Anonymous Referee 2

GENERAL COMMENTS

The manuscript proposes a framework to assess the quality of hydrometeorological forecasts for flash flood events and applies it to the event that affected the Aude basin in October 2019.

Conceptually, the proposed framework consists of determining the so-called hydrological focus time and hydrological focus area as the relevant temporal and spatial domains over which the hydrometeorological forecasts are evaluated in terms of the forecasted rainfall accumulations and hydrographs at different points of the river network using existing approaches.

The topic is relevant and the application of the methodology for the analysed event produces interesting results. However, the writing and organization of the manuscript need to be significantly improved to make it ready for publication. Also, some further discussion about the hypotheses made and the applicability of the methodology would make the manuscript more interesting.

Consequently, the manuscript requires major revisions before I can recommend its publication in Natural Hazards and Earth System Sciences.

We thank anonymous referee number 2 for the useful comments about this initial version of the manuscript. We provide hereafter our detailed answers and explanations about the modifications introduced in the revised version of the manuscript (which is already available). Thanks to this revision, we think the manuscript is now much easier to follow.

MAJOR COMMENTS

1. The text should be thoroughly revised to improve its clarity, provide a description of all the tools used, avoid repetitions (some aspects appear in several parts of the manuscript), reconsider figures with little discussion (e.g., Fig. 3, Fig. 7), make the text more synthetic (specially sections 4.3 and 5), describe and present all the elements in a sequential way (avoid jumping back and forth), and expand the captions to clearly describe all the figure elements.

   We achieved a general revision of the manuscript to improve its clarity, avoid repetitions, and provide details regarding all the unclear aspects.

2. Organization of the manuscript: Right now the manuscript does not read smoothly. In particular, I think that the readability would improve that Appendix A should be included as a subsection. This could be a rough organization of the manuscript:

   - Introduction
   - Methodology for an event-scale evaluation of hydro-meteorological ensemble forecasts: with the presentation of the 3 steps and the definition of HFA and HFT.
   - Case study, data and models
   - Application of the methodology to evaluate the Ens-QPF products during the event: describing how the methodology has been applied, including the contents of Appendix A.
   - Results
   - Discussion and conclusions: combining current sections 5 and 6.

   The content of Appendix A is important for a good understanding of the manuscript, and thus we agree it should be highlighted. However, the content of this appendix appears quite long to us to be incorporated in the text, and this would also complicate the structure of the manuscript. We preferred to add references to this appendix and to provide more details in the text about its content (section 4.2).

3. The proposed methodology adapts well to the spatio-temporal hydrometeorological features of the analysed event (which shows a quasi-triangular hyetograph in the catchment and mostly single-peak hydrographs). However, I miss some

discussion about how it could be applied to longer, more complex events; e.g., with multiple rainfall periods and multiple hydrograph peaks, or showing high variability of the magnitude of the floods within the affected area. In the latter case, I would like the authors to discuss the possibility of using more than one threshold to assess the quality of the hydrometeorological forecasts; in such a case, would the HFA and HFT be threshold dependent?

The reviewer raises an important question. The HFT and HFA need to encompass the flood event (or the various related flood peaks), and their definition depends on the spatio-temporal settings of the event. We agree that the presented event has relatively simple spatio-temporal features. For events with complex characteristics, particularly multiple flood peaks, it can be considered that each rising or peak phase of the flood event could be examined separately. However, in some cases the different phases of the floods may be difficult to separate: in such situations, each run of forecast may be examined separately along the event. These explanations have been added in the discussion section: "Particularly, in case of events with multiple flood peaks, the fact that the method focuses only on the first threshold exceedance can be seen as a limitation. In such a situation, each rising phase of the flood event could be examined separately, even if in some cases the different phases of the floods may be difficult to separate. An alternative could be to analyze the anticipation of threshold exceedances for each run of forecast during the event, independently of the times the thresholds are exceeded for the RS hydrographs.".

In case of variable flood magnitudes, we agree that several thresholds could be examined for the same event. Evaluating multiple flood events based on the same threshold would also be an interesting extension of the proposed approach. In both cases, it would require an adaptation of the HFA to ensure a balance between the sub-basins for which the considered threshold is or is not exceeded in the RS simulation. It can be noted that other evaluation criterions would less sensitive to the limits of the HFA (such as the Critical Success Index). A specific discussion of these aspects has been added in the text: "Providing an evaluation for multiple flood events, or for multiple thresholds for the same flood, may also be interesting complements to the proposed approach. This may nevertheless complicate the definition of appropriate HFT/HFA, which are event specific. The HFA will also have to be adapted to the considered threshold. To avoid changing the limits of the HFA, an option could be to use a score that does not account for Correct Rejections, and therefore would be less sensitive to the extent of the HFA, such as the Critical Success Index."

MINOR COMMENTS

1. Abstract: the final part of the abstract could be more informative about the results obtained in the study and the conclusions

   The following development has been included in the abstract to provide more details about the final results of the work: "The results show that, provided that the larger ensemble percentiles are considered (75% percentile for instance), these products correctly retrieve the area where the larger rainfall accumulations were observed, but have a tendency to overestimate its spatial extent. The hydrological evaluation indicates that the discharge threshold exceedances are better localized and anticipated if compared to a naive zero-future rainfall scenario, but at the price of a significant increase of false alarms. Some differences in the performances between the three ensemble rainfall forecast products are also identified".

2. Motivation of the study. The introduction provides an interesting description of the topic of flash flood forecasting systems and some of their limitations. However, I miss a better connection between the general context description and the presentation of the objective of the study that clearly states the motivation of the study and justifies the proposed analysis strategy.

   We have shortened and modified the text to better focus on the link between the general context of flash flood forecasting and the presentation of the objectives of the study.

3. Page 31, line 685: “to summary” could be “to summarize”.

   This has been corrected.

4. Page 31, line 695: at this point the acronym “RS” has not yet been defined.

   This has been changed.

5. Page 31, lines 695-696: The following sentence is not fully clear: “The drastic reduction of the number of considered time steps is compensated by the common consideration of the large number of outlets hit by the event.”
   The sentence has been rephrased in the following way: "The use of many outlets (1174 in this study) compensates the small number of considered time steps for the elaboration of the ROC curves.".

6. Page 31, lines 704-706: Please, check the writing.
   We checked the sentence and modified it: "For each outlet, the discharge threshold is then compared with the RS hydrograph for all time steps of the HFT".

7. Section 4 and Appendix A: Given that RS stands for “Reference Scenario” (page 16, line 360), the expressions “reference RS”, “reference RS simulation” or similar need to be corrected.
   The RS acronym is defined in section 2.2 as "reference simulation". We decided to keep this definition. The necessary corrections have been done in the text (page 16, section 4 and appendix), where only "RS hydrograph" or "RS scenario" are now used.

8. Page 31, lines 708-810: “All the discharge forecasts issued before and covering this date (according to the maximum forecast lead time, i.e 6 runs) are then selected. For a given forecast probability (ensemble percentile), a hit is counted in the contingency table if at least one of the six runs exceed the discharge threshold at any lead time (fig A1 - left) left), and a miss is counted if none of the six forecast hydrographs exceed the threshold at any lead time (fig A1 - right)”. This sentence assumes that the reader is aware about the temporal resolution and lead times of the precipitation ensemble forecasts and how they have been applied to produce discharge forecasts. However, the first reference to Appendix A appears in page 6 (line 161), where none of this information has been provided.
   The text has been adapted to mention that the six runs of forecasts correspond to the specific lead time (6h) and forecast refresh time (1h) of this study: "All the runs of forecasts issued before and covering this date are then selected (i.e. 6 runs here according to the maximum forecast lead time of 6 hours and the 1-hour time step between successive runs)".
   Also, in this sentence, the way the probabilistic discharge forecasts are treated should be described better. If I understand well, the rainfall-runoff model is run with each member of the ensemble of precipitation forecasts to generate an ensemble of hydrographs (one per rainfall forecast member); and from these the ROC analysis is based on setting probability thresholds to obtain the associated time series of discharge forecasts. Because these are not necessarily obtained from a single run of the rainfall-runoff model, I would not use the term “hydrograph” when referring to them (page 31, line 711).
   This is right, the hydrologic model is run for each member of the rainfall forecast to generate a hydrological forecast ensemble. From this ensemble, a probability is applied to obtain a time series of discrete discharge forecasts. Examining successively different probabilities finally leads to the ROC curve. This information has been added in the text (section 4.2) : "The hydrological model is first run for each member of the rainfall forecast to generate a hydrological forecast ensemble. From this ensemble, at each hydrological outlet in the HFA time series of forecast discharges are obtained for several probability thresholds.".

9. Fig A1. I would expect the oldest forecast to end at the evaluation time, and the newest forecast to be issued 1 hour before the reference time. In the figure, I cannot see this. Also, explain (at least in the figure caption) what the term “anticipation” used in the Figure shows.
   Since the hydrological model is run at a 15-min time step, and the runs of forecasts are issued every hour, the end of the first considered run can slightly exceed the evaluation time (exceedance of the threshold by the RS hydrograph), and the last considered run can be issued less than 1 hour before the evaluation time. The definition of the anticipation has been provided in the caption.

10. Page 32, line 715: One could think that, if a correct negative occurs in the time range between t-6h and t, but the discharge forecasts exceed the threshold in a different time step, this situation should be classified as a false alarm. I would like to know the authors’ opinion about this aspect and how it affects the presented results should be included in the manuscript.
    It is right that all the forecasts runs issued during the event are not analyzed. We chose to focus systematically on the 6 runs covering the most critical phase of the event (i.e. the threshold exceedance, or in this case the maximum of the

RS hydrograph). Other choices could have been done, and the content of the contingency table may change if additional runs of forecasts were considered. We therefore added the following sentences in the discussion section: "It can be noted that only the runs of forecasts covering the most critical phase of the event (i.e. the time of the threshold exceedance, or the maximum of the RS hydrograph) were considered to build the contingency tables. The results obtained could differ if other runs of forecasts and/or other phases of the event were considered. Particularly, in case of events with multiple flood peaks, the fact that the method focuses only on the first threshold exceedance can be seen as a limitation. In such a situation, each rising phase of the flood event could be examined separately, even if in some cases the different phases of the floods may be difficult to separate. An alternative could be to analyse the anticipation of threshold exceedances for each run of forecast during the event, independently of the times the thresholds are exceeded for the RS hydrographs."

11. Page 32, lines 717-719: “as many values (. . .) as the number of outlets in the HFA”. By combining the results obtained in the different sub-catchments, one could be masking the quality of the forecasts in the most affected areas with those where the event did not event reach the threshold. This could be quite serious in moderate or very local events. Similarly, how would the method be applied in more complex events (e.g. with multiple flow peaks over a few days or affecting sub-catchments of different catchments)?
    We agree that the analysis should avoid masking the performance of the forecasts in the most affected area. The risks of false alarms in the areas where the threshold was not exceeded should also not be masked. This is exactly why the choice of the HFA and of the considered threshold are very important: this helps including in the analysis both affected areas and areas where false alarms may be observed according to the issued rainfall forecasts. The representation of the results in the form of maps (Figure 9) and not only as a ROC curve also avoids an overly global view of results (and masking the exact location of errors). The use of a different score, such as the Critical Success Index, can also be an alternative to limit the sensitivity to the spatial extent of the HFA. Regarding the events with complex spatio-temporal features, see our response to general comment n°3.

12. Page 7, line 190: “The Aude River basin is located in southwestern France”. It could be more appropriate to use “southern France”.
    We changed southwestern to southern.

13. Page 7, line 199: I do not fully understand what is meant by “to be compared to the local 100-year percentile of 200 mm in 6-hours (Ayphassorho et al., 2019)”.
    The sentence was rephrased as follows: "The maximum accumulated rainfall amounts over short durations were also extreme: up to 60 mm in one hour and 213 mm in six hours recorded at Villegailhenc (Figure 2), while the local 100-year rainfall accumulation is 200 mm in six hours (Ayphassorho et al., 2019)".

14. Figure 2, caption: Please, describe how the rainfall accumulation map was obtained. Could you please verify that this is a 47-h rainfall accumulation map as the caption suggests? Also, it could be interesting to include the location of the 31 stream gauges in the Aude catchment mentioned in section 3.2 (lines 219-220).
    The rainfall accumulation map was computed for the October 14 00:00 to October 15 23:00 period. The rasters of hourly Antilope J+1 quantitative precipitation estimates (combining radar and ground measurements) were just added to obtain the map. This information has been included in the caption as follows: "The Aude River basin, its river network, and the observed rainfall accumulations observed from 14 October 2018 00:00 to 15 October 2018 23:00, according to the ANTILOPE J+1 quantitative precipitation estimates (see Section 3.2).".

15. Page 9, line 226, (title of Section 3.3). For consistency, use “AROME” everywhere within the text.
    The modification was done.

16. Page 9, line 240: The sentence “The number of members in the "pepi" product is 18 (respectively 13) for a lead time of 1h (respectively 6h).” needs some rephrasing to guarantee its clarity. Is the 1-h leadtime pepi product used in this study?
    We modified the sentence to provide more details:"The resulting "pepi" product provides forecasts for a maximum lead time of 6 hours. It combines 12 members from the last available AROME-EPS run, and 1 to 6 members from AROME-NWC, depending on the considered lead time. The resulting number of members varies between 13 (for 6-hour lead

time) and 18 (for the 1-hour lead time).". All the lead times of the three ensemble products are combined to compute ROC curves and threshold exceedance anticipation maps (see Appendix A).

17. Section 3.3: I suggest finding alternative notation for the terms “pepi” and “pertDpepi” that describes better these two sets of ensemble forecasts. What do these terms stand for? What are their spatial resolution and rainfall accumulation window?
    Pepi stands for the contraction of AROME-EPS (AROME-PE in French) and AROME-NWP (AROME-PI in French), and pertDpepi for the PERT method (Vincendon, 2011) applied on pepi ensemble. We agree this notation is not fully explicit for English-speaking people, but nevertheless this does not affect the understanding of the manuscript in our opinion. The spatial resolution is 0.025° and the spatial window covers the metropolitan territory of France (see next comment).

18. Pages 9 and 10, lines 231-247: the description of rainfall ensemble forecasts needs to be rewritten to guarantee that it is clear how the forecasts from these 3 products have been applied in the study (not only what are the maximum lead times, but also if some spin-up time has been established, how the hourly frequency has been handled in the case of the AROME-EPS…). Also, information about the spatial resolution of the grids and about the rainfall accumulation windows needs to be provided.
    For an improved clarity, the paragraph describing the spatial resolutions of the ensemble forecast products has been moved just below the description of the ensembles, and further details have been added, including the spatial window and resolution.

19. Page 10, lines 245-247: “The spatial shift applied to this product represents an ideal distance because i) it captures the main uncertainties due to the localization of the rainfall event, and ii) it is a shift that does not combine too incompatible areas.” Is there any reference to support such a statement? How could this be verified?
    This description has been rewritten (also in response to reviewer 3 about lines 619-621) as follows: "The shift scale of 20 km represents a typical forecast location error scale: according to Vincendon et al. (2011), 80% of location errors are less than 50 km. The value of 20 km has been empirically tuned to produce the largest possible ensemble spread on a set of similarly intense precipitation cases, without noticeably degrading the ensemble predictive value as measured by user-oriented scores such as the area under the ROC curves."

20. Page 10, lines 246-247: “it is a shift that does not combine too incompatible areas.” Please, clarify.
    We rephrased this sentence, see answer to comment n°19.

21. Figure 3: What is shown in a reliability diagram needs to be clearly described to facilitate the interpretation of this figure by the non-expert reader (for the ROC curve, at least, mention that this interpretation can be found in Appendix A). Also, the text in Fig 3a needs to be clearer (ensure the readability of all numbers).
    We preferred here to remove this figure which is not completely in line with the objectives of the paper, as mentioned by referee n°3: evaluation of QPFs at large temporal and spatial scales and for a relatively low threshold of rainfall intensity (5 mm/h). Moreover, the figure is based on scores which differ to the ones developed in the paper (or at least which are computed differently).

22. Page 10, line 254. I suppose that “≈ 2 km2” should be “≈ 2 x 2 km2”. Is this the original resolution of the EPS grids? How were the different resolutions between observations (≈ 1 x 1 km2)” and the forecasts treated to do the evaluation (e.g. Figure 3)? Were the observations upscaled to the forecasts grid? Or the forecasts interpolated to the observations grid?
    The ensemble forecasts are actually provided on a 0.025° by 0.025° grid. For the comparison with observations (rank diagrams), the forecast values have been disaggregated on the 1 km x 1 km grid. These explanations have been added in the text.

23. Page 11, line 271. Please, specify that KGE stands for the Kling-Gupta efficiency, and provide a reference.
    We detailed the acronym in the text and added a reference for the KGE criteria.

24. Page 11, lines 271-273: "The KGE calibration (validation) values obtained were of 0.80 (0.71), which indicates good model performance, except for one validation outlet, where a low KGE value of 0.1 was obtained (Figure 4a)." It is unclear where the reported KGE values (0.80 – 0.71) were calculated. At the downstream-most level-gauges? Are these the average KGE values at all the gauge stations? Besides the validation gauge with KGE 0.1, Figure 4a shows the KGE is, approximately, between 0.6-1 at the calibration gauges and between 0.3 and 0.8 at the validation gauges.
    The values provided in the text correspond to averaged values for the 16 calibration and 15 validation stream gauges. This explanation has been added in the text.

25. Figure 4b. The reference to the "HyMex estimates" is only provided in section 3.2. The reference to the section or to the work of Lebouc et al. (2019) could be added in the figure caption or in the description of CINECAR.
    We included a reference to Lebouc et al. (2019) in the caption of the Figure 4b.

26. Page 11, line 291: What is "ANTILOPE J+1"?
    ANTILOPE J+1 corresponds to the QPEs obtained with the ANTILOPE algorithm (Laurantin et al, 2008), by combining the radar data and rain gauge observations available the next day (J+1). This information has been added in the "Observed hydrometeorological data" section (Section 3.2).

27. Page 11, lines 296-297: "with some few exceptions that can be explained by the spatial averaging". What does it mean? Is not the same averaging applied to the 3 ensemble forecasts and over the same area?
    The spatial averaging is the same for the three ensemble products. Since there are several time steps (15-10 11h - Figure 5a) where the dispersion is surprisingly lower for pepi and pertDpepi than for AROME-EPS, we think this lower dispersion could be explained by the spatial averaging over the chosen spatial window: the members added in pepi an pertDpepi, even if different from the members included in AROME-EPS, may have a similar averaged value. We removed this confusing sentence, which is not essential.

28. Figure 5. The range of the y axis for the two panels should be the same. In the figure caption, it would be useful to state that the Aude catchment is 6074 km2.
    We modified the y-axis with the same range for the two panels, and modified the caption of Figure 5.

29. Page 14, lines 310 – 317. The selection of the HFT seems to be quite subjective. Why is it based on a threshold of the Aude average rainfall intensity of 2 mm/h? The discussion about the analysis of the results being dominated by periods of low rainfall intensities would also apply to the fact that several parts of the catchment registered low rainfall. Similarly, the decision of taking the Aude catchment as the HFA is arbitrary. How much these decisions could have an effect on the obtained results? Could the HFT and HFA be obtained based on more objective criteria? For instance, considering the spatio-temporal structure of the observed and forecasted 1-h rainfall accumulations as depicted by the space-time correlogram or variogram? Discussion about these questions would be very interesting.
    We agree the selection of the HFT and HFA can be relatively subjective. We therefore mentioned in the text that other thresholds could have been selected. We think the HFT and HFA should just be determined by answering the following question: when and where floods could have been observed according to observed AND forecast rainfall fields? Even if several choices are possible, answering this question should lead to set a HFT and a HFA accounting respectively for timing and spatial errors of rainfall forecasts. The thresholds we applied were determined in this way. For the HFT, the threshold of 2 mm/h can be seen as relatively low, but it corresponds to a spatial average and thus reflects significantly larger point rainfall intensities. For the HFA, almost all the Aude river basin is covered with rainfall forecasts accumulations exceeding 150 mm, at least for the 75% and 95% quantiles. We consider these thresholds (2 mm/h of spatial averaged intensity and 150 mm of point rainfall accumulation) both result in significant risks of flooding. Extending the HFA outside the limits of the Aude river catchment would even have been possible (but not essential, since the Aude catchment already largely exceeds the area of the actually observed intense rainfall cell). Examining the spatio-temporal structure of observed or forecast rainfall could be an alternative, but in our opinion this would also necessitate to set thresholds and would also result in some subjectivity.

30. Page 14, lines 329-331. The text gives the impression that some members clearly overestimate the rainfall in the catchment. Although Fig. 5 shows that this is the case by a few mm/h, there are no individual members showing average rainfall accumulations over the catchment similar to those of the 75%- and 95%-percentiles. Instead, the maps of Fig. 6 (second and third rows) are most likely the result of different members showing the largest accumulations in different locations in the catchment. Consequently, to a good extent what is referred in the text as "false alarms" are mostly location errors.

    We agree that figure 6 is neither representative of individual members, nor of actual forecast rainfall accumulations, since it combines several runs of forecasts. It is rather a cumulated representation of the areas affected with high rainfall intensities, for the successive runs of forecasts and for a fixed lead time and quantile. We added a sentence in the beginning of this paragraph to remind this: "Note that the forecast panels do not correspond to rainfall accumulation for one unique run of forecast, but rather to a cumulated representation of the areas affected with high forecast rainfall intensities, for the successive forecasts issued during the event." We also completely agree that the spatial extent of the large rainfall rates in the second and third rows result from the location errors of the successive runs of forecast, and therefore added the following sentence in the paragraph: " .. the area of high intensities spreads and becomes larger than the area seen in the observed field of rainfall accumulations. This may be attributed to the location errors of some members in the successive runs of forecasts." We also modified the text of lines 329-331 to avoid any confusion : "Even if not entirely hit by the observed heavy precipitation event, the Aude River basin is almost entirely covered with repeated high forecast rainfall intensities during the event. This led to the choice of keeping the entire Aude River basin as HFA. Considering this whole area will help in evaluating the risks of false alarms attributed to rainfall forecast location errors when forecasting floods."

31. Page 15, line 339. "largest" could be replaced by "highest".

    This has been modified.

32. Fig. 6. It would be very useful to provide the values of the event accumulation in the catchment for each panel. My impression is that the 75% percentiles show significantly larger catchment accumulations than those observed, and probably a lower percentile would be closer.

    The values of event accumulation are represented in the observed panel (there is no quantile in the event observed accumulations as we one have one observation, not members). As mentioned in our answer to comment n°30, the right panels do not correspond to actual forecast rainfall accumulations, since they combine several runs of forecasts. Providing forecast rainfall accumulations for one single run of forecast would be possible, but this would not represent the event accumulation because of the limited lead time of 6 hours. .

33. Page 16, line 346-347: "As a consequence, to produce effective hydrological forecasts based on a good estimate of the rainfall rates. . . , users would need to work based on a high ensemble percentile value (the 75% percentile in the present case . . .)" I find this sentence misleading, as it could give the impression that this is the rainfall that has been used in the analysis (which would be contradictory with what is described in Appendix A, page 32, line 347, "for each considered forecast percentile"). Also, the discussion about how using a high percentile might generate false alarms could fit better in the discussion.

    As mentioned in appendix A, all percentiles from 5% to 95% are used to calculate the ROC curves presented in Figure 8. This phrase "As a consequence, to produce effective hydrological forecasts based on a good estimate of the rainfall rates. . . , users would need to work based on a high ensemble percentile value (the 75% percentile in the present case . . .)" just explains why a specific interest is given to the 75% percentile in section 4.2 (Figures 9 and 10). The description of figure 6 has been grouped and modified, and we believe there is no risk of misunderstanding anymore. The sentence about the general decision principle has also been moved to the discussion section.

34. Caption of Fig. 7. Mention the hourly rainfall thresholds for the presented ranked histograms.

    We completed the caption with the different thresholds used.

35. Discussion about Fig. 6 appears before and after the discussion about Fig. 7. Please, combine them (one option could be that Fig. 7 appears before Fig. 6).

We modified the text to group and simplify the description of Figure 6. Because Figure 7 is dependent on HFT and HFA, we prefer to show it after Figures 5 and 6.

36. Page 16, lines 360-361: “Hourly rainfall accumulations were uniformly disaggregated to run the model at a 15-min time resolution.” Why is this necessary?
The Cinecar model runs at a 15-min time resolution, whereas the rainfall observations and forecasts are provided at an hourly time step. We modified the sentence: "Hourly rainfall accumulations were uniformly disaggregated to fit the 15-min time resolution of the model."

37.) Page 16, lines 368-369 (“This means that one unique result (either a hit, a miss, a false alarm or a correct rejection) is obtained for each of the 1174 sub-basins”). Please, specify that this is for each probability value (see also comment 33).
The sentence has been modified : "This means that one unique result (either a hit, a miss, a false alarm or a correct rejection) is obtained for each of the 1174 sub-basins and for each ensemble percentile."

38. Page 16, line 371: By highlighting the 75% percentile in the ROC curve, it gives the impression that this result is obtained with the rainfall of Fig. 6 (see also comment 33), whereas this is the result obtained from setting a 75% on the forecasted discharges.
The ROC curves presented in figure 8 correspond to the hydrological forecasts (obtained with the Cinecar model). This is already mentioned in the caption of the figure, but we adapted the text to avoid any risk of misunderstanding. The 75% percentile has been highlighted here since it leads to balanced proportions of detections and false alarms.

39. Page 18, lines 385-386: “This is clearly the dominant effect for the 75% percentile of the pertDpepi ensemble product and the 2018 event.” Please, refer to Fig. 9.
The reference has been added in the text.

40. Figure 9, caption: “Maps of anticipation (0-6h) of the 10-year return period discharge threshold”. If I understand correctly, this is not what the figure shows.
The maps presented in this figure show the ability to anticipate a flow threshold exceedance for the 75% percentile for the RF0 scenario and the three hydrological ensemble forecasts. This is a spatial representation of the content of the contingency table for each outlet: hit (dark green), miss (dark red), correct rejection (light green) or false alarm (light red). We modified the caption to be more explicit: "Maps illustrating the detailed anticipation results (hits - misses - false alarms - correct rejections) of the 10- year return period discharge threshold, for the hydrological forecasts based on .." .

41. Page 19, lines 387-388: I would expect that the first point of the ROC for the 3 ensemble forecasts should be almost identical to that of RF0 scenario (which is almost the case). My interpretation is that the skill shown by the RF0 point (particularly the hits shown in Fig. 9) is due to the catchments’ response to past rainfall. Do you agree?
Yes we agree. By definition of the RF0 scenario, the False Alarm Rate equals 0 and the Probability of Detection corresponds to the combined effect of past precipitation and propagation times. Since the first point of the ROC curves for the ensemble forecasts corresponds to the 5% percentile, one would expect a POD that is at least slightly better than for the RF0 forecast. This is exactly what is observed.

42. Page 19, line 389: “All ensemble forecasts lead to an increase of the number of hits (9)”. Should “(9)” be “Fig. 9”?
The word "Figure" was missing. We corrected it.

43. Sections 4.2 and 4.3. The results of Section 4.2 were obtained with the CINECAR model, and those of Section 4.3 with GRSDi. If no comparison between models is provided, what is the advantage of using 2 different models? At least some discussion about the 4.2-Hydrological anticipation capacity of GRSDi should be provided.
The purpose of this work was not to compare the performance of the two hydrological models. Since the reference RS scenario corresponds to the simulated hydrographs, the evaluation of the hydrological ensemble forecasts do not highly depend on the model used. The advantage of using Cinecar for the evaluation of the 10-year discharge threshold (section 4.2) is that this model is highly distributed and enables to draw detailed anticipation maps including small ungauged basins (Figure 9). For hydrographs (section 4.3), we preferred here to use the GRSDi model since this model was not

exclusively calibrated on the 2018 event (even if the calibration period - October 2008 to October 2018 - includes the 2018 event), and thus the hydrological forecasts can be compared to both the RS scenario and the observed hydrographs to illustrate the importance of hydrological modeling errors in a real world forecasting situation (the GSRDi model is close to the models used for operational flood forecasting in France). A paragraph including these explanations has been integrated in section 3.4. More details about the application of the GRSDi model can be found in Peredo et al. (2022).

44. Page 20-21, lines 424 – 434. Please, add the reference to Figures 11 and 12.
    References have been added.

45. Page 21, line 441: it should be clarified how both the “spread” and the “skill score” have been calculated. Also, in the y axis of Figs. 11a-16a, it seems that the units spread / skill are mm. Is this correct?
    The spread/skill score has been calculated by following the methodology proposed by Fortin et al. (2014). While the methodology to obtain the skill score seems to be well known (root-mean-square error RMSE of the ensemble mean), the authors also explain that it is not the case for the spread and has previously led to some mistakes. They demonstrate that the ensemble skill should match the average of the standard deviation of ensemble forecasts, and they explain that the ensemble spread is in fact the square root of the average of the squared values of the standard deviations. They propose the following equation (equation 15 in Fortin et al. 2014).

$$spread = \sqrt{\left(\frac{R+1}{R}\right)}\left(\overline{s_t^2}\right)^{\frac{1}{2}} = \sqrt{\left(\frac{R+1}{R}\right)\frac{1}{T}\sum_{T=1}^{T} s_t^2} \qquad (1)$$

    Where R is the number of ensemble members, s is the standard deviation of ensemble forecasts, T the number of analyzed time steps, and t is a time step.

    We added a short explanation for the computation of the spread/skill score (section 2.3), and corrected the y-axis caption (adimensional), we thank the referee for helping us notice this mistake.

46. Figures 11-14: Some of the discharge forecasts show obvious biases with respect to the reference (simulated discharge). Some interpretation about this could be interesting. How do these biases affect the spread / skill results and their interpretation?
    Large bias between forecasted and simulated discharges should result in large skill values and therefore low spread/skill scores. Spread/skill scores significantly lower than the target value of 1 are observed only in the case of figure 14. But as already mentioned in the text, the reason in this case seems to be rather a timing error than a real systematic bias. In the case of figure 12, the large bias corresponds to a modelling error (difference between observed and simulated hydrograph) and not to a forecasting error (also mentioned in the text). Thus, in this case, the spread/skill score is logically not affected.

47. Figure 14 (panels b and c). The legend hides part of the results (observed and simulated discharges).
    We have modified the figure legends in order to avoid any covering of the discharge values.

48. Page 27: The title of section 5.2 could be more concise.
    We changed the title to "Performance of the three ensemble forecast products" instead of "What should be concluded about the comparative performance of the three forecast ensemble products evaluated?"

49. The study focuses on the evaluation of flash-flood hydrometeorological forecasts at the event scale. It could be interesting to add some discussion about how/if the method could be applied to evaluate the performance of the forecasting system on a multi-event framework. Also, it could be interesting to include some discussion about the applicability of the method to other regions and countries.
    Applying the method to other regions or countries prone to flash-floods should not cause any difficulty if the required data and models are available. A development on the question of applying the method in a multi-event and/or multi-site framework has been added in the discussion section: see our answer to general comment n°3.

50. The Introduction states that “We adopt the point of view of end-users, who aim at providing resources and assistance for evacuations and rescue operations at a regional scale.” However, I have not found any analyses or results supporting this statement beyond a few statements in sections 5 and 6 that are quite general.
    We reformulated this paragraph as follows: "In this approach, the evaluation is mainly focused on the capacity of the hydrometeorological forecasts to anticipate the exceedance of predefined discharge thresholds and to accurately localize the affected streams within the region of interest. These are two essential qualities of hydrometeorological forecasts that are needed to plan rescue operations in real time."