Assessing the ability of a new seamless short-range ensemble rainfall product to anticipate flash floods in the French Mediterranean area

Juliette Godet[1], Olivier Payrastre[1], Pierre Javelle[2], and François Bouttier[3]

[1]GERS-LEE, Univ Gustave Eiffel, IFSTTAR, F-44344 Bouguenais, France
[2]RECOVER, INRAE, Université d'Aix-Marseille, Aix En Provence, France
[3]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Correspondence: Juliette Godet (juliette.godet@univ-eiffel.fr)

Abstract. Flash floods have dramatic economic and social consequences, and efficient adaptation policies are required to reduce their impacts, especially in a context of global change. Developing more efficient flash flood forecasting systems can largely contribute to these adaptation requirements. The aim of this study was to assess the ability of a new seamless short range ~~(0-6h)~~ ensemble quantitative precipitation forecast (QPF) product, called PIAF-EPS and recently developed by Météo-France, to predict flash floods when used as input of an operational hydrological forecasting chain. For this purpose, eight flash flood events that occurred in the French Mediterranean region between 2019 and 2021 were reanalysed, using a similar hydrological modeling chain to the one implemented in the French "Vigicrues-Flash" operational flash flood monitoring system. The hydrological forecasts obtained from PIAF-EPS were compared to the forecasts obtained with different deterministic QPFs from which PIAF-EPS is directly derived ~~(i.e. the AROME-NWC numerical weather prediction model, and the deterministic PIAF product)~~. The verification method applied in this work uses scores calculated on contingency tables, and combines the forecasts issued on each $1km^2$ pixel of the territory. This offers a detailed view of the forecast performances, covering the whole river network and including the small ungauged rivers. The results confirm the added value of the ensemble PIAF-EPS approach for flash flood forecasting, in comparison to the different deterministic scenarios considered.

1 Introduction

The year 2022, and particularly the summer season, was marked by several deadly and catastrophic flash floods in Pakistan, Kentucky (USA), Iran, Sierra Leone, Bangladesh, Australia, and unfortunately many other countries. Very few parts of the world seem to be spared from flash floods. According to the World Meteorological Organization (WMO, 2020), floods are the deadliest natural hazards, and flash floods account for 85% of the flooding events and have the highest mortality rate within the category (5000 victims annually). In France, the Mediterranean region is particularly prone to severe flash floods. Even though an intensification of extreme rainfall events in response to anthropogenic influence was diagnosed (Ribes et al., 2019), the consequences of climate change on flash floods remain unclear in this region, particularly because of the compensating effect of the expected decrease in soil moisture (Tramblay et al., 2019). However, the increase in the vulnerability to these episodes may lead to an increase in the global risk associated with flash floods in the future years.

In this context, developing flash flood forecasting is of crucial interest to limit the death toll and optimize the emergency response. Several operational flash flood warning systems have recently been developed worldwide, and they generally have similar features. The observed or forecast rainfall can be directly compared to reference thresholds to estimate the flash flood likelihood. It is the case for instance in the Flash Flood Guidance system in the US (Clark et al., 2014), or the ERIC/ERICHA system in Europe (Raynaud et al., 2015; Corral et al., 2019). Rainfall data can also be used as input of highly distributed hydrological models, which may bring additional information about the intensity and temporal dynamics of the floods, and may be particularly interesting for decision-making (Zanchetta and Coulibaly, 2020). The FLASH system in the USA (Gourley et al., 2017) or the Vigicrues Flash service in France (Javelle et al., 2016; Piotte et al., 2020) are following this second principle. The operational systems using hydrological models are still often based on radar Quantitative Precipitation Estimates (QPEs), without involving Quantitative Precipitation Forecasts (QPFs). This choice increases the quality of detection and limits the risks of false alarms, but also highly limits the anticipation, that cannot exceed the (limited) response times of the small catchments where flash-floods do occur.

The development of convection permitting Numerical Weather Prediction (NWP) models has paved the way for the use of QPFs as input of flash-flood warning systems, with the objective to extend anticipation lead times up to 24-48h (Collier, 2007; Hapuarachchi et al., 2011; Zanchetta and Coulibaly, 2020). Convection permitting models offer an interesting capacity to describe heavy precipitation events, and offer space and time resolutions which are suited to the hydrological models used in flash-flood warning systems. However, the current QPF products still show spatial and temporal uncertainties in the description of intense rainfall cells, that may significantly exceed the typical scales of small river basins (Roberts and Lean, 2008; Clark et al., 2016; Armon et al., 2020). This may highly limit the capacity to issue relevant flash-flood warnings, without appropriate strategies to represent or reduce uncertainties (Silvestro et al., 2011; Vincendon et al., 2011; Furnari et al., 2020). Even if ensemble approaches have been widely used as input of flash flood forecasting chains (Vié et al., 2012; Alfieri and Thielen, 2012; Davolio et al., 2013, 2015; Hally et al., 2015; Nuissier et al., 2016; Amengual et al., 2017; Furnari et al., 2020; Sayama et al., 2020; Amengual et al., 2021), uncertainties in QPFs can still hardly be reduced for lead times exceeding 6-8h, even with enhanced assimilation schemes in NWP models (Davolio et al., 2017; Lagasio et al., 2019).

Efficient flash flood forecast strategies can also be developed for short lead-times (< 6 hours, i.e. the nowcasting range), with a high update frequency (typically 5 min to one hour between two runs of forecasts) to take regular benefit from the last available observations (Lovat et al., 2022). For such applications, the QPF products can be either derived from adapted versions of convection permitting NWP models (Auger et al., 2015; Benjamin et al., 2016), or by extrapolating the last radar observations (Berenguer et al., 2011; Silvestro and Rebora, 2012; Imhoff et al., 2022). Simple Lagrangian radar extrapolations can easily outperform NWP models for lead times up to 2-3h (Mandapaka et al., 2012), however they are not suited to larger lead times because they cannot reproduce the physical changes occurring in the atmosphere. For that reason, up-to-date short-range QPF approaches now combine both information sources through blending techniques, to offer a seamless transition between observed and forecast rainfall fields (Poletti et al., 2019; Lovat et al., 2022; Scheufele et al., 2014). However, despite all these efforts to create seamless short-range QPFs products, the forecast uncertainties still remain significant and need to be quantified through ensemble approaches (Bowler et al., 2006; Seed et al., 2013; Descamps et al., 2015; Osinski and Bouttier, 2018; Bouttier and Raynaud, 2018).

The objective of this paper is to assess the potential of a new seamless short-range ensemble QPF product, called PIAF-EPS ("PIAF" meaning Prévision Immédiate Agrégée Fusionnée, and "EPS" meaning Ensemble Prediction System) and recently developed by Meteo-France, for flash flood nowcasting purposes.~~The objective of this paper is to assess the potential of a new seamless short-range ensemble QPF product, called PIAF-EPS and recently developed by Météo-France, for flash flood forecasting purposes.~~ This ensemble aims to represent very short-range ~~(0–3 hours)~~ forecast uncertainties. It can be frequently updated at a very small numerical cost, to keep it consistent with the latest nowcasting data based on radar images. The aim is to confirm the benefits of using such an ensemble seamless product as input of flash flood nowcasting chains, compared to other short-range deterministic QPF products from which PIAF-EPS is directly derived. For this purpose, a reanalysis of eight flash flood events observed in the French Mediterranean region between 2019 and 2021 is proposed, using a similar hydrological modeling chain as the one implemented in the French "Vigicrues Flash" operational flash flood monitoring system. Since the selected flash floods mainly occurred on small rivers, the proposed evaluation framework does not focus only on a couple of gauged outlets, but offers a comprehensive coverage of the small rivers hit by the studied rainfall events. This is achieved by comparing the hydrological forecasts obtained using QPFs with simulated discharges (i.e. based on QPEs), at each pixel of the hydrological model grid (1-km resolution), and following a methodology adapted from Charpentier-Noyer et al. (2023).

The paper is organised as follows: Section 2 describes the hydrometeorological forecasting chains compared in the study; Section 3 provides details about the case studies used for the evaluation as well as the chosen verification method; finally, Section 4 presents and discusses the verification results.

2 The short-range hydrometeorological forecasting chains

2.1 General structure of the chains

The forecasting chains applied in this study are directly inspired from the French Vigicrues Flash operational flash flood monitoring service (Javelle et al., 2016; Piotte et al., 2020). They are presented on figure 1. The chains evaluate the severity of the floods, by comparing the simulated or forecasted hydrographs to reference discharge quantiles. The simulated/forecast hydrographs and the reference discharges are obtained using a fully distributed rainfall runoff model, detailed in section 2.4. This hydrological model is forced with the PANTHERE (Projet Aramis Nouvelles Technologies en Hydrométéorologie Extension et Renouvellement) rainfall QPEs,~~The simulated/forecast hydrographs and the reference discharges are obtained using a fully distributed rainfall runoff model (SMASH model). In the operational version of Vigicrues Flash, this hydrological model is forced with the PANTHERE rainfall QPEs,~~ derived from a network of about 30 radars over mainland France and its vicinity (Tabary et al., 2013). As mentioned in the introduction, the choice of using radar QPEs without QPFs would tend~~s~~ to limit the false alarms emitted by the chain~~system~~, but would~~it~~ also drastically limit~~s~~ its anticipation capacity.

In this paper, we thus combined QPEs with~~compared~~ different QPFs products as input of the chain, with the objective of increasing the current anticipation levels. The common time step for QPE-QPF and the hydrological model is 15 minutes. All the QPFs mentioned are available up to 6-hour forecast lengths but their refresh times depend on the considered product. For the present study, we decided to consider QPFs only for 0-3h forecast ranges and with a common refresh time of one hour. The QPF products include the new PIAF-EPS ensemble product, and three deterministic products used as reference. Two of these reference QPFs are directly involved in the generation of the PIAF-EPS ensemble (see figure 2), i.e. the deterministic version of PIAF, and the AROME-NWC numerical weather prediction model (AROME: Applications de la Recherche à l'Opérationnel à Méso-Echelle, NWC: Nowcasting). The third reference QPF corresponds to a naive constant rain scenario.

The next sections present each of the components involved in the forecasting chains applied in this study.

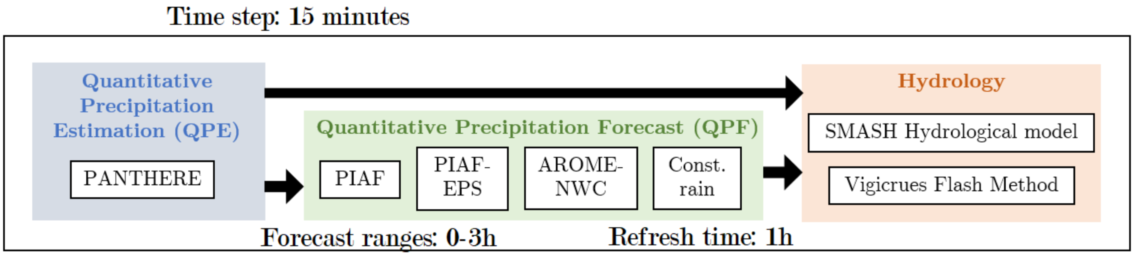

Figure 1. General structure of the forecasting chains

2.2 The three deterministic QPFs: AROME-NWC, deterministic PIAF and naive constant rain scenario

The first QPF product used as input of the chain corresponds to the AROME-NWC system, documented in (Auger et al., 2015). It is a rapid refresh version of the AROME convection-permitting numerical weather prediction system. It is updated every hour by a 3D-Var data assimilation system with a ten-minute observation cutoff (i.e. the initial state of each forecast is prepared using observations collected up to 10 minutes after its validity time), from which 6-hour forecasts are produced at 1.3km resolution with a 20-minute delivery time. Each 3D-Var analysis updates the model state by multivariately blending tens of thousands of observations from various meteorological networks (including radar winds and reflectivities, satellite radiances, GPS data, in situ surface and aircraft reports, etc). More information about the AROME-NWC 3D-Var can be found in Auger et al. (2015).

The second QPF product involved is the deterministic rain nowcast system called PIAF ("Prévision Immédiate Agrégée Fusionnée" in French, (Moisselin et al., 2019). Each PIAF forecast blends rainfall fields between radar QPF products, and AROME-NWC numerical predictions, as explained hereafter. The radar QPF product is derived from the PANTHERE radar QPEs. Rainfall accumulations are estimated every 5 minutes at 1km resolution, and extrapolated in time using an optical flow technique that persists the apparent motion of reflectivity from recent radar images. The blending between radar extrapolations and AROME-NWC follows equation $PIAF = \alpha \times radarQPF + (1-\alpha) \times AromeNWC$, where α is a forecast range~~lead~~

~~time~~-dependent weighting factor. At short forecast ranges~~lead times~~, α is equal to 1 so PIAF is equivalent to extrapolated radar QPF, which tends to be better than AROME-NWC. At longer forecast ranges~~lead times~~, typically beyond 1 to 2 hours, α smoothly decreases towards zero so that the PIAF converges to the latest available AROME-NWC precipitation forecast, which consistenly outperforms radar QPF at longer forecast ranges~~lead times~~. The speed at which α decreases is case-dependent: it is determined by a simple online machine learning procedure (Auer et al., 2002; Devaine et al., 2013) that minimizes the average forecast errors over the past six hours, as measured by a Gerrity score over large subdomains. In a nutshell, this algorithm produces a smooth transition (as a function of forecast range~~lead time~~) between the latest available radar extrapolation and AROME-NWC forecast; compared to climatologically optimal weights, this transition occurs earlier if AROME-NWC performed better than average during the six preceding hours (relative to radar extrapolation). Evaluations of the deterministic PIAF precipitation forecasts (Moisselin et al., 2019; Lovat et al., 2022) indicate that they statistically outperform both radar QPF products and AROME-NWC forecasts for forecast ranges~~lead times~~ between 0 and 3 hours.

Finally, we considered a third "naive" QPF scenario, corresponding to a constant future rain. Despite its very simplistic principle, this scenario may give valuable information since flash floods are often caused by quasi-stationary storm systems (Gaume et al., 2009).

2.3 The new PIAF-EPS ensemble QPF product

PIAF-EPS is a new experimental short range ensemble rainfall product, which is built by adding perturbations to the deterministic PIAF nowcast. The ensemble generation is original, and inspired by previously proposed stochastic nowcasting schemes, e.g. Bowler et al. (2006) and Seed et al. (2013). The perturbation tuning parameters have been kept to a minimum, in order to facilitate future operational deployment and maintenance of the proposed system. The perturbation technique is an adaptation to nowcasting ranges of the ’pertDpepi’ method used by Peredo et al. (2021) and Charpentier-Noyer et al. (2023). It is illustrated in figure 2. Each PIAF forecast (available every 5 min) is used to generate 16 perturbed members using equiprobable perturbations of the precipitation field: spatial perturbations and amplitude perturbations.

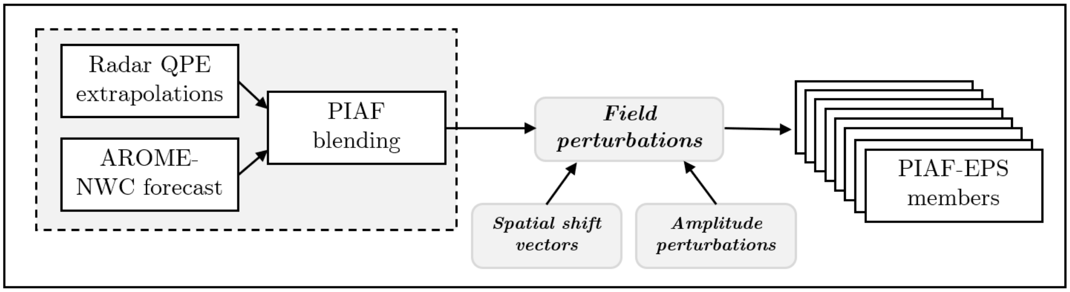

Figure 2. Illustration of the generation of the PIAF-EPS ensembles

The 16 spatial perturbations are pseudorandom~~subrandom~~ shifts that approximate (together with the unperturbed forecast) a 17-member, isotropic Gaussian sample in the 2D space. The shifting vectors are computed following the recommendations and dataset of Wang et al. (2019), which is a Dirac mixture algorithm involving the Cramér-von Mises method. It is a deterministic 2D distribution that is on average a better approximation of a Gaussian than a Monte Carlo sample, given the small ensemble size. The vector directions are constant in time for each ensemble member. The vector amplitudes are scaled as a function of lead time so that the amplitude of the spatial shifts grows linearly from zero to 30km over three hours, after which it is kept constant. This setting was based on a visual examination of spatial prediction errors for a set of high-impact precipitation events (independent from the ones used for the evaluations in this study).

The 16 amplitude perturbations are multiplications by 2D patterns of the spatially shifted fields. Each pattern is an independent realization of a 2D random field that has Gaussian autocorrelations in space, and a serial time autocorrelation from a clipped AR(1) autoregressive process. The autocorrelation scales are set to approximately (40km, 6h), respectively. Thus, the amplitude perturbations are independent between members, and they slowly evolve in time. The standard deviation of the perturbation amplitude grows linearly in time for the first forecast hour, after which it is kept constant; it has been tuned to produce reliable average standard deviations of the precipitation spread (as measured by the spread/skill ratio of the whole ensemble) over a large forecast tuning sample (one month, independent from the cases evaluated in this study). Likewise, a small bias correction (amplification of the highest precipitation intensities) of the forecasts with respect to precipitation observations has been applied using the same tuning sample. An example of the pertubations is given in figure 3.

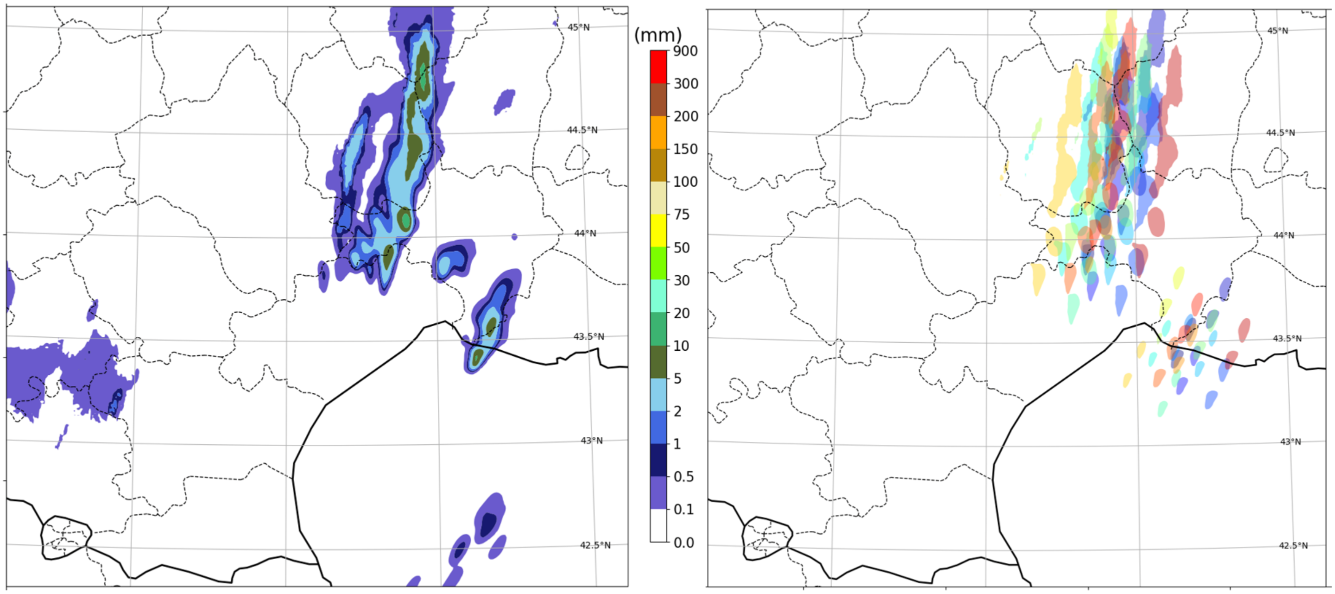

Figure 3. Example of PIAF-EPS ensemble forecast perturbations. Left : deterministic PIAF forecast of 15-minute rainfall accumulation (forecast start: 19 Sept 2020 at 06utc, forecast range: 2 hours). This is used as member zero of the ensemble. Right: same field in members 1 to 16, the shading represents rainfall areas above 5mm, with one colour for each member.

The unperturbed PIAF forecast is used as a 17th ensemble member, which makes the ensemble slightly non-equiprobable, but minimizes the risk of corrupting a good deterministic forecast by applying too large ensemble perturbations. The justification is that, in a few high impact cases, experience shows that intense Mediterranean precipitation can be quite precisely predicted by numerical models thanks to the influence of local orographic features. Further improvement to our (purely statistical) ensemble generation technique would be needed to automatically reduce the perturbation amplitudes in such cases, which is left for a future study.

2.4 The SMASH hydrological model and the Vigicrues Flash method

The rainfall-runoff part of the forecasting chains is based on the SMASH (Spatially-distributed Modelling and Assimilation for Hydrology) model. SMASH is a highly distributed, continuous and conceptual hydrological model developed at INRAE and Hydris Hydrologie (Jay-Allemand et al., 2020). The general principle of the model is presented in Figure 4. SMASH is inspired by the GR (Génie Rural) reservoir-based family of models (Perrin et al., 2003). For each pixel of the territory, the model includes a production reservoir (capacity c_p), a transfer reservoir (capacity c_{tr}), and an adapted cell-to-cell routing model, represented by a routing reservoir (capacity c_r).

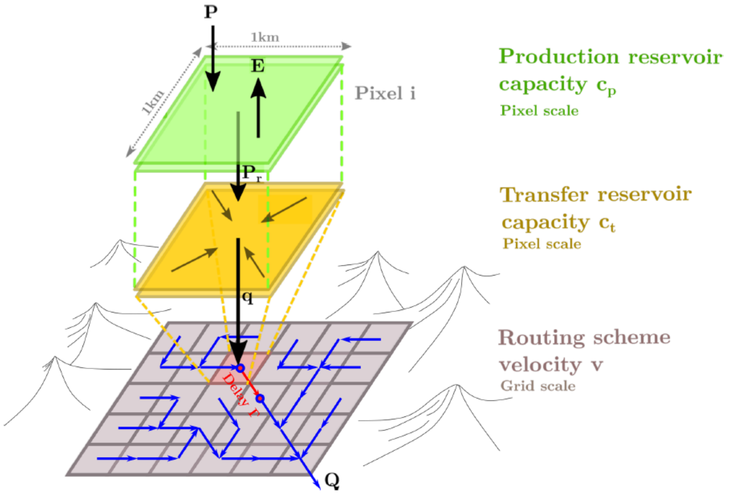

Figure 4. General outlines of SMASH (Jay-Allemand, 2020). P represents the local rainfall over one cell; E is the potential evapotranspiration; P_r is the effective rainfall; q is the elementary discharge; and Q is the total routed discharge

The version of SMASH used in this study is the one that is currently operational in the Vigicrues Flash system. This version is working on a 1-km grid, at a 15 min time resolution. It is a "lag 0" version, which means that there is no cell to cell routing scheme (or in other words, that the routing velocity is infinite) : the discharge on a cell is the sum of the instantaneous discharges of all the upstream cells. This method does not provide realistic hydrographs, but this is not considered as a problem, since the warning thresholds are defined based on a "climatological" run of the same model (see next paragraph).

According to the Vigicrues Flash method, the forecasted hydrographs obtained with SMASH are compared with reference discharge quantiles corresponding to defined return periods. These reference values are obtained by running the SMASH model for a long and continuous period, and by adjusting a Gumbel distribution to the corresponding annual maximum series. For this study, a 15-year long simulation period was used, which is the longest period that can be simulated based on an homogeneous PANTHERE QPE product. T=2, 5 and 10-year discharge quantiles were obtained on each $1km^2$ pixel of the studied area (see section 3.1). Figure 5b illustrates the discharge quantiles obtained for the return period T=2-year. In the West zone, the effect of relief in the Cévennes mountainous area is clearly distinguishable, logically resulting in higher rainfall amounts and higher flood quantiles. However, the results appear less consistent in the East zone, firstly because several very intense events occurred in the 2006-2021 simulation time window (sensibility to sampling), and secondly because the quality of the radar rainfall is questionable in this area. Indeed, a V-shaped band can be clearly observed, which is probably the result of bad calibration of the Collobrières radar. However this does not alter the methodology and results proposed in this study, since only simulated (and not observed) discharges are used to assess the quality of the forecast results.

3 Case studies and verification method

3.1 Study area and selected events

The South of France and particularly the Mediterranean region has experienced a large number of catastrophic flash flood events this last decade, both in terms of economical damage and casualties. The study has been focused on the most recent events that hit this area, since the ensemble PIAF-EPS forecasts can be released only from February 2019 (it would be labour intensive to process older cases, because of technical constraints in the archiving system, and they would be less and less relevant to current operational forecasting systems because the AROME and PIAF systems are frequently upgraded, typically once a year)~~(technically, it would be a costly process to go further in the archives, because of the constant evolution of the systems)~~. Eight heavy precipitation events which occurred between 2019 and 2021 in the South-Eastern region of France (see Figure 5a) were selected. Figure 6 shows the maps of rainfall accumulations for each event and Table 1 provides additional information including the duration, the maximum rainfall accumulation (spatial maximum), and the intensity and geographical extent of the hydrological responses simulated by the SMASH model~~and the importance of simulated hydrological reactions~~.

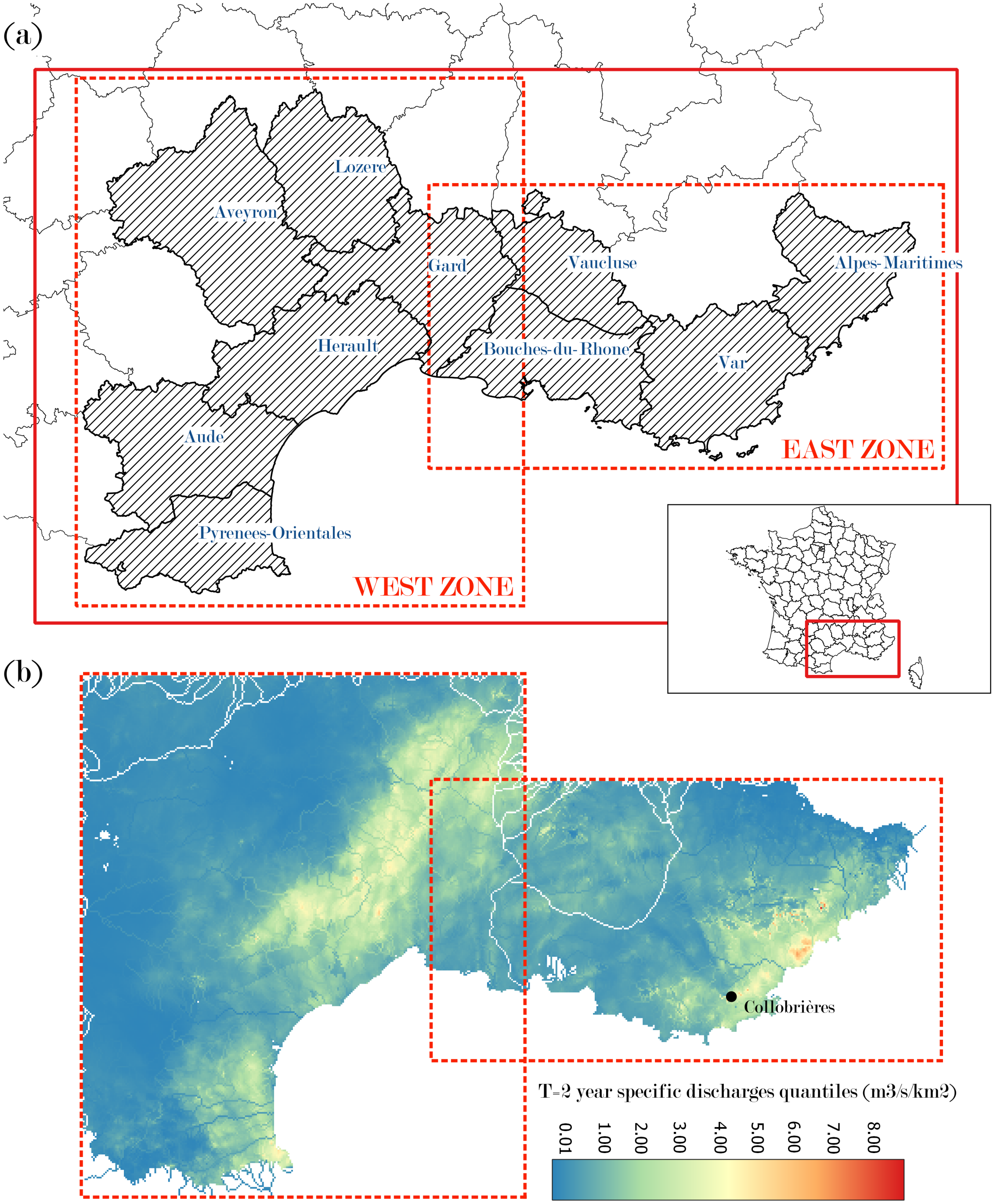

Figure 5. Study area and French departments affected by the events (a) and T = 2years specific discharge quantiles estimated on the study area based on a 15-year SMASH simulation (b)

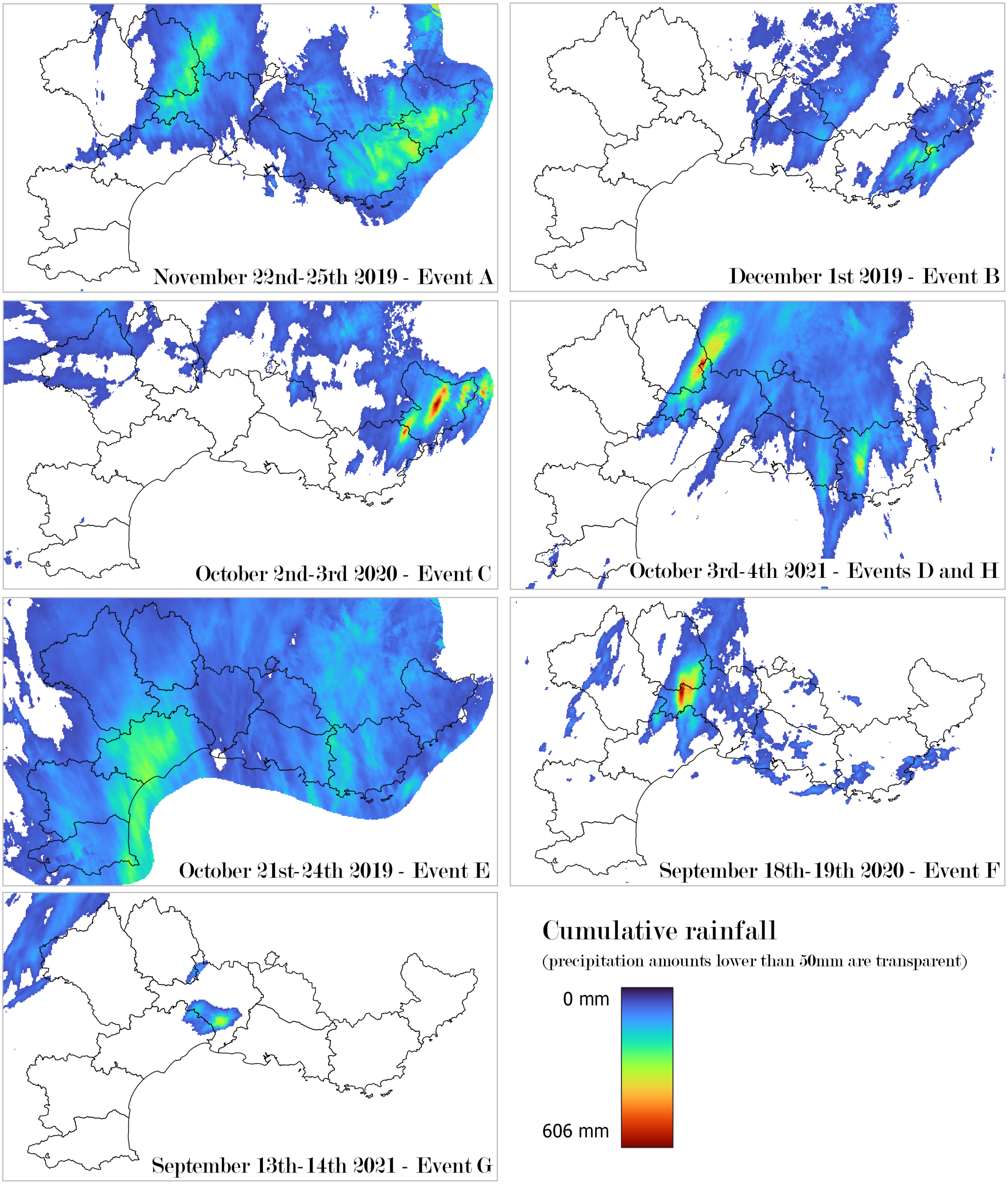
November 22nd-25th 2019 - Event A
December 1st 2019 - Event B
October 2nd-3rd 2020 - Event C
October 3rd-4th 2021 - Events D and H
October 21st-24th 2019 - Event E
September 18th-19th 2020 - Event F
September 13th-14th 2021 - Event G
Cumulative rainfall
(precipitation amounts lower than 50mm are transparent)
0 mm
606 mm

Figure 6. Maps of rainfall accumulations for each of the selected events. These maps were drawn using the ANTILOPE QPE (Laurentin, 2008), i.e. the best reanalysed QPE merging radar estimations and rain gauge observations

Date~~Duration~~	Max. cumulative rainfall	$\frac{S_{Q\geq Q_{T=2y}}}{S_{Zone}} \times 100$	$\frac{S_{Q\geq Q_{T=5}}}{S_{Zone}} \times 100$	$\frac{S_{Q\geq Q_{T=10y}}}{S_{Zone}} \times 100$	Zone	Label
22-25/11/2019	380 mm	22.06	7.92	3.86	East	A
01/12/2019	360 mm	11.92	6.53	4.77	East	B
02-03/10/2020	600 mm	2.42	1.28	0.97	East	C
03-04/10/2021	370 mm	4.31	2.00	1.33	East	D
21-24/10/2019	300 mm	11.03	6.91	5.39	West	E
18-20/09/2020	590 mm	1.35	0.77	0.53	West	F
13-14/09/2021	330 mm	0.36	0.27	0.25	West	G
03-04/10/2021	540 mm	3.94	1.76	1.38	West	H

Table 1. Description of the flash flood events: date, maximum rainfall accumulation (spatial maximum), percentages of the study area where the SMASH simulated peak discharge exceeds the 2-year, 5-year and 10-year discharge thresholds, affected zone (by reference to figure 5) and assigned label

The rainfall accumulations presented in Figure 6 show that the selected events have very different features. Some events, such as events A and E, show a wide spread of rainfall. For these events, the larger rainfall accumulations appear homogeneous over areas covering one or several departments. The other events are much more localized and have a larger variability of rainfall accumulations. Some of them show locally very intense rainfall cells (events C, D and F). The rainfall accumulation map for the October 2021 event shows that two separated zones were affected by the heavy rains, and the study of the QPE over time revealed that both zones were not affected at the same time : the heavy rainfall hit the Lozère department first, on October 3rd, and then the Bouches-du-Rhône and Var departments, on October 4th. It was therefore decided to separate this event into two distinct events, labelled D and H.

For most of the 8 selected events, the larger hydrological responses occurred in small ungauged catchments. However, for events A, B and C, post-event studies could estimate the maximum peak discharge values (Lebouc and Payrastre, 2020; Brigode et al., 2021; Payrastre et al., 2022). For events A and B, it was estimated that peak discharges reached locally a 5 to 15 $m^3/s/km^2$ range, in the small basins hit by the larger rainfall accumulations. These two events are hydrologically interesting because they happened close in time and in the same area: the maximum peak discharges were probably observed during event B because of larger soil saturation and higher rainfall intensities observed on short time steps (Brigode et al., 2021). For event C (Storm Alex) which hit the same department (Alpes Maritimes), Payrastre et al. (2022) estimated that despite significantly higher rainfall accumulations, the peak discharges were globally similar to those observed during the November-December 2019 flash flood events, except on some upstream basins were the estimated peak discharges reached values in the 15 to 20 $m^3/s/km^2$ range. Considering these discharge values, events B and C are among the most intense floods observed up to now in the Var and Alpes Maritimes departments.

Because of the limited information about the actual intensity and location of the flood responses during the eight selected events, it is difficult to compare more thoroughly the characteristics of the flood events. The comparison of simulation results obtained with the SMASH model~~Vigicrues Flash chain~~ can, nevertheless, bring additional information. Table 1 gives the percentage of surface where several discharge thresholds (2, 5, and 10-year return periods) were exceeded by the reference simulation.~~The column $S_{Q>Q_{T=10y}}$ in table 1 gives the area in kilometers where the 10-year discharge quantile was exceeded by the PANTHERE simulation.~~ It refers to figure A1 in appendix A, which gives an idea of the extent of the flood responses exceeding these thresholds~~several discharge thresholds (2, 5, and 10-year return periods)~~ for each event. Again, we can see here that the considered flood events show very different characteristics in terms of spatial extent: from very localized events (F and G), up to more generalized flood responses (events A, B and E), according to the SMASH simulations.

3.2 Verification method

As mentioned in the introduction, the objective is to assess the benefits of forcing the flash flood nowcasting chain with the PIAF-EPS forecasts. The whole river network of the study area, including small ungauged rivers, should be considered. As a consequence, the verification process has been applied at each $1km^2$ pixel of the SMASH model, to provide an as detailed evaluation as possible. Since no discharge observation is generally available at this $1km^2$ scale, the discharge simulated by the SMASH hydrological model forced with the observed PANTHERE QPE was used as reference. This reference also allows ignoring the errors due to the hydrological model, the performance of which have not been assessed in this study. Therefore, in the following, Q_{sim} is assigned~~assimilated~~ to the discharge calculated by the SMASH model forced with PANTHERE QPE, while Q_{for} is assigned~~assimilated~~ to the discharge forecasts obtained by forcing the model with the different rainfall forecasts.

The verification process aims to evaluate if the exceedances by Q_{sim} of discharge thresholds Q_t, are well anticipated by Q_{for}, i.e. when rainfall forecasts are used as input of the chain, instead of the PANTHERE QPEs. Since Q_{for} may correspond either to ensemble forecasts (case of PIAF-EPS QPF) or to deterministic forecasts (case of other QPFs), we selected verification scores that can be applied on both deterministic and probabilistic forecasts.

The verification is based on the filling of contingency tables, which are commonly used for assessing the ability of forecasting systems to detect binary events (Mason, 1982). The contingency tables are filled by comparing the reference (Q_{sim}) and forecasted (Q_{for}) discharges to a threshold (Q_t), resulting in the four outcomes presented in table 2. Then, the probability of detection ($POD = \frac{a}{a+c}$) and the probability of false detection ($POFD = \frac{b}{b+d}$) can be calculated. In the case of ensemble forecasts, they can be plotted for different forecast probabilities to create a ROC curve (Mason, 1982).

	$Q_{for} \geq Q_t$	$Q_{for} < Q_t$
$Q_{sim} \geq Q_t$	HIT (a)	MISS (c)
$Q_{sim} < Q_t$	FALSE ALARM (b)	CORRECT REJECTION (d)

Table 2. Content of a contingency table. a, b, c, and d correspond to the number of forecasts in each category.

Contingency tables are usually filled by combining a continuous temporal sequence of forecasts, at one single site and for one unique lead time. However, we followed here the principle proposed by Charpentier-Noyer et al. (2023) to build the contingency tables by aggregating the forecasts issued during the most critical phase of the event (i.e. the forecasts issued during the flood rising limb, just before the threshold exceedance by Q_{sim}, independently of the lead time). The detailed methodology is explained in appendix B.1. It includes a selection process of the forecasts considered to fill the contingency table, called stratification. Some adaptations have been introduced in this process in order to create a forecast-based stratification, rather than an observation-based stratification, following the recommendations of Bellier et al. (2017). The details are provided in appendix B.2.

Here we considered the forecasts obtained for each pixel of the SMASH model to build the contingency tables. Three discharge thresholds were considered at each pixel : $Q_t = \{Q_{T=2y}, Q_{T=5y}, Q_{T=10y}\}$ (see section 2.4 and appendix A). The contingency tables obtained for each threshold may be visualised directly on a map (see figure 7), or summarized using synthetic scores such as the POD or POFD. However, the POFD score is sensitive to the extent of the verification area, which directly determines the number of correct rejections (see figure 7). The choice of the verification area was already identified as an important issue by Charpentier-Noyer et al. (2023) who suggested to pay a particular attention to the choice of the "HFA" (Hydrological Focus Area). For that reason, we chose to summary the contingency tables based on the Critical Success Index ($CSI = \frac{a}{a+b+c}$) instead of POD/POFDs. The CSI score does not take into account the correct rejections, thus is much less sensitive to the choice of the verification spatial window. This choice avoided the issue of defining an appropriate verification zone for each of the considered events.

It is important to mention that although the verification method is based on classical statistical scores, it cannot characterize the performances of the QPFs for long temporal series. We are only assessing here the ability of the different QPFs to correctly predict several specific events of high intensity. The obtained results, even if providing interesting information, cannot be extrapolated to future events because of the too limited number of events considered in this study.

4 Results and discussion

4.1 Maps of contingency tables : presentation on a result sample

As explained in section 3.2, the contingency tables filled for each event can be represented on maps. One map can be extracted for each discharge threshold $Q_t = \{Q_{T=2y}, Q_{T=5y}, Q_{T=10y}\}$, and each QPF product. In the case of the PIAF-EPS ensemble forecast, one map is obtained for each percentile of the forecast ensemble.

These maps allow to observe and compare the performance of the forecasts. For example, figure 7 shows the maps obtained in the case of event F and a threshold $Q_t = Q_{T=2y}$. For this event, the constant rain and the AROME-NWC forecasts showed poor performance ~~(CSI=0.12 and CSI=0.13 respectively)~~, the first one correctly emitting warnings on the affected region but emitting many false alarms elsewhere, and the other one missing most of the area affected by the event. The CSI scores summarizing the content of the contingency tables (see section 3.2) are 0.12 and 0.13 respectively for these two forecasts. The improvements observed for the deterministic PIAF forecast (CSI=0.21) confirm that, in this case, the blending of radar QPFs and AROME-NWC was effective. Moreover, the 60% percentile of the PIAF-EPS forecast, which has the highest CSI score among the other percentiles, shows even better results than the PIAF forecast (CSI=0.27), by reducing notably the area affected by false alarms.

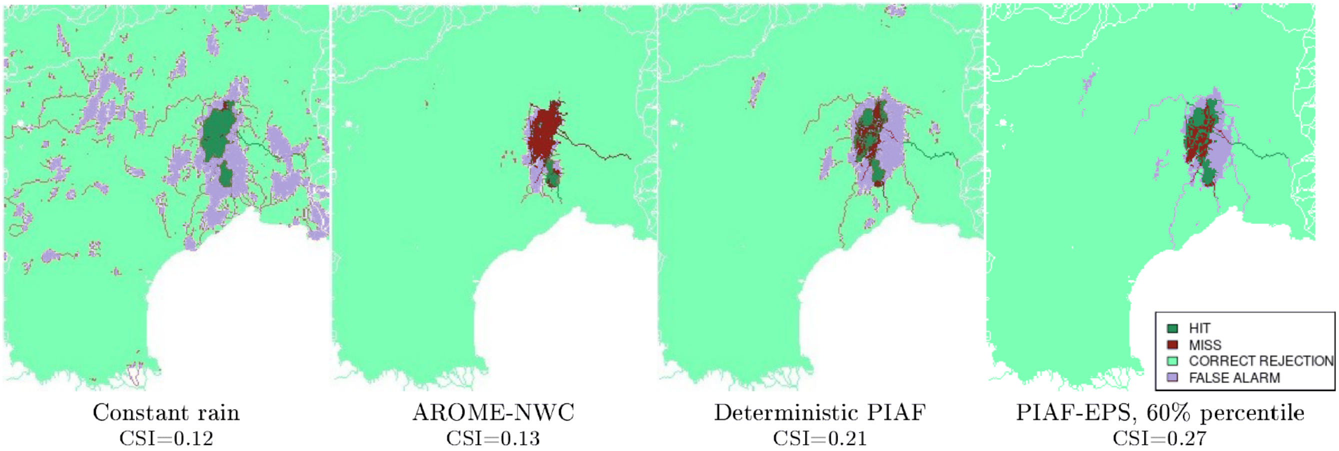

Figure 7. Maps of contingency tables for each forecast product, event F, $Q_t = Q_T = 2y$

Figure 8 shows the maps of the contingency tables obtained for another single event (event B) and for the different PIAF-EPS percentiles. This figure illustrates the evolution of the forecast performance depending on the considered percentile of the ensemble forecast. Logically, for low percentiles, the number of correct detections is very small and much lower than the number of missed warnings, resulting in low values of CSI. For intermediate percentiles, the number of correct detections increases with respect to the number of missed warnings and false alarms, which results in an increase of the CSI values. However, for larger percentiles, the number of false alarms increases up to the point that it outweighs the increase of correct detections, resulting in a decrease of the CSI scores.

This evolution of detection performance depending on the percentile of the ensemble forecasts directly explains the expected "hill" shape of the CSI curves presented on figure 9. The "best" ensemble percentiles can be identified by selecting the maximum on these curves. In the case of event B, figures 8 and 9 show an optimal HIT/MISS/FALSE ALARM balance for percentiles around 50-60%.

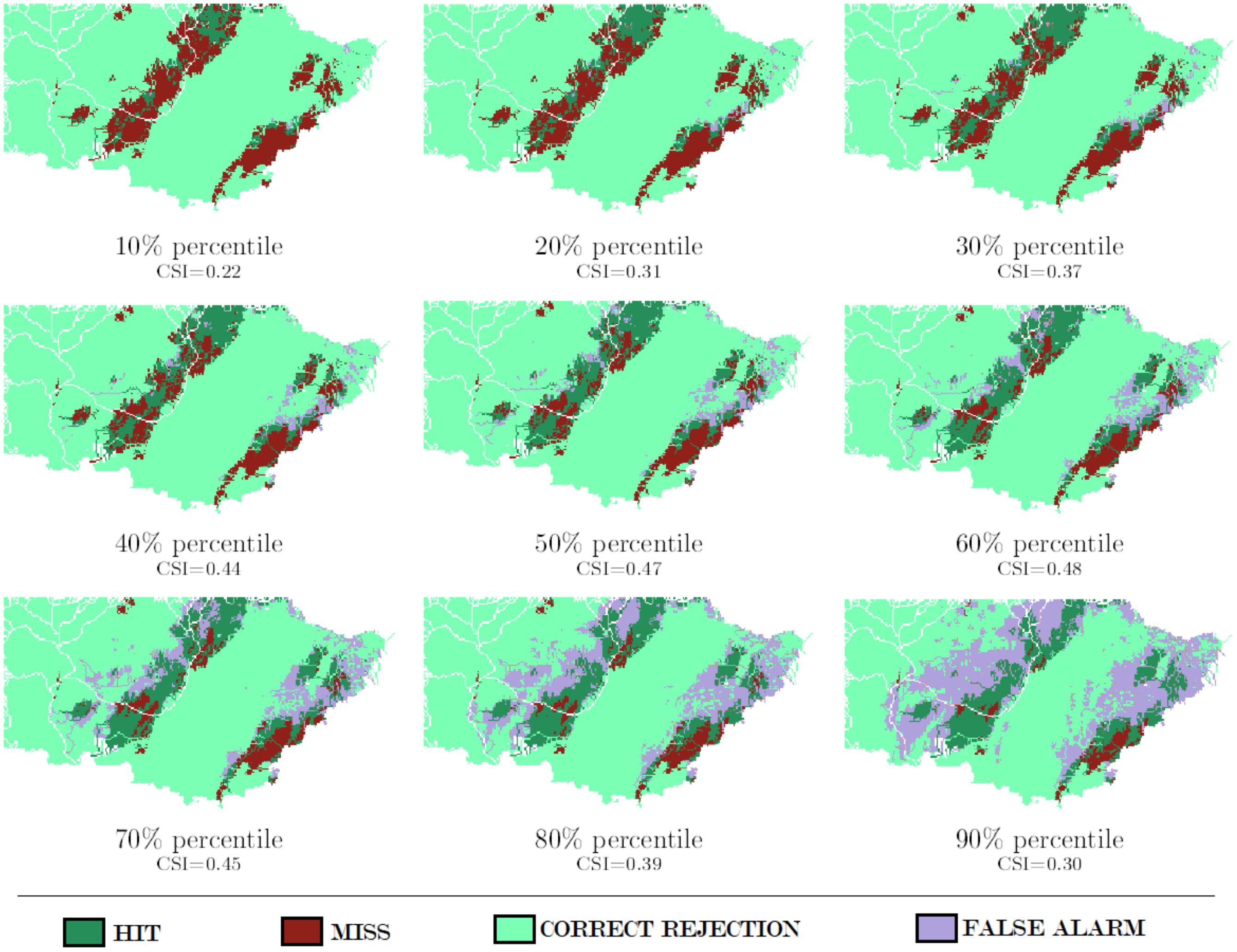

Figure 8. Map results of contingency tables for the $Q_{T=5y}$ (event B).

4.2 Analysis of CSI scores at the event scale

Figure 9 summarizes the CSI scores obtained for each event, each discharge threshold and each of the considered forecast products. One unique CSI score is computed for the reference deterministic forecasts, whereas a CSI curve is obtained in 290 the case of the ensemble PIAF-EPS forecast. These CSI scores provide a synthetic view enabling to compare the respective performances of the different forecast approaches. For a more detailed analysis, the corresponding maps of contingency tables are presented on figure D1 in appendix D: the best performing (i.e. maximal CSI) reference deterministic forecast is compared to the best performing PIAF-EPS percentile, for the $Q_{T=10y}$ threshold and for each event. These maps are complementary to the CSI values presented on figure 9, since they allow to visualize the geographical differences between the different QPF products, resulting in CSI differentials.

Before comparing the different forecast approaches, two generic observations can be made on figure 9. First, the performance of all forecasts tends to decrease as the discharge thresholds increase. This is in agreement with the theory developed by Schaefer (1990), according to which the CSI is biased by the frequency of the forecast event. Typically, the rarer an event is, the higher its return period is, and the lower is the CSI. Here it leads to a CSI decrease of 0.1 to 0.2, when comparing the 2-year and 10-year discharge thresholds. Second, the CSI curves for the PIAF-EPS forecasts do not always have the expected "hill" shape with a maximum at intermediate percentiles. Especially, for events D, E, G and H, the maximum CSI is reached at low percentiles. It means that, for those events, PIAF-EPS had a tendency to over-estimate the discharge probabilities. This is consistent with the rank diagrams plotted for each event in appendix C (figure C1), which generally show a slight positive discharge bias for the same events (D,E,G,H). A larger bias can even be observed for the forecast discharges exceeding the 2-year return period threshold, even if in this specific case the bias can be increased by the stratification process (see appendix C). ~~However, in this case the positive bias might be caused by the forecast-based stratification: since only forecasts with high discharge values are considered, they tend to be higher than the observations (see appendix C). But this stratification does not affect the global rank diagrams.~~

Furthermore, figure 9 shows that the intermediate percentiles of PIAF-EPS (i.e. 40% to 60%) outperform almost systematically the naive forecast approach (constant rain), and the deterministic PIAF forecast. This confirms that adding spatial and amplitude perturbations to the PIAF QPFs, to obtain the PIAF-EPS ensemble QPF product, resulted in better performances of the flash-flood forecasts, at least for the 8 intense flash-floods considered in this study. The results are more mixed concerning AROME-NWC. For 5 out of 8 events, AROME-NWC shows equivalent or lower CSI values than the deterministic PIAF, and it is outperformed by the PIAF-EPS forecasts in these cases, at least for intermediate percentiles. Conversely, there are 3 events for which AROME-NWC leads to significantly higher CSI values than deterministic PIAF (events E, G and H). For two of these three events (events E and G), PIAF-EPS largely compensates the poor performance of deterministic PIAF, leading to CSI values that are similar to AROME-NWC (events E and G). Lovat et al. (2022) have already shown that, depending on the lead time, AROME-NWC can outperform deterministic PIAF forecasts. The relationship between the CSI values and the lead times can hardly be investigated in this work, as the verification method applied looks at all the forecast runs emitted in a specific time window, regardless of their respective lead times. An explanation for the poor performance of deterministic PIAF on these events could be a sudden stationarization of the rain cells. In such a situation, the Lagrangian radar QPE extrapolation becomes a very poor rain predictor, because it cannot account for rapid changes in the speed of high precipitation areas. The results obtained here suggest that PIAF-EPS can at least partly handle the inherent uncertainty of these situations where the blending with radar QPE extrapolations limits the quality of the deterministic PIAF forecast.

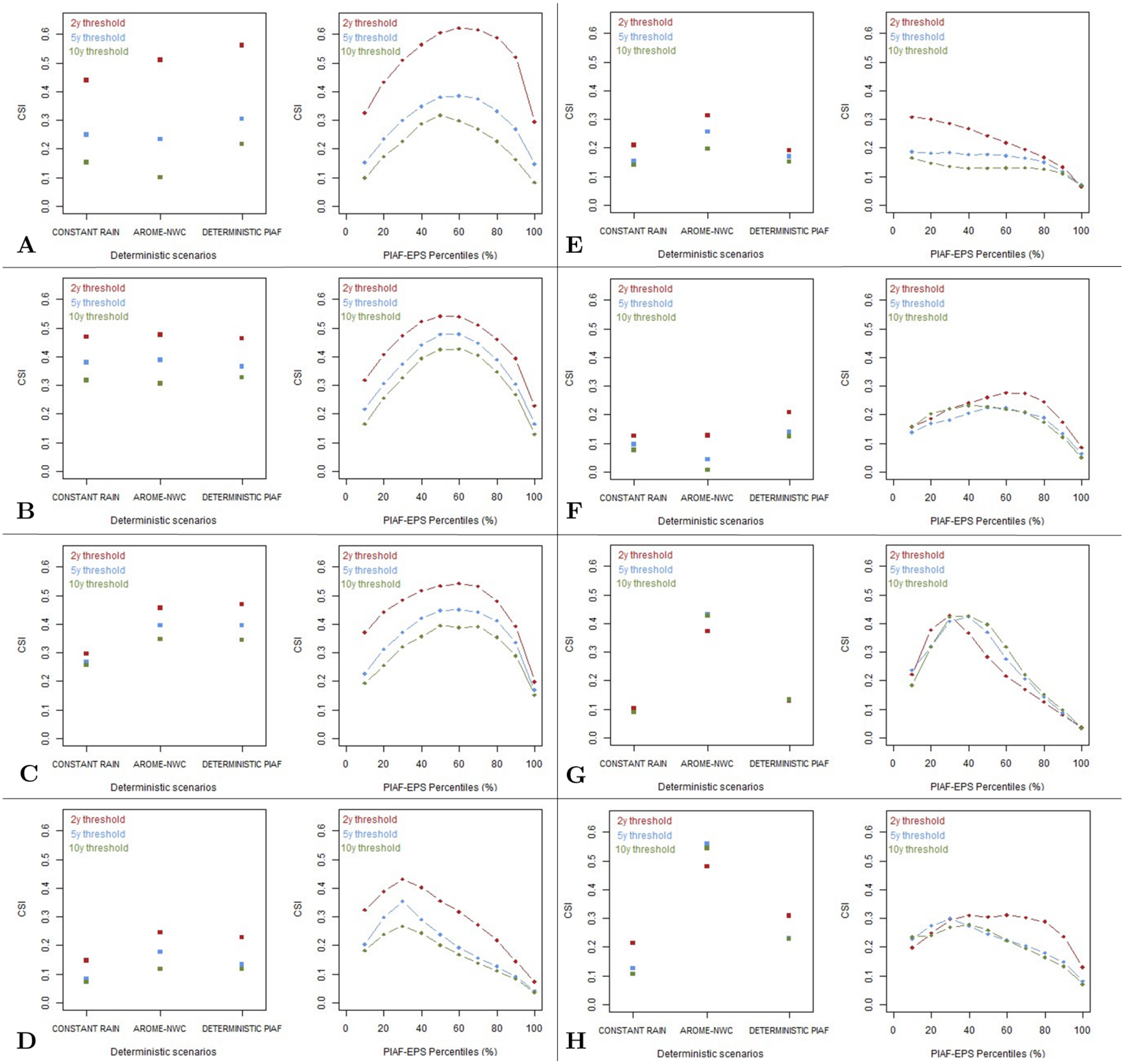

Figure 9. CSI curves of each event, for the various forecast products and thresholds

4.3 Averaged CSI scores

Averaged CSI scores were also calculated in order to provide a more synthetic view of the forecast performance, by aggregating them over all studied events. There are two possible ways of calculating global CSI scores:

$CSI_1 = \frac{1}{N_{evt}} \sum_{n=1}^{N_{evt}} \frac{a_n}{a_n+b_n+c_n}$ and $CSI_2 = \frac{\sum_{n=1}^{N_{evt}} a_n}{\sum_{n=1}^{N_{evt}} a_n+b_n+c_n}$

Since the studied events have very different spatial extents (see figure A1 and table 1), we chose to use the CSI_1 formula where the averaging implies that all the events have the same weight. CSI_2 would have given much more relative weight to the large-scale events.~~CSI_1 is the mean CSI for all the events, where the averaging implies that all the events have the same weight. We chose not to use the CSI_2 formula, because the studied events have very different spatial extents (see figure A1). If we used the CSI_2, we would give much more relative weight to the largest-scale events.~~

The global CSI curves obtained from CSI_1 are presented in figure 10 and indicate that globally, PIAF-EPS 30-60% percentiles outperform all the deterministic reference forecasts. This identification of "best" percentiles can be useful for end-users (WMO, 2012), particularly if they remain relatively stable depending the considered events. In this paper, we assess the value of these "best" percentiles in a slightly overoptimistic way, since we use all events to derive the optimum CSI values (i.e. it is not an "out of sample" optimization), because our sample is very small. This methodological weakness does not invalidate our conclusions, however, since the CSI curves are rather stable from case to case in our sample. The global CSI_1 scores also confirm that AROME-NWC globally performs better than PIAF for the eight studied events. However, this conclusion is largely influenced by events G and H, for which the CSI differences between AROME-NWC and deterministic PIAF are the largest.

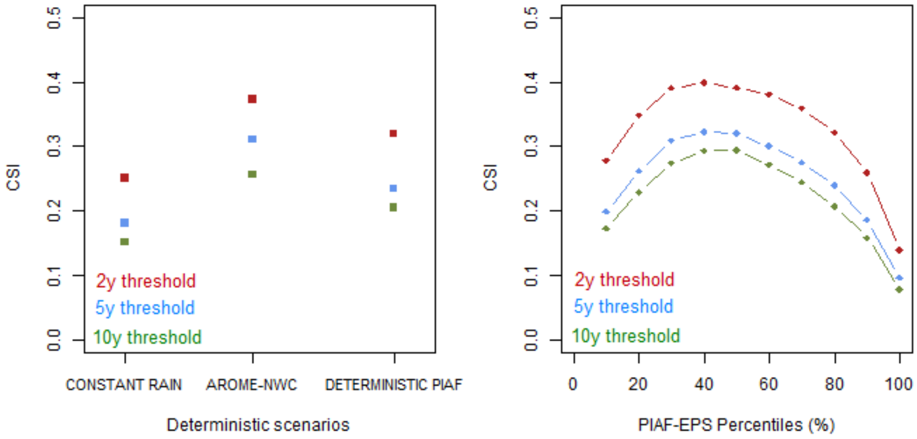

Figure 10. Averaged CSI scores calculated for the eight flood events

Finally, the averaged CSI values obtained with the intermediate percentiles (40%-50%) of PIAF-EPS are in the 0.3-0.4 range, depending on the return period of the discharge threshold considered. The CSI scores can reach up to 0.6 for some specific events. These CSI values may appear relatively low in the perspective of operational decision making. However, the real added value of these forecasts for decision making can only be evaluated by considering the balance between the gains associated with the hits, and the costs related to false alarms. Moreover, other studies dealing with flash-flood nowcasting found similar

CSI values of 0.20 (Clark et al., 2014) and 0.38 (Gourley et al., 2017), even though these CSI values were obtained in very different contexts (using observations over the whole U.S.), and thus cannot be directly compared with the values of this paper.

4.4 Anticipation lead times

The CSI scores presented above~~methodology used in this work can~~ assess the ability of the rainfall products to predict discharge threshold exceedances, regardless of anticipation. However, maximizing the anticipation times of good forecasts is another desirable property in an operational forecasting context.~~Maximizing the anticipation times of good forecasts is another desirable property in an operational forecasting context, but anticipation has not been evaluated in our work.~~ An estimation of the anticipation times associated with the HITs in contingency tables was proposed~~involved in the contingency tables values was attempted~~ by Charpentier-Noyer et al. (2023) , by computing the difference $t_{sim} - t_{run}$, where t_{sim} represents the first threshold exceedance by the reference simulation, and t_{run} corresponds to the starting time of the first forecast that identifies this threshold exceedance event. Anticipation times for each QPF (50th percentile for PIAF-EPS, as intermediate percentiles where identified as optimal in the previous section) and for the 5-year threshold are presented in figure 11. Firstly, the results show that the anticipation times can reach up to 6 hours. This is due to the choice of counting a HIT when t_{sim} falls within the interval $]t_{run}; t_{run} + T + 3h]$ (see appendix B2), T being the forecast range ($0 < T \leq 3$ hours). Anticipation times exceeding the forecast length of 3 hours, even if helpful in anticipating threshold exceedances, result from unrealistic forecasts where the threshold crossing is forecasted too early. It is thus logical to observe that the constant rain scenario has the highest number of anticipation times exceeding 3 hours, and it is rather satisfying to note that PIAF-EPS has the least occurrences in this anticipation range. The comparison of histograms in the 0-3h range of anticipation times confirms that PIAF and PIAF-EPS yield a larger number of HITs globally. Additionally, it shows that this increase of HITs is primarily obtained in the 0-2h range of anticipation times compared to AROME-NWC. This logic is clear as it corresponds to the forecast range where radar extrapolations are involved in building PIAF and PIAF-EPS. Furthermore, it suggests that PIAF-EPS brings additional HITs mainly in the 0-1h range of anticipation times when compared to PIAF. However, drawing systematic conclusions is complicated, as we are only examining one ensemble percentile here, and we are considering all events, while important differences may exist within each event.~~. Unfortunately, our present use of a forecast-based stratification problem following Bellier et al. (2017), would make it more difficult to investigate the notion of anticipation. As explained in annex A.2, a correct detection is counted if the moment t_{sim} of first threshold occurrence by the reference simulation (SMASH model forced by PANTHERE QPE) is in the interval $]t_{run}; t_{ldt} + 3h]$, (t_{run}, t_{ldt}) being respectively the start time and the lead time of the first forecast that identifies a threshold exceedance. The anticipation time is defined as $t_{sim} - t_{run}$, thus it is possible to obtain anticipation times larger than three hours even though the maximum lead time available in PIAF forecasts is three hours. The notion of anticipation is therefore tricky to define, because of our choice of the interval $]t_{run}; t_{ldt} + 3h]$ which was intended to introduce some tolerance on the time window considered for counting the correct detections.~~

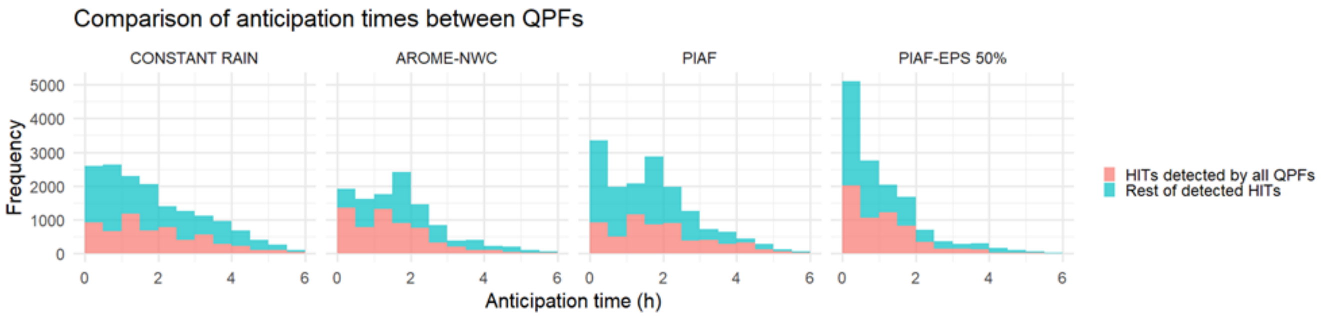

Figure 11. Anticipation times aggregated for all events and for each QPF (50th percentile for PIAF-EPS), for the 5-year threshold

5 Conclusions

The development of efficient tools and methods for flash-flood forecasting is of crucial importance to limit the often catastrophic consequences of flood hazards. The objective verification of newly developed forecasting methods and products is a key step before their integration into operational forecasting suites. In the current study, the potential of the experimental PIAF-EPS short-range ensemble rainfall product for flash flood forecasting purposes has been assessed. Eight heavy precipitation events that occurred between 2019 and 2021 in the South-Eastern region of France were reanalysed using a hydrological forecasting suite similar to the one that is currently operational in the French national flash flood warning system, Vigicrues Flash. An original verification process, directly derived from Charpentier-Noyer et al. (2023), was performed on each $1km^2$ pixel of the area. This allowed us to plot maps to precisely visualize the forecasts performance, and to summarize it as CSI scores.

The hydrological forecasts based on PIAF-EPS have been compared to those obtained with deterministic PIAF and AROME-NWC rainfall forecasts, since PIAF-EPS is directly obtained from these two deterministic products. A naive constant rainfall scenario was also used as a reference. The results showed that PIAF-EPS systematically outperformed the constant rainfall and the deterministic PIAF forecasts. As indicated in previous studies, it was also observed that PIAF does not always outperform AROME-NWC, because the forecast quality depends on the lead time and on the performance of radar QPE extrapolations. Over the eight events considered in this study, it was observed that the PIAF-EPS performance is generally similar to, or better than AROME-NWC.

In a nutshell, the results obtained confirm the added value of using the PIAF-EPS products for anticipating flash floods in the Mediterranean area. We argue that statistical scores such as the CSI provide valuable indications of performance despite not being applied on long data series, but on only eight particularly intense flash-floods. Indeed, when assessing the performance of such a new forecasting product, it is essential to carefully check its behaviour on some high-impact events, as a complement to more generic statistical evaluations. The results presented here should nevertheless be complemented with more robust statistical evaluations over longer periods of time and on a larger number of high precipitation events, bringing a more generic overview of the quality of the forecast ensembles.~~We recommend that the results presented here should be confirmed with more robust statistical evaluations over longer periods of time and on a large number of high precipitation events. It would also be interesting to extend this study by an assessment of how anticipation lead-times evolve depending on the QPF product considered.~~

405 ## Appendix A: Hydrological presentation of the studied events

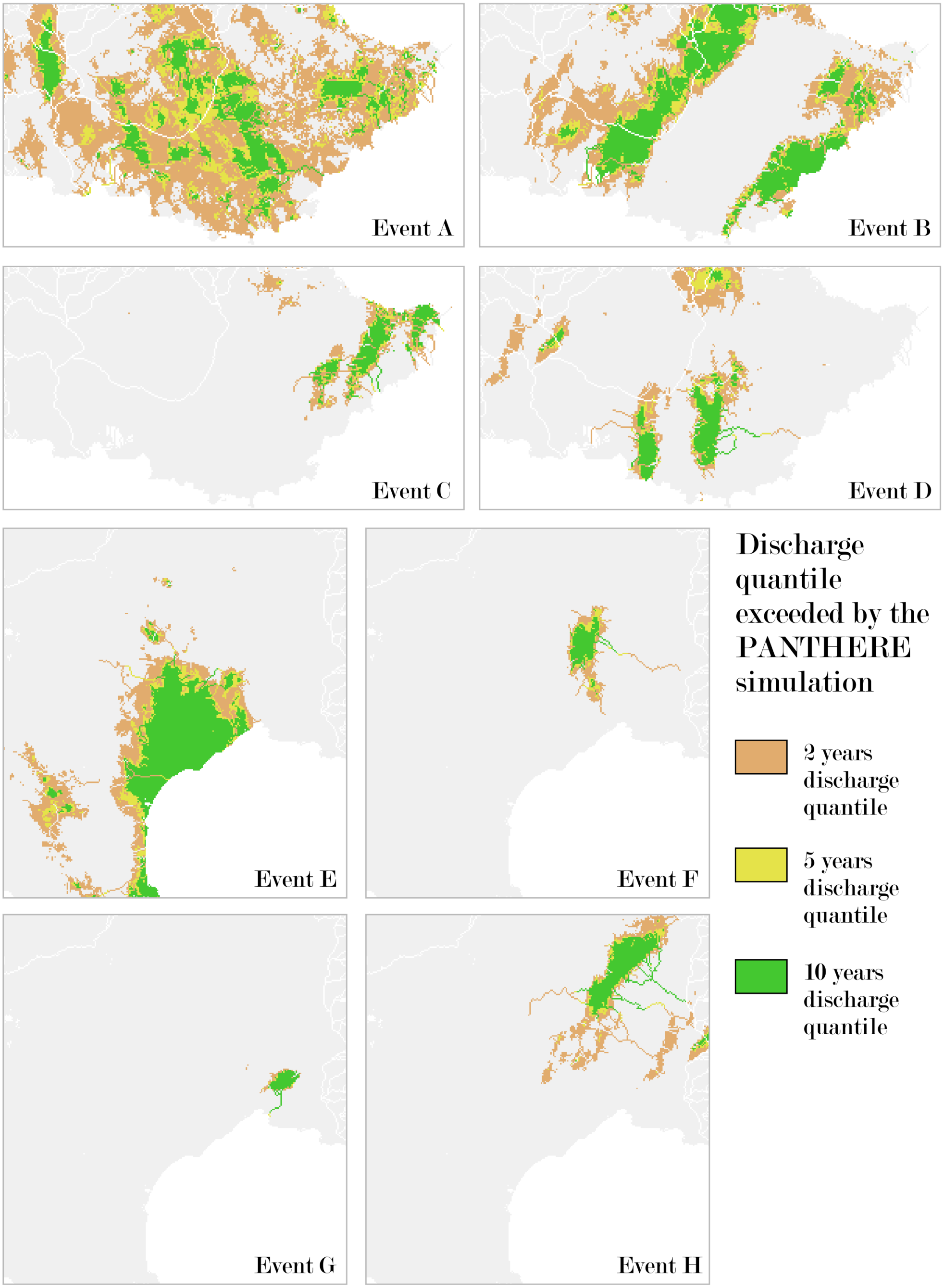
Event A
Event B
Event C
Event D
Event E
Event F
Event G
Event H
Discharge quantile exceeded by the PANTHERE simulation
2 years discharge quantile
5 years discharge quantile
10 years discharge quantile

Figure A1. Discharge quantiles exceeded by the PANTHERE simulation Q_{sim} during each event

Appendix B: Methodology for filling the contingency tables

The contingency tables are filled for each pixel of the zones. For an ensemble forecast product, each percentile is considered separately, to be treated as a deterministic product. For each pixel, we look at the hydrograph $Q_{sim}(t)$ simulated by SMASH with the PANTHERE QPE as input, and at the hydrographs $Q_{for}(t)$ forecasted by SMASH when forced with a QPF product (constant rain scenario, AROME-NWC, PIAF or PIAF-EPS percentiles). Let t_{run} be the forecast start time, let T be the forecast range ($0 < T \leq 3h$), and let Q_t be the considered discharge threshold.

B1 Detailed method of Charpentier-Noyer et al. (2023) : observation-based stratification

The method, applied on each spatial entity and for each forecast probability, is as follows:

- If there exists t such that $Q_{sim}(t) \geq Q_t$, then the date t_{sim} corresponding to the first threshold exceedance by Q_{sim} is selected. A sample S consisting of all the pairs (t_{run}, T) such that $t_{run} < t_{sim} \leq t_{run} + T$ is constructed.
  - If there exists $(t_{run}, T) \in S$ such that $Q_{for}(t_{run}, T) \geq Q_t$, a HIT is counted. The t_{run} corresponding to the first threshold exceedance is selected to calculate the anticipation time : $t_{sim} - t_{run}$.
  - If $Q_{for}(t_{run}, T) < Q_t \quad \forall (t_{run}, T) \in S$, a MISS is counted.
- If $Q_{sim}(t) < Q_t \quad \forall t$, the date t_{sim} corresponding to the global maximum of Q_{sim} is selected and a sample S consisting of all the pairs (t_{run}, T) such that $t_{run} < t_{sim} \leq t_{run} + T$ is constructed.
  - If there is $(t_{run}, T) \in S$ such that $Q_{for}(t_{run}, T) \geq Q_t$, a FALSE ALARM is counted.
  - If $Q_{for}(t_{run}, T) < Q_t \quad \forall (t_{run}, T) \in E$, a CORRECT REJECTION is counted.

B2 Adapted method : forecast-based stratification

The detailed process used to build the contingency table is as follows:

- If there exists (t_{run}, T) such that $Q_{for}(t_{run}, T) \geq Q_t$, the pair (t_{run}, T) corresponding to the first threshold exceedance by Q_{for} is selected.
  - If $Q_{sim}(t) < Q_t \quad \forall t$, a FALSE ALARM is counted. See figure B1a.
  - If there is t such that $Q_{sim}(t) \geq Q_t$, the first threshold exceedance occurring at t_{sim} :
    - If $t_{sim} \in]t_{run}; t_{run} + T + 3h]$, a HIT is counted. The anticipation time is equal to $t_{sim} - t_{run}$. The time interval $]t_{run}; t_{run} + T + 3h]$, and particularly the $+3h$ part, was chosen in order to create a tolerant window for the HIT counting. It seemed coherent to use a time tolerance equal to the maximum lead time. See figure B1b.
    - If $t_{sim} > t_{run} + T + 3h$, a FALSE ALARM is counted. See figure B1c.

- If $t_{sim} \leq t_{run}$, a MISS is counted. See figure B1d.

– If $Q_{for}(t_{run}, T) < Q_t \quad \forall (t_{run}, T)$:

  – If $Q_{sim}(t) < Q_t \quad \forall t$, a CORRECT REJECTION is counted. See figure B1e.

  – If there exists t such that $Q_{sim}(t) \geq Q_t$, the first threshold exceedance occurring at t_{sim}, a MISS is counted. See figure B1f.

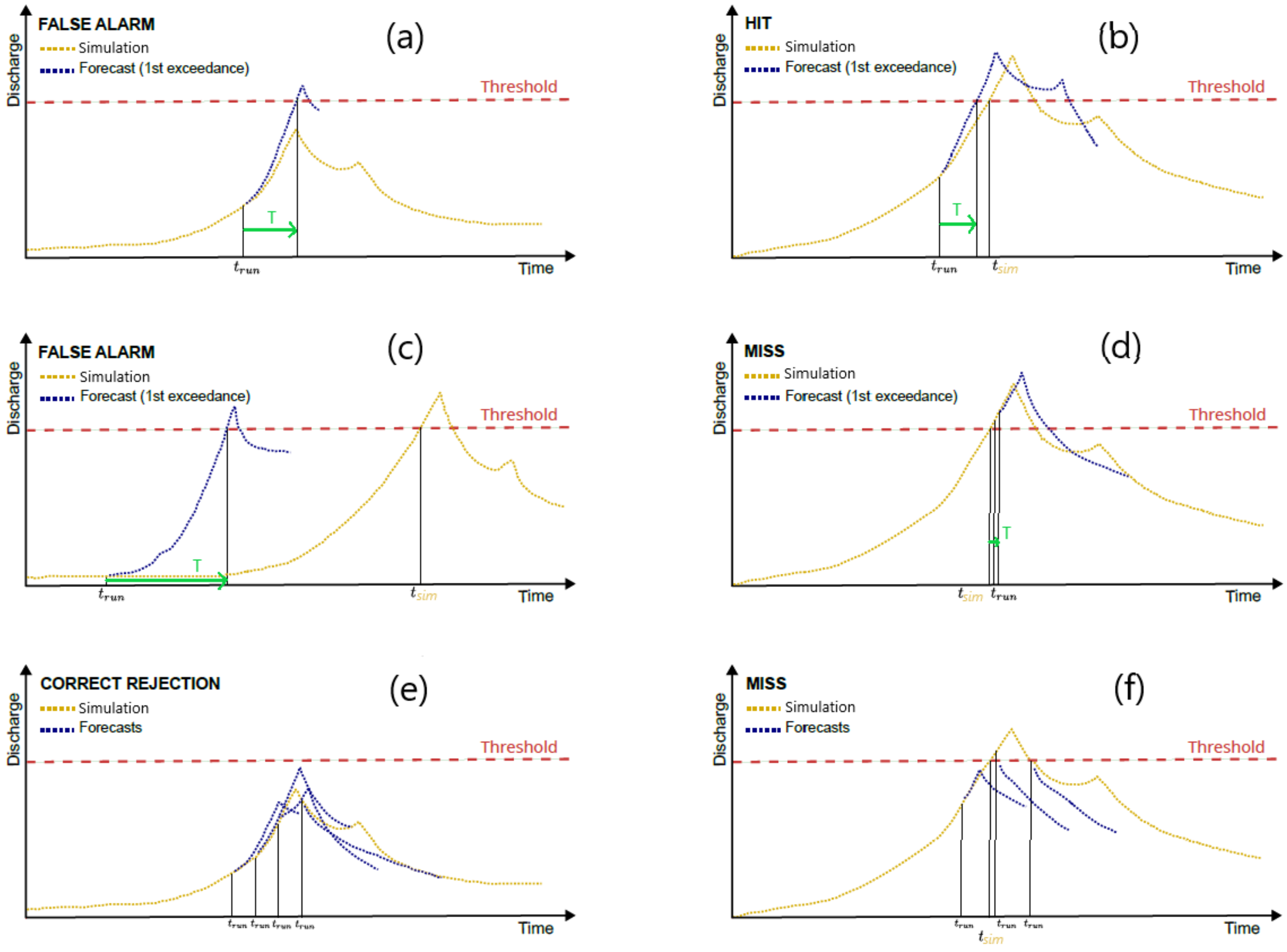

Figure B1. Six possible cases in the new methodology (inspired by Charpentier-Noyer et al., 2022): (a) forecasted threshold exceedance not present in the simulated hydrograph (False Alarm), (b) threshold exceedance correctly forecasted (Hit), (c) threshold exceedance anticipated but anticipation largely exceeding the forecast lead time (false alarm), (d) threshold exceedance detected by one forecast, but right after the simulation (miss), (e) absence of threshold exceedance both in the simulation the forecasts (correct rejection), (f) threshold exceedance undetected by all the forecasts (miss)

Appendix C: Rank diagrams

Rank diagrams, also called Talagrand diagrams (Candille and Talagrand, 2005), are one of the most common tools for assessing the reliability of meteorological ensemble forecasts. The general idea of this tool is to count the number of times the observation value is included in a given interval of the ensemble forecast quantiles. As a consequence, if the observation value is often close to low quantiles, it means that the forecast model has a tendency to overestimate. On the other hand, if the observation is more often close to high quantiles, then the model tends to underestimate. Logically, a perfect diagram would be perfectly flat, which would mean that the observation is uniformly distributed among the ensemble forecast quantiles. However this never happens in reality. The rank diagram is useful to quickly detect biases in an ensemble forecast. It can detect positive or negative biases, but also under and over dispersion of the ensemble forecasts. Traditionally, the rank diagram is applied to rainfall ensemble forecasts. However, in this study it was applied to discharge ensemble forecasts, at each pixel of the SMASH model computation grid. The rank diagrams presented here combine all the forecasts issued during each event, and are computed for a fixed lead time (one hour).

In order to distinguish the roles of high and low discharges in the rank diagram form, it was decided to build separate rank diagrams for each category. However it is necessary to take precautions concerning the criteria that distinguish those categories. Indeed, Bellier et al. (2017) have shown that a sample stratification based on the observations can introduce bias. A sample stratification based on forecasts is recommended in most of the cases. Therefore, the following categories were chosen:

- $Q_{med} \leq \frac{1}{2} Q_{T=2y}$ for low discharges.
- $\frac{1}{2} Q_{T=2y} < Q_{med} \leq Q_{T=2y}$ for medium discharges.
- $Q_{med} > Q_{T=2y}$ for high discharges.

Where Q_{med} is the median discharge of the hydrological ensemble forecasts and $Q_{T=2y}$ is the 2-year return period quantile, according to the historical run of the SMASH model.

Note that even if based on forecast discharges, this stratification can still cause bias: when only the areas and time steps with high forecast discharges are considered, the overall probability that the considered forecasts exceed the observed discharges tends logically to be higher, and conversely. However, this stratification effect does not affect the global rank diagrams.

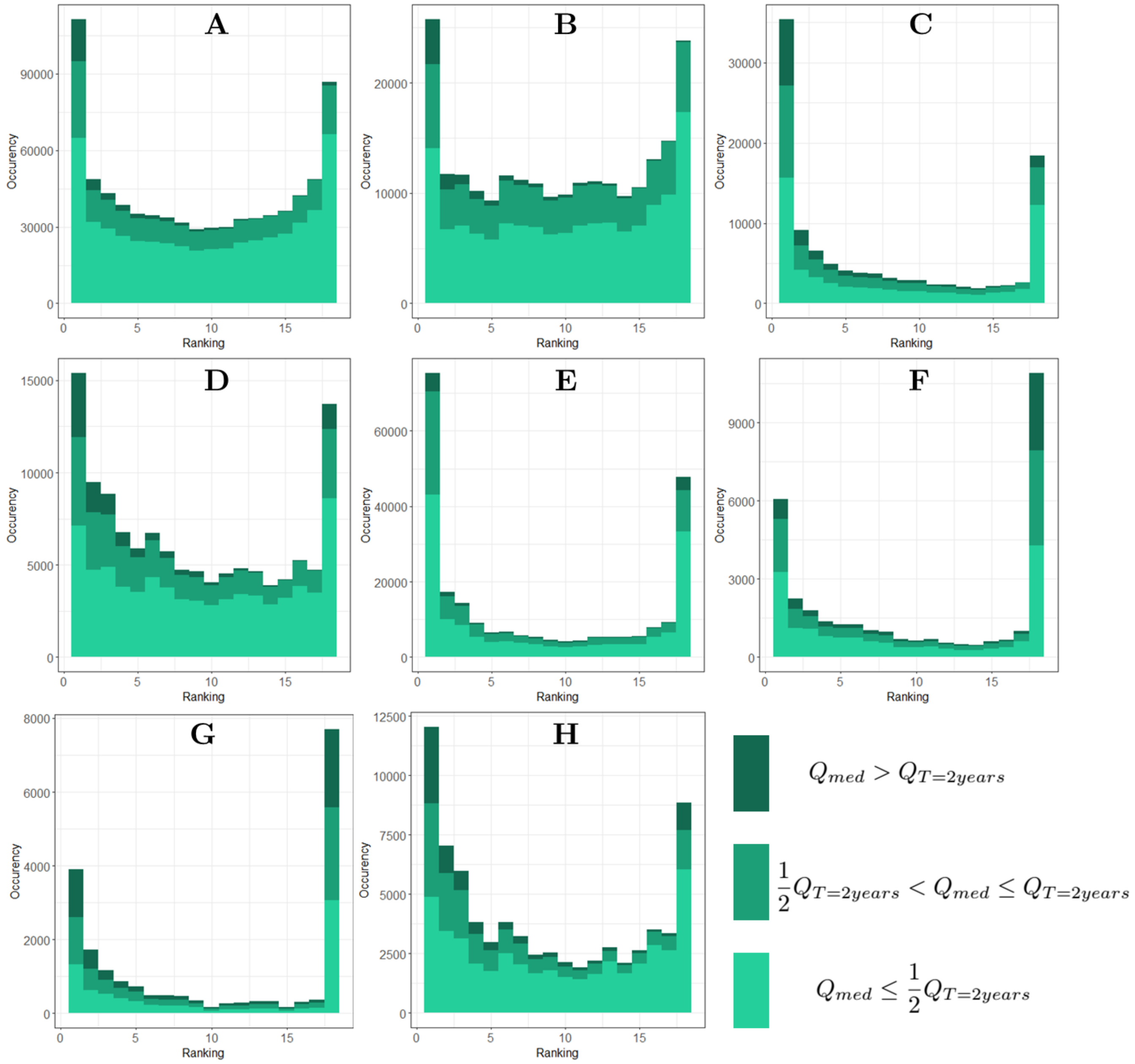

Figure C1. Rank diagrams obtained for each event, 1hour-lead-time

Appendix D: Maps of contingency tables for each event

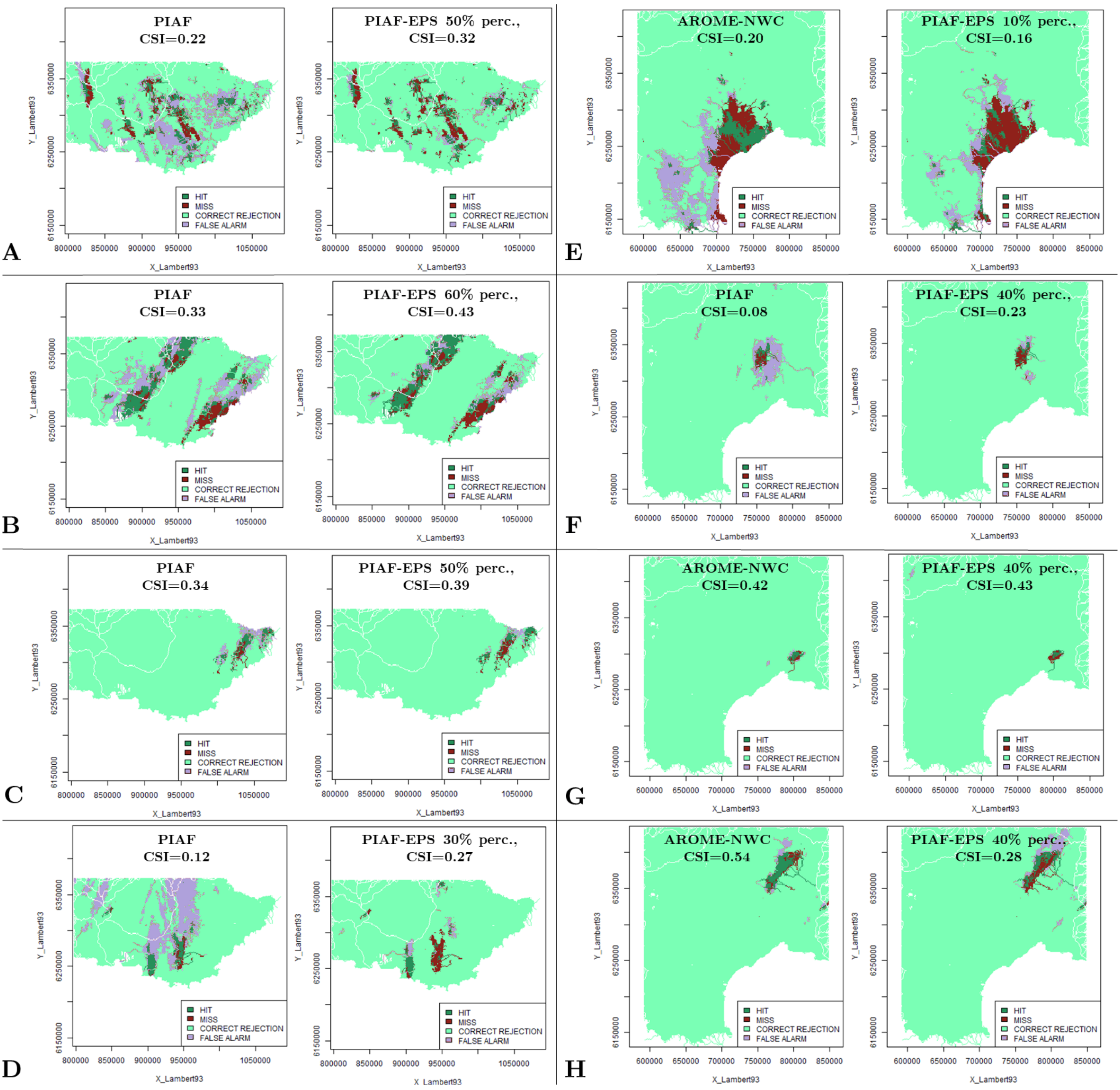

Figure D1. Best deterministic forecast *(left)* and best PIAF-EPS percentile *(right)* in terms of CSI, for each event, $Q_t = Q_{T=10y}$

Data availability. All hydrological data are provided in open access on the french platform data.gouv.fr (https://doi.org/10.57745/IHKGRE): discharges obtained from SMASH model forced with PANTHERE QPE, PIAF, PIAF-EPS, AROME-NWC and constant future rain, for each of the eight event. The discharge quantiles corresponding to the 2, 5 and 10-year return periods are also provided.

Author contributions. The production and analysis of the results were performed by JG, under the supervision of PJ, OP and FB. The paper was written by JG and OP, except for the parts about PIAF and PIAF-EPS which were written by FB. All co-authors replied to the reviewers' comments.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. Rainfall observations and forecasts data were provided by Météo-France. The authors would like to thank the editor and the two anonymous referees for their constructive comments which helped to improve the quality of the paper.

Financial support. This research was performed within the framework of the MUFFINS project (ANR-21-CE04-0021-01).

References

Alfieri, L. and Thielen, J.: A European precipitation index for extreme rain-storm and flash flood early warning, Meteorological Applications, 22, 3–13, https://doi.org/10.1002/met.1328, publisher: Wiley, 2012.

Amengual, A., Carrió, D. S., Ravazzani, G., and Homar, V.: A Comparison of Ensemble Strategies for Flash Flood Forecasting: The 12 October 2007 Case Study in Valencia, Spain, Journal of Hydrometeorology, 18, 1143–1166, https://doi.org/10.1175/JHM-D-16-0281.1, publisher: American Meteorological Society, 2017.

Amengual, A., Hermoso, A., Carrió, D. S., and Homar, V.: The sequence of heavy precipitation and flash flooding of 12 and 13 September 2019 in eastern Spain. Part II: A hydro-meteorological predictability analysis based on convection-permitting ensemble strategies, Journal of Hydrometeorology, https://doi.org/10.1175/jhm-d-20-0181.1, publisher: American Meteorological Society, 2021.

Armon, M., Marra, F., Enzel, Y., Rostkier-Edelstein, D., and Morin, E.: Radar-based characterisation of heavy precipitation in the eastern Mediterranean and its representation in a convection-permitting model, Hydrology and Earth System Sciences, 24, 1227–1249, https://doi.org/10.5194/hess-24-1227-2020, publisher: Copernicus GmbH, 2020.

Auer, P., Cesa-Bianchi, N., and Gentile, C.: Adaptive and Self-Confident On-Line Learning Algorithms, Journal of Computer and System Sciences, 64, 48–75, https://doi.org/10.1006/jcss.2001.1795, 2002.

Auger, L., Dupont, O., Hagelin, S., Brousseau, P., and Brovelli, P.: AROME-NWC: A new nowcasting tool based on an operational mesoscale forecasting system, Quarterly Journal of the Royal Meteorological Society, 141, 1603–1611, https://doi.org/10.1002/qj.2463, 2015.

Bellier, J., Zin, I., and Bontron, G.: Sample Stratification in Verification of Ensemble Forecasts of Continuous Scalar Variables: Potential Benefits and Pitfalls, Monthly Weather Review, 145, 3529–3544, https://doi.org/10.1175/MWR-D-16-0487.1, aDS Bibcode: 2017MWRv..145.3529B, 2017.

Benjamin, S. G., Weygandt, S. S., Brown, J. M., Hu, M., Alexander, C. R., Smirnova, T. G., Olson, J. B., James, E. P., Dowell, D. C., Grell, G. A., Lin, H., Peckham, S. E., Smith, T. L., Moninger, W. R., Kenyon, J. S., and Manikin, G. S.: A North American Hourly Assimilation and Model Forecast Cycle: The Rapid Refresh, Monthly Weather Review, 144, 1669 – 1694, https://doi.org/https://doi.org/10.1175/MWR-D-15-0242.1, 2016.

Berenguer, M., Sempere-Torres, D., and Pegram, G. G. S.: SBMcast – An ensemble nowcasting technique to assess the uncertainty in rainfall forecasts by Lagrangian extrapolation, Journal of Hydrology, 404, 226–240, https://doi.org/10.1016/j.jhydrol.2011.04.033PANIST, publisher: Elsevier BV, 2011.

Bouttier, F. and Raynaud, L.: Clustering and selection of boundary conditions for limited-area ensemble prediction, Quarterly Journal of the Royal Meteorological Society, 144, 2381–2391, https://doi.org/10.1002/qj.3304, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3304, 2018.

Bowler, N. E., Pierce, C. E., and Seed, A. W.: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP, Quarterly Journal of the Royal Meteorological Society, 132, 2127–2155, https://doi.org/10.1256/qj.04.100, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1256/qj.04.100, 2006.

Brigode, P., Vigoureux, S., Delestre, O., Nicolle, P., Payrastre, O., Dreyfus, R., Nomis, S., and Salvan, L.: French Riviera floods: hydrometeorological comparison of 2015 and 2019 extremes events, LHB-HYDROSCIENCE JOURNAL, 107, https://doi.org/10.1080/27678490.2021.1976600, 2021.

Candille, G. and Talagrand, O.: Evaluation of probabilistic prediction systems for a scalar variable, Quarterly Journal of the Royal Meteorological Society, 131, 2131–2150, https://doi.org/10.1256/qj.04.71, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1256/qj.04.71, 2005.

Charpentier-Noyer, M., Peredo, D., Fleury, A., Marchal, H., Bouttier, F., Gaume, E., Nicolle, P., Payrastre, O., and Ramos, M.-H.: A methodological framework for the evaluation of short-range flash-flood hydrometeorological forecasts at the event scale, Natural Hazards and Earth System Sciences, 23, 2001–2029, https://doi.org/10.5194/nhess-23-2001-2023, publisher: Copernicus GmbH, 2023.

Clark, P., Roberts, N., Lean, H., Ballard, S. P., and Charlton-Perez, C.: Convection-permitting models: a step-change in rainfall forecasting, Meteorological Applications, 23, 165–181, https://doi.org/10.1002/met.1538, publisher: Wiley, 2016.

Clark, R. A., Gourley, J. J., Flamig, Z. L., Hong, Y., and Clark, E.: CONUS-Wide Evaluation of National Weather Service Flash Flood Guidance Products, Weather and Forecasting, 29, 377–392, https://doi.org/10.1175/WAF-D-12-00124.1, publisher: American Meteorological Society Section: Weather and Forecasting, 2014.

Collier, C. G.: Flash flood forecasting: What are the limits of predictability?, Quarterly Journal of the Royal Meteorological Society, 133, 3–23, https://doi.org/10.1002/qj.29, publisher: Wiley, 2007.

Corral, C., Berenguer, M., Sempere-Torres, D., Poletti, L., Silvestro, F., and Rebora, N.: Comparison of two early warning systems for regional flash flood hazard forecasting, Journal of Hydrology, 572, 603–619, https://doi.org/https://doi.org/10.1016/j.jhydrol.2019.03.026, 2019.

Davolio, S., Miglietta, M. M., Diomede, T., Marsigli, C., and Montani, A.: A flood episode in northern Italy: multi-model and single-model mesoscale meteorological ensembles for hydrological predictions, Hydrology and Earth System Sciences, 17, 2107–2120, https://doi.org/10.5194/hess-17-2107-2013, publisher: Copernicus GmbH, 2013.

Davolio, S., Silvestro, F., and Malguzzi, P.: Effects of Increasing Horizontal Resolution in a Convection-Permitting Model on Flood Forecasting: The 2011 Dramatic Events in Liguria, Italy, Journal of Hydrometeorology, 16, 1843–1856, https://doi.org/10.1175/JHM-D-14-0094.1, publisher: American Meteorological Society, 2015.

Davolio, S., Silvestro, F., and Gastaldo, T.: Impact of Rainfall Assimilation on High-Resolution Hydrometeorological Forecasts over Liguria, Italy, Journal of Hydrometeorology, 18, 2659–2680, https://doi.org/10.1175/JHM-D-17-0073.1, publisher: American Meteorological Society Section: Journal of Hydrometeorology, 2017.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., and Cébron, P.: PEARP, the Météo-France short-range ensemble prediction system, Quarterly Journal of the Royal Meteorological Society, 141, 1671–1685, https://doi.org/10.1002/qj.2469, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.2469, 2015.

Devaine, M., Gaillard, P., Goude, Y., and Stoltz, G.: Forecasting electricity consumption by aggregating specialized experts, Machine Learning, 90, 231–260, https://doi.org/10.1007/s10994-012-5314-7, 2013.

Furnari, L., Mendicino, G., and Senatore, A.: Hydrometeorological Ensemble Forecast of a Highly Localized Convective Event in the Mediterranean, Water, 12, 1545, https://doi.org/10.3390/w12061545, publisher: MDPI AG, 2020.

Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbuc, M., Bateman, A., Blaškovičová, L., Blöschl, G., Borga, M., Dumitrescu, A., Daliakopoulos, I., Garcia, J., Irimescu, A., Kohnova, S., Koutroulis, A., Marchi, L., Matreata, S., Medina, V., Preciso, E., Sempere-Torres, D., Stancalie, G., Szolgay, J., Tsanis, I., Velasco, D., and Viglione, A.: A compilation of data on European flash floods, Journal of Hydrology, 367, 70–78, https://doi.org/10.1016/j.jhydrol.2008.12.028, 2009.

Gourley, J. J., Flamig, Z. L., Vergara, H., Kirstetter, P.-E., Clark, R. A., Argyle, E., Arthur, A., Martinaitis, S., Terti, G., Erlingis, J. M., Hong, Y., and Howard, K. W.: The FLASH Project: Improving the Tools for Flash Flood Monitoring and Prediction across the United

States, Bulletin of the American Meteorological Society, 98, 361–372, https://doi.org/10.1175/BAMS-D-15-00247.1, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 2017.

Hally, A., Caumont, O., Garrote, L., Richard, E., Weerts, A., Delogu, F., Fiori, E., Rebora, N., Parodi, A., Mihalović, A., Ivković, M., Dekić, L., Verseveld, W. v., Nuissier, O., Ducrocq, V., D'Agostino, D., Galizia, A., Danovaro, E., and Clematis, A.: Hydrometeorological multi-model ensemble simulations of the 4 November 2011 flash flood event in Genoa, Italy, in the framework of the DRIHM project, Natural Hazards and Earth System Sciences, 15, 537–555, https://doi.org/10.5194/nhess-15-537-2015, publisher: Copernicus GmbH, 2015.

Hapuarachchi, H. A. P., Wang, Q. J., and Pagano, T. C.: A review of advances in flash flood forecasting, Hydrological Processes, 25, 2771–2784, https://doi.org/10.1002/hyp.8040, publisher: Wiley, 2011.

Imhoff, R. O., Brauer, C. C., van Heeringen, K. J., Uijlenhoet, R., and Weerts, A. H.: Large-Sample Evaluation of Radar Rainfall Nowcasting for Flood Early Warning, Water Resources Research, 58, e2021WR031 591, https://doi.org/10.1029/2021WR031591, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021WR031591, 2022.

Javelle, P., Organde, D., Demargne, J., Saint-Martin, C., Saint-Aubin, C. d., Garandeau, L., and Janet, B.: Setting up a French national flash flood warning system for ungauged catchments based on the AIGA method, E3S Web of Conferences, 7, 18 010, https://doi.org/10.1051/e3sconf/20160718010, publisher: EDP Sciences, 2016.

Jay-Allemand, M.: Estimation variationnelle des paramètres d'un modèle hydrologique distribué, These de doctorat, Aix-Marseille, https://www.theses.fr/2020AIXM0400, 2020.

Jay-Allemand, M., Javelle, P., Gejadze, I., Arnaud, P., Malaterre, P.-O., Fine, J.-A., and Organde, D.: On the potential of variational calibration for a fully distributed hydrological model: application on a Mediterranean catchment, Hydrology and Earth System Sciences, 24, 5519–5538, https://doi.org/10.5194/hess-24-5519-2020, publisher: European Geosciences Union, 2020.

Lagasio, M., Silvestro, F., Campo, L., and Parodi, A.: Predictive Capability of a High-Resolution Hydrometeorological Forecasting Framework Coupling WRF Cycling 3DVAR and Continuum, Journal of Hydrometeorology, 20, 1307–1337, https://doi.org/10.1175/JHM-D-18-0219.1, publisher: American Meteorological Society Section: Journal of Hydrometeorology, 2019.

Laurentin: Hourly rainfall analysis merging radar and rain gauge data, pp. 2–8, 2008.

Lebouc, L. and Payrastre, O.: Reconstitution des débits de pointe des crues des 23-24 novembre et 1er décembre 2019 dans les départements du Var et les Alpes-Maritimes, Research Report, IFSTTAR - Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux, https://hal.archives-ouvertes.fr/hal-02933695, 2020.

Lovat, A., Vincendon, B., and Ducrocq, V.: Hydrometeorological evaluation of two nowcasting systems for Mediterranean heavy precipitation events with operational considerations, Hydrology and Earth System Sciences, 26, 2697–2714, https://doi.org/10.5194/hess-26-2697-2022, publisher: Copernicus GmbH, 2022.

Mandapaka, P. V., Germann, U., Panziera, L., and Hering, A.: Can Lagrangian Extrapolation of Radar Fields Be Used for Precipitation Nowcasting over Complex Alpine Orography?, Weather and Forecasting, 27, 28–49, https://doi.org/10.1175/WAF-D-11-00050.1, publisher: American Meteorological Society Section: Weather and Forecasting, 2012.

Mason, I.: A Model for Assessment of Weather Forecasts, Australian Meteorological Magazine, 30, 291–303, 1982.

Moisselin, J.-M., Cau, P., Jauffret, C., Bouissières, I., and Tzanos, R.: Seamless approach for precipitations within the 0-3 hours forecast-interval, http://hdl.handle.net/20.500.11765/10588, accepted: 2019-05-22T10:44:08Z Publisher: Agencia Estatal de Meteorología, 2019.

Nuissier, O., Marsigli, C., Vincendon, B., Hally, A., Bouttier, F., Montani, A., and Paccagnella, T.: Evaluation of two convection-permitting ensemble systems in the HyMeX Special Observation Period (SOP1) framework, Quarterly Journal of the Royal Meteorological Society, 142, 404–418, https://doi.org/10.1002/qj.2859, publisher: Wiley, 2016.

Osinski, R. and Bouttier, F.: Short-range probabilistic forecasting of convective risks for aviation based on a lagged-average-forecast ensemble approach, Meteorological Applications, 25, 105–118, https://doi.org/10.1002/met.1674, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/met.1674, 2018.

Payrastre, O., Nicolle, P., Bonnifait, L., Brigode, P., Astagneau, P., Baise, A., Belleville, A., Bouamara, N., Bourgin, F., Breil, P., Brunet, P., Cerbelaud, A., Courapied, F., Devreux, L., Dreyfus, R., Gaume, E., Nomis, S., Poggio, J., Pons, F., Rabab, Y., and Sevrez, D.: Tempête Alex du 2 octobre 2020 dans les Alpes-Maritimes : une contribution de la communauté scientifique à l'estimation des débits de pointe des crues, LHB Hydroscience Journal, p. 2082891, https://doi.org/10.1080/27678490.2022.2082891, 2022.

Peredo, D., Ramos, M.-H., Marchal, H., and Bouttier, F.: Challenges of event-based evaluation of flash floods: example of the October 2018 flood event in the Aude catchment in France., https://events.ecmwf.int/event/222/contributions/2255/attachments/1291/2358/Hydrological-WS-Peredo.pdf, 2021.

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

Piotte, O., Montmerle, T., Fouchier, C., Belleudy, A., Garandeau, L., Janet, B., Jauffret, C., Demargne, J., and Organde, D.: Les évolutions du service d'avertissement sur les pluies intenses et les crues soudaines en France, La Houille Blanche, pp. 75–84, https://doi.org/10.1051/lhb/2020055, number: 6 Publisher: EDP Sciences, 2020.

Poletti, M. L., Silvestro, F., Davolio, S., Pignone, F., and Rebora, N.: Using nowcasting technique and data assimilation in a meteorological model to improve very short range hydrological forecasts, Hydrology and Earth System Sciences, 23, 3823–3841, https://doi.org/10.5194/hess-23-3823-2019, publisher: Copernicus GmbH, 2019.

Raynaud, D., Thielen, J., Salamon, P., Burek, P., Anquetin, S., and Alfieri, L.: A dynamic runoff co-efficient to improve flash flood early warning in Europe: evaluation on the 2013 central European floods in Germany, Meteorological Applications, 22, 410–418, https://doi.org/https://doi.org/10.1002/met.1469, 2015.

Ribes, A., Thao, S., Vautard, R., Dubuisson, B., Somot, S., Colin, J., Planton, S., and Soubeyroux, J.-M.: Observed increase in extreme daily rainfall in the French Mediterranean, Climate Dynamics, 52, 1095–1114, https://doi.org/10.1007/s00382-018-4179-2, 2019.

Roberts, N. M. and Lean, H. W.: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events, Monthly Weather Review, 136, 78 – 97, https://doi.org/https://doi.org/10.1175/2007MWR2123.1, 2008.

Sayama, T., Yamada, M., Sugawara, Y., and Yamazaki, D.: Ensemble flash flood predictions using a high-resolution nationwide distributed rainfall-runoff model: case study of the heavy rain event of July 2018 and Typhoon Hagibis in 2019, Progress in Earth and Planetary Science, 7, https://doi.org/10.1186/s40645-020-00391-7, publisher: Springer Science and Business Media LLC, 2020.

Schaefer, J. T.: The Critical Success Index as an Indicator of Warning Skill, Weather and Forecasting, 5, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2, publisher: American Meteorological Society Section: Weather and Forecasting, 1990.

Scheufele, K., Kober, K., Craig, G. C., and Keil, C.: Combining probabilistic precipitation forecasts from a nowcasting technique with a time-lagged ensemble, Meteorological Applications, 21, 230–240, https://doi.org/10.1002/met.1381, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/met.1381, 2014.

Seed, A. W., Pierce, C. E., and Norman, K.: Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme, Water Resources Research, 49, 6624–6641, https://doi.org/10.1002/wrcr.20536, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wrcr.20536, 2013.

Silvestro, F. and Rebora, N.: Operational verification of a framework for the probabilistic nowcasting of river discharge in small and medium size basins, Natural Hazards and Earth System Sciences, 12, 763–776, https://doi.org/10.5194/nhess-12-763-2012, publisher: Copernicus GmbH, 2012.

Silvestro, F., Rebora, N., and Ferraris, L.: Quantitative Flood Forecasting on Small- and Medium-Sized Basins: A Probabilistic Approach for Operational Purposes, Journal of Hydrometeorology, 12, 1432–1446, https://doi.org/https://doi.org/10.1175/JHM-D-10-05022.1, publisher: American Meteorological Society, 2011.

Tabary, P., Augros, C., Champeaux, J.-L., Chèze, J.-L., Faure, D., Idziorek, D., Lorandel, R., Urban, B., and Vogt, V.: Le réseau et les produits radars de Météo-France, La Météorologie, 8ème série, 15–27, https://doi.org/10.4267/2042/52050, 2013.

Tramblay, Y., Mimeau, L., Neppel, L., Vinet, F., and Sauquet, E.: Detection and attribution of flood trends in Mediterranean basins, Hydrology and Earth System Sciences, 23, 4419–4431, https://doi.org/10.5194/hess-23-4419-2019, publisher: Copernicus GmbH, 2019.

Vincendon, B., Ducrocq, V., Nuissier, O., and Vié, B.: Perturbation of convection-permitting NWP forecasts for flash-flood ensemble forecasting, Natural Hazards and Earth System Sciences, 11, 1529–1544, https://doi.org/10.5194/nhess-11-1529-2011, publisher: Copernicus Publ. / European Geosciences Union, 2011.

Vié, B., Molinié, G., Nuissier, O., Vincendon, B., Ducrocq, V., Bouttier, F., and Richard, E.: Hydro-meteorological evaluation of a convection-permitting ensemble prediction system for Mediterranean heavy precipitating events, Natural Hazards and Earth System Sciences, 12, 2631–2645, https://doi.org/10.5194/nhess-12-2631-2012, publisher: Copernicus GmbH, 2012.

Wang, D., Stapor, P., and Hasenauer, J.: Dirac mixture distributions for the approximation of mixed effects models$**$We acknowledge financial support by the China Scholarship Council (201706060200) and the German Ministry for Education and Research (BMBF) in the FitMultiCell project (grant number 031L0159A)., IFAC-PapersOnLine, 52, 200–206, https://doi.org/10.1016/j.ifacol.2019.12.258, 2019.

WMO: Guidelines on Ensemble Prediction Systems and Forecasting, WMO, WMO, Geneva, 2012.

WMO: Climate and water (2020) - Floods, https://public.wmo.int/en/resources/world-meteorological-day/previous-world-meteorological-days/climate-and-water/floods, 2020.

Zanchetta, A. D. L. and Coulibaly, P.: Recent Advances in Real-Time Pluvial Flash Flood Forecasting, Water, 12, 570, https://doi.org/10.3390/w12020570, number: 2 Publisher: Multidisciplinary Digital Publishing Institute, 2020.