Authors' response

We would like to thank the editor and two reviewers for the detailed and thoughtful comments and the time that you spent on the article. All your invaluable comments have inspired us to rethink carefully the problem that we tried to investigate. With our revision we hope that you will find it a better article.

Regarding the manuscript with track changes, we did not use the MS-word track change function since the changes are really a lot so that the manuscript became hard to be understood with the complex track change indications. Instead, we highlighted the changes by ourselves using different colors. If this is a problem, please tell us.

In addition, we have modified the abstract according to the comments from the editor and two reviewers to show more clearly about the object of our Part 2 paper. We tried to answer the rest concerns from the editor in the responses below. We hope it can satisfy you.

Finally, we would like to change the order of the authors as indicated in the revised manuscript. We hope that you can accept the changes.

Reviewer 1:

(1) The paper describes the application of a large (1600 members) ensemble of high-resolution rainfall forecasts for flood forecasting in a small (72 km^2) catchment where the lead time required to respond to a flood warning is longer than the characteristic response time of the catchment. This is an important issue for unban catchments where hydrological predictions need to be based on rainfall forecasts and not observed rainfall.The temporal resolution should always be included when discussing resolution. I assume that the ensemble had 10-minute resolution since this is used by the hydrological model.

Reply: Since 4D-EnVAR-NHM outputted data every hour due to limited data storage, we also applied the hourly data to the rainfall-runoff model.

(2) Section 2 – details of the rainfall event. I looked up the 2016 paper for more details of the meteorological situation, but found very little extra. It would be very helpful to understand better the meteorological situation. I am assuming that, since this case is in Japan and summer, the situation was mostly orographic triggering of severe convection in a very moist airmass. This implies that the model rainfall forecasts are closely forced by the topography where the storms are initiated in the near vicinity of the catchment and are likely to be slow moving? This is important because advection nowcasts will not be able to provide accurate nowcasts in these circumstances.

Reply: We have added a paragraph in the revised manuscript to analyze the meteorological situation in more details, which we reproduce here:

For the details of the 2011 Niigata-Fukushima heavy rainfall, see our Part 1 paper (Kobayashi et al., 2016). An additional note is that the torrential rain of the 2011 Niigata-Fukushima heavy rainfall occurred over the small area along the synoptic scale stationary front (for surface weather map, see Fig. 1 of Kobayashi et al. 2016). Saito et al (2013) conducted two 11-member downscale ensemble forecasts with different horizontal resolutions (10 and 2 km) for this event using JMA-NHM and JMA's global ensemble EPS perturbations. They found that the location where intense rain concentrates variable to small changes of model setting, thus the position of the heavy rain was likely controlled mainly by horizontal convergence along the front, rather than the orographic forcing.

(3) I really missed some radar rainfall images, say the 10 (or 30)-min rain rates at the times of the three peaks in the hydrograph, just so that we can get a feeling for the space-time structure of the rainfall fields. Actually, the spatial and temporal correlation functions would also be interesting, at least to me as a rainfall person.

Reply: For supplement information, we would like to show here the radar images corresponding to the times when the three peaks occurred in the hydrograph (Figure S1). We do not intend to add this figure into the manuscript.

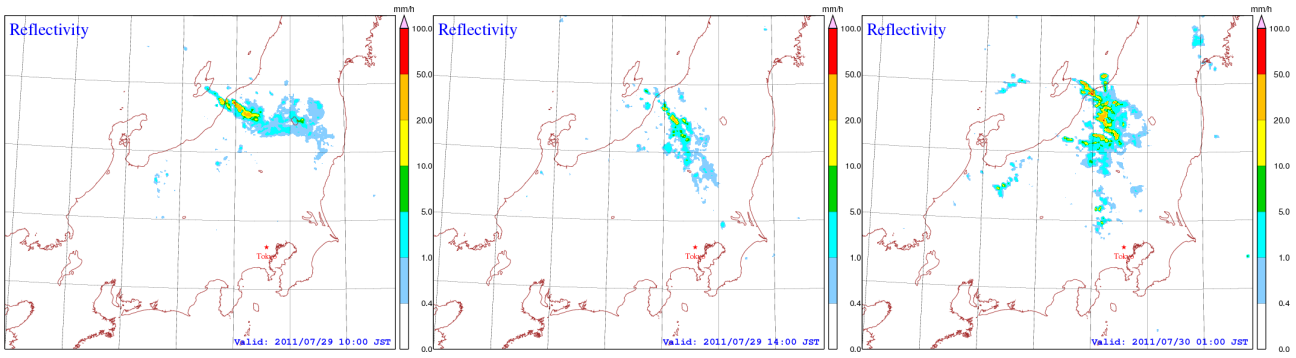

Figure S1. Reflectivity from radar composition at the time of the first (left), second (center), and third (right) peaks of the hydro-graph.

(4) Section 6 – Results. It would be very good to extend the results to include a basic analysis of the rainfall forecasts before going to the hydrological verification. In particular, how reliable are the probability of precipitation estimates for the extreme rain rates, especially as a function of ensemble size? This is very important if we are running an ensemble prediction system to predict the probability of extreme rainfall. The paper should include some results that show the skill of

the model, say the reliability diagram for high rain rates, as a function of lead time. Did subsequent model runs reproduce the second and third maxima in the hydrograph?

Reply: We have added two paragraphs to Section 3 describing the verification results of the rainfall forecasts. We agree with the reviewer's comment that verification scores for rainfall forecasts with respect to different lead times should be included in the paper. However, it was very costly to run a high-resolution (2 km grid spacing) ensemble forecast using 1600 members even for a specific time. Due to this reason, we could only run deterministic forecasts for all other initial times and use the Fraction Skill Score to measure forecast performance of the deterministic forecasts at different lead times. We had also run an additional experiment using only 50 ensemble members to compare with the case using 1600 members. Reliability diagrams are then plotted for these ensemble forecasts by the two experiments, even though we only run the ensemble forecasts at a specific time. Of course, the FSSs are also calculate for this additional experiment. Here are the paragraphs that we have added to the revised manuscript:

"Due to limited computational resource, ensemble forecasts with 1600 members were only employed for the target time of 0000 JST July 29th, 2011. However, deterministic forecasts were run for all other initial times to examine impact of number of ensemble members on analyses and the resulting forecasts. Figure 1 shows the verification results for the 3-hour precipitation forecasts as measured by the Fraction Skill Score (FSS) (Duc et al., 2013). Here we aggregate the 3-hour precipitation in the first and second 12-hour forecasts to increase samples in calculating the FSS. By this way, robust statistics are obtained but at the same time dependence of the FSS on the leading times can still be shown. Note that an additional experiment with 4D-EnVAR-NHM using 50 ensemble members, which is called 4DEnVAR50 to differentiate with the original one 4DEnVAR1600, was run. It is very clear from Figure 1 that 4DEnVAR1600 outperforms 4DEnVAR50 almost for all precipitation thresholds, especially for intense rain. Also for high rain-rate, compared to JNoVA, 4DEnVAR1600 forecasts are worse than JNoVA forecasts for the first 12-hour forecasts, which can be attributed to the fact that 4D-EnVAR-NHM did not assimilate satellite radiances and surface precipitation like JNoVA. However, it is interesting to see that 4D-EnVAR-NHM produces forecasts better than JNoVA for very intense rains for the next 12-hour forecasts.

To check reliability of the ensemble forecasts, reliability diagrams are calculated and plotted in Figure 2 for 4DEnVAR1600 and 4DEnVAR50. Since JNoVA only provided deterministic forecasts, reliability diagram is irrelevant for JNoVA. Note that we only performed ensemble forecasts initialized at the target time of 0000 JST July 29th, 2001 due to lack of computational resource to run 1600-member ensemble forecasts at different initial times. Therefore, the same strategy of aggregating 3-hour precipitation over the first and second 12-hour forecasts in calculating the FSS in Figure 1 is applied to obtain significant statistics. Clearly, Figure 2 shows that 4DEnVAR1600 is distinctively more reliable than 4DEnVAR50 in predicting intense rain. While 4DEnVAR50 cannot capture intense rain, 4DEnVAR1600 tends to overestimate areas of intense rain. The tendency of

overestimation of 4DEnVAR1600 becomes clearer if we consider the forecast ranges between 12 and 24 hours. However, for the first 12 hours, 4DEnVAR1600 slightly underestimates areas of light rains. This also explains why the FSSs of 4DEnVAR1600 are smaller than those of 4DEnVAR50 for small rainfall thresholds in Figure 1."

Since we only run deterministic forecasts for other initial times, we show here the forecast results for other lead times as supplement information (Figure S2). Again, we do not intend to add this figure into the manuscript. It turns out that it is more difficult to forecast the second and third peaks in the hydrograph.

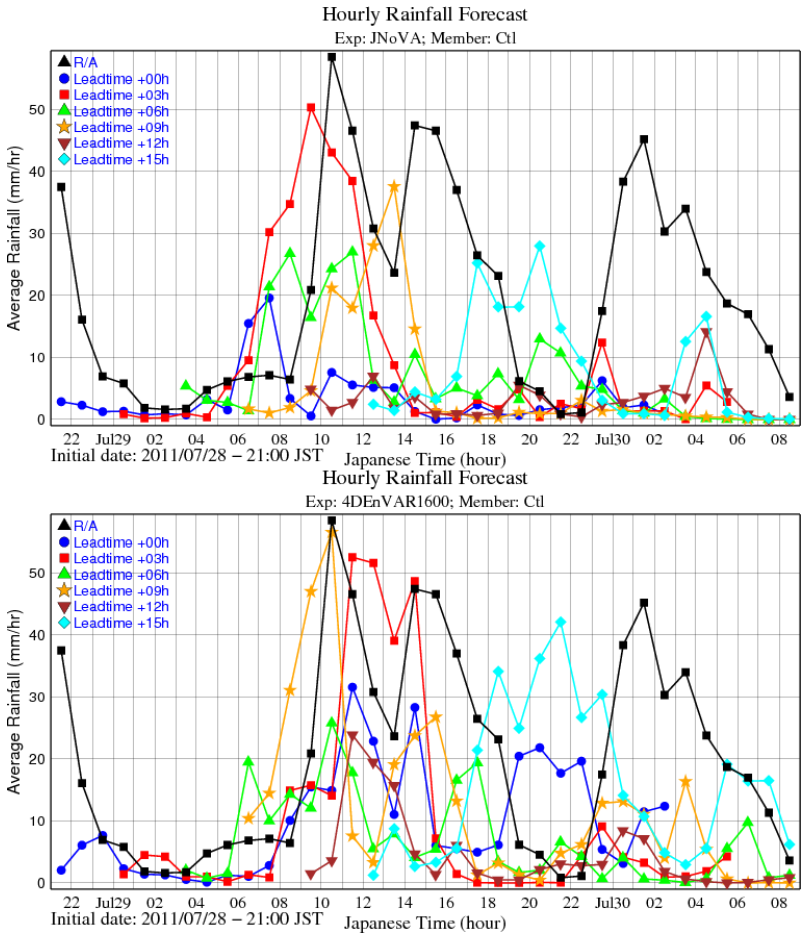

Figure S2. Time series of one-hour accumulated rainfall over the catchment by deterministic forecasts of JNoVA (top) and 4DEnVAR1600 (bottom) at different lead times.

(5) I really liked Figures 7, the probability of the inflow exceeding a critical threshold, and 10, the probability of an emergency operation, as examples of probabilistic products that meets the needs of an end-user. Once again, it would be interesting to see these products for a range of lead times. Regarding Figure 7, moving the forecasts around in time did not improve the results, but what about moving the ensemble in space? Generally, I find that the NWP rainfall forecasts that I work with have limited skill at scales that are below around 100 km. I assume that the rainfall in this

case is strongly influenced by the topography so you would not want to shift the rainfall fields too much, but it would still be interesting to move them around by a few tens of km.

Reply: As explained above, our computational resource can only afford running 1600-member ensemble forecast for a specific time. Although it's desirable to know the forecasts as plotted Figure 7 for other lead times, limited computation resource prevented us to employ this. In design the plot in Figure 7, we introduced the idea of using spatial and temporal uncertainty in verification from the FSS into the hourly discharges. It is clear that hourly discharges have strong correlation with hourly precipitation. Then it is reasonable to consider temporal uncertainty in hourly precipitation. Since the rainfall over the catchment here is not rainfall at any specific grid point but rainfall over many grid points (more than 70 km^2 in our problem). Therefore, temporal uncertainty is more relevant to hourly catchment rainfall rather than spatial uncertainty. Also, computation with spatial uncertainty is more complicated in this case since we must consider all directions of displacement vectors in a two-dimensional space, which have more degree of freedom that just one direction in the one-dimensional space of temporal uncertainty. Therefore, we do not consider spatial uncertainty in plotting Figure 7 and 10.

Nevertheless, please kindly note that the supplemental information for the spatial uncertainty was then written to reply the reviewer 2 (1) comment.

(6) The conclusion that it is difficult to select a "set of best ensemble members" based on past performance is significant, if a little discouraging.

Reply: The second reviewer has also shown interest on this problem. In the revised version, we have tried to established a theoretical framework to support the idea of "the best ensemble members". From this theory, we have explained why the best ensemble members based on past performance can vary considerably in time and this makes selection of best ensemble members difficult.

Reviewer 2:

(1) 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.

Reply: First, we agree with the reviewer that impact of large weather ensemble forcing on hydrological forecasts are worth to examine in details. However, this interesting research topic itself is not the main topic that we would like to pursue this time in this paper as the Part 2 in a series on the hydrological forecast for the Kasahori dam. Our purpose in using the 1600-member ensemble forecast was only that we have found the rainfall forecast for the large area covering the small catchment of the Kasahori dam was significantly improved when using this ensemble forecast and we have hoped that this helps, as a result, to improve the streamflow forecast for the Kasahori dam. The impression that the topic of our research is on impact of large ensemble forcing on hydrological forecasts will not occur if, instead of the 1600-member ensemble forecast, we used a 50-member ensemble forecast in the paper. In fact, we have tried to use a 50-member ensemble forecast with the same data assimilation system 4D-EnVAR-NHM, however the performance of the rainfall forecast was not better than the operational forecast that we used in Part 1. Of course, the improvement rooted in the use of very large ensemble members and this problem involved to the way that this large ensemble members mitigate site-effects of localization in 4D-EnVAR-NHM by removing vertical localization and imposing a unique horizontal localization length scale for all variables at all vertical levels. We have described these special characteristics of 4D-EnVAR-NHM in the text.

Likewise, as the information from hydrological aspects, we would like to show Figures S3 and S4 below (Figures 6 and 7 in the revised manuscript this time). Figure S3 shows the comparisons of the hydrographs of (a) 11 discharge simulations in Part 1 (cited from Kobayashi et al., 2016), (b) same 11 member but with a positional shift in Part 1 (Kobayashi et. al. 2016), (c) 50 discharge simulations with 4D-EnVAR-NHM and (d) 1600 discharge simulations with 4D-EnVAR-NHM. In addition, Figure S4 shows the histogram of NSE based on the simulated and observed discharges for the 4 cases.

First, Figure S4 (c) and (d) are showing the large improvement of 1600 ensemble discharge simulation in terms of NSE compared with 50 ensemble simulations. Second, Figure S4 (a) and (c) are showing that the performance of the discharge forecast with 50 member 4D-EnVAR-NHM was not necessarily better than the discharge forecast with the forecast in Part 1 in terms of NSE. Likewise, Figure S4 (b) shows that the discharge forecasts could be improved by positional shift of the rainfall field in Part 1, though this positional shift still needs further statistical verification with more rainfall events.

With these analysis, the explanation below is added in pages 8 and 9 of the manuscript.

"Figure 6 shows the comparisons of the hydrographs of (a) 11 discharge simulations in Part 1, (b) same 11 member but with a positional shift in Part 1, (c) 50 discharge simulations with 4D-EnVAR-NHM and (d) 1600 discharge simulations with 4D-EnVAR-NHM. Note that the duration of the 4D-EnVAR-NHM ensemble weather simulation is 30 hours from 0000 July 29th to 0700 July 30th JST, but the ensemble flood simulation is carried out only for 24 hours from 0300 July 29th to 0300 July 30th, 2011 JST since we consider that JMA-NHM uses the first 3 hours to adjust its dynamics. The result in Figure 6 (d) shows that, except for the third peak, the 1600 ensemble inflows can encompass the observed rainfall within the range, which was not realized in Part 1 with 11 downscale ensemble rainfalls of 2 km resolution (Figure 6 (a)). In other words, the extreme rainfall intensity of the event can be reproduced by the ensemble members with 1600 4D-EnVAR-NHM. Likewise, comparing Figure 6 (c) and (d), the simulated discharges by 50 ensemble rainfalls of 4D-EnVAR-NHM encompass the observation within the range less than those of 1600 ensemble members. In addition, Figure 7 shows the histogram of NSE based on the simulated and observed discharges for the 4 cases. Looking at Figure 7 (c) and (d), clearly, the 1600 ensemble discharge simulations outperform 50 ensemble simulations. The NSE>0 was around 17.75% (284 members) in 1600 ensembles, while it was 0% in 50 ensembles. On the other hand, Figures 7 (a) and (c) show that the performance of the discharge forecast with 50 member 4D-EnVAR-NHM was not necessarily better than the 11 discharge forecast in Part 1 in terms of NSE. Likewise, Figures 6 (b) and 7 (b) show that the discharge forecasts could be improved by the positional shift of the rainfall field in Part 1, though this positional shift still needs further statistical verification with more rainfall events."

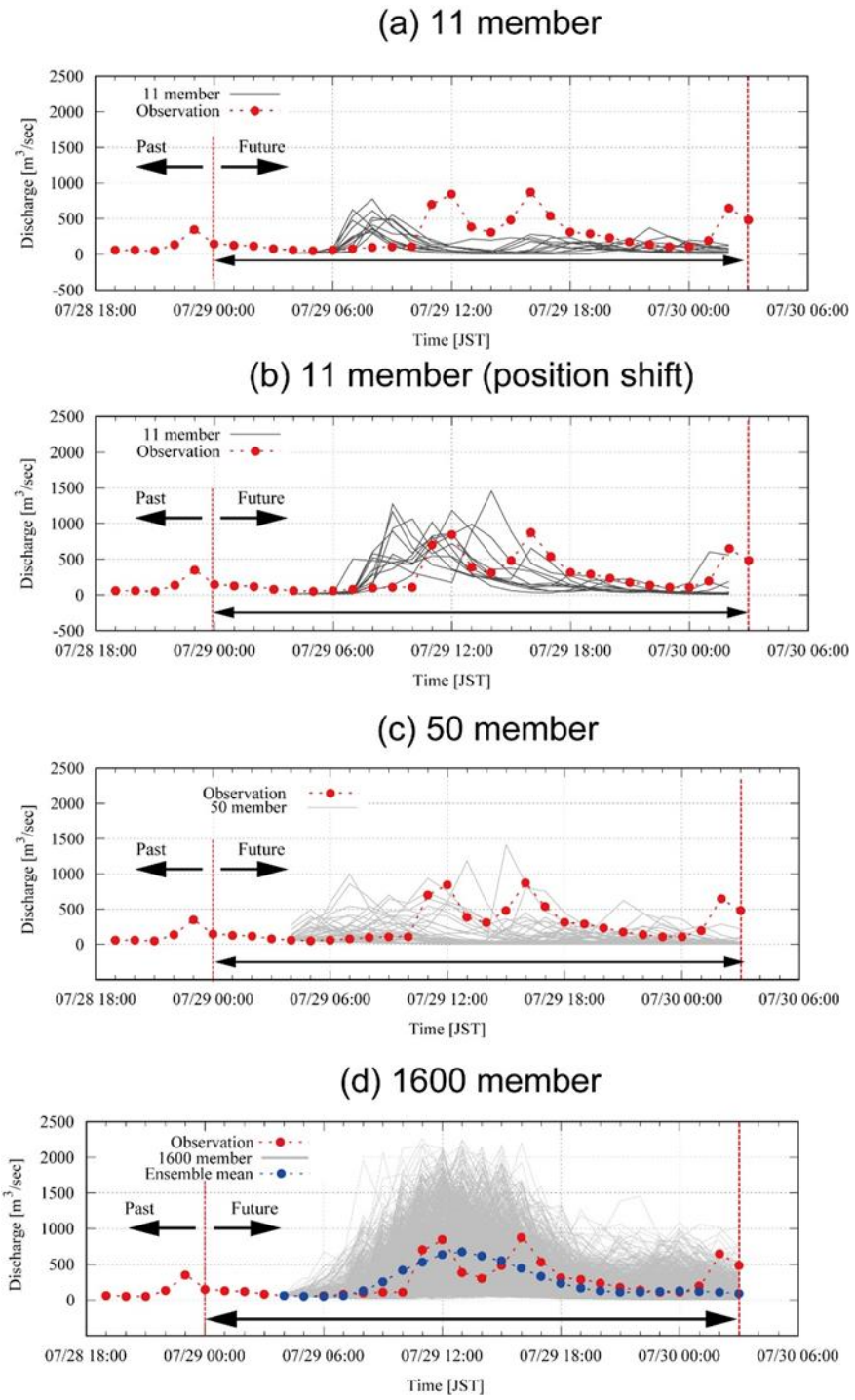

Figure S3. Hydrographs of (a) 11 discharge simulations in Part 1 (Kobayashi et al., 2016), (b) same 11 member but with a positional shift in Part 1, (c) 50 discharge simulations with 4D-EnVAR-NHM and (d) 1600 discharge simulations with 4D-EnVAR-NHM.

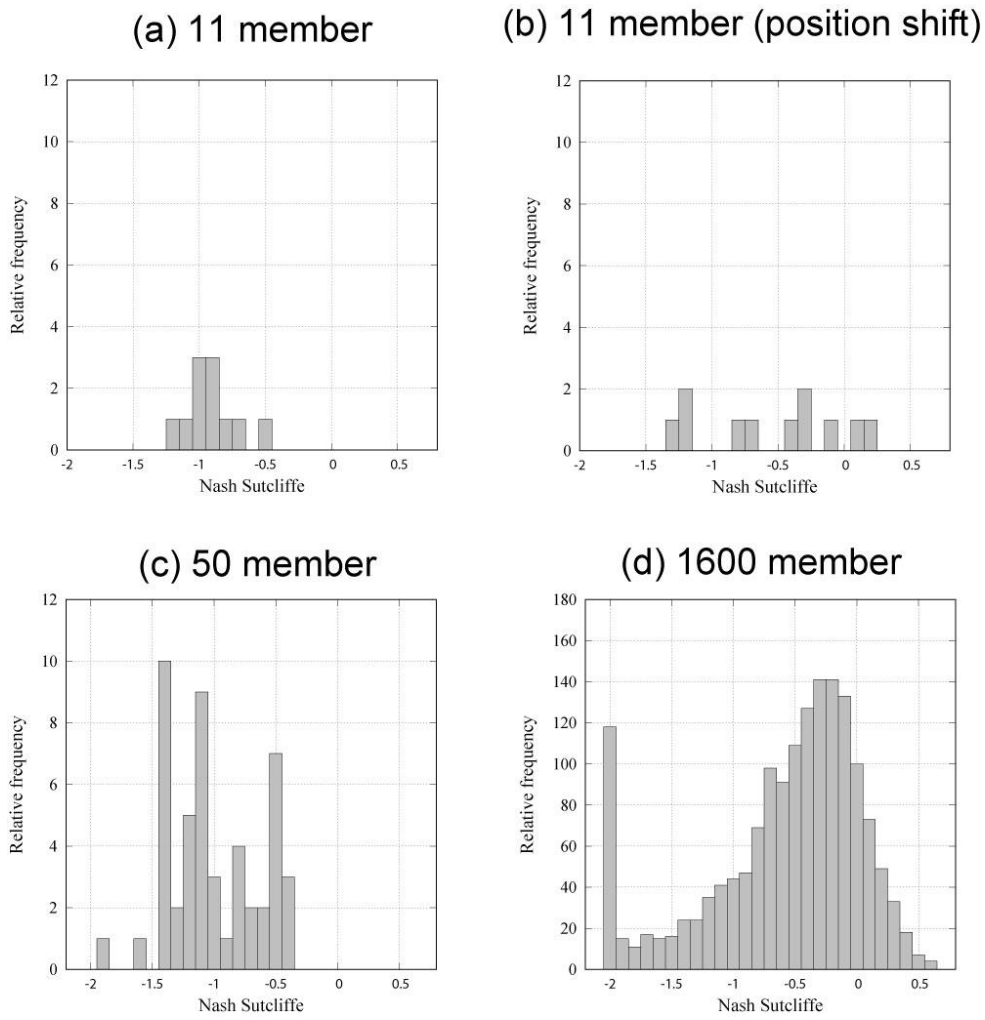

Figure S4. Histograms of NSE based on the simulated and observed discharges to Kasahori dam: (a) 11 discharge simulations in Part 1 (Kobayashi et al., 2016), (b) same 11 member but with a positional shift in Part 1 (Kobayashi et. al. 2016), (c) 50 discharge simulations with 4D-EnVAR-NHM and (d) 1600 discharge simulations with 4D-EnVAR-NHM.

(2) - 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 gauge, 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.

Reply: As we have explained in the previous answer, our research focused on the hydrological forecast for the Kasahori dam but not on impact of large ensemble forcing. Thus, it is inappropriate here to verify the hydrological ensemble forecast at all gauging locations in the domain covered by the meteorological forecast. We agree with the reviewer that if the topic is on impact of large ensemble forcing, conclusion should be relied on verification at all gauging locations. However, in this case verification for rainfall forecast over the whole domain is enough to draw conclusion on

performance of the ensemble forecast. We have already provided this kind of verification in Section 3 of the manuscript.

(3) - 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.

Reply: We agree with the reviewer that this section lacks a rigorous theory on guiding to choose the best ensemble members that supports our conclusions. We have added this theory on the revised manuscript that we would like to copy here:

"Clearly, the flood forecasting becomes very useful if we could just select the best ensemble members in advance. Logically, this is impossible since we only know the best members after knowing the observations which enable us to compute verification scores like NSE. This raises the question whether or not the best ensemble members can be inferred from the partial information provided by the observations at the first few hours. It is easy to see that the answer should be negative due to nonlinearity of the model and the presence of model error: the best marching at the first few hours is almost certainly not the best marching over all forecast ranges. However, it is obvious that the observations at the first few hours have a certain value which can help to reduce uncertainty in the ensemble forecast if we could incorporate this information into the forecast.

This procedure has already been well-known under the name "data assimilation" in which we assimilate the observations at the first few hours to turn the prior probabilistic density function (pdf) given by the short-range forecasts into the posterior pdf given by the analysis ensemble (Reich and Cotter, 2015). Thus, if we know the observations at the first few hours, we should assimilate these data to replace the short-range ensemble forecasts by the ensemble analyses at these hours, then run the model initialized by the new ensemble to issue a new ensemble forecast. As a result, we should replace the definition of the best members based on verification scores to a more appropriate one based on the posterior pdf. Here, we identify the best members with the most likely members. Clearly, if we assume the posterior pdf is unimodal, the best members should be the members clustering around the mode of this pdf, which is also the analysis. However, it is not clear how to identify the best members if this pdf is multimodal.

To overcome this problem, we will use the mathematical framework settled up by particle filter (Doucet et al., 2001). Let us denote the short-range forecasts by $\mathbf{x}_1$ to $\mathbf{x}_K$ where K is the number of ensemble members. The short-range ensemble forecast therefore yields an empirical pdf given by the sample ($\mathbf{x}_i$, $w_i^{pre} = 1/K$) with w_i^{pre} denoting the equal weight for the i-th member

$$p_X(\mathbf{x}) = \sum_{i=1}^{K} w_i^{pre}\, \delta(\mathbf{x} - \mathbf{x}_i) = \sum_{i=1}^{K} \frac{1}{K}\, \delta\,(\mathbf{x} - \mathbf{x}_i). \quad (3)$$

Using this prior pdf as the proposal density, the posterior pdf has the following form

$$p_X(\mathbf{x}|\mathbf{y}) = \sum_{i=1}^{K} w_i^{post}\, \delta(\mathbf{x} - \mathbf{x}_i) = \sum_{i=1}^{K} \frac{p_Y(\mathbf{y}|\mathbf{x}_i)}{\sum_{j=1}^{K} p_Y(\mathbf{y}|\mathbf{x}_j)} \delta(\mathbf{x} - \mathbf{x}_i). \quad (4)$$

Here, $p_Y(\mathbf{y}|\mathbf{x}_i)$ denotes the likelihood of the observations y conditioned on the forecast $\mathbf{x}_i$, and the weight w_i^{post} are the relative likelihoods. Moreover, it can be shown that $p_Y(\mathbf{y}|\mathbf{x}_i)$ is the observation evidence for the ith member (Duc and Saito, 2018). Then applying the model M as the transition model, the predictive pdf is given by

$$p_X(\mathbf{x}|\mathbf{y}, \mathrm{M}) = \sum_{i=1}^{K} w_i^{post}\, \delta(\mathbf{x} - \mathrm{M}(\mathbf{x}_i)). \quad (5)$$

This equation shows that the contribution of each member to the predictive pdf is unequal, which differs from the prior pdf (3). While the members with large values of w_i^{post} dominate the predictive pdf, those with very small values of w_i^{post} can be ignored. This suggests that the best members can be identified with the largest values of w_i^{post}. Thus, if we sort w_i^{post} in the descending order, the first N weights are corresponding to the first N best ensemble members. In this case, the predictive pdf (5) is approximated by

$$p_X(\mathbf{x}|\mathbf{y}, \mathrm{M}) = \sum_{i=1}^{N} \frac{p_Y(\mathbf{y}|\mathbf{x}_i)}{\sum_{j=1}^{N} p_Y(\mathbf{y}|\mathbf{x}_j)}\, \delta(\mathbf{x} - \mathrm{M}(\mathbf{x}_i)). \quad (6)$$

Note that by introducing the notion of the best ensemble members, a substantial change occurs, that is we now work with a unequal weighted sample $(\mathbf{x}_i,\ w_i^{post})$. This should be taken into account in computing statistics like ensemble mean from the best ensemble members.

If the likelihoods have the Gaussian form

$$p_Y(\mathbf{y}|\mathbf{x}_i) \propto \exp\left[-\frac{1}{2}(\mathbf{y} - \mathrm{h}(\mathbf{x}_i))^T \mathbf{R}^{-1} (\mathbf{y} - \mathrm{h}(\mathbf{x}_i))\right], \quad (7)$$

where h is the observation operator, and R is the observation error covariance, it is easy to see that the largest weights are corresponding to the smallest weighted root mean square errors (WRMSE)

$$WRMSE_i = (\mathbf{y} - \mathrm{h}(\mathbf{x}_i))^T \mathbf{R}^{-1} (\mathbf{y} - \mathrm{h}(\mathbf{x}_i)). \quad (8)$$

Therefore, if R is a multiple of the identity matrix I, the WRMSEs become the RMSEs, which in turn are equivalent to the NSEs. This shows that selection of the best members based on verification scores over the first few hours is in fact selection of the best members based on the relative likelihoods in the posterior pdf. It can also be understood as model selection based on observation evidence (Mackay, 2003)."

The theory can be traced back to the theory of data assimilation in which the mathematical form of particle filter can be served as a foundation for choosing the best ensemble members. Relied on available observations, a weight can be assigned for each ensemble members, which in fact represents the likelihood of the observations conditioned on this forecast. This reduces to verification scores for each member depending on the form of the likelihood. We would like to remark here that the theory is applied for any number of ensemble members.

Thus, statistical features that define the best ensemble members are the relative likelihood $w_i^{post} = \frac{p_Y(\mathbf{y}|\mathbf{x}_i)}{\sum_{j=1}^{K} p_Y(\mathbf{y}|\mathbf{x}_j)}$. These likelihoods vary with the lead time, therefore at different time steps, we will have different good ensemble members. The observation $\mathbf{y}$ here denotes all kinds of observations that means the relative likelihoods encompass not only rainfall but also other meteorological variables available at the first few hours. However, if we have already known the meteorological observations, it is better to run the next assimilation cycle, and using the new analysis ensemble to produce the new ensemble forecast. It is more practical to assume that we only know observations of discharge and rainfall at the first few hours.

(4) - 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.

Reply: We have removed the call for machine learning methods in the manuscript since this is not relevant now. On the other hand, we have added the text below to the page 13 of the revised manuscript.

"Herein lies the problem that, NSEs are quite sensitive to spatial and temporal displacement errors in rainfall. In principle, it is possible to introduce those errors into NSEs in a way similar to FSSs. However, it should be cautious in introducing such errors into NSEs before investigated well, although this type of approach has been used frequently in meteorology community. How to incorporate them qualitatively is also a problem to be addressed."

We have tested to incorporate the temporal displacement errors of rainfall into discharge NSE. One result is shown in Figure S5 below. The figure shows NSE considering temporal shift of ±1, 2 hr. It shows clearly that the number of discharge simulation members with NSE >0 increases, e.g. if we consider time-lag of +1hr. Nevertheless, as this analysis needs more elaboration, we would like to show them temporally as your information in this reply letter only.

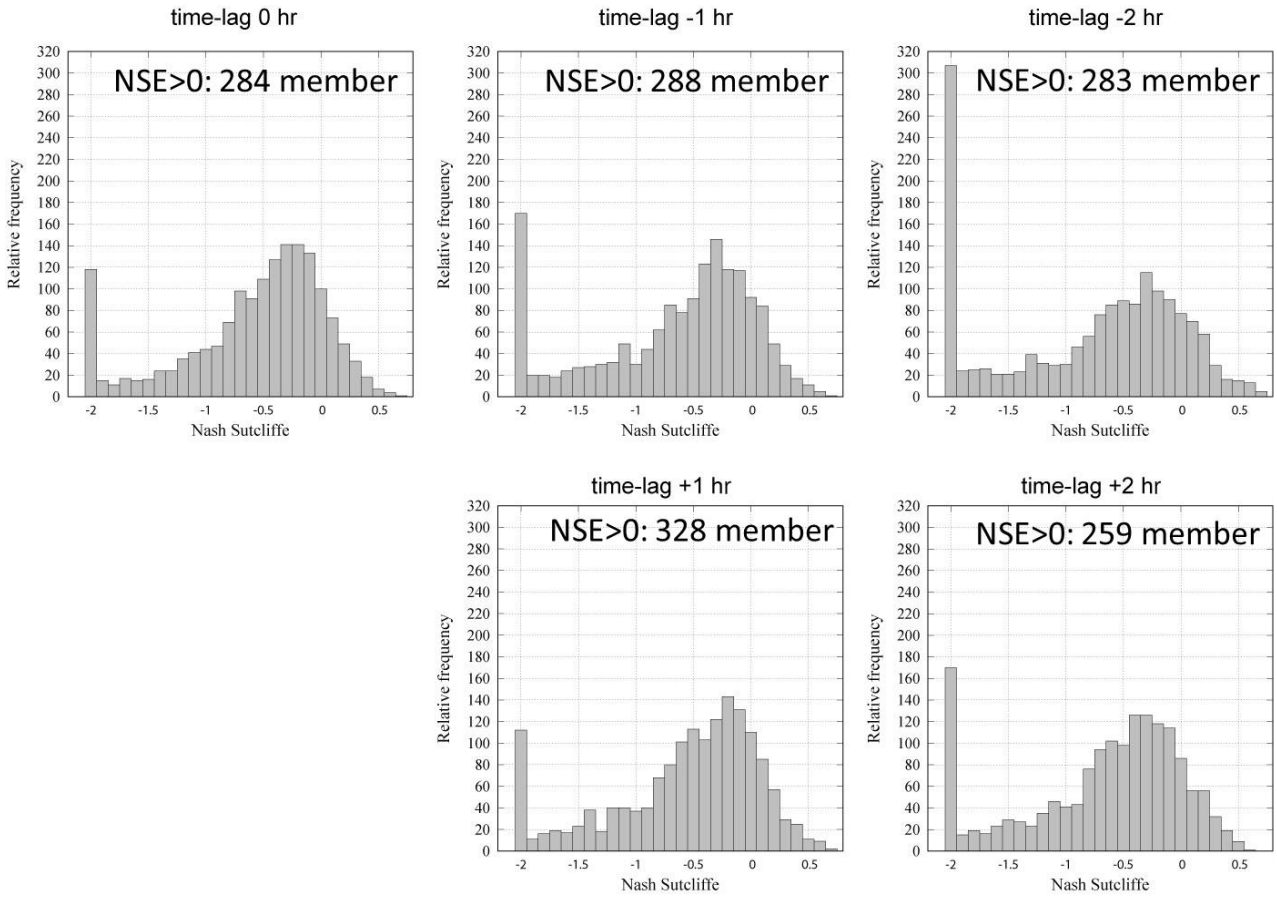

Figure S5. NSE considering temporal displacement errors.

(5) - 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.

Reply: When we select the best ensemble members based on the observations at the first few hours, we have already replaced the original ensemble forecast given by the sample ($\mathbf{x}_i$, $w_i^{pre} = 1/K$, i=1,..,K) by the new sample consisting of a small number of the best ensemble members. The old manuscript did not describe the form of this resulting pdf, and mistakenly assumed that the new ensemble forecast is populated from an equally weighted sample ($\mathbf{x}_j^{best}$, $w_j^{best} = 1/N$, j=1,..,N). Furthermore, we lost useful information when moving from a large ensemble to a small ensemble. And, we have somehow agreed with the point of view of the reviewer on this problem.

Under the theory on selection of the best ensemble members that we have introduced in the revised version, the predictive pdf yielded by the best ensemble members is now given its explicit form in (6). The theory shows that the notion of the best ensemble members still has a certain application if many members have very small relative likelihood $w_i^{post} = \frac{p_Y(\mathbf{y}|\mathbf{x}_i)}{\sum_{j=1}^{K} p_Y(\mathbf{y}|\mathbf{x}_j)}$, and the predictive pdf is dominated by a small number of ensemble members. The weights w_i^{post} strongly depend on our

model on observation errors.

(6) - 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.

Reply: The first reviewer has also shown the same concern on rainfall verification. In the revised manuscript we have added reliability diagrams for rainfall forecasts. Ensemble spreads have in deed given in the plots for 1600 forecasts of the rainfall and discharge at the catchment (Figures 2 and 6 in the original manuscript), but under the form of inter-quartiles in the box-and-whisker diagrams. The use of inter-quartile is more robust than the normal spread since the latter is more sensitive to outliers. Another probabilistic score (the Brier score) has been shown implicitly in Figures 7 and 10 (in the original manuscript) for discharge and accumulated volume when we predefined some thresholds. Instead of plotting all forecast probabilities and the corresponding observations, the Brier scores can be computed from all lead times. However, it is more informative to plot all pairs of the forecast probabilities and the observations.

(7) - 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.

Reply: The number (i.e. 26 to 38) corresponds to the number of (NSE>0, NSE>0.25, NSE>0.9 etc.). Thus, it is not same each other. But we decided to make the number all 50 and replotted the figures considering your comments

(8) - 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.

Reply: We have removed this figure in the revised manuscript since the theory on selection of the best members introduced in the revised manuscript makes this figure irrelevant now.

(9) - 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.

Reply: Yes, we can use radar rainfall available at the first few hours to run our hydrological model.

However, in this case, we can no longer select the best members from forecasts for the first few hours. In this study one of the topics is to know whether we can infer the best members from the observations at the first few hours. If the answer is negative, we agree that your approach is the most appropriate way to improve forecast performance. But of course, in this way we need to run all 1600 members.

(10) - 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.

Reply: We have added several studies focusing on rainfall forecasts at cloud-resolving scales around mountainous areas, to show the importance of cloud-resolving ensemble weather simulation especially as the input to the rainfall-runoff model. Those are written in the revised manuscript as follows:

"Likewise in Europe, Hohenegger (2008) carried out the cloud-resolving ensemble weather simulations of the August 2005 Alpine flood. Their cloud resolving EPS of 2.2 km grid space included the explicit treatment of deep convection and was the result of downscaling of COSMO-LEPS (10km resolution driven by ECMWF EPS). Their conclusion was that despite the overall small differences, the 2.2 km cloud resolving ensemble produces results as good as and even better than its 10km EPS, though the paper did not deal with the hydrological forecasting. Another paper which dealt with cloud resolving ensemble simulations can be found in Vie et al. (2011) for Mediterranean heavy precipitation event. Their ensemble weather simulation model resolution was 2.5 km by AROME from Meteo-France which uses ALADIN forecast for lateral boundary condition (10km resolution), thus the deep convection was explicitly resolved. We can recognize from these researches that the European researchers especially around mountain region have been farsighted from early days for the importance of these cloud resolving ensemble simulations."

Likewise, we have also added some clear explanation about the scope of our papers Part 1 and 2 in series. Those are written in the revised manuscript as follows.

"Since the new EPS produced better forecasts of the rainfall field, in this study, as a Part 2 version of Kobayashi et al. (2016), we applied those 1600 ensemble rainfalls to the ensemble inflow simulations to Kasahori Dam without the positional lag correction. 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 and this would, as a result, help to improve the streamflow forecast for the Kasahori dam. In the series of Part 1 and 2, we intentionally have chosen a rainfall-runoff model whose specification is quite close to those runoff models used in many governmental practices of Japanese flood forecasting to see the usefulness of 1600 ensemble rainfalls."

(11) - 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.

Reply: These sentences are removed.

(12) - 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 selfplagiarism if not cited properly.

Reply: We have considered that many readers do not have time to read Part 1, thus we added basic information of the catchment and dam by citing Part 1 paper. But now the paper is restructured considering your comments and keeps balance with Part 1.

Changes in manuscript:

We summarize our changes in the revised version:

- Section 1 Introduction now emphasizes this study is a continuation of the Part 1 in a series on the hydrological forecast for the Kasahori dam. The main difference is that in the Part 2 we could access a better ensemble forecast for rainfall over the domain around the Kasahori dam. An interesting feature of this ensemble forecast is a large number of ensemble members which suggested us many interesting problems in verification. Some new references have also been updated in this section.
- Section 2 on the heavy rainfall event has been rewritten. We have mainly referred to the Part 1 for necessary information. A text has been added to briefly explain the mechanism that causes heavy rain in the area around the Kasahori dam.
- Section 3 on meteorological ensemble forecast has been divided into two subsections: the subsection 3.1 uses the original content of Section 3 in the old manuscript; and the subsection 3.2 devotes to a new content on rainfall verification for the new ensemble forecast.
- Section 4 is in fact the original Section 5 in the old manuscript. We have removed the original Section 4 on the Kasahori dam catchment since all information is available from the Part 1.
- Section 5 on the results has been divided into two subsections: the subsection 5.1 presents probabilistic forecasts for hydrological variables; and the subsection 5.2 describe a theory on selection of best members based on past performance and its application in our case. Whereas the theory is the new content, the application is mainly based on the content in the old manuscript. With this theory, a call for machine learning methods has been irrelevant now and all texts involving this topic has been removed.
- Section 6 Conclusion has been slightly modified by adding discussion on the possibility of

introducing spatial and temporal uncertainty into NSE.

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[1], Le Duc[2,6], Apip[3], Tsutao Oizumi[2, 6] and Kazuo Saito[4,5,6]

[1]Research Center for Urban Safety and Security, Kobe University, 1-1 Rokkodai-machi, Nada-ku, Kobe, 657-8501, Japan
[2]Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokohama, Japan
[3]Research Centre for Limnology, Indonesian Institute of Sciences (LIPI), Bogor, Indonesia
[4]Japan Meteorological Business Support Center, Tokyo, Japan
[5]Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Japan
[6]Meteorological Research Institute, Tsukuba, Japan

Correspondence to: Kenichiro Kobayashi (kkobayashi@phoenix.kobe-u.ac.jp)

Abstract. This paper is a continuation of the authors' previous paper (Part 1) on the feasibility of ensemble flood forecasting for a small dam catchment (Kasahori dam; approx.70 km2) in Niigata Japan using a distributed rainfall-runoff model and rainfall ensemble forecasts. The ensemble forecasts were given by an advanced data assimilation system, a four-dimensional ensemble variational assimilation system using the Japan Meteorological Agency non-hydrostatic model (4D-EnVAR). A noteworthy feature of this system was the use of a very large number of ensemble members (1600), which yielded a significant improvement in the rainfall forecast compared to Part 1. The ensemble flood forecasting using the 1600 rainfalls succeeded in indicating the necessity of emergency flood operation with the occurrence probability and enough lead time (e.g., 12 hours). Then, dynamical selection of the best ensemble members using the Nash Sutcliffe Efficiency (NSE) with different evaluation periods are discussed. As the result, it is recognized that the selection based on NSE does not provide an exact discharge forecast with several hours lead time, but it can provide some trend in the near future.

1 Introduction

Flood simulation driven by ensemble rainfalls is gaining more attention in recent years, because ensemble simulation is expected to provide flood forecasting with the probability of occurrence. In the Japanese case, it is considered that the ensemble rainfall simulation with a high resolution (2 km or below) is desirable since extreme rainfall often takes place due to mesoscale convective systems and the river catchments are not as large as continental rivers; even the largest Tone River Basin, is around 17000 km^2.

A good review of ensemble flood forecasting using medium term global/European ensemble weather forecasts (2-15 days ahead) by numerical weather prediction (NWP) models can be found in Cloke and Pappenberger (2009). In much of their review, the resolution of NWP model is relatively coarse (over 10 km), the number of ensembles is moderate (10-50) and the

target catchment size is often large (e.g., Danube River Basin). They basically reviewed global/European ensemble prediction systems (EPS) but also introduced some researches on regional EPS nested into global EPS (e.g., Marsigli et al. 2001). They stated that "One of the biggest challenges therefore in improving weather forecasts remain to increase the resolution and identify the adequate physical representations on the respective scale, but this is a source hungry task".

Short-term flood forecasting (1-3 day) based on ensemble NWPs is gaining more attention in Japan.~~., as evidenced by a project for high resolution weather/flood forecasting using the K supercomputer in Kobe, Japan (Saito et al., 2013b, hereinafter the K Project) and a successor project for the preparation toward the use of a next generation exascale computer (hereinafter the Post K Project; https://www.jamstec.go.jp/pi4/en/sub_00.html). In the K Project,~~ Kobayashi et. al. (2016) dealt with an ensemble flood (rainfall-runoff) simulation of a heavy rainfall event occurred in 2011 over a small dam catchment (Kasahori Dam; approx. 70 km^2) in Niigata, central Japan, using a rainfall-runoff model with a resolution of 250 m. Eleven-member ensemble rainfalls by the Japan Meteorological Agency nonhydrostatic model (JMA-NHM; Saito et al. 2006) with horizontal resolutions of 2 km and 10 km were employed. The 10 km EPS was initiated by the JMA operational mesoscale analysis and employed the modified Kain–Fritsch convective parameterization scheme, while its downscaling, the 2 km EPS, did not use the convective parameterization. The results showed that, although the 2 km EPS reproduced the observed rainfall much better than the 10 km EPS, the resultant cumulative and hourly maximum rainfalls still underestimated the observed rainfall. Thus, the ensemble flood simulations with the 2 km rainfalls were still not sufficiently valid. To improve the ensemble rainfalls in quantity and timing, the cumulative rainfalls of each 2 km ensemble member were calculated, then the rain distribution was shifted within 30 km from the original position to where the catchment-averaged cumulative rainfall for the Kasahori Dam maximized (i.e., positional lag correction of the rainfall field). Using this translation method, the magnitude of the ensemble rainfalls and likewise the inflows to the Kasahori Dam became comparable with the observed inflows.

Other applications of the 2 km EPS, which permit deep convection on some level, can be found in for example Xuan et al. (2009). They carried out an ensemble flood forecasting at the Brue catchment, with an area of 135 km^2, in southwest England, UK. The resolution of their grid based distributed rainfall-runoff model (GBDM) was 500 m and the resolution of their NWP forecast by the PSU/NCAR mesoscale model (MM5) was 2 km. The NWP forecast was the result of downscaling of the global forecast datasets from the European Centre for Medium-range Weather Forecasts (ECMWF). In the downscaling, four step nesting were carried out with the inner-most domain covering a region around 100 km x 100 km. The duration of the ensemble weather forecasting was 24 hours. Fifty members of the ECMWF EPS and one deterministic forecast were downscaled. Since the original NWP rainfall of a grid average still underestimates the intensity compared with rain-gauges, they introduced a best match approach (location correction) and a bias-correction approach (scale-up) on the downscaled rainfall field. The results showed that the ensemble flood forecasting of some rainfall events are in good agreement with observations within the confidence intervals, while those of other rainfall events failed to capture the basic flow patterns.

Likewise in Europe, Hohenegger (2008) carried out the cloud-resolving ensemble weather simulations of the August 2005 Alpine flood. Their cloud resolving EPS of 2.2 km grid space included the explicit treatment of deep convection and was the result of downscaling of COSMO-LEPS (10km resolution driven by ECMWF EPS). Their conclusion was that despite the

overall small differences, the 2.2 km cloud resolving ensemble produces results as good as and even better than its 10km EPS, though the paper did not deal with the hydrological forecasting. Another paper which dealt with cloud resolving ensemble simulations can be found in Vie et al. (2011) for Mediterranean heavy precipitation event. Their ensemble weather simulation model resolution was 2.5 km by AROME from Meteo-France which uses ALADIN forecast for lateral boundary condition (10km resolution), thus the deep convection was explicitly resolved. We can recognize from these researches that the European researchers especially around mountain region have been farsighted from early days for the importance of these cloud resolving ensemble simulations.

While in Japan, Yu et al. (2018) have also used a post-processing method using the spatial shift of NWP rainfall fields for correcting the misplaced rain distribution. Their study areas are Futatsuno (356.1 km^2) and Nanairo (182.1 km^2) dam catchments of the Shingu River Basin, in Kii Peninsula, Japan. The resolution of the ensemble weather simulations were 10 km and 2 km by JMA-NHM, which is similar to the downscaling EPS in Kobayashi et al. (2016) but for a different heavy rainfall event in west central Japan caused by a typhoon. The data have a 30-hour forecast time. The results showed that the ensemble forecasts produced better results than the deterministic control run forecast, although the peak discharge was underestimated. Thus, they also carried out a spatial shift of the ensemble rainfall field. The results showed that the flood forecasting with the spatial shift of the ensemble rainfall members was better than the original one, likewise the peak discharges more closely approached the observations.

~~As part of our review, we found several pieces of research which increased the resolution of EPSs (up to e.g. 2 km), while short range flood forecasting of relatively small catchment (several 10–100 km2) were dealt with. Nevertheless, the results showed that 2 km resolution EPS were not necessarily sufficient to represent the observed rainfall field both in timing and location, and thus the post-processing, such as the location correction of the rainfall field and scaling of the peak discharges, were required.~~

Recently, as a further improvement upon the 2 km downscale ensemble rainfall simulations used by Kobayashi et al. (2016), Duc and Saito (2017) developed an advanced data assimilation system with the ensemble variational method (EnVAR) and increased the number of ensemble members to 1600. Since the new EPS produced better forecasts of the rainfall field, in this study, as a Part 2 version of Kobayashi et al. (2016), we applied those 1600 ensemble rainfalls to the ensemble inflow simulations to Kasahori Dam without the positional lag correction. 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 and this would, as a result, help to improve the streamflow forecast for the Kasahori dam. In the series of Part 1 and 2, we intentionally have chosen a rainfall-runoff model whose specification is quite close to those runoff models used in many governmental practices of Japanese flood forecasting to see the usefulness of 1600 ensemble rainfalls. The organization of this paper is as follows. In Section 2, an additional comment for the 2011 Niigata-Fukushima heavy rainfall is given. ~~the 2011 Niigata-Fukushima heavy rainfall is briefly presented.~~ Section 3 describes the new mesoscale EPS, its forecast and rainfall verification results. Section 4 describes ~~and 5 introduce the Kasahori Dam catchment and~~the rainfall-runoff model for explaining the

changes in the model parameters. Results are shown in Section 5. In Section 6, concluding remarks and future aspects are presented.

2 The 2011 Niigata–Fukushima heavy rainfall

~~A severe rainstorm with two rainfall peaks occurred on 27–30 July 2011 over Niigata and Fukushima prefectures in north central Japan. Niigata Prefecture (Niigata, 2011) reported that the cumulative rainfall from the onset of the rainfall to 1300 JST (0400 UTC) on 30 July 2011 reached 985 mm at the Kasahori Dam Observatory. There were 68 rainfall observatories managed by JMA, the Ministry of Land, Infrastructure and Transport and Tourism (MLIT), and the Niigata Prefecture, where the cumulative rainfall exceeded 250 mm. During the rainfall event, JMA announced "record-setting, short-term, heavy rainfall information" on 30 occasions. The hourly rainfall recorded from 2000 to 2100 JST on 29 July at the Tokamachi Shinko Observatory reached 120 mm. Six people were killed and more than 13000 houses were damaged by dike breaks, river flooding, and landslides. A detailed description of this rainfall event has been published by JMA as a special issue of the JMA Technical Report (JMA, 2013).~~

For the details of the 2011 Niigata-Fukushima heavy rainfall, see our Part 1 paper. An additional note is that the torrential rain of the 2011 Niigata-Fukushima heavy rainfall occurred over the small area along the synoptic scale stationary front (for surface weather map, see Fig. 1 of Kobayashi et al. 2016). Saito et al (2013) conducted two 11-member downscale ensemble forecasts with different horizontal resolutions (10 and 2 km) for this event using JMA-NHM and JMA's global ensemble EPS perturbations. They found that the location where intense rain concentrates variable to small changes of model setting, thus the position of the heavy rain was likely controlled mainly by horizontal convergence along the front, rather than the orographic forcing. ~~Two different types of rainfall are introduced in the following text. The descriptions are as follows:~~

~~(a) Radar Composite (1 km resolution): The echo intensity, which can be converted to rainfall intensity, is observed by 20 meteorological radar stations of JMA and is available with 10 min temporal resolution.~~

~~(b) Radar-AMeDAS (1 km resolution): The rainfall intensity observed by the radar is corrected using rain gauge data (ground observation data). The data is available with 30 min temporal resolution.~~

3 Mesoscale ensemble forecast

3.1 Ensemble prediction system

An advanced mesoscale EPS was developed and employed to prepare precipitation data for the rainfall-runoff model. The EPS was built around the operational mesoscale model JMA-NHM for its atmospheric model as the downscale EPS conducted by Saito et al. (2013). In this study, a domain consisting of 819 × 715 horizontal grid points and 60 vertical levels was used for all ensemble members. This domain had a grid spacing of 2 km and covered the mainland of Japan. With this high resolution, convective parameterization was switched off. Boundary conditions were obtained from forecasts of the JMA's global model.

Boundary perturbations were interpolated from forecast perturbations of the JMA's operational one-week EPS as in Saito (2013).

To provide initial conditions and initial perturbations for the EPS, a four-dimensional, variational-ensemble assimilation system (4D-EnVAR-NHM) was newly developed, in which background error covariances were estimated from short-range ensemble forecasts by JMA-NHM before being plugged into cost functions for minimization to obtain the analyses (Duc and Saito, 2017). If the number of ensemble members is limited, ensemble error covariances contain sampling noises which manifest as spurious correlations between distant grid points. In data assimilation, the so-called localization technique is usually applied to remove such noise, but at the same time it removes significant correlations in error covariances. In this study, we have chosen 1600 members in running the ensemble part of 4D-EnVAR-NHM to retain significant vertical correlations, which have a large impact in heavy rainfall events like the Fukushima-Niigata heavy rainfall. That means only horizontal localization is applied in 4D-EnVAR-NHM. The horizontal localization length scales were derived from the climatologically horizontal correlation length scales of the JMA's operational four-dimensional, variational assimilation system JNoVA by dilation using a factor of 2.0.

Another special aspect of 4D-EnVAR-NHM is that a separate ensemble Kalman filter was not needed to produce the analysis ensemble. Instead, a cost function was derived for each analysis perturbation and minimization was then applied to obtain this perturbation, which is very similar to the case of analyses. This helped to ensure consistency between analyses and analysis perturbations in 4D-EnVAR-NHM when the same background error covariance, the same localization, and the same observations were used in both cases. To accelerate the running time, all analysis perturbations were calculated simultaneously using the block algorithm to solve the linear equations with multiple right-hand-side vectors resulting from all minimization problems. The assimilation system was started at 0900 JST July 24th, 2011 with a 3-hour assimilation cycle. All routine observations at the JMA's database were assimilated into 4D-EnVAR-NHM. The assimilation domain was the same as the former operational system at JMA. To reduce the computational cost, a dual-resolution approach was adopted in 4D-EnVAR-NHM where analyses had a grid spacing of 5 km, whereas analysis perturbations had a grid spacing of 15 km. The analysis and analysis perturbations were interpolated to the grid of the ensemble prediction system to make the initial conditions for deterministic and ensemble forecasts.

3.2 Rainfall verification

Due to limited computational resource, ensemble forecasts with 1600 members were only employed for the target time of 0000 JST July 29th, 2011. However, deterministic forecasts were run for all other initial times to examine impact of number of ensemble members on analyses and the resulting forecasts. Figure 1 shows the verification results for the 3-hour precipitation forecasts as measured by the Fraction Skill Score (FSS) (Duc et al., 2013). Here we aggregate the 3-hour precipitation in the first and second 12-hour forecasts to increase samples in calculating the FSS. By this way, robust statistics are obtained but at the same time dependence of the FSS on the leading times can still be shown. Note that an additional experiment with 4D-EnVAR-NHM using 50 ensemble members, which is called 4DEnVAR50 to differentiate with the original one 4DEnVAR1600,

was run. It is very clear from Figure 1 that 4DEnVAR1600 outperforms 4DEnVAR50 almost for all precipitation thresholds, especially for intense rain. Also for high rain-rate, compared to JNoVA, 4DEnVAR1600 forecasts are worse than JNoVA forecasts for the first 12-hour forecasts, which can be attributed to the fact that 4D-EnVAR-NHM did not assimilate satellite radiances and surface precipitation like JNoVA. However, it is interesting to see that 4D-EnVAR-NHM produces forecasts better than JNoVA for very intense rains for the next 12-hour forecasts.

To check reliability of the ensemble forecasts, reliability diagrams are calculated and plotted in Figure 2 for 4DEnVAR1600 and 4DEnVAR50. Since JNoVA only provided deterministic forecasts, reliability diagram is irrelevant for JNoVA. Note that we only performed ensemble forecasts initialized at the target time of 0000 JST July 29th, 2001 due to lack of computational resource to run 1600-member ensemble forecasts at different initial times. Therefore, the same strategy of aggregating 3-hour precipitation over the first and second 12-hour forecasts in calculating the FSS in Figure 1 is applied to obtain significant statistics. Clearly, Figure 2 shows that 4DEnVAR1600 is distinctively more reliable than 4DEnVAR50 in predicting intense rain. While 4DEnVAR50 cannot capture intense rain, 4DEnVAR1600 tends to overestimate areas of intense rain. The tendency of overestimation of 4DEnVAR1600 becomes clearer if we consider the forecast ranges between 12 and 24 hours. However, for the first 12 hours, 4DEnVAR1600 slightly underestimates areas of light rains. This also explains why the FSSs of 4DEnVAR1600 are smaller than those of 4DEnVAR50 for small rainfall thresholds in Figure 1.

As examples of the forecasts, Figure 3 shows the accumulated precipitation at the peak period (1200-1500 JST July 29th, 2011) as observed and forecasted by the 4D-EnVAR prediction system. For comparison, the deterministic forecast initialized by the analysis from JNoVA using the same domain has also been given. Note that the forecast range corresponding to this peak period is from 12 to 15 hours. Clearly, the deterministic forecast initialized by 4D-EnVAR-NHM outperformed that by the JNoVA, especially in terms of the location of the heavy rain, although the forecast by 4D-EnVAR-NHM tended to slightly overestimate the rainfall amount as verified with the reliability diagrams in Figure 2. This over-estimation can also be observed in the coastal area near the Sea of Japan. Note that a significant improvement was also attained against the former downscale EPS used in Part 1 (see Fig. 9 of Kobayashi et al. 2016) .

Since it is not possible to examine all 1600 forecasts, the ensemble mean forecast is only plotted in the bottom right of Figure 3. Again, the location of the heavy rain corresponds well with the observed location, as in the case of the deterministic forecast, but the ensemble mean precipitation is smeared out as a side effect of the averaging procedure. Therefore, to check the performance of the ensemble forecast we plot one-hour accumulated precipitation over the Kasahori Dam catchment in time series under box-and-whisker plots in Figure 4. It can be seen that while the deterministic forecast could somehow reproduce the three-peak curve of the observed rainfall, ensemble members tended to capture the first peak only. Note that some members showed this three-peak curve, such as the best member, but their number was much less than the number of ensemble members.

~~**4 Kasahori Dam catchment**~~

~~Figure 5 (left) shows the Shinanogawa and Aganogawa river catchments, where severe floods occurred in the 2011 Niigata - Fukushima heavy rainfall. The Kasahori Dam catchment exists in the Shinanogawa river catchment. Figure 5 (right) shows an enlarged view of the Kasahori Dam catchment (catchment area 72.7 km^2, MLIT, 2012). The land use of the Kasahori Dam catchment is mostly occupied by forest, and as such, the applied rainfall-runoff model assumed the entire area was forest.~~

~~The basic operation of the Kasahori Dam is summarized as follows.~~

~~1. The reservoir water level is lowered to the normal water level for the rainy season (elevation level (EL) 194.5 m).~~

~~2. If a flood risk due to extreme rainfall is expected by weather monitoring/prediction, the water level is further lowered to the preliminary release water level (EL 192.0 m).~~

~~3. When the inflow exceeds 140 m^{3}s^{-1}, the threshold value for the onset of flood control operations, the gate opening is fixed such that the outflow amount is determined only by the water pressure in the dam. This is, in a broad sense, a natural regulation operation. The gate opening is not adjusted until the water level reaches EL 206.6 m.~~

~~4. When the reservoir water level reaches EL 206.6 m, an emergency (Tadashigaki in Japanese) operation is taken, and the outflow is set equal to the inflow.~~

~~Note that the dam has been under renovation to increase its flood control capacity after the flood event in July 2011, but we do not address the changes due to the dam renovation here. We consider the dam operational rules at the time of the 2011 flood event.~~

4 Distributed Rainfall-Runoff Model

The distributed rainfall–runoff (hereinafter DRR) model used in Part 1 was applied again in this paper. See Kobayashi et al. (2016) for the details. The DRR model applied was originally developed by Kojima et al. (2007) and called CDRMV3. ~~the details of which can be seen in Apip et al. (2011). In the DRR model, the surface and river flows are simulated using a 1D kinematic wave model. The subsurface flow is simulated using the q–h relationship by Tachikawa et al. (2004). The details of the model can be seen in the paper by Kobayashi et al. (2016).~~ As described in the previous section, we intentionally have chosen a rainfall-runoff model whose specification is close to those runoff models used by national/local governments since the purpose is more to investigate the usefulness of 1600 ensemble rainfalls.

The parameters of the DRR model were recalibrated in this study using the hourly Radar-Composite of JMA, since Radar precipitation data is in general the primary source for real time flood forecasting. ~~Radar-Composite data can be obtained in Japan at 10-minutes intervals.~~ The recalibrated equivalent roughness coefficient of the forest, the Manning coefficient of the river, and the identified soil-related parameters are described in Table 1 with the parameters in Part 1. The simulated hydrograph and observations are shown in Figure 5. The duration of the calibration simulation is from 0100 July 28th to 0000 July 31th, 2011 JST.

The Nash Sutcliffe Efficiency (hereinafter NSE: Nash and Sutcliffe, 1970), which is used for the assessment of model performance, is calculated as follows:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}\left\{Q_0^i - Q_s^i\right\}^2}{\sum_{i=1}^{N}\left\{Q_0^i - Q_m\right\}^2} \quad (1)$$

$$Q_m = \frac{1}{N}\sum_{i=1}^{N} Q_0^i \quad (2)$$

where N is the total number of time steps (1 h interval), Q_0^i is observed dam inflow (discharge) at time i, Q_s^i is simulated dam inflow (discharge) at time i, Q_m is the average of the observed dam inflows.

In the calibration simulation in Figure 5, the NSE is 0.754. The 2nd peak is not captured well in the simulation because the Radar-Composite basically could not capture the strong rainfall intensity of the 2nd peak. Nevertheless, we consider that the model can reproduce the discharge on some level if rainfall is properly captured by the observations. Thus, the DRR model is used in the following ensemble simulations.

5 Results

In this section, the results of the ensemble flood simulations are shown focusing on two aspects:

(1) We examined whether the ensemble inflow simulations can show the necessity of starting the flood control operations and emergency operations with sufficient lead time (e.g. 12 h).

(2) We also examined if we could obtain high accuracy ensemble inflow predictions several hours (1-3 h) before the occurrence, which could contribute to the decision for optimal dam operation.

Item (1) provides us with the scenario that we can prepare for any dam operations with enough lead time. Likewise, it may enable us to initiate early evacuation of the inhabitant living downstream of the dam. Item (2) is the target that has been attempted by researchers of flood forecasting. If we could forecast the inflow almost correctly several hours before the occurrence, it could help the dam administrator with the decision for actual optimal dam operations.

5.1 Probabilistic forecast

Item (1) is considered first herein. Figure 6 shows the comparisons of the hydrographs of (a) 11 discharge simulations in Part 1, (b) same 11 member but with a positional shift in Part 1, (c) 50 discharge simulations with 4D-EnVAR-NHM and (d) 1600 discharge simulations with 4D-EnVAR-NHM. Note that the duration of the 4D-EnVAR-NHM ensemble weather simulation is 30 hours from 0000 July 29th to 0700 July 30th JST, but the ensemble flood simulation is carried out only for 24 hours from 0300 July 29th to 0300 July 30th, 2011 JST since we consider that JMA-NHM uses the first 3 hours to adjust its dynamics. The result in Figure 6 (d) shows that, except for the third peak, the 1600 ensemble inflows can encompass the observed rainfall within the range, which was not realized in Part 1 with 11 downscale ensemble rainfalls of 2 km resolution (Figure 6 (a)). In other words, the extreme rainfall intensity of the event can be reproduced by the ensemble members with

1600 4D-EnVAR-NHM. Likewise, comparing Figure 6 (c) and (d), the simulated discharges by 50 ensemble rainfalls of 4D-EnVAR-NHM encompass the observation within the range less than those of 1600 ensemble members. In addition, Figure 7 shows the histogram of NSE based on the simulated and observed discharges for the 4 cases. Looking at Figure 7 (c) and (d), clearly, the 1600 ensemble discharge simulations outperform 50 ensemble simulations. The NSE>0 was around 17.75% (284 members) in 1600 ensembles, while it was 0% in 50 ensembles. On the other hand, Figures 7 (a) and (c) show that the performance of the discharge forecast with 50 member 4D-EnVAR-NHM was not necessarily better than the 11 discharge forecast in Part 1 in terms of NSE. Likewise, Figures 6 (b) and 7 (b) show that the discharge forecasts could be improved by the positional shift of the rainfall field in Part 1, though this positional shift still needs further statistical verification with more rainfall events.

Figure 8 shows the 95 % confidence limits and inter-quartile limits of the 1600 ensemble members. The results show that the 3rd peak of the observations was not covered by the 95 % confidence interval, although the rest of the observations can be reproduced within the 95 % confidence interval. It is considered also that the ensemble mean and median values capture the overall trend of the observations on some level.

Figure 9 shows the probability that the inflow discharge is beyond 140 $m^3 s^{-1}$ (hereinafter expressed as "$q > 140$", where q is the discharge), the threshold value for starting the flood control operations. The figure considers the temporal shift of the ensemble rainfalls, i.e., temporal uncertainty due to the imperfect rainfall simulation. In the figure, 0-hour uncertainty means that we only considered discharges at time t to calculate probability, while 1-hour uncertainty means that we considered the discharges at t-1, t, t+1 to calculate probability and 2-hour means that we considered the discharges at t-2, t-1, t, t+1, t+2 to calculate probability. The 3- and 4-hour uncertainties were calculated in the same way. It becomes clear from the figure that the starting time of $q > 140$ is likely at t = between 0800 and 0900 July 29th JST, where all curves cross, while the ending time is likely at t = 1800 JST, where all curves cross again. Before and after the cross points there are jumps in the probabilities. In other words, the forecast can indicate that the situation of $q > 140$ would take place after 8–9 hours from the beginning of forecasting with the probability of around 50 %. We consider that this is a very valuable information for the users of the ensemble forecast.

On the other hand, the emergency operation was undertaken in the actual flood event. In the emergency operation, the dam outflow has to equal the inflow to avoid dam failure as the water level approaches overtopping of the dam body. As written in Part 1, when the reservoir water level reaches EL 206.6 m, an emergency operation is undertaken, and the outflow is set to equal the inflow. As the Height-Volume (H-V) relationship of the dam reservoir was not known during the study, we judged the necessity of the emergency operation by whether the cumulative dam inflow was beyond the flood control capacity of 8700000 m^3. Actually, the flood control capacity had not been previously filled during regular operations more than the estimation given herein, since the dam can release the dam water by natural regulation. However, again, since we do not know some of the relationships to calculate the dam water level, the judgement is done based on whether the cumulative dam inflow exceeds the flood control capacity.

Figure 10 shows the cumulative dam inflows of all the ensemble simulations starting from 0300 July 29th, 2011 JST, as well as the mean and observed cumulative inflows with the flood control capacity. The figure shows that the mean of the ensembles was roughly similar to the observations. ~~Figure 11 shows the 38 best ensemble members selected based on NSE > 0.25, as well as the mean of all ensemble members, mean of the best ensemble members, and observations and flood control capacity. Figure 11 shows that the ensemble mean of the best 38 members resembles the observations for the first 12 hours better than the mean of all ensemble members, but the accuracy deteriorates for the last 12 hours. The difference between the observations and ensemble mean of all members is about 20 % after 24 hours.~~ Figure 11 shows the probability that the cumulative dam inflow exceeds the flood control capacity of 8700000 m^3. The figure indicates that, for instance, the cumulative inflow would exceed flood control capacity after 12 hours from the start of the forecast with the probability of around 45 %. In the actual event, the cumulative inflow based on observations and assuming no dam water release, would exceed the flood control capacity between 1200 and 1300 July 29th, 2011 JST. Around that interval, the exceedance probability of the forecast is 35–55 %. Until around this time, the forecast shows a slight delay in the estimate of the cumulative dam inflow. In the end, the forecast shows that the flood control capacity will be used up with the probability of more than 90 % with regard to this flood event. Thus, we consider this information is very useful as it can inform the inhabitant downstream of the dam to evacuate.

5.2 Selection of the best members

Figure 12 shows all ensemble members, the 50 best ensemble members out of 1600 ensembles selected based on NSE > 0.224, and observations. ~~The 38 best ensemble members are the same as in Figure 11.~~ The figure shows that the selected 50 members reproduce the observations well. In some of the selected members, even the 3rd peak is reproduced. In the case where the 3rd peak is reproduced, the inflow hydrographs are beyond the 95 % confidence interval. Figure 13 shows the catchment average rainfalls of the 50 best ensemble inflow simulations. The black line is the observed gauge rainfall, the blue line is the Radar-AMeDAS (operational precipitation analysis of JMA based on radar and rain gauge observations), the green line is the Radar-Composite, while the grey lines are the 50 rainfalls for the best ensemble discharges. As mentioned, the rainfall-runoff model parameters are calibrated using Radar-Composite since the Radar-Composite is the primary source for the flood forecasting. Therefore, the rainfalls from the best 50 ensemble inflow simulations resemble those of the Radar-Composite.

Clearly, the flood forecasting becomes very useful if we could just select the best ensemble members in advance. Logically, this is impossible since we only know the best members after knowing the observations which enable us to compute verification scores like NSE. This raises the question whether or not the best ensemble members can be inferred from the partial information provided by the observations at the first few hours. It is easy to see that the answer should be negative due to nonlinearity of the model and the presence of model error: the best marching at the first few hours is almost certainly not the best marching over all forecast ranges. However, it is obvious that the observations at the first few hours have a certain value which can help to reduce uncertainty in the ensemble forecast if we could incorporate this information into the forecast.

This procedure has already been well-known under the name “data assimilation” in which we assimilate the observations at the first few hours to turn the prior probabilistic density function (pdf) given by the short-range forecasts into the posterior pdf

given by the analysis ensemble (Reich and Cotter, 2015). Thus, if we know the observations at the first few hours, we should assimilate these data to replace the short-range ensemble forecasts by the ensemble analyses at these hours, then run the model initialized by the new ensemble to issue a new ensemble forecast. As a result, we should replace the definition of the best members based on verification scores to a more appropriate one based on the posterior pdf. Here, we identify the best members with the most likely members. Clearly, if we assume the posterior pdf is unimodal, the best members should be the members clustering around the mode of this pdf, which is also the analysis. However, it is not clear how to identify the best members if this pdf is multimodal.

To overcome this problem, we will use the mathematical framework settled up by particle filter (Doucet et al., 2001, Tachikawa et al., 2011). Let us denote the short-range forecasts by $\mathbf{x}_1$ to $\mathbf{x}_K$ where K is the number of ensemble members. The short-range ensemble forecast therefore yields an empirical pdf given by the sample $(\mathbf{x}_i, w_i^{pre} = 1/K)$ with w_i^{pre} denoting the equal weight for the i-th member

$$\mathrm{p}_X(\mathbf{x}) = \sum_{i=1}^{K} w_i^{pre}\, \delta(\mathbf{x} - \mathbf{x}_i) = \sum_{i=1}^{K} \frac{1}{K} \delta(\mathbf{x} - \mathbf{x}_i). \quad (3)$$

Using this prior pdf as the proposal density, the posterior pdf has the following form

$$\mathrm{p}_X(\mathbf{x}|\mathbf{y}) = \sum_{i=1}^{K} w_i^{post}\, \delta(\mathbf{x} - \mathbf{x}_i) = \sum_{i=1}^{K} \frac{\mathrm{p}_Y(\mathbf{y}|\mathbf{x}_i)}{\sum_{j=1}^{K} \mathrm{p}_Y(\mathbf{y}|\mathbf{x}_j)} \delta(\mathbf{x} - \mathbf{x}_i). \quad (4)$$

Here, $\mathrm{p}_Y(\mathbf{y}|\mathbf{x}_i)$ denotes the likelihood of the observations $\mathbf{y}$ conditioned on the forecast $\mathbf{x}_i$, and the weight w_i^{post} are the relative likelihoods. Moreover, it can be shown that $\mathrm{p}_Y(\mathbf{y}|\mathbf{x}_i)$ is the observation evidence for the ith member (Duc and Saito, 2018). Then applying the model M as the transition model, the predictive pdf is given by

$$\mathrm{p}_X(\mathbf{x}|\mathbf{y}, \mathrm{M}) = \sum_{i=1}^{K} w_i^{post}\, \delta(\mathbf{x} - \mathrm{M}(\mathbf{x}_i)). \quad (5)$$

This equation shows that the contribution of each member to the predictive pdf is unequal, which differs from the prior pdf (3). While the members with large values of w_i^{post} dominate the predictive pdf, those with very small values of w_i^{post} can be ignored. This suggests that the best members can be identified with the largest values of w_i^{post}. Thus, if we sort w_i^{post} in the descending order, the first N weights are corresponding to the first N best ensemble members. In this case, the predictive pdf (5) is approximated by

$$\mathrm{p}_X(\mathbf{x}|\mathbf{y}, \mathrm{M}) = \sum_{i=1}^{N} \frac{\mathrm{p}_Y(\mathbf{y}|\mathbf{x}_i)}{\sum_{j=1}^{N} \mathrm{p}_Y(\mathbf{y}|\mathbf{x}_j)} \delta(\mathbf{x} - \mathrm{M}(\mathbf{x}_i)). \quad (6)$$

Note that by introducing the notion of the best ensemble members, a substantial change occurs, that is we now work with a unequal weighted sample $(\mathbf{x}_i, w_i^{post})$. This should be taken into account in computing statistics like ensemble mean from the best ensemble members.

If the likelihoods have the Gaussian form

$$\mathrm{p}_Y(\mathbf{y}|\mathbf{x}_i) \propto \exp\left[-\frac{1}{2}(\mathbf{y} - \mathrm{h}(\mathbf{x}_i))^T \mathbf{R}^{-1} (\mathbf{y} - \mathrm{h}(\mathbf{x}_i))\right], \quad (7)$$

where h is the observation operator, and $\mathbf{R}$ is the observation error covariance, it is easy to see that the largest weights are corresponding to the smallest weighted root mean square errors (WRMSE)

$$WRMSE_i = (\mathbf{y} - \mathrm{h}(\mathbf{x}_i))^T \mathbf{R}^{-1} (\mathbf{y} - \mathrm{h}(\mathbf{x}_i)). \quad (8)$$

Therefore, if **R** is a multiple of the identity matrix **I**, the WRMSEs become the RMSEs, which in turn are equivalent to the NSEs. This shows that selection of the best members based on verification scores over the first few hours is in fact selection of the best members based on the relative likelihoods in the posterior pdf. It can also be understood as model selection based on observation evidence (Mackay, 2003).

One of the practical problems with the assimilation process lies in high cost in computational resource, especially when 1600 members are used to sample the prior and posterior pdfs. To be more practical, we assume that only rainfalls and discharges at the first few hours are available. As a first step, we attempted to select some of the best members out of the 1600 members several hours in advance of the event based only on NSEs for the discharges. Figure 14(a) shows a result where we selected the best 50 ensemble members (NSE > -0.04) for the first 9 hours from the start of the forecast. In this case, we had a 3-hour lead time towards the observed peak discharge, and the selected 50 members cover the observed discharge after the first 9 hours on some level. The result shows that the ensemble inflow simulations selected can indicate the possibility of rapid increases in the discharge after the 9 hours with a three-hour lead time.

Likewise Figure 14 (b) shows the selected best 50 members (NSE > -0.33) for the first 10 hours (two hours ahead of the observed peak discharge). It is apparent that the result is worse than the previous first 9-hour selection. The ensemble inflow simulations after the 10 hours do not cover the observation well in this case. Figure 14(c) shows the selected best 50 members (NSE > 0.86) for the first 11 hours (1 hour ahead of the observed peak discharge). In this case, the ensemble inflows after the 11 hours could cover the observed peak discharge 1 hour later on some level, although it only has a one-hour lead time.

~~Nevertheless, overall it is recognized that we cannot select the best members in advance only by judgement based on NSE of the discharge. Figure 16(a) shows a scatter plot of NSE of the catchment average rainfall vs NSE of the discharge. Clearly, the figure shows that catchment average rainfalls with similar NSEs produce discharges with different NSEs. In detail, the catchment average rainfall with NSE of around 0 produces discharges with NSE close to 0.5 and -0.5. We consider that the spatial distribution of the rainfall field caused these differences even though the amount of the catchment average rainfalls are the same. Even if the catchment area is small, different patterns in the rainfall field bring different discharge simulations with different NSEs. As a reference, Figure 16(b) shows the Root Mean Square (RMS) of the simulated and observed discharge vs simulated and observed rainfall. It is apparent that RMS cannot be used for the decision in regard to the best discharge simulations as the catchment average rainfalls with the same RMS also produce both favorable and less favorable discharges. The rainfall pattern chosen based only on NSE or RMS does not reflect the variety of rainfall patterns. We consider that selection directly from the rainfall data, and comparing them with Radar based on e.g. Self-Organizing Map (SOM), Support Vector Machine (SVM), pattern recognition, machine learning, etc., would be more promising to better cluster the ensemble rainfalls. However, we have not addressed that aspect in this study and this remains for future work.~~ It seems that the selection method based on NSE does not provide us an exact discharge forecast with several hours lead time, although it can provide us some trend in the near future. This can be traced back to the use of the Gaussian form (7) to model the likelihood $\mathrm{p}_Y(\mathbf{y}|\mathbf{x}_i)$. The resulting WRMSE (8), or equivalently NSE, is quite sensitive to spatial and temporal displacement errors of rainfall. Part

1 of this study is an illustration for impact of spatial displacement errors on forecast performance while Figures 9 and 11 here show the case of temporal displacement errors. Thus, it is expected that if we can introduce spatial and temporal uncertainty in modelling the likelihood $p_Y(\mathbf{y}|\mathbf{x}_i)$, the predictive pdf (6) could yield a more useful ensemble forecast. However, this requires a lengthy mathematical treatment that is worth to explore in details in a separate study.

6 Concluding Remarks and Future Aspects

The study used 1600 ensemble rainfalls produced by 4D-EnVAR which contain various rainfall fields with different rainfall intensities. No post processing such as the location correction of the rainfall field and/or rescaling of rainfall intensity was employed. The ensemble flood forecast using the 1600 ensemble rainfalls in this study has shown that the extremely high amount of observed inflow discharge can be reproduced within the confidence interval, which was not possible by the 11 member downscale ensemble rainfalls used in Part 1, although the accuracy of each simulation of 1600 ensembles is, at best, around NSE = 0.6. We can calculate the probability of occurrence (e.g. the necessity of emergency dam operations) with the 1600 ensemble rainfalls. Thus, the result of the study shows that the ensemble flood forecasting can inform us that, after 12 hours for example, emergency dam operations would be required with the probability of around 45 %, and that the probability would be more than 90 % for the entire flood event, etc. We consider that this kind of information is very useful. For instance, a warning of dam water release can be issued to the inhabitant in the downstream with enough lead time, if the result obtained in this study is applicable to other locations and events.

On the other hand, the accuracy of each discharge simulation is, at best, around NSE = 0.6 out of all the 1600 ensemble members. Likewise, discharge simulations with similar NSEs until X hours before the onset of forecasting produce different future forecasts after the X-th hour. In other word, we cannot select the best discharge simulation from the NSE only until X hours. Herein lies the problem that, NSEs are quite sensitive to spatial and temporal displacement errors in rainfall. In principle, it is possible to introduce those errors into NSEs in a way similar to FSSs. However, it should be cautious in introducing such errors into NSEs before investigated well, although such type of approach has been used frequently in meteorology community. How to incorporate them qualitatively is also a problem to be addressed. ~~similar NSEs of the catchment average rainfall with different rainfall distribution, even in the small catchment areas, produce different NSEs of the discharges. Thus, we cannot select one best ensemble discharge simulation from the rainfall NSEs. by comparing the simulated rainfall field with observed Radar fields, etc. using some methods, such as SOM, SVM, pattern recognition, machine learning, etc.,~~ Thus, in this sense, the dynamical selection of the best rainfall field from rainfall simulations considering both spatial and temporal displacement errors is required, although this was not addressed here and remains for future work.

Acknowledgments. A part of this work was supported by the Ministry of Education, Culture, Sports, Science and Technology as the Field 3, the Strategic Programs for Innovative Research (SPIRE) and the FLAGSHIP 2020 project (Advancement of meteorological and global environmental predictions utilizing observational “Big Data”). Computational results were obtained

using the K computer at the RIKEN Advanced Institute for Computational Science (project ID: hp140220, hp150214, hp160229, hp170246, and hp180194). JMA-NHM is available under collaborative framework between MRI and related institute or university. Likewise, the DRR model is available under collaborative framework between Kobe, Kyoto Universities and related institute or university. The JMA's operational analyses and forecasts, radar rain gauge analyses, and radar composite analyses can be purchased at http://www.jmbsc.or.jp/. The rain gauge data were provided by MLIT, Niigata Prefecture and JMA.

References

Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: A review, J. Hydrol, 375, 613-626, doi:/10.1016/j.jhydrol.2009.06.005, 2009.

Doucet, A., de Freitas, N., and Gordon, N. J.: Sequential Monte Carlo Methods in Practice. Springer-Verlag, New York, 2001.

Duc, L. & Saito, K.: A 4DEnVAR data assimilation system without vertical localization using the K computer. Japan Geoscience Union meeting, Chiba, Japan, 2017.

Duc, L. and Saito, K.: Verification in the presence of observation errors: Bayesian point of view. Quart. J. Roy. Meteor. Soc., 144, 1063–1090, doi:10.1002/qj.3275, 2018

Hohenegger, C., Walser, A., Langhans, W. and Schaer, C.: Cloud-resolving ensemble simulations of the August 2005 Alpine flood, Quarterly Journal of the Royal Meteorological Society 134: pp. 889-904, DOI: 10.1002/qj.252 , 2008.

Japan Meteorological Agency: Report on "the 2011 Niigata-Fukushima heavy rainfall event", typhoon Talas (1112) and typhoon Roke (1115). Tech. Rep. JMA, 134, http://www.jma.go.jp/jma/kishou/books/gizyutu/134/ALL.pdf (last access: 01 October 2018), 253pp, 2013, (in Japanese).

Kobayashi, K., Otsuka, S., Apip and Saito, K.: Ensemble flood simulation for a small dam catchment in Japan using 10 and 2km resolution nonhydrostatic model rainfalls. Nat. Hazards Earth Syst. Sci., 16, 1821-1839 doi:10.5194/nhess-16-1821-2016, 2016.

Kojima, T., Takara, K. and Tachikawa, Y.: A distributed runoff model for flood prediction in ungauged basin. Predictions in Ungauged Basins: PUB Kick-off (Proceedings of the PUB Kick-off meeting held in Brasilia, 20–22 November 2002). IAHS Publication, 309, 267 – 274, 2007.

Mackay, D. J. C.: Information Theory, Inference, and Learning Algorithms. Cambridge Univ. Press, 640 pp, 2003.

Marsigli, C., Montani, A., Nerozzi, F., Paccagnella, T., Tibaldi, S., Molteni, F. and Buizza, R.: A strategy for high–resolution ensemble prediction. Part II: limited-area experiments in four alpine flood events, Q. J. R. Meteorol. Soc., 127, 2095-2115, doi:10.1002/qj.49712757613, 2001.

Ministry of Land, Infrastructure, Transport and Tourism (MLIT): Digital national land information download service. http://nlftp.mlit.go.jp/ksj/ (last access: 01 October 2018), 2012, (in Japanese).

Nash, J.E. and Sutcliffe, V.: River flow forecasting through conceptual models part I - A discussion of principles, Journal of Hydrology, Volume 10, Issue 3, Pages 282-290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970

Niigata Prefecture: Niigata/Fukushima extreme rainfall disaster survey documentation (as of 22 August 2011), http://www.pref.niigata.lg.jp/kasenkanri/1317679266491.html, (last access: 01 October 2018), 2011, (in Japanese).

Reich, S., and Cotter, C.: Probabilistic Forecasting and Bayesian Data Assimilation. Cambridge Univ. Press, 308 pp, 2015.

Saito, K., Fujita, T., Yamada, Y., Ishida, J., Kumagai, Y., Aranami, K., Ohmori, S., Nagasawa, R., Kumagai, S., Muroi, C., Katao, T., Eito, H. and Yamazaki, Y.: The operational JMA nonhydrostatic meso-scale model. Mon. Wea. Rev., 134, 1266-1298, doi:10.1175/MWR3120.1, 2006.

Saito, K., Origuchi, S., Duc, L., and Kobayashi, K.: Mesoscale ensemble forecast experiment of the 2011 Niigata-Fukushima heavy rainfall, Technical Report of the Japan Meteorological Agency, 134, 170–184, http://www.jma.go.jp/jma/kishou/books/ gizyutu/134/ALL.pdf, (last access: 22 July 2015), 2013 (in Japanese).

~~Saito, K., Tsuyuki, T., Seko, H., Kimura, F., Tokioka, T., Kuroda, T., Duc, L., Ito, K., Oizumi, T., Chen, G., Ito, J. and SPIRE Field 3 Mesoscale NWP Group: Super high resolution mesoscale weather prediction, J. Phys.: Conf. Ser., 454, 012073. doi:10.1088/1742-6596/454/1/012073, 2013b.~~

Tachikawa, Y., Sudo, M, Shiiba, M., Yorozu, K. and Sunmin, K.: Development of a real time river stage forecasting method using a particle filter, The Journal of Japan Society of Civil Engineers, Ser. B1,Vol.67, No.4, I_511-I_516,2011 (in Japanese).

Vie, B., Nuissier, O. and Ducrocq, V.: Cloud-resolving ensemble simulations of Mediterranean heavy prediction events: Uncertainty on Initial conditions and lateral boundary conditions, Special Thorpex collection, Americal meteorological society, https://doi.org/10.1175/2010MWR3487.1, 2011

Xuan, Y., Cluckie, I. D. and Wang, Y. : Uncertainty analysis of hydrological ensemble forecasts in a distributed model utilizing short-range rainfall prediction. Hydrol, Earth Syst. Sci., 13, 293-303, doi:10.5194/hess-13-293-2009, 2009.

Yu, W., Nakakita, E., Kim, S. and Yamaguchi, K.: Assessment of ensemble flood forecasting with numerical weather prediction by considering spatial shift of rainfall fields, KSCE J. Civ. Eng., 1–11 (2018), 10.1007/s12205-018-0407-x, 22(9), 3686-3693. doi:10.1007/s12205-018-0407-x, 2018.

List of Table

Table 1. The equivalent roughness coefficient of the forest, the Manning coefficient of the river, and identified soil-related parameters of Part 2 (this paper) and Part 1 (Kobayashi et al., 2016).

	Forest [m(-1/3)/s]	River [m(-1/3)/s]	D [m]	Ks [ms-1]
This paper	0.170	0.005	0.234	0.0008
Part 1	0.150	0.004	0.320	0.0005

List of figures

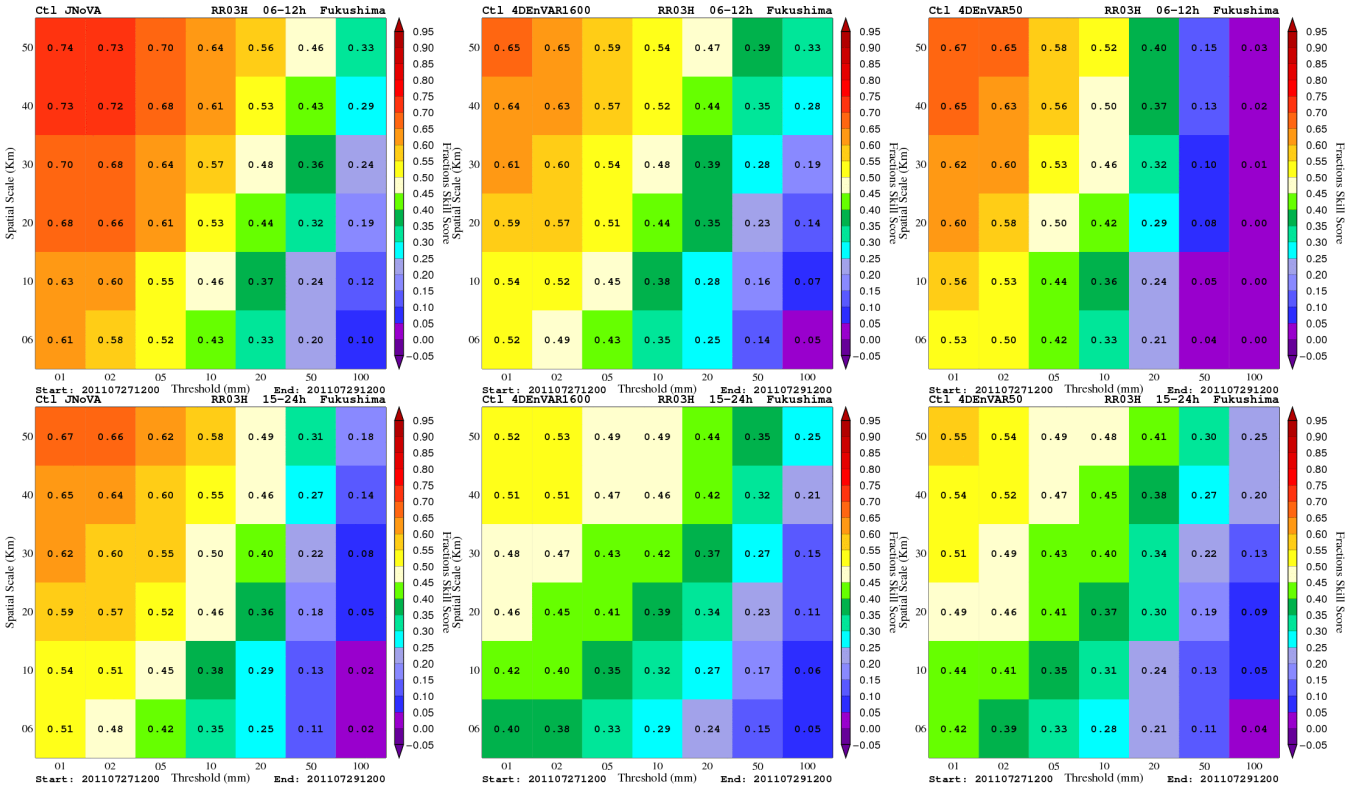

Figure 1. Fraction skill scores of 3-hour precipitation at Fukushima-Niigata from deterministic forecasts initialized by analyses from JNoVA (left), 4D-EnVAR-NHM using 1600 (center) and 50 members (right). These scores are averaged over the period from 2100 JST July 27th to 2100 JST July 29th, 2011. To obtain robust statistics, precipitation is aggregated over the first 12-hour forecasts (valid between 03-12-hour forecast) and the next 12-hour forecasts (valid between 12-24-hour forecasts) as shown in the top and bottom rows, respectively. Note that the first 3-hour precipitation is discarded due to the spin-up problem.

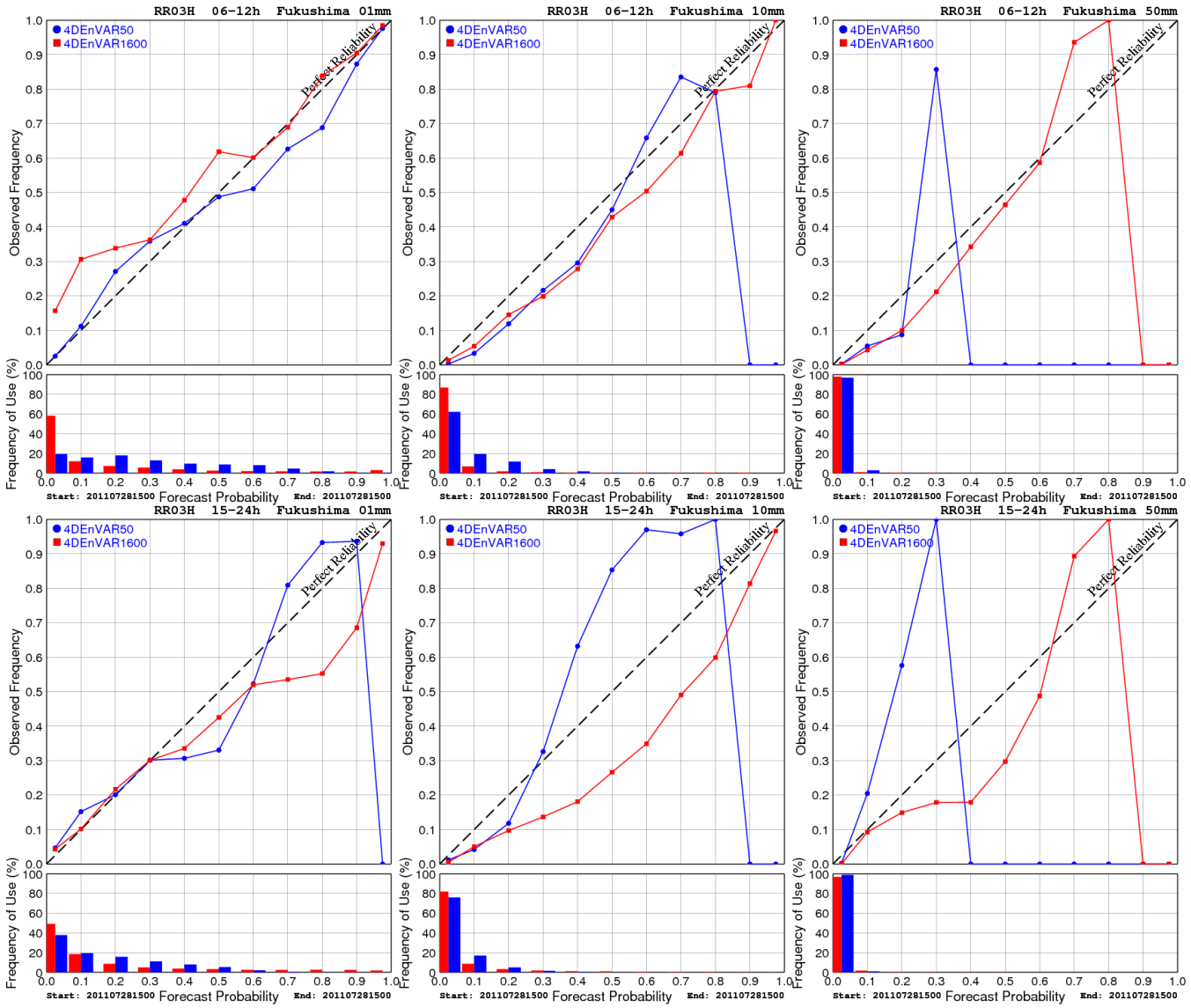

Figure 2. As Figure 1 but for reliability diagrams of 3-hour precipitation from ensemble forecasts initialized by analysis ensembles of 4D-EnVAR-NHM using 1600 and 50 members. Three precipitation thresholds of 01 mm (left), 10 mm (center), and 50 mm (right) are chosen. Note that the ensemble forecasts were only run for the time 0000 JST July 29th, 2011.

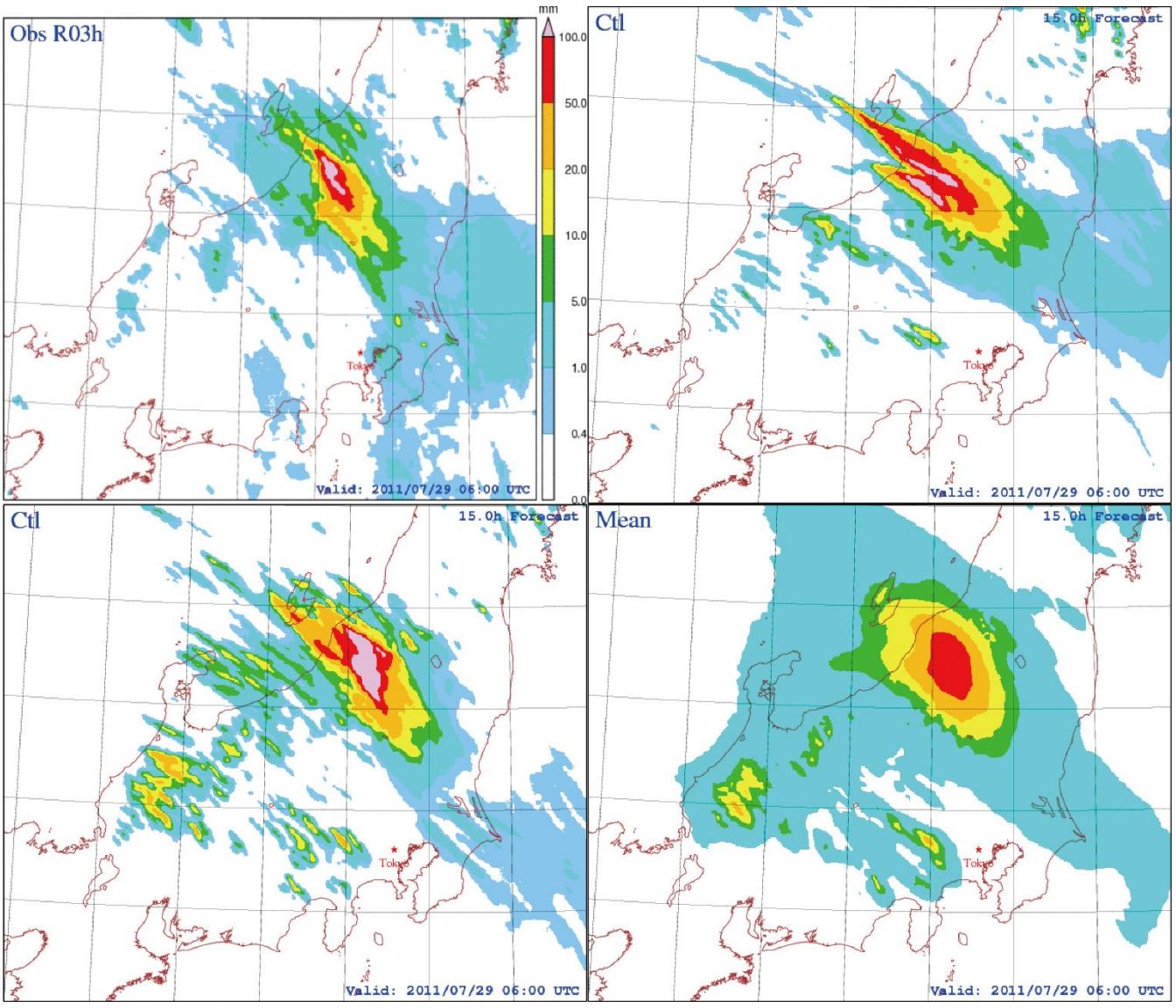

Figure 3.Three-hour accumulated precipitation for 1200-1500 JST July 29th, 2011 at Fukushima-Niigata as observed by Radar-AMeDAS (R/A; top left), forecasted by NHM initialized by the analysis of JNoVA (top right), forecasted by NHM initialized by the analysis of 4D-EnVAR-NHM (bottom left), and the ensemble mean forecast of NHM initialized by the analysis ensemble of 4D-EnVAR-NHM (bottom right). All forecasts were started at 0000 JST July 29th, 2011.

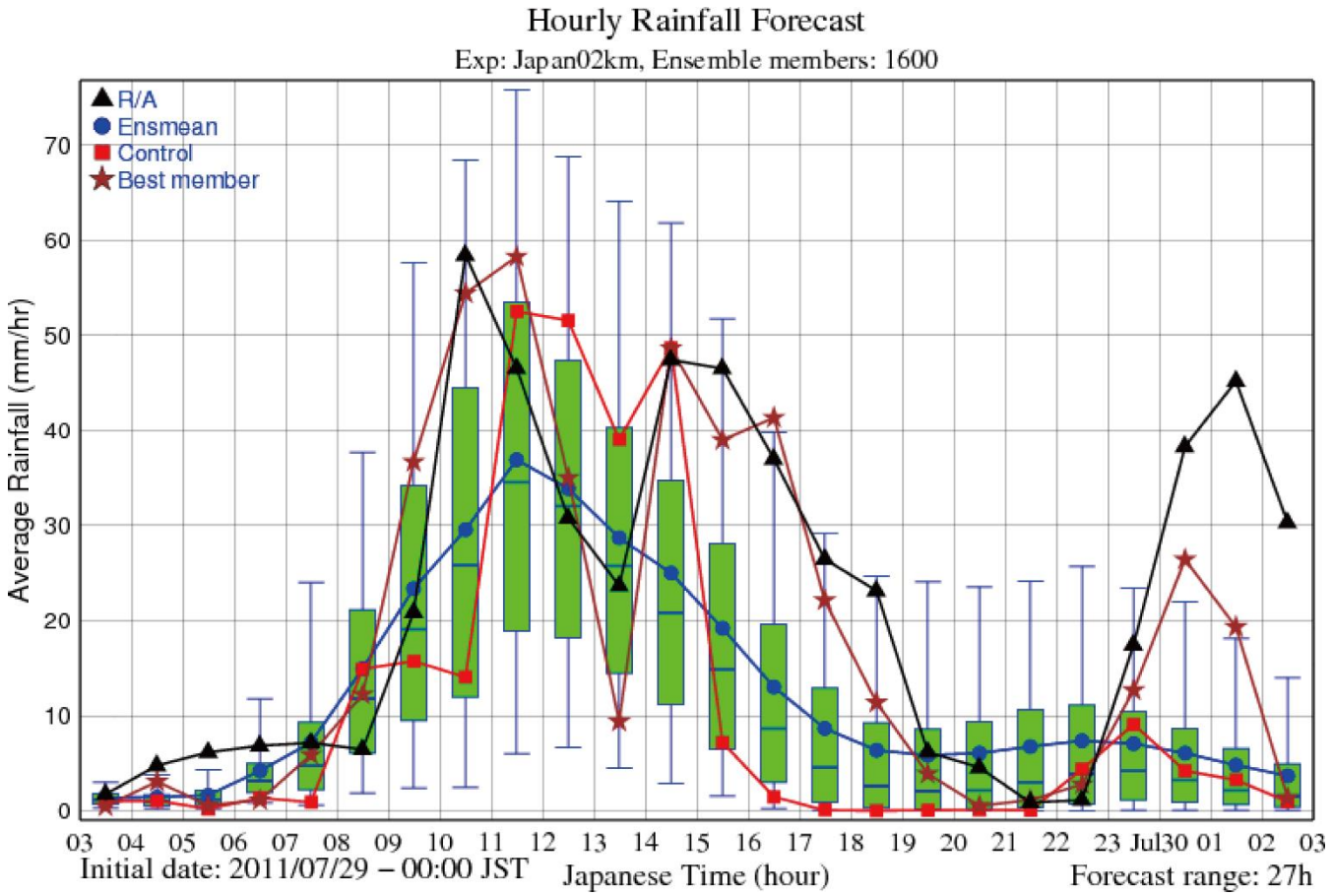

Figure 4. Time series of one-hour accumulated rainfall over the catchment as forecasted by all ensemble members. The two whiskers in each box-and-whisker diagram show the inter-quartile and 5th and 95th percentile of forecasted precipitation. The observation, control forecast, ensemble mean forecast, and best member forecast are also plotted for comparison. Here, the best member is defined as the member that has the minimum distance between its time series and the observed time series.

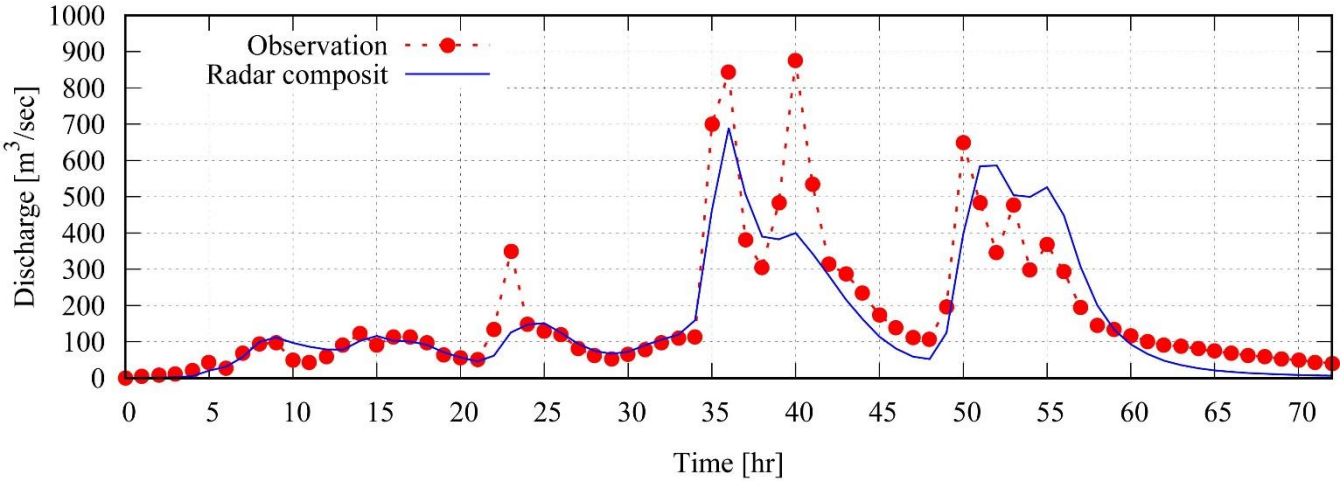

Figure 5. The observed inflow and simulated dam inflow using Radar-Composite.

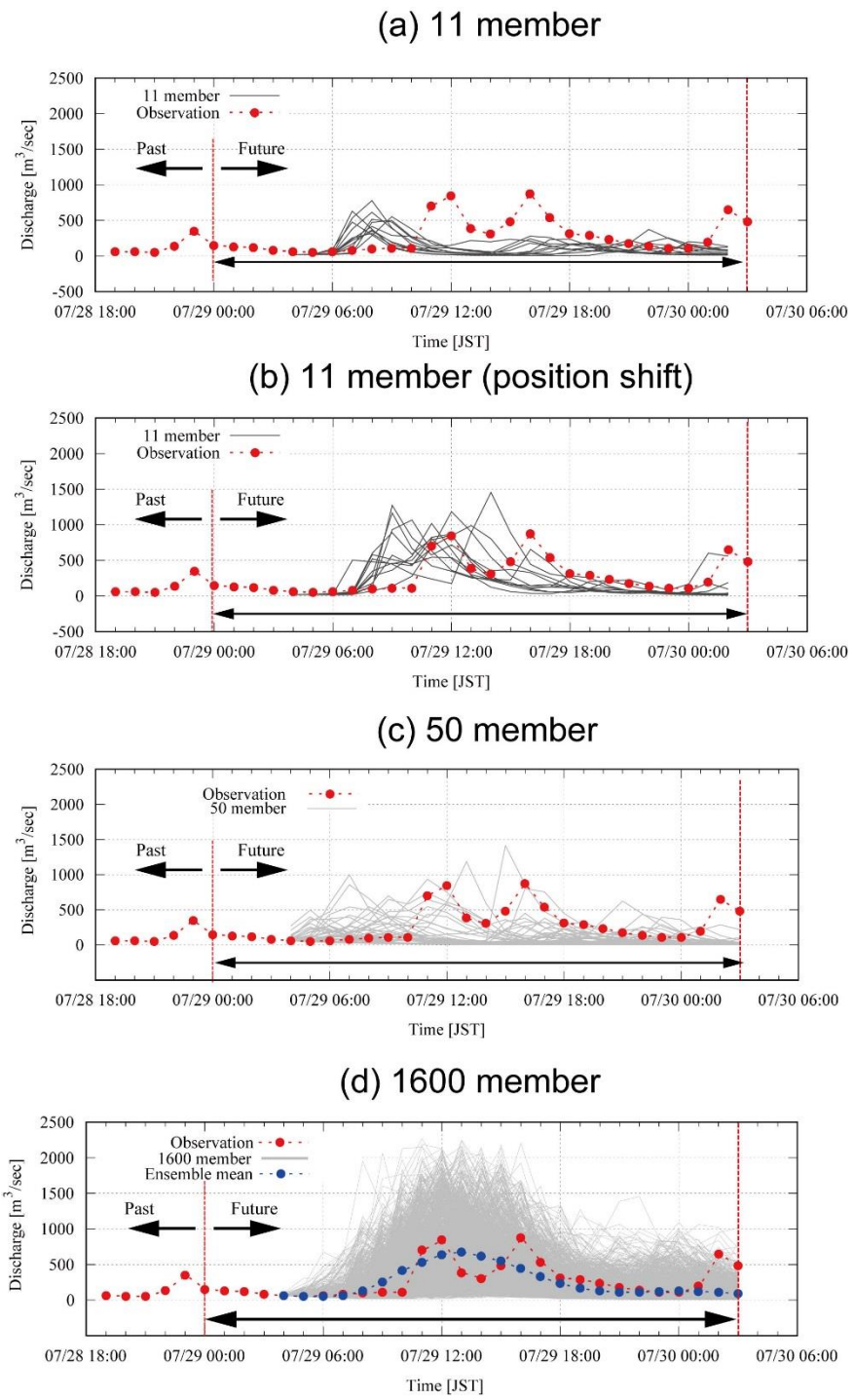

Figure 6. Hydrographs of (a) 11 discharge simulations in Part 1 (Kobayashi et al., 2016), (b) same 11 member but with a positional shift in Part 1, (c) 50 discharge simulations with 4D-EnVAR-NHM and (d) 1600 discharge simulations with 4D-EnVAR-NHM.

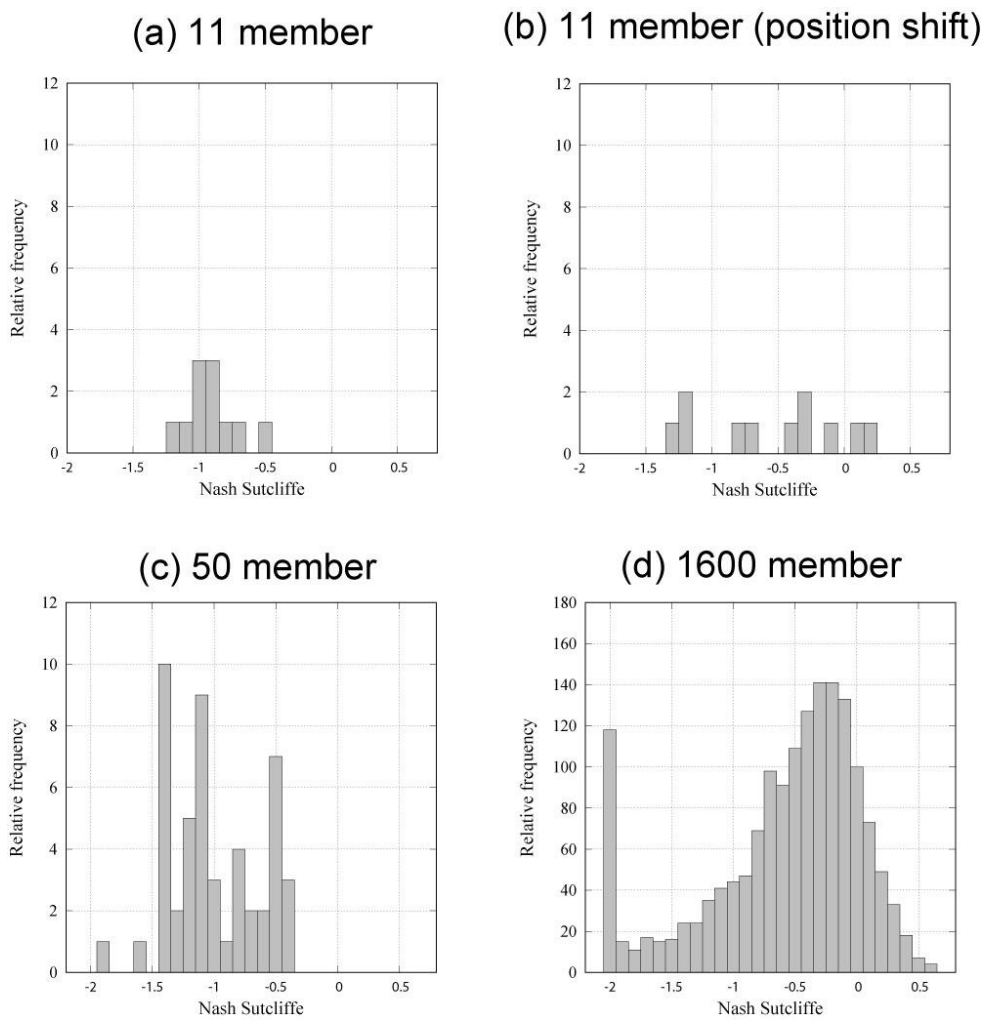

Figure 7. Histograms of NSE based on the simulated and observed discharges to Kasahori dam: (a) 11 discharge simulations in Part 1 (Kobayashi et al., 2016), (b) same 11 member but with a positional shift in Part 1 (Kobayashi et. al. 2016), (c) 50 discharge simulations with 4D-EnVAR-NHM and (d) 1600 discharge simulations with 4D-EnVAR-NHM.

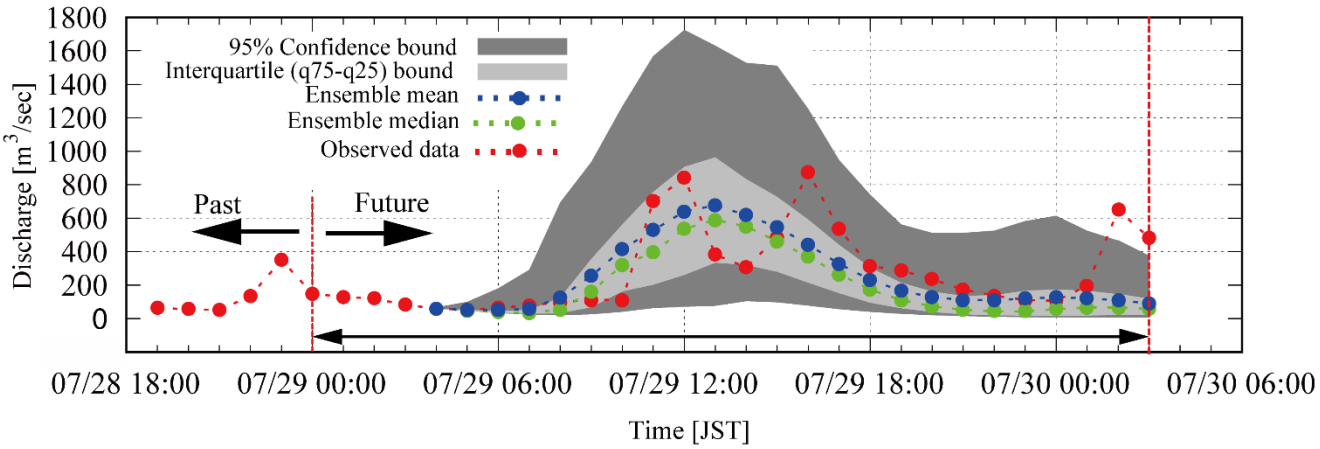

Figure 8. The 95% confidence limits and inter-quartile limits of the 1600 ensemble members.

5

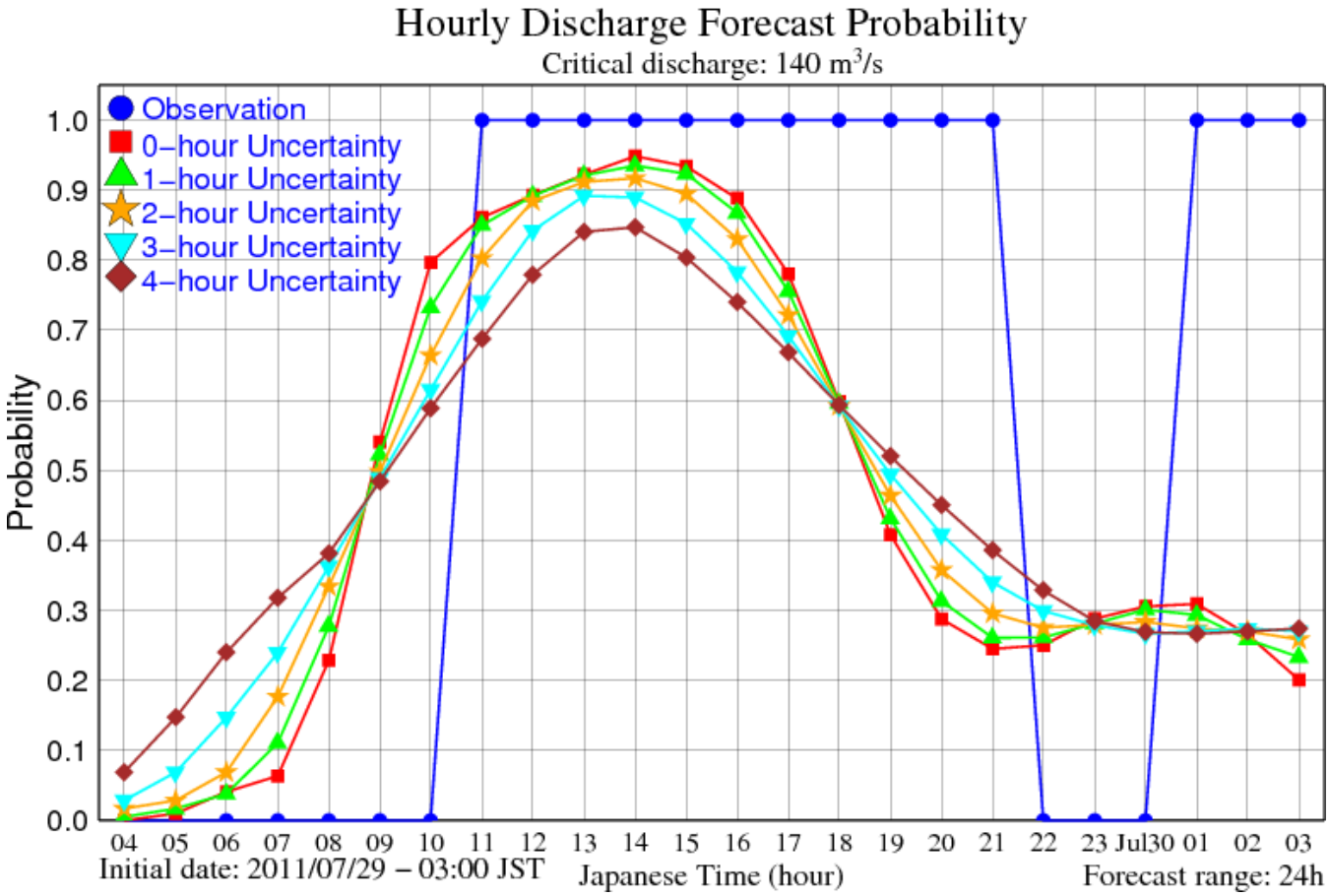

Figure 9. Probability that the simulated inflow is beyond 140 m^3/s considering temporal uncertainty.

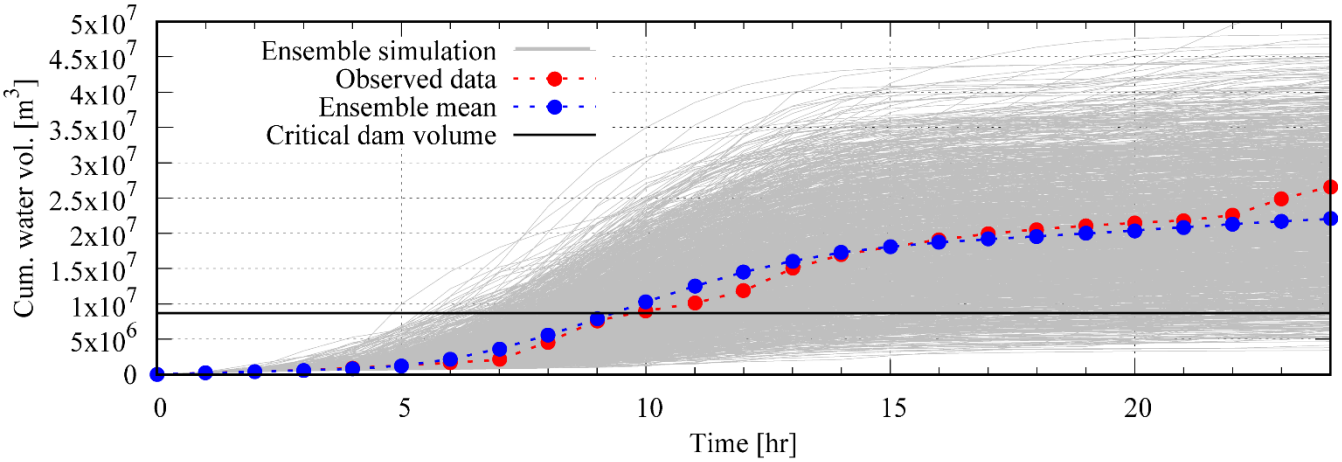

Figure 10. Cumulative dam inflow by the ensemble simulations, mean of simulation and observations, as well as critical dam volume.

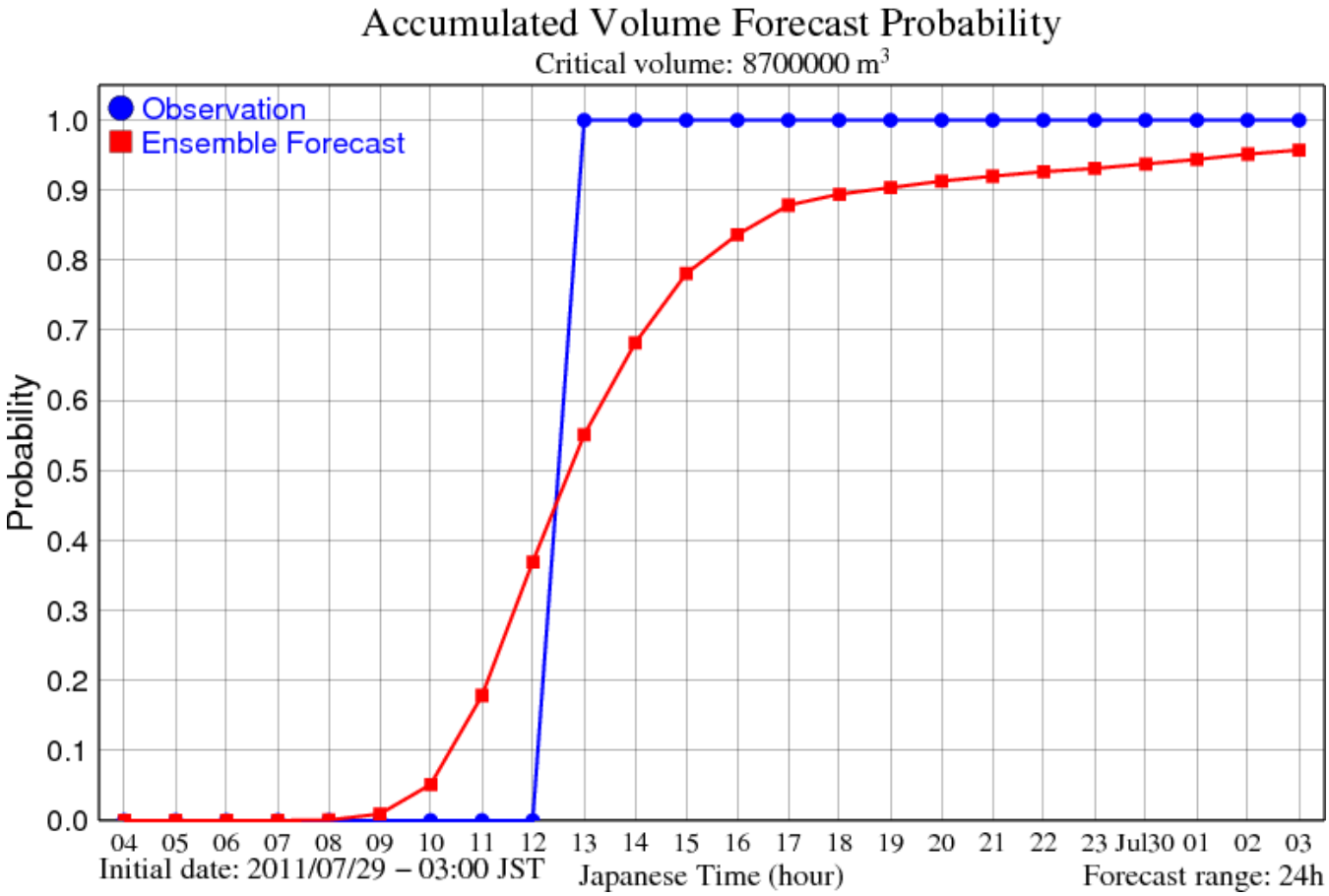

Figure 11. Probability that the dam needs emergency operation.

5

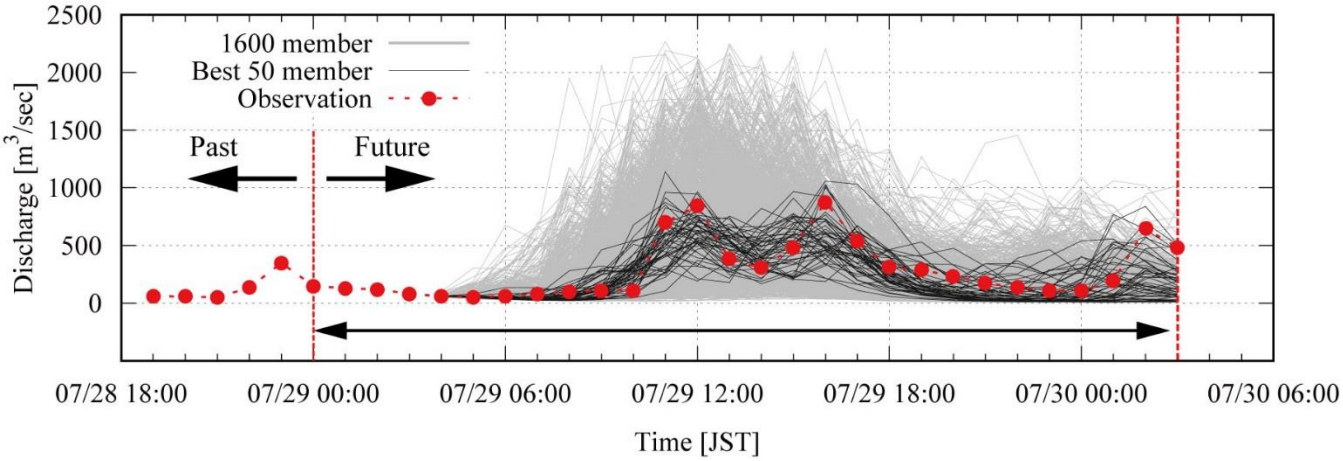

Figure 12. Hydrographs of all 1600 ensemble members, the 50 best ensemble members (NSE>0.224), and observations.

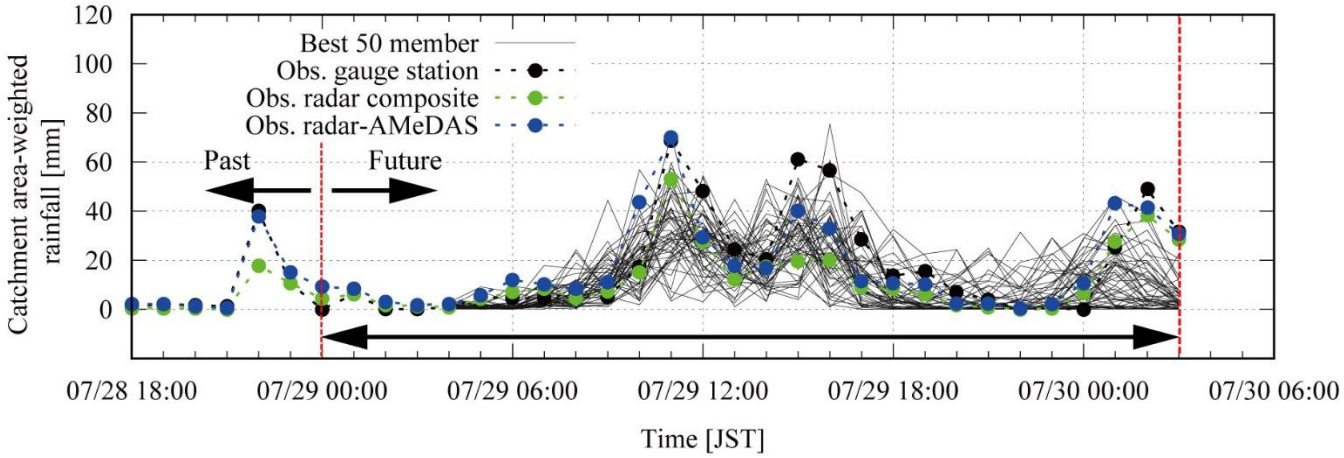

Figure 13. Rainfall intensity of the 50 best ensemble inflow simulation members, of Radar AMeDAS, of Radar-Composite, and ground observations.

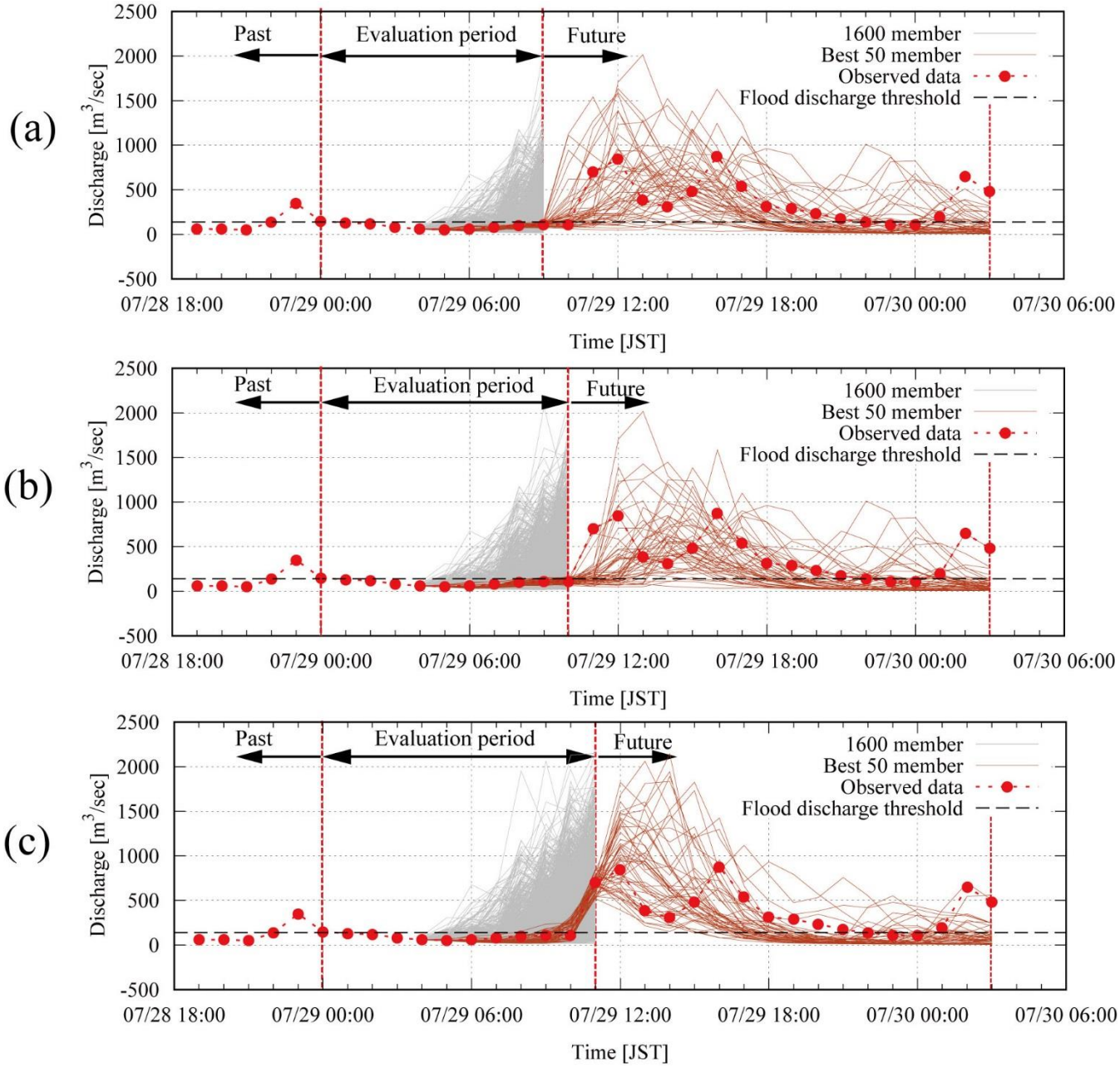

Figure 14. (a) best 50 ensemble members (NSE >-0.04) selected from first 9-hour forecast, (b) best 50 ensemble members (NSE >-0.33) selected from first 10-hour forecast, and (c) best 50 ensemble members (NSE > 0.86) selected from first 11-hour forecast.