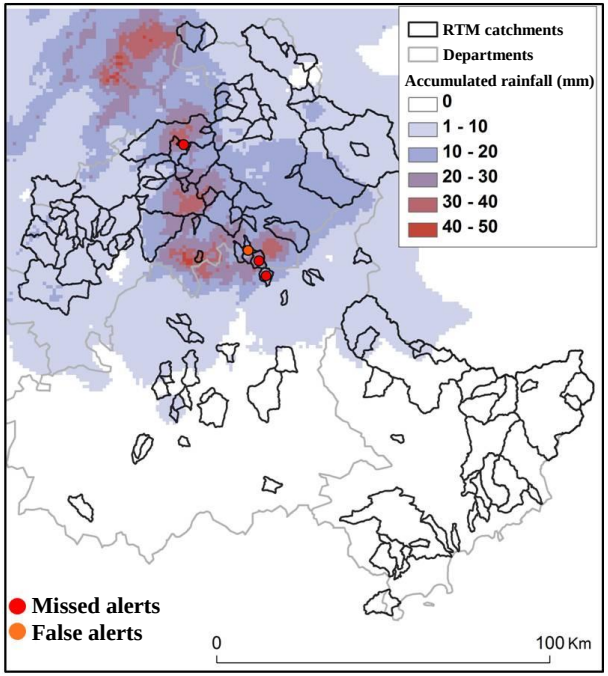

Fig. 8. Cumulated rainfall during the 27 July 2003 event in the French Alpes-de-Haute-Provence, Hautes-Alpes and Alpes-Maritimes departments. A comparison of the alerts from the new AIGA method in relation to the three RTM damage reports, with a very low triggering threshold (N = 27).

HESSD
11, 4365–4401, 2014

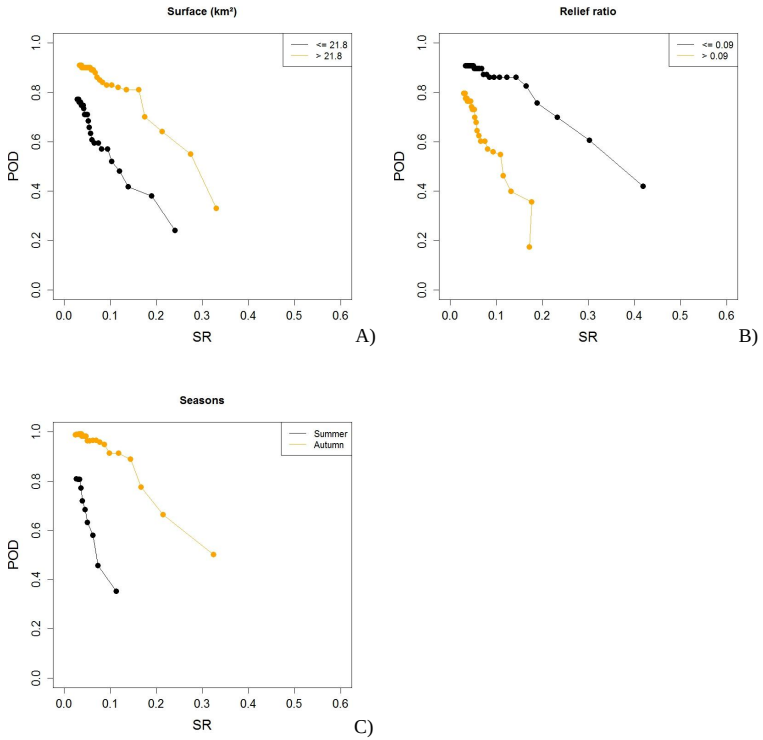

Fig. 9. Demonstration of the new AIGA method's problems linked to catchment characteristics and seasonal rainfall data. **(A)** Surface: performance comparison between large (orange) and small (black) catchments. **(B)** Relief ratio (slope): performance comparison between steep (orange) and "flat" (black) catchments. **(C)** Seasons: performance comparison between autumn (orange) and summer (black).

HESSD
11, 4365–4401, 2014