Nat. Hazards Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/nhess-2018-343-RC2, 2019

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "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" *by* Kenichiro Kobayashi et al.

Anonymous Referee #2

Received and published: 11 March 2019

This manuscript presents the ensemble streamflow forecast forced by a large number of rainfall ensemble at a high resolution. It is a follow-up study of a NHESS paper published in 2016. The main difference is the number of ensemble from 11 to 1600 while the catchment setup is nearly identical. The key idea, use of a large number of rainfall ensemble, is of interest and worth testing since lack of diversity in ensemble has been commonly used as a good excuse for explaining poor performance of hydrologic forecasting. Now, there are 1,600 rainfall ensemble. Can such a large number of rainfall ensemble significantly improve the streamflow forecasting? What factors play

an important role in the improved forecast? These are critical questions hydrological and meteorological communities have been pursuing for a long time. Unfortunately, these questions couldn't be properly answered in the manuscript. For convincing potential readers with new evidence, the experimental setup and analysis methods need significant changes. However, the required changes are too enormous. I have nothing but suggest reject & resubmit. If this manuscript is accepted despite my suggestion, I request the following comments would be addressed before publication.

- Validation at multiple streamflow gauges: Impact of large ensemble forcing should be estimated on multiple gauging locations. If findings are based on the results from a single streamflow gage, feasibility of flood forecasts cannot be claimed while any conclusions can be considered site-specific. In my view, new locations for ensemble verification don't have be limited to dam reservoirs. Any streamflow gages affected by the extreme rainfall are encouraged to be included.

- Selection of a proper size of ensemble: The later part of the result section is about how to find good rainfall ensemble members among 1600 for better streamflow forecasts. The conclusion is vague while all additional efforts are left as future research. Anything additional should be done to draw meaningful findings on this topic. For example, what statistical features do good ensemble have? In addition to rainfall, other meteorological variables may be examined together for analyzing good ensemble. How different or similar ensemble are selected at each time step compared to the previous steps? These questions also should be addressed for multiple gauges, not for a single gauge. If necessary, some machine learning approaches the authors mentioned, e.g. SOM and SVM, may be used in the current manuscript rather remaining them in the future topics.

- To give up the selection of the best ensemble: In the last part of the result section, although it was failed, the authors discussed the possibility of selecting the best discharge simulation using the best rainfall ensemble from 1600. I highly disagree with this idea because ensemble approaches were introduced to overcome the limitation of

deterministic approaches.

- Probabilistic verification: Although this study is about ensemble forecasting, all measures are deterministic, no probabilistic measures are not used for verification of probabilistic forecasts. Since ensemble forecasts aim at providing not only better averages from ensemble but also predictive uncertainty, adequacy of ensemble spread is critical to assessing probability of flooding risks and there are common metrics used in hydrological and meteorological communities for assessing reliability, discrimination, resolution, and sharpness of ensemble. Such metrics should be estimated and discussed.

- Please elaborate why the different number of ensemble was used for each analysis. For figures 11, 13(a), and 13(b), the number of selected ensemble varies from 38 to 26.

- Figure 14. It is negative that any meaningful findings come from simulation results whose NSE values are less than -1. This figure is comparing NSE ranging from 1 to -7.

- Given that the authors also admitted the accuracy of radar rainfall is better than that of NWP, why didn't you use radar rainfall as input for hydrologic modeling in the past time steps? If NWP ensemble are used only for forecasting steps, as most operational models are doing, generally forecast performance is expected to be better.

- Review in Introduction: A simple summary of several papers should be avoided. Previous papers should be used to show how research questions or gaps the current study is dealing with are addressed and remain unsolved.

- The summary of K Project should be removed and, if required, moved to Acknowledgement section because the exascale computing is far from the scope of this journal, despite its importance to the motivation or institutional support to this study.

- Section 4 and several figures on the catchment are nearly identical to Section 3 and

associated figures in the 2016 NHESS paper, which should be considered as self-plagiarism if not cited properly.

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2018-343, 2019.