Reply to Reviewer 2

The authors present a first attempt to develop a flood impact forecasting procedure that is fully integrated in a continental scale flood early warning system. They demonstrate this system by benchmarking various components against a flood events in May 2014 in Bosnia-Herzegovina, Croatia and Serbia. The paper builds on two directions of several previous works of the various authors: (1) the EFAS system that has previously been used for forecasting peak flows; and (2) the impact assessment module that has been used in several past risk studies for current and future conditions. In my opinion, this is a laudable effort – the need for such studies has been clearly vocalized in many past papers, and in many scientific and policy-related fora. I greatly appreciate the effort undertaken not simply to present the framework, but to try to benchmark it for an actual event. Of course, 1 event remains a limited benchmarking, but I believe that the benchmarking has been carried out in a way much more thorough to past studies in large scale risk modelling. The novelty here is not in the models themselves, which have been developed in pervious papers, but bringing them together for impact forecasting. The paper is well written and clear, and provides enough level of detail on the already developed models, without too much repetition.

We thank Reviewer 1 for his/her positive comments on our work.

I believe that the paper therefore is an important first step forward in this direction, and therefore merits publication in NHESS, subject to the authors being able to address the following issues:

1) L119-121: "In case thresholds are exceeded persistently over several forecasts, flood warnings for the affected locations are issued to the members of the EFAS consortium." Please explain this statement better: which thresholds? And what is meant by "over several forecasts"?

The thresholds mentioned are the local discharge values corresponding to 1, 2, 5 and 20-year return periods, calculated from the EFAS reference simulation (L 113-117). The persistence criterion is that the 5 year threshold must be exceeded over 3 consecutive forecasts. To clarify this part, lines 117-121 will be rephrased as follows:
"The reference simulation is also used to estimate discharge values for the return periods corresponding to 1, 2, 5 and 20-year at every point of the river network. All flood forecasts are compared against these thresholds and the threshold exceedance calculated. In case the 5 year threshold is consistently exceeded over 3 consecutive forecasts, flood warnings for the affected locations are issued to the members of the EFAS consortium. The persistence criterion has been introduced to reduce the number of false alarms and focus on large fluvial floods caused mainly by widespread severe precipitation, combined rainfall with snow-melting or prolonged rainfalls of medium intensity".

2) L161-162: "We first identify the maximum discharge predicted over the full forecasting period, calculated using the median discharge from ensemble forecasts at each river grid cell". It is not clear to me from this sentence how this works. Do you take the maximum discharge

across the entire ensemble for each lead time? (e.g. for lead time 1 day take the max discharge of all the ensemble members at 1 day lead) Or is something else meant here? Please clarify.

We first consider the median of the ensemble forecast, and then select the maximum discharge of the median over the full forecasting period (10 days). In the revised paper we will rewrite this sentence accordingly.

3) It is stated that the flood protection standards of Jongman et al. (2014) are used, and integrated with information from literature review and local authorities where available. In terms of transparency and reproducibility, I recommend a list (e.g Supplementary Information or in Appendix) showing the regions in which the values from Jongman et al were replaced, and which values were used.

Following the Reviewer's suggestion, the revised paper will include an appendix with a list of the updates and additions to the flood protection level map developed by Jongman et al.
The list will show the regions where values have been updated, the old and new values, and the source of information.

4) In the validation of the inundation maps, the authors have chosen only to report the hit rates. I find this problematic, as a (theoretical) model that greatly overestimates flood extent would tend to have very high hit rates. Therefore, in itself it only tells half the story. I believe that it would be more prudent to also report the false alarm ratios. This is especially important, since in Table 3 it is shown that the simulations show a much larger flooded area than the observed datasets, which could be leading to the high hit rates.

We agree with the Reviewer on that presenting the results also in term of overestimation is necessary. To this end, in the revised version Table 3 will include overestimation (or underestimation) ratios between simulations and all the available observations, to provide a more objective presentation of the results.
However, regarding the results in Table 4 we believe that it is more correct not to compute false hit ratio because, as discussed in the manuscript, we know that the available satellite flood maps underestimated the actual flood extent. As such, false alarm ratio scores would be low without being supported by reliable observations, giving an incorrect view of the performance.

5) With regards the validation of the flood risk (I think it would be better called "flood impacts"), expressed as affected population, on lines 414-415 it is stated that: ". . .results from the reference simulation match well figures reported for all the flooded counties of Croatia except for the Vukovar-Srijem County." This is a very subjective statement: how is "match well" defined? For example, in the Osjek-Baranja Country, the observed dataset reports 200 people, whilst the simulated dataset suggests 1300 – i.e. a difference of 550%. I realise that the definitions used in the simulated/observed datasets are different, and so the direct comparison is difficult, but it would be more transparent to report the differences openly than disguise relatively large differences with ambiguous language.

We agree on that the evaluation of results requires the use of a more precise language. In the revised manuscript, we will present both absolute and relative differences between observations and simulations, in order to provide a more objective discussion of results, and we will avoid ambiguous terms.
Also, we will carefully revise the use of terms "flood risk" and "flood impact" in the paper (see also the reply to Reviewer 1 for a more detailed discussion on this point).

6) One of the reasons given for the large difference in simulated damage between the reported and simulated dataset is that the damage curves applied have not yet been calibrated for Bosnia-Herzegovina, Croatia and Serbia. If this is the case, is it even useful to include this information in the warning?

The operational rapid risk assessment includes damage estimation because of specific requests of EFAS end users, therefore we deemed correct to show the results for the case study here presented, even if available data for validation are limited to Croatia and no country-specific damage functions are available. Moreover, from this point of view the test area is representative of the majority of European countries, which have not specific damage functions. Even considering the mentioned issues, we think that the application can provide useful information on the performance of the modelling framework.

7) In the conclusion, it is stated that the "Comparison of reported and simulated flooded areas suggests that the methodology enables to identify areas at risk well in advance. . ." Whilst the results do indeed show some encouraging skill, I think the phrase "well in advance" seems like oversell. The 12th May forecast for the 14th May flood showed little sign of flooding. The impacts were rather clear on the 13th May, giving a good confidence warning 1 day in advance. It is of course subjective whether 1 day is "well in advance" – it depends on the actions that planners need to take.

We apologize for not having been precise on presenting the performance regarding lead time. In fact, the timing of peak flow was rather variable across the Sava river basin, due to its extent. While in the Kolubara river the highest discharges occurred on 14th and 15th May, peak flows in other tributaries were reached later (between 14th and 16th for Bosna River, on 16th for Drina, 17th May for Sana River), and on the main branch of the Sava River the flood peaks occurred after 17th May. Thus, for the majority of affected areas the lead time was at least 2 days, if we consider the EFAS forecast issued on 13th May. In the revised paper we will evaluate the performance considering these additional details, and discussing emergency actions that could be taken base on available lead time.

Minor comments:

a) L60: the authors refer to a paper by Ward et al., 2016 to support the claim that "flood impact forecasts are increasingly being requested by end users of early warning systems". This facet is already discussed in Ward et al (2015), which would seem a more prudent paper to cite.

We agree with the Reviewer, in the revised manuscript the reference will be replaced as suggested.

b) L131: “we decided create” to “we decided to create” ; L222: wide spread to widespread; L368: “time o image” to “time of image”.

These typos will be corrected.

c) L179: Batista e Silva et al. (2012) → Batista and Silva et al. (2012)

The reference is actually correct, first author’s surname is “Batista e Silva”.