Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/nhess-2022-182-RC2, 2022

Comment on nhess-2022-182

Anonymous Referee #2

Referee comment on "A methodological framework for the evaluation of short-range flash-flood hydrometeorological forecasts at the event scale" by Maryse Charpentier-Noyer et al., Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2022-182-RC2, 2022

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.*

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.

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.

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?

MINOR COMMENTS

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

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.

3) Page 31, line 685: "to summary" could be "to summarize".

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

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."

6) Page 31, lines 704-706: Please, check the writing.

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.

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.

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).

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.

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.

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 subcatchments 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)?

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

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)".

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).

15) Page 9, line 226, (title of Section 3.3). For consistency, use "AROME" everywhere within the text.

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?

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?

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 stablished, 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.

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?

20) Page 10, lines 246-247: "it is a shift that does not combine too incompatible areas." Please, clarify.

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).

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?

23) Page 11, line 271. Please, specify that KGE stands for the Kling-Gupta efficiency, and provide a reference.

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.

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.

26) Page 11, line 291: What is "ANTILOPE J+1"?

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?

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.

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.

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.

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

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.

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.

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

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).

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?

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).

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.

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.

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.

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?

42) Page 19, line 389: "All ensemble forecasts lead to an increase of the number of hits (9)". Should "(9)" be "Fig. 9"?

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.

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

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?

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?

47) Figure 14 (panels b and c). The legend hides part of the results (observed and simulated discharges).

48) Page 27: The title of section 5.2 could be more concise.

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.

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.