Referee report

nhess-2018-343 Submitted on 13 Nov 2018 Companion Paper nhess-2015-271

Ensemble flood simulation for a small dam catchment in Japan using nonhydrostatic model rainfalls. Part 2: Flood forecasting using 1600 member 4D-EnVAR predicted rainfalls

Kenichiro Kobayashi, Le Duc, Apip, Tsutao Oizumi, and Kazuo Saito

General comments:

The authors provide an interesting study about using ensemble forecasts with a very large number of members and subsequent reduction of the members for a more "practical" application in real time flood forecasting. The temporal and spatial error of NWP is less investigated than the error in the predictor. This paper addresses such issues and is of scientific interest.

The study is based on only one flood event, which makes conclusions very difficult. It is a major drawback, but an inherent problem in using a relatively new technique for forecasting very rare events. Therefore, I consider studies about single events useful for the flood forecasting community even if there should not be a significant advance of science. However, this must be made clearer.

Other authors have already done a lot of research in ensemble calibration or member reduction techniques for application in forecasting. E.g., methods based on Bayesian theory and regression techniques were used to assign weights to the members of an ensemble forecast based on prior information. Here, the authors should extend their literature study in their introduction (Reich and Cotter is "just" a text book citation). They could also add this aspect to the objectives of their work.

How stable is the selection of members with time? It seems to be relatively stable for their case study, but might happen that the selection must be updated very often. I have doubts if a single event study can provide a solution ready for operational forecasting. In the discussion, they should give more information or judgement if the proposed solution is transferrable to other events and make limitations clearer.

Specific comments:

Abstract: when introducing the „dynamical selection" of the best ensemble members (a kind of sub-ensemble), the authors should not refer to the criterion (NSE), which can be questioned, but mention the techniques applied. Furthermore, they should make clear that only a single extreme event was used for the study.

P 1 L 25ff: ensemble forecasts do not necessarily give the probability of occurrence of a flood – that is a common misunderstanding. A good ensemble gives information about the range of uncertainty ("frames the future development"), i.e. the observation should be within the uncertainty band. Most ensemble forecasts assume the same probability for each member and use frequency evaluations. Probability is obtained by data assimilation and prost-processing, as the authors have added in their revision. The authors could carefully revise their usage of probability in the text and check where frequency is a more appropriate term.

P 3 L 27: "The main theme of this Part 2 paper is that the 1600 ensemble rainfall forecasts can significantly improve the rainfall forecast over the large area around Kasahori dam": They should not give a theme with statements of their result, but first put the research question and objectives here. Also, after the revision, they have added work regarding the selection of best members in an operational case. It could be mentioned now in the objectives if not the title.

Section 2 is a very short – the content might be moved to section 1.

P 5 L 31: The rationale of the FSS should be briefly explained. Please explain the meaning of high and low values (just add sth. like “can have values between x and y, where y indicates the best possible score…”). Equations can be referred to by the citation. The reference (Duc et al., 2013) is not appropriate. The FSS is relatively new, so the original source must be cited here instead of own work of the authors using the FSS, which others proposed earlier.

P 5 L 34 “Note that an additional experiment with 4DEnVAR-NHM using 50 ensemble members”: how were the 50 members produced? Please give information (or a citation) about the differences in the ensemble generation mechanism of the 50- and 1600-member ensembles.

P 6 L 26: “but the ensemble mean precipitation is smeared out as a side effect of the averaging procedure”: then, the averaging procedure is maybe not a good solution. This is a well-known effect of ensembles of large size. Instead of averaging, other authors have used ensemble size reduction techniques. Finally, the authors did that but do not introduce at this stage of the manuscript (see comment above).

P 7 l 26 ff: The calibration of a hydrological model for a single event is questionable. Furthermore, using radar data instead of observations rises questions about the quality of the radar data, as can be seen later (fig. 13). Calibrating against “wrong” inputs produces higher uncertainty of the hydrological model, because it does not represent the physical processes well. The observed runoff is not a product of the radar data but a product of the observed rain. Parameters could get non-behavioral values in order to fit with the wrong rain input. As the authors assume a perfect hydrological model (without considering its uncertainty), it should be calibrated against the most perfect input data available. I think that the hydrological model is not valid here. However, the calibration and their discussion could be updated with a reasonable effort and the overall study is not about hydrology. If the radar data are used in operational service, but an error is known, the input data must be improved or post-processing can be applied, e.g. bias corrections. Research went a lot further in these topics.

P 8 L 7: In figure 4, it can be seen that the observation captured all three peaks quite well. Observed data show a consistent behavior. The hydrological model shows weakness in simulating double-peak flood waves. This would not necessarily prohibit it’s use for the study. The authors could add observed areal rainfall in Fig. 5 for better interpretation. From fig. 13, it is clear that there is an underestimation of rainfall by the radar products compared to the gauge stations. It would be good to use observed rain input to simulate stream flow as the reference for comparisons.

P 8 L 29: “observed rainfall within the range” – I think that Fig. 6 only shows runoff ensembles, so instead of “rainfall” they should use “runoff” here.

P 9 L 2 ff: see comment for fig. 7.

P 10 L 17 ff.: The authors should add if the NSE is of stream flow. Why not choose the best 50 members of the rain forecast?

Fig. 1: what is the meaning of the spatial scale? I did not find an explanation in the text or the caption.

Fig. 2: Y axis is “Observed Relative Frequency” and should be labelled accordingly. The reliability diagram is not intuitive, in my opinion not even appropriate for a single event situation – even if recommended by one reviewer. Usually, there are not such pairs like 90 % forecast probability and 0% observed frequency. For small data sets, other authors have used bootstrapping to refine the probability distributions. However, from a single event, one cannot conclude the reliability of a certain forecast technique. The authors may re-consider their usage of reliability diagrams in that

context. Fig. 1 gives a good idea of the performance of the different systems for different rain intensities. I propose to leave out the reliability diagrams here, or replace with a more suitable performance measure for the single event. Maybe just give the Brier score or other metrics, as mentioned by reviewer #2.

Fig. 4: does R/A stand for “observation”? Is that radar composite or ground based? The plot is not easy to understand. It would be better to draw the 5-95 percentiles as light gray, the iqr as darker gray areas and the observed/derived lines as such, in colors. As done in fig. 8.

Fig. 6: Using the same style as figure 8 would be more informative. Then, figure 8 could be left out.

Fig. 7: the NSE is not good in characterizing the performance of stream flow ensembles of a single event. It is usually applied for calibrating hydrological models against observations, and more common for long time series. I propose to leave out fig. 7, and remove the corresponding text. It does not add to the findings but rises questions.

Fig. 10: again, the style of fig. 8 would be better here.