Anonymous Referee 3

The revised version of the manuscript is characterized just by minor revisions with respect to the first submission. Several essential issues still remain unsolved. Actually, authors have not fulfilled the majority of main concerns highlighted in the first review process. The current contents of the paper represent a post-event analysis; there is no novelty and originality of contents (meteo-hydrological forecasting tools, statistical forecast verification, outcomes). Many themes are discussed throughout the manuscript, but none of them is deeply investigated by introducing innovative tools or results with respect to past studies. Challenging purposes are proposed, but the described outcomes do not allow supporting fully statements written throughout the manuscript. Even, the discussion of results brings to face with questions that remain unsolved. Many parts of the text describe reasonings and outcomes characterised by a weak significance or obvious conclusions (in particular, Sections 5 and 6; for instance, several times the outcome of the discussion of results is to state that ensembles are useful with respect to the scenario of using zero rainfall as forecast). In my opinion, the current manuscript resembles a technical or internal report that may be of interest for local forecasters and end-users (in particular, Sections 2, 3 and 4), but its soundness for researchers and readers of an international peer-reviewed journal is weak.

Main general concerns are recalled below.

We thank referee n°3 for this feedback about the revised version of the paper. Despite our point to point answers to the remarks formulated in the first turn of reviews, it seems that there is still some misunderstanding about the scope and the novelty brought by this paper. We provide further explanation on these aspects in our detailed answers below. The novelty and originality of the paper lies in our dealing with short-range flash-flood forecasts. This requires specific QPFs ensemble products (high temporal and spatial resolution, high refresh rate, seamless forecasts, ...) that have, to our knowledge, never been used in any comparable study. These original forecasting tools are essential for addressing the large spatio-temporal variability and limited predictability of the heavy precipitation events generating these floods. The paper involves two new experimental ensemble products that aim to address these specific requirements, and have never been evaluated before. As of today, there is no consensus in the numerical weather prediction (NWP) community about the best approach to cater for the needs of flash-flood numerical prediction models at these scales. Such new experimental rain products are still not operationally available, and they are not available on enough past high precipitation events for hydrologists to evaluate the performance and make decisions on their added value (i.e., changing (or not) the configuration of operational systems). The usefulness of this paper lies in its comparison of the relative merits of several NWP approaches that could be implemented operationally in the near future, which is a question of strategic importance and societal significance in many meteorological and hydrological prediction institutes. To that end, we have presented a first evaluation focused on the capacity to provide efficient forecasts for one single and intense flash flood event.

The novelty brought by the proposed evaluation framework lies in the way the conventional metrics are combined and adapted, to obtain an as detailed and meaningful evaluation as possible of the hydrological forecasts for the considered event. We think that this kind of event-based evaluation can bring interesting complements to the conventional large-scale and statistically more representative evaluation of ensemble forecasts, which remains of course necessary. By definition, high impact flash flood cases are rare, and state-of-the-art numerical prediction tools are computationally expensive and continuously evolving. Thus, it is often impossible to properly assess the performance of these tools in a statistical way: event-based evaluations are unavoidable. Both approaches (statistical and event-based) should not be opposed but rather considered as complementary, and this opinion seems to be increasingly shared by the community (see for instance p.3-4 of this report of a ECMWF worshop on model uncertainty). The reviewer rightly points out that statistical evaluation of ensemble predictions of precipitation has already been performed in many studies, so there would be no point in bloating the article with yet another one. We believe that it is more productive in focusing the paper on the more original aspects of event-based evaluation challenges. Event based evaluations have been shown to be very useful communication tools to exchange with and get feedback from end users (Dasgupta et al., 2023). Our paper does not claim that event-based evaluations provide a complete picture of the performance of forecasting systems, we merely claim that such evaluations are important tools for designing of these systems and documenting their performance in rare, high-impact flood cases.

Our study aims at opening a discussion on this issue of event based evaluation. The introduction section of the paper has been adapted to make this objective appearing more clearly. We are aware that finding a solution is complex and might take some time (if ever) to be achieved. We believe however that the research community needs to address this issue. Our paper tries to contribute to this discussion: how can we better evaluate the quality of new hydrometeorological forecasting systems

that target to improve flash-flood forecasts? How can one decide which system is the best when we only have reforecasts for a single flood event to evaluate at a given river basin? We believe that these questions are important to the forecasting community, and they are clearly of interest to forecasting offices in many regions of the world that are affected by similar flash flood events. Kilometric-scale ensemble numerical weather prediction systems are only beginning to be used for nowcasting purposes, and their application to flash flood prediction is a timely question that will interest many readers, given the growing economical and human impacts of flash-floods worlwide.

The declared aim to focus on the needs of civil protection authorities appears incomplete: some verifications on QPFs (in particular, Figs 5 and 6) and discharge forecasts (Figs 10-15) were discussed just for very short lead times (the 1-h lead time in Figs 5-6, the 3-h lead time in Figs 10-15), neglecting longer lead times which are more proper and useful for the aim of warnings by authorities in charge of decisions in case of flood. Lead times shorter than 6-12 hours do not allow issuing timely warning and take effectively actions for safety and emergency services (bearing in mind also the time to collect observed data, run the hydrological models, analyze result and issue warnings).

The rainfall forecasts evaluated here correspond to short range forecasts limited to 6 hours lead-time. We fully agree that warnings issued with larger lead-times are very useful for preparedness actions. But lead times shorter than 6-12 hours are a reality for many operational forecasters and emergency managers dealing with flash floods (as is the case of the Mediterranean flash flood we are analysing in our paper), because of the fast evolution and limited predictability of the triggering heavy precipitation events. Many flood events have a too low predictability for warnings to be usefully issued more than a few hours in advance (Davolio et al., 2017; Carrio et al., 2022), as (again) demonstrated by several catastrophic events in the Mediterranean area in 2022. For this reason, several research contributions dealing with flash floods have focused in the last years on very short range (<6h) forecasts, based on radar advection and NWP blending approaches, and/or radar data assimilation in NWP models (Bowler et al., 2006; Berenguer et al., 2011; Silvestro and Rebora, 2012; Davolio et al., 2017; Poletti et al., 2019; Zanchetta and Coulibaly, 2020; Lovat et al., 2022).

Civil protection authorities which are facing flash-floods also express the need for short-range (<6h) forecasts, issued with high refreshment frequency, to help in localizing more accurately in space and time the areas at risk during the development of these events. Delivering flash flood forecasts with up to 6h lead time would represent a significant improvement in comparison with currently existing flash flood monitoring or now-casting systems, which often still rely on radar rainfall observations (Gourley et al., 2017). Particularly, the need for short range forecasts has been confirmed within the PICS project (https://pics.ifsttar.fr/en), through exchanges with an end-users group including the varied authorities involved in flash-flood crisis management in France (Javelle et al., 2021). The experimental short range-rainfall products studied in this paper have been specifically released to address this demand. This is the reason why some studied products are not yet operational, and also why this study specifically focuses on short lead-times.

Nevertheless, since the studied forecasts are released for up to 6-hour lead times, it is possible to illustrate the results obtained for a wider range of lead-times from 1 hour to 6 hours, for the rainfall forecasts (fig. 5-6) and the forecast hydrographs (fig. 10-15). We propose to include all these results in the revised version of the paper: the figures 5-6 and 10-15 have been all focused on the intermediate 3-hours lead-time, and the 1-hour and 6-hour lead times have been presented in a new appendix (Appendix A) for rainfall forecasts, and in appendix C for forecast hydrographs (corresponding to appendix B in the former version of the manuscript).

Another desired goal of the manuscript is to draw lessons from the analysis of the selected case study for the users of hydrological forecasts. But, two of the proposed ensemble systems are not routinely run (it seems that the experimental phase about these tools cover just the year 2018). It is not clear the sense of investigating performance of these products with respect to the aim of end-users if they have not at disposal such forecasting tools in the operational practice.

In the operational practice of flood forecasting, new tools and forecasting systems are rarely tested in real-time before being tested first, and evaluated, in research. This is because decisions involving flood forecasts and warnings/alerts impact human lives and hence have to be carefully taken. In many National flood forecasting services, forecasters are legally responsible for the warnings they launch (or not) and their consequences, in particular when human life losses are involved. This explains why the systems being evaluated in our paper are designed for operational use but are evaluated within a research project first, before implementation in real-time (PICS Project, https://pics.ifsttar.fr/en, which involved both research labs and operational flood forecasting services). The complexity of this evaluation is the question at the heart of this paper. The tested products have been proposed specifically to address the demand for short-range (6h) seamless ensemble forecasts, with a 1h refreshment rate.

Two of these products are experimental, and are generated in an original way, by merging different runs of two convection permitting NWP models, including one ensemble run and several (time lagged) deterministic runs. Since they are experimental and computationally expensive, these products have been released only for selected events of year 2018 and are not routinely run yet. Non-real time evaluation, as we have done here, is a prerequisite for establishing confidence that the tested systems are worth running in real time for preoperational evaluation, which will be the next step, but it is beyond the scope of our study. The first evaluation we propose in this paper aims to verify if such products could be useful for flash flood forecasting, and can be expected to have value for end users. Such preliminary analyses are necessary before the experimental products can be adapted and tested on larger periods to build statistically more robust conclusions, and finally be integrated in operational workflows. A new development has been added in the introduction section of the paper to clarify these aspects.

The way to build the ensembles based on time lagging (i.e., "pepi") and spatial shift (i.e., "pertDpepi") could be questionable about some characteristics that appear as unsolved (maybe, additional investigations would be useful to improve the characteristics of the two ensembles). On the one hand, it is not clear the physical meaning underlying the way to build the ensemble based on AROME-EPS and AROME-NWC (i.e., "pepi"). The main and only reason seems to obtain an ensemble with a certain number of members to run the hydrological model. But, the gain in performance of "pepi" are mainly due to NWC and the time lagging, the members associated to EPS do not give an added value. On the other hand, the spatial scale of the shift for "pertDpepi" appears as not optimal for the investigated study area (a sensitivity analysis of the extension of the spatial shift should be proposed).

We fully agree that the precipitation ensembles used could still be improved. Further improving the atmospheric ensembles would be a big endeavor that is beyond the scope of this paper. This paper just presents a first hydrological evaluation of three products, among which two are experimental and were specifically designed to address the need for seamless very short range rainfall forecasts to better anticipate flash floods. A sentence has been added in section 3.3 to better explain the origin of these new products. The point of using these precipitation ensembles in this study is to show that they clearly impact forecast performance for the considered event, as demonstrated by the ROC evaluations in fig 8. There, the ROC curves are well above the diagonal for all ensembles, which proves that the ensembles carry predictive value, regardless of the ensemble perturbation details, for this event. Regarding the way of building the ensembles:

- The AROME-EPS ensemble is a state-of-the-art weather forecasting ensemble and its perturbations are physically based as explained in the linked journal papers (Bouttier et al., 2012; Raynaud and Bouttier, 2016);
- The PEPI and pertDpepi ensembles are based on lagging and statistical field perturbation, which is the currently dominant approach to generate short-range precipitation ensembles (**?**Lu et al., 2007; Bowler et al., 2006).

Our study is the first that compares these both types of ensembles from a hydrological point of view, so it gives novel information about their respective merits, even if this first evaluation is focused here on one single event and can obviously not be extrapolated.

It should also be noted that these ensembles (pepi, pertDpepi) are not limited to run over the Aude river catchment only. They are designed to run over the whole French territory (as it is usually the case with NWP models in meteorology). Therefore, perturbations and other steps implemented when building the ensembles cannot be calibrated to satisfy one event in one river basin. This additional complexity justifies the event-based evaluation we are proposing in this paper.

The verification metrics are commonly used over large datasets, in order to highlight statistical characteristics of the forecast product. The use of verification metrics to analyze a single event has poor significance (rank histogram, ROC curves and spread-skill relationship in Figs. 6, 7 and panels a) in Figs 10-15, respectively).

We agree that the ensemble verification metrics are generally used over large datasets to reach statistical significance. An analysis focused on one single event cannot reach this objective. On the other hand, analyses focusing on single events are also needed (see for instance the conclusions of ECMWF workshop on model uncertainty). They enable to delve into the precise characteristics of forecasts, and to illustrate the information obtained for the most intense and critical events. This is the reason why they are interesting for end-users. We think both approaches are complementary and should not be opposed. What we propose here is only a single event analysis. We do not intend for the metrics we use to deliver statistically significant conclusions, they are just used here for characterizing the detailed behaviour of the ensembles for the analyzed event, and they are combined and adapted in an original way to this purpose. We added several sentences in the manuscript to avoid

any misleading on this point, and to remind that the conclusions are valid only for the considered event and should not be extrapolated to future events.

Even though the focus is to evaluate a rare flood event, a statistical analysis of discharges performed over a long dataset is significant to test false alarms due to potential overestimation of QPFs (especially, for events on very small areas like many of the investigated sub-basins). Otherwise, a deceptive reliability about the forecasting tool could be induced in the users of those forecasts. The statistical analysis in terms of discharge forecast should consider at least the whole period covered by QPFs (in the first submission, the performance of the meteo ensembles were shown for the whole year 2018), not just a flood event.

We agree that a statistical analysis over a long continuous period would be essential to characterize the risks of false alarms, and the actual statistical performance of the forecasts. The period of such an analysis should probably largely exceed the year 2018 and the Aude area, to include a significant number of flood events exceeding the 10-year discharge threshold (which is a relevant threshold in our opinion, since often corresponding to the observation of first significant inundations and damages). This is not yet possible, since the experimental rainfall ensembles we used here have been released only for a few intense rainfall events of 2018 (Aude river in October, Ardèche and Cèze rivers in August and Argens rivers in October), among which only the October flood in the Aude river exceeds a 10-year return period. The evaluation of the rainfall forecasts presented in the initial version of the manuscript was pooling all these three events, for the whole product geographic window covered by the rainfall forecast products, but its statistical significance also remains limited. The event analysis provided in the paper enables to examine in detail the hydrological forecasts obtained for one interesting intense flood event. The performance observed on such an event can help to decide if it is worth conducting a continuous analysis on a larger period of time, to draw robust statistical conclusions.

The selected case study has been already investigated by some past studies, at least by the meteorological point of view (with the same or similar QPF forecasting tools).

The event has been studied from a meteorological point of view by Caumont et al. (2021), including the performances of QPFs based on AROME and AROME EPS. Lovat et al. (2022) studied the performance of deterministic hydrological forecasts for this event, using AROME-NWC and PIAF short-range QPFs. Other studies focused on this event to improve a hydrological model (Peredo et al., 2022), or to evaluate automated flood mapping approaches (Hocini et al., 2021). But none of these studies have focused yet on short range ensemble hydrological forecasting, and the experimental pepi and pertDpepi QPF products have never been evaluated yet from a hydrological forecasting perspective.

The proposed framework for the evaluation of meteo-hydrological model coupling does not represent an innovation, a new approach. That analysis collects different approaches commonly used in the worldwide operational practice for the verification of accuracy of ensemble precipitation forecasts and to convey and analyze the information provided in terms of discharge forecast by a meteo-hydrological forecasting chain. Roughly summarizing, the forecast is verified over a study area of interest for local end-users (identified as “HFA”) within a time window useful to issue warnings (identified as “HFT”).

The novelty of the proposed approach lies in two aspects in our opinion: 1 - it proposes a new combination of well known evaluation metrics in order to provide a synthetic and as informative as possible analysis of the considered event, and 2 - the common contingency table and ROC curve approach has been adapted here (see appendix) to obtain one unique ROC curve summarizing the performance for all the lead-times, and enabling them to examine separately the anticipation times. To our knowledge, similar approaches have not been proposed yet in the literature: if we are wrong, we would be grateful to obtain the corresponding references.

The analysis related to anticipation times for the hydrological forecasts (i.e., Fig 9) is questionable and misleading. Outcomes depends on concurrently by the accuracy of rainfall forecast for the event study as well as by the characteristic of the basin (namely, the response times of the considered catchments to rainfall). Many statements describe reasoning of weak significance that lead to obvious outcomes or are strictly valid for the selected runs of QPFs. The representativeness and significant level of contents conveyed by Fig.9 are weak to gain insight about the proposed tools for flood forecasts (in particular when aimed to warning purposes).

We completely agree that the anticipation times presented on Fig.9 depend both on the accuracy of rainfall forecasts and the response times of the considered basins. This is the reason why the RF0 (zero future rainfall) forecast is presented on this figure: this reference run illustrates the part of anticipation that is only due to the basins’ response times. The comparison of the anticipation times obtained with RF0 and with the QPF products shows the gains in anticipation associated with the QPF products. This important role of the reference RF0 forecast is explained in section 2.2 of the manuscript. The conclusion here is

that the gain in anticipation is significant with the QPFs for the studied event, which is in our opinion important information to characterize the added value of forecasts for this specific event. This conclusion should obviously not be extrapolated to future events: we have reminded this in the text to avoid misleading interpretation. We do not see in the comments of this figure any other questionable or misleading points.

Anonymous Referee 4

We thank referee n°4 for this very positive evaluation and his suggestions to improve the manuscript.

- Line 245 (typos): spatial is spatial We corrected this typing mistake.
- Lines 300-305: referring to figure 4, is it a) referring to 1-hour lead time and b) to 6-hour, or viceversa? When you write "except at the end of the rainfall event, on the 15th October between 7:00 and 11:00 utc, where all ensemble forecasts overestimate the rainfall rates, particularly for the 1-hour lead time forecast", it seems to me that this is more evident in figure 4 b), that is 6-hour lead time forecast. And when you write "for the 6-hour lead time forecast a time shift of 2 hours is observed during the rising phase ", it seems to me more evident in figure 4a), that is 1-hour lead time.

  The paper now includes the results for a 3-hour lead time in figure 4, and for the 1-hour and 6-hour lead times in appendix A. We changed the analysis to mention that the overestimation at the end of the rainfall event is present rather for the 3-hour and 6-hour lead times, and that only the 6-hour lead-time shows a relatively systematic time shift of 1 to 2 hours during the whole event.
- Figure 7: I suggest to add in the caption the definition of POD and FAR (they are defined in appendix A)

  We added the definitions in the caption of Fig.7

References

Berenguer, M., Sempere-Torres, D., and Pegram, G. G.: 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.2011.04.033, 2011.

Bouttier, F., Vié, B., Nuissier, O., and Raynaud, L.: Impact of stochastic physics in a convection-permitting ensemble, Monthly Weather Review, 140, 3706–3721, https://doi.org/10.1175/MWR-D-12-00031.1, 2012.

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, 2006.

Carrio, D. S., Jansa, A., Homar, V., Romero, R., Rigo, T., Ramis, C., Hermoso, A., and Maimo, A.: Exploring the benefits of a Hi-EnKF system to forecast an extreme weather event. The 9th October 2018 catastrophic flash flood in Mallorca, ATMOSPHERIC RESEARCH, 265, https://doi.org/10.1016/j.atmosres.2021.105917, 2022.

Caumont, O., Mandement, M., Bouttier, F., Eeckman, J., Brossier, C. L., Lovat, A., Nuissier, O., and Laurantin, O.: The heavy precipitation event of 14-15 October 2018 in the Aude catchment: A meteorological study based on operational numerical weather prediction systems and standard and personal observations, Natural Hazards and Earth System Sciences, 21, 1135–1157, https://doi.org/10.5194/nhess-21-1135-2021, 2021.

Dasgupta, A., Arnal, L., Emerton, R., Harrigan, S., Matthews, G., Muhammad, A., O'Regan, K., Pérez-Ciria, T., Valdez, E., van Osnabrugge, B., Werner, M., Buontempo, C., Cloke, H., Pappenberger, F., Pechlivanidis, I. G., Prudhomme, C., Ramos, M. H., and Salamon, P.: Connecting hydrological modelling and forecasting from global to local scales: Perspectives from an international joint virtual workshop, John Wiley and Sons Inc, https://doi.org/10.1111/jfr3.12880, 2023.

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, 2017.

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, 2017.

Hocini, N., Payrastre, O., Bourgin, F., Gaume, E., Davy, P., Lague, D., Poinsignon, L., and Pons, F.: Performance of automated methods for flash flood inundation mapping : a comparison of a digital terrain model (DTM) filling and two hydrodynamic methods, Hydrology and Earth System Sciences, 25, 2979–2995, https://doi.org/10.5194/hess-25-2979-2021, 2021.

Javelle, P., Payrastre, O., Boudevillain, B., Bourgin, F., Bouttier, F., Caumont, O., Charpentier-Noyer, M., Ducrocq, V., Fleury, A., Garambois, P.-A., Gaume, E., Hocini, N., Janet, B., Jay-Allemand, M., Lague, D., Lovat, A., Moncoulon, D., Naulin, J.-P., Nicolle, P., Peredo, D., Perrin, C., Pons, F., Ramos, M.-H., Ruin, I., and Terti, G.: Flash flood impacts nowcasting within the PICS project (2018-2022): End-users involvement and first results, pp. null–null, Periodica Polytechnica Budapest University of Technology and Economics, https://doi.org/10.3311/floodrisk2020.17.3, 2021.

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, 2022.

Lu, C., Yuan, H., Schwartz, B. E., and Benjamin, S. G.: Short-range numerical weather prediction using time-lagged ensembles, Weather and Forecasting, 22, 580–595, https://doi.org/10.1175/WAF999.1, 2007.

Peredo, D., Ramos, M.-H., Andréassian, V., and Oudin, L.: Investigating hydrological model versatility to simulate extreme flood events, Hydrological Sciences Journal, 0, 1–18, https://doi.org/10.1080/02626667.2022.2030864, 2022.

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, 2019.

Raynaud, L. and Bouttier, F.: Comparison of initial perturbation methods for ensemble prediction at convective scale, Q. J. R. Meteorol. Soc., 142, 854–866, https://doi.org/10.1002/qj.2686, 2016.

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, 2012.

Zanchetta, A. D. L. and Coulibaly, P.: Recent Advances in Real-Time Pluvial Flash Flood Forecasting, WATER, 12, https://doi.org/10.3390/w12020570, 2020.