RC1: 'Comment on egusphere-2023-907', Anonymous Referee #1, 23 Jun 2023

General comments:

The research work presented in the paper "Assessing the ability of a new seamless short-range ensemble rainfall product to anticipate flash floods in the French Mediterranean area" aims at the evaluation of the added value of a seamless short range (0-6h) ensemble quantitative precipitation forecast (QPF) product, called PIAF-EPS for flash floods forecasting.

The work shows the results in terms of the skills of the system comparing the PIAF-EPS system with different deterministic QPFs calculated for eight intense rainfall events occurred between 2019 and 2021 in the French Mediterranean region.

The paper is quite-well written and readable even if some adjustment can be made in the order of the paragraph to make it clearer; the abstract and conclusions are satisfactory; the scientific methods and assumptions valid and well outlined. As a general comment I think the work of the paper is interesting from the point of view of the method of the research and for the analysis of the results. There are some parts less clear than other that need improvements and maybe further insight in some sections. Corrections and comments can be found in details in the pdf attached.

⇨ First of all, we thank Referee #1 for this positive evaluation of our work and for the useful comments provided which will help to improve the overall quality of the manuscript. We provide below detailed answers showing how we plan to adapt the manuscript according to these suggestions.

Specific comments:

Line 106: Can you please explain in few words how this procedure works?

⇨ We will add the following explanation line 107 : "*…over large subdomains. In a nutshell, this algorithm produces a smooth transition (as a function of forecast range) 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)*."

Line 114: It is not clear if 6 hours of rainfall forecast are always used as input of the hydrological model for every forecasting model considered

⇨ To clarify this point, line 85 will be modified as follows:
"In this study, the common time step for QPE-QPF and the hydrological model is 15-min. All the QPFs mentioned are available up to 6-hour lead times but refreshment periods depend on the considered product. For the present study, we decided to consider QPFs only for 0-3h lead times and refreshed only each hour. The QPF

products include…"

Fig 5: Why the first event is not A event? and the second is not B?

- ⇨ *The East and West geographical zones were considered separately, for computation times purposes. The East zone was considered first, which is why Event A is the November 2019 event. Events B, C and D chronologically occur on the East zone, and event E is the first event occurring on the West zone. We will add a column "Zone" to table 1 to clarify this choice, and we will stick to a "A-B-C-D-E-F-G-H" order in all the paper.*

Table 1: Maybe it is useful to add another column that specifies the percentage on the total area analysed to evidence the spatial distribution of the event

- ⇨ *Three columns will be added to table 1, corresponding to the percentages of zone Q>QT=Xyears, relatively to the total area analysed, for each threshold.*

Line 251: How CSI is explained is calculated in paragraph 4.3, at least say that, otherwise the readers try to go back to find how it is calculated

- ⇨ The CSI scores presented l249-251 do not refer to the averaged CSI described in section 4.3, but to the CSI calculated on a single event, for a defined threshold and ensemble percentile, as described in section 2.4, line 237. We will add a reference to this section in the text to avoid any confusion: "*For this event, the constant rain and the AROME-NWC forecasts showed poor performance, 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 2.4) are 0.12 and 0.13 respectively for these two forecasts. The improvements …"*

Line 265: You should explain before explaining figure 6 in which you are using 60% percentile and the reader doesn't know why this percentile

- ⇨ *The sentence starting in line 254 will be modified as follows: "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."*

Fig. 7: Please put again here the legend for the colors of HIT, MISS, FA and CR

- ⇨ *The legend will be added in the revised manuscript.*

Line 286: this is not clear, please try to rephrase

- ⇨ This development about the effects of stratification certainly needs further explanations to be completely clear. However, it does not bring crucial information for the interpretation of the results, and thus we propose to move it into appendix C

where the stratification of rank diagrams is developed and includes the reference to Bellier et al. (2017) for more details.

Modification proposed for line 286: "*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 stratification (see appendix C).*"

Modification proposed for the end of appendix C (l.419): "*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.*"

Line 309: Please rephrase: You wrote the explanation of CSI1 and then you state that you are not choosing CSI2 because ... It should be more linear and clearer if you explain CSI1 then state that you chose it because ...

⇨ The sentence will be modified as follows: *"Since the studied events have very different spatial extents (see figure A.1 and table 1), we chose to use the CSI1 formula where the averaging implies that all the events have the same weight. CSI2 would have given much more relative weight to the large-scale events."*

Line 328: I think this is a big limit of the work since it's really crucial in operational use of the chain, can you anticipate how you will work on that?

⇨ As mentioned in the text, the anticipation times can be computed by taking the difference between the time Tsim of the first threshold crossing by the reference simulation, and the time Trun of the first forecast that detects the threshold crossing. In the following, T is the forecast range : while T is limited to 3 hours, we chose here to count a HIT when Tsim is in the]Trun; Trun + T + 3h] interval, i.e. even when Tsim slightly exceeds Trun+3hours. This results in anticipation times which can reach up to 6 hours, which may appear confusing for the reader. This is the reason why we chose not to present these anticipation times in the first version of the manuscript.
The histograms of anticipation times we obtain for the 5years threshold and the 50th PIAF-EPS percentile are presented on the figure below.

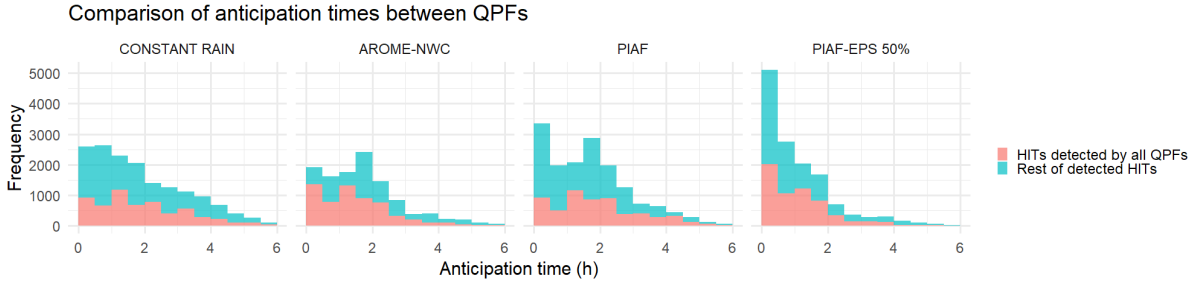

Fig 13: Anticipation times aggregated for all events for each QPF (50th percentile for PIAF-EPS), for the 5years threshold

Since this figure brings interesting information about the differences in anticipation, we propose to add this figure in the manuscript with an explanation for the reason of anticipation times exceeding 3 hours:

Instead of lines 331-339: *"… contingency tables values was attempted by Charpentier-Noyer et al. (2022), and can be obtained by 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 (50^{th} percentile for PIAF-EPS) and for the 5-year threshold are presented in figure 13. First, 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 ($0<T\leq3$ 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. 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."

Fig. B1: Not really clear the miss example because of the style of the lines both for image d and f

- ⇨ *We will try to improve the presentation of figure B1, but since it is difficult to present all the information considered on this figure, we propose to add some explanations in the figure caption: (a) forecasted threshold exceedance but not recorded 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)*

RC2: 'Comment on egusphere-2023-907', Anonymous Referee #2, 08 Aug 2023

General comments:

This article presents a comparison between different QPF-forced hydrologic predictions of flash flooding events, featuring a recently developed ensemble technique at Meteo France. Selected events are used to assess the ability of these predictions to detect the occurrence of streamflow values exceeding defined frequency-based thresholds. QPE-forced hydrologic simulations are used as reference data. Metrics based on contingency table data such as the Critical Success Index are used to assess skill. While I don't see any new scientific insights into QPF-forced hydrologic forecasts, this is a good contribution to the literature of tools with operational implementation. The article's methods are based on robust work previously published by others. The manuscript is written well and in a concise manner. My only concern is that the authors do not provide enough details about the featured technique that is central to the study. My recommendation is that following some minor revisions, the paper could be accepted.

⇨ We thank Referee #2 for the careful reading of our manuscript and the relevant comments and suggestions. We provide point to point answers below, including details about the way we plan to adapt the manuscript.

- My main comment is that the PIAF-EPS methodology is not described with enough details. The workflow schematic in Figure 2 does not provide any information on how the perturbations are computed, which seems to be an important aspect of the technique. Authors should include an example of these perturbations, so readers can get a sense of what they look like.

⇨ We propose to add the following figure, line 137: *"An example of the perturbations is given in figure 3."*

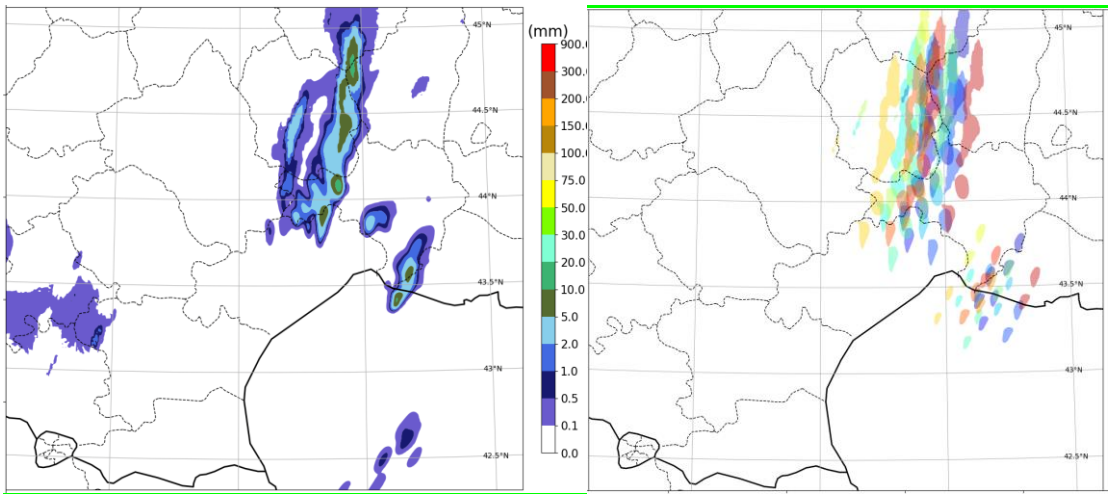

Figure 3: example of PIAF-EPS ensemble forecast perturbations. Left : deterministic PIAF forecast of 15-minute rainfall accumulation (forecast start: 19 Sept 2020 at 06utc, forecast range: 2 hours). This is used as member zero of the ensemble. Right: same field in members 1 to 16, the shading represents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

- Many acronyms not spelled out the first time they appear in the text. At least some of them are spelled out later in the document, but they should be spelled out as soon as they are used for the first time, so the reader is not left wondering about it.

⇨ Thank you for noticing this. We will carefully check that acronyms are spelled at first occurrence.

Specific comments:

Line 61: "…for flash flood nowcasting purposes" Is it appropriate to say "flash flood nowcasting"? I have only seen nowcasting being used to describe QPE extrapolation.

⇨ We agree to rephrase as "*for flash flood forecasting purposes*"

Line 81: "SMASH" (L80) Is "PANTHERE" an acronym? If so, please spell it out.

⇨ SMASH acronym is detailed in section 2.4, and PANTHERE is indeed an acronym. As a consequence, the sentences will be modified as follows: "*The simulated/forecast hydrographs and the reference discharges are obtained using a fully distributed rainfall runoff model, detailed in section 2.4. In the operational version of Vigicrues Flash, this hydrological model is forced with the PANTHERE (Projet Aramis Nouvelles Technologies en Hydrométéorologie Extension et Renouvellement) rainfall QPEs, derived from a network of about 30 radars over mainland France and its vicinity (Tabary et al., 2013).*"

Line 94: What is "…a ten-minute observation cutoff"? Do you mean only the first ten minutes worth of observations are assimilated? How many observations (how many data points) are actually assimilated? Also, specify what data are assimilated (radar, satellite, rain gauge?).

⇨ We will insert the following explanation on line 94: "*...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)*." Also, additional information will be inserted at line 95: "*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).*"

Line 96 – 97: Spelling out this acronym (PIAF) should occur earlier in the document, as soon as it is first used. Same with all other acronyms.

⇨ Line 60 (first occurrence of "PIAF") will be modified as follows: *"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."

Line 92 – 112: Use of "Lead time". Consider replacing the term "Lead time" with something like forecast length, or simply referring to a particular forecast by its length. For example, the 3h forecast, to refer to a forecast that goes out 3 hours into the future. The term "Lead time" implies skill associated to a particular forecast length, and not a configurable parameter.

⇨ Agreed, we will replace "lead time" by "forecast range" which is universally used in the meteorological community ("forecast length" is rather used for the total forecast duration, which is a different thing). In this study, the forecast length is 3 hours and the forecast range can take values between 15 minutes and 3 hours.

Line 119: With "equiprobable", do you mean perturbations are "drawn" from a uniform probability distribution?

⇨ No, the meaning is "equiprobable: having the same degree of logical or mathematical probability" (www.merriam-webster.com). The distributions are clarified a few lines after: 2D Gaussian sample for spatial perturbations, and clipped AR(1) autoregressive process for the amplitude perturbations. Sentence will be clarified as follows: "*using a priori equiprobable perturbations of the precipitation field, as explained below*:"

Line 121: What do you mean with "subrandom"?

⇨ We will correct as "pseudorandom"

Line 126: A better term to replace "lead time" here would be forecast length.

⇨ We will replace by "... as a function of forecast range"

Line 172 – 173: How costly? How often is the system changing?

⇨ We will rephrase *as "(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)"*

Line 176: "…the importance of hydrological **reaction** response"?

⇨ We propose to rephrase this sentence as follows to be more clear and explicit: "*.. and the intensity and geographical extent of the hydrological responses simulated by the SMASH model"*

Table 1: "Duration" does not seem appropriate. Use "Date"? Also, I see the order of the events in here is by date, but labels "A-H" are all over the place. Not a big deal, but this is very odd order to follow here and Figure 5. It feels like the labels' purpose was to make it

easier for the events to be organized/classified, but the way they are presented in this table seems to defeat said purpose?

- ⇨ *The table will be re-organized, and "Duration" will be replaced by "Date".*

Line 211 – 212: The word "assimilated" is misleading. Do you recursively use reference streamflow to improve model states and/or parameters? If not, then I strongly recommend using a different term here.

- ⇨ *"assimilated" will be replaced by "assigned".*

Line 255: Why was the 60th percentile chosen for the comparison? I could not see anything in the previous texts that would give indication that a particular percentile was to be used.

- ⇨ The sentence starting in line 254 will be modified as follows: *"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."*

Line 360 – 361: More than confirming, a robust study on a large enough sample dataset should inform how truly valuable and applicable is the ensemble-based technique, particularly in real-time.

- ⇨ We agree, such additional study on a large and continuous sample would bring additional information, not only confirmation. The sentence will be modified as follows: "*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."*