

Response to Dr. Seed

We would like to thank Dr. Seed for his constructive comments. Here we would like to present our answer following each comment.

5

Reviewer comments:

The paper describes the application of a large (1600 members) ensemble of high-resolution rainfall forecasts for flood forecasting in a small (72 km²) 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 urban 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 NHM-4DnVAR stored data every hour due to limited data storage, we applied the hourly data to the rainfall-runoff model.

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 that we reproduce here:

“As mentioned in our Part 1 paper (Kobayashi et al. 2016), this torrential rain 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 (2013a) 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 varies with a small change of model settings, thus the position of the heavy rain was likely controlled mainly by horizontal convergence along the front, rather than the orographic forcing.”

10

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.

15

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 (see Figure S1. we do not intend to add these figures into the manuscript).

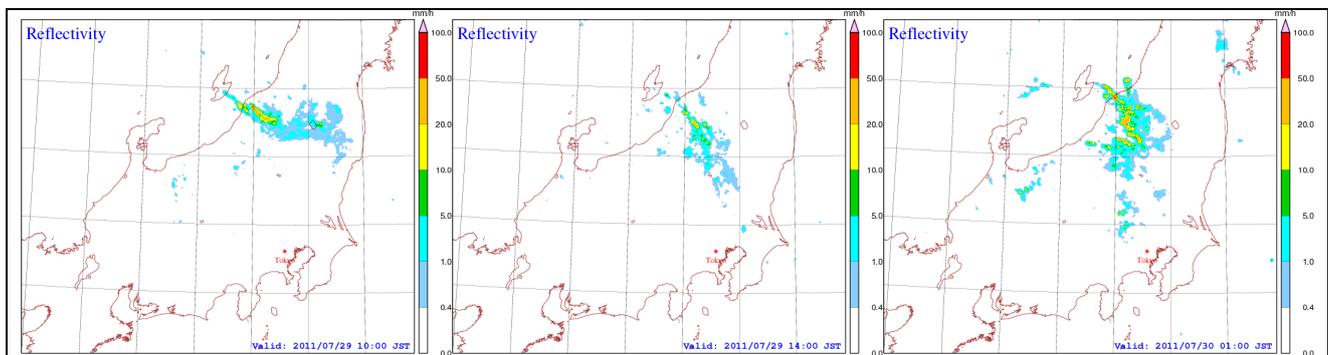


Figure S1. Reflectivity from radar composition at the time of the first (left), second (center), and third (right) peaks of the hydro-graph.

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?

10

Reply: We have added two paragraphs to Section 3 describing the verification results of the rainfall forecasts. We agree with the 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 FSS is also calculate for this additional experiment. Here are the paragraphs that we have added to the revised manuscript:

20

“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

25

experiment with the 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, 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 radiances like JNoVA. However, it is interesting to see that 4D-EnVAR-NHM produces forecasts better than JNoVA 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 (see Figure S2. we do not intend to add these figures into the manuscript). It turns out that it is more difficult to forecast the second and third peaks in the hydrograph.

25

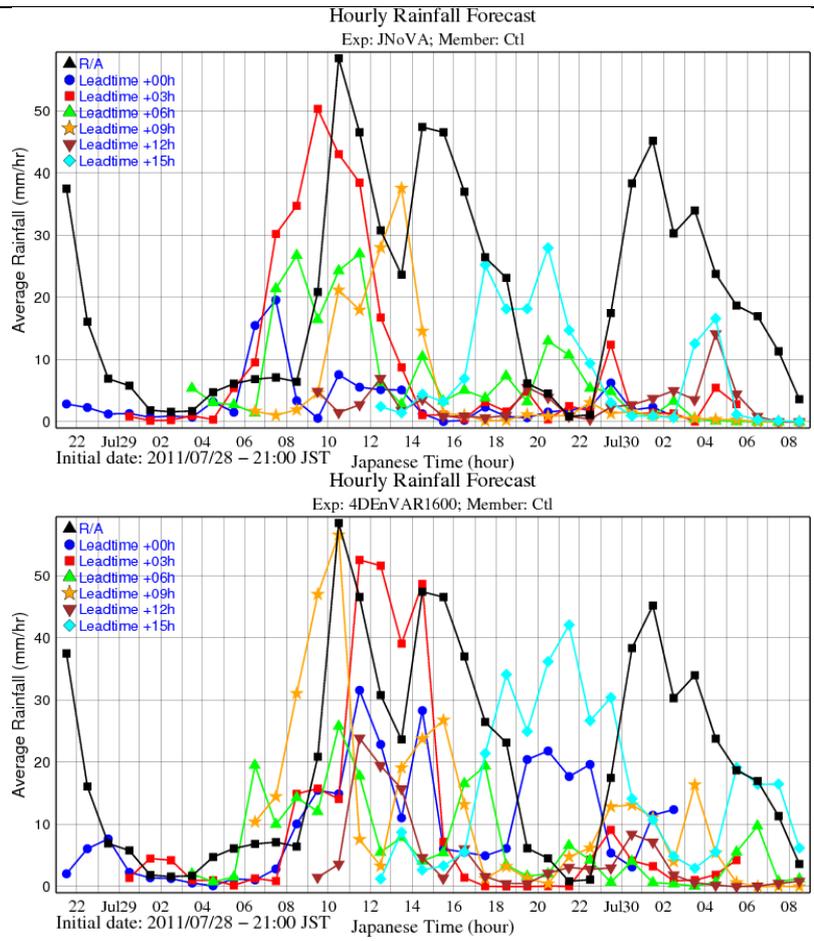


Figure S2. Time series of one-hour accumulated rainfall over the catchment by deterministic forecasts of JNoVA (top) and 4DnVAR1600 (bottom) at different lead times.

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.

5

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.

5 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 (Figure 9 in supplement) 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
10 rainfall over the catchment here is not rainfall at any specific grid point but rainfall over many grid points (more than 70 km² 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 than just one direction in the one-dimensional space of temporal uncertainty.
15 Therefore, we do not consider spatial uncertainty in plotting Figure 7 and 10 (Figure 9 and 12 in supplement).

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

20

The revised manuscript is attached in the next pages. The yellow highlight indicates where the revision is made.

25

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.

5 Kenichiro Kobayashi¹, Apip², Le Duc^{3,6}, Tsutao Oizumi^{3,6} and Kazuo Saito^{4,5,6}

¹Research Center for Urban Safety and Security, Kobe University, 1-1 Rokkodai-machi, Nada-ku, Kobe, 657-8501, Japan

²Research Centre for Limnology, Indonesian Institute of Sciences (LIPI), Bogor, Indonesia

³Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokohama, Japan

10 ⁴Japan Meteorological Business Support Center, Tokyo, Japan

⁵Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Japan

⁶Meteorological Research Institute, Tsukuba, Japan

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

Abstract. This paper elaborated the feasibility of flood forecasting using a distributed rainfall-runoff model and huge number
15 of ensemble rainfalls with an advanced data assimilation system. Specifically, 1600 ensemble rainfalls simulated by a four-
dimensional ensemble variational assimilation system with the **Japan Meteorological Agency** nonhydrostatic model (4D-
EnVAR-NHM) were given to the rainfall-runoff model to simulate the inflow discharge to a small dam catchment (Kasahori
dam; approx. 70 km²) in Niigata, Japan. The results exhibited that the ensemble flood forecasting can indicate the necessity of
flood control operation and emergency flood operation with the occurrence probability and a lead time (e.g. 12 hours). Thus,
20 the ensemble flood forecasting may be able to inform us the necessity of the early evacuation of the inhabitant living
downstream of the dam e.g. half day before the occurrence. On the other hand, the results also showed that the exact forecasting
to reproduce the discharge hydrograph several hours before the occurrence is yet difficult, and some optimization technique is
necessary such as the **dynamical** selection of the good ensemble members.

1 Introduction

25 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 maximum Tone River Basin is
around 17000 km².**

30 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”.

5 Short-term flood forecasting (1-3 day) based on ensemble NWP 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 km²) 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 used. 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 km², 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.

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 km²) and Nanairo (182.1 km²) 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, in the Post K Project, 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 using the K supercomputer. Since the new EPS produces better forecasting 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 organization of this paper is as follows. In Section 2, the 2011 Niigata–Fukushima heavy rainfall is briefly presented. Section 3 describes the new mesoscale EPS and its forecast. Sections 4 and 5 introduce the Kasahori Dam catchment and the rainfall-runoff model. Results are shown in Section 6. In Section 7, concluding remarks and future aspects are presented.

20 **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).

As mentioned in our Part 1 paper (Kobayashi et al. 2016), this torrential rain 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 (2013a) 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 varies with a small change 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:

- 5 (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

- 10 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. (2013a). 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.
- 15 Boundary perturbations were interpolated from forecast perturbations of the JMA's operational one-week EPS as in Saito (2013a). 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
- 20 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 can remove significant correlations in error covariances. In this study, we have chosen 1600 members in running the ensemble part of the 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 the 4D-EnVAR-NHM. The horizontal localization length scales were derived from the
- 25 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 the 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
- 30 analysis perturbations in the 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 the 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 the 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.

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 the 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, 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 radiances like JNoVA. However, it is interesting to see that 4D-EnVAR-NHM produces forecasts better than JNoVA 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 the 4D-EnVAR-NHM outperformed that by the JNoVA, especially in terms of the location of the heavy rain, although the forecast by the 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 by Saito et al (2013a) (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 km², 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 m³ 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.

5 Distributed Rainfall-Runoff Model

The distributed rainfall–runoff (hereinafter DRR) model used in Kobayashi et al. (2016) was applied again in this paper. 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).

The parameters of the DRR model were recalibrated in this study using the hourly Radar-Composite precipitation data of JMA, since Radar 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. The simulated hydrograph and observations are shown in Figure 6. The duration of the calibration simulation is from 0100 July 28th to 0000 July 31th, 2011 JST.

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

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

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

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 6, 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.

6 Results

In this chapter, 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.

Item (1) is considered first herein. Figure 7 shows the results of the inflow simulations to the Kasahori Dam driven by the 1600 ensemble rainfalls. The duration of the 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 0400 July 30th, 2011 JST since we consider that the NHM uses the first 3 hours to adjust its dynamics. The results show that, except for the third peak, the 1600 ensemble inflows can encompass the observed rainfall within the range, which was not realized in the previous research of Kobayashi et al. (2016) with 11 ensemble rainfalls of 2 km resolution. In other words, the extreme rainfall intensity of the event can be reproduced by the ensemble members with the 4D-EnVAR-NHM.

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 $140 \text{ m}^3 \text{ s}^{-1}$ (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 the previous section, 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 $\text{NSE} > 0.25$, 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 12** shows the probability that the cumulative dam inflow exceeds the flood control capacity of 8700000 m³. 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 %.

5 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 residence downstream of the dam to evacuate.

10 Hereafter, the focus is put on Item (2). **Figure 13** shows all ensemble members, the 38 best ensemble members out of 1600 ensembles selected based on $NSE > 0.25$, and observations. The 38 best ensemble members are the same as in **Figure 11**. The figure shows that the selected 38 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 14** shows the catchment average rainfalls of the 38 best ensemble inflow simulations. The black line is the

15 observed gauge rainfall, the blue line is the Radar-AMeDAS, the green line is the Radar-Composite, while the grey lines are the 38 ensemble rainfalls. 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 38 ensemble inflow simulations resemble those of the Radar-Composite.

It is apparent that the flood forecasting becomes very useful if we could just select the 38 ensemble members in advance.

20 Thus, 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 NSE.

Figure 15(a) shows a result where we selected the best 46 ensemble members based on $NSE > 0.0$ 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 46 members cover the observed discharge after the first 9 hours on some level. The result shows that the ensemble inflow

25 simulations selected can indicate the possibility of rapid increases in the discharge after the 9 hours with a three-hour lead time. Likewise **Figure 15(b)** shows the selected best 26 members based on $NSE > 0.0$ 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 15(c)** shows the selected best 30 members based on $NSE > 0.9$ for the first 11 hours (1 hour ahead of the observed peak discharge). In this case, the ensemble inflows

30 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. We conclude that the selection method used here 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.

7 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 the previous **study** by Kobayashi et al. (2016), although the accuracy of each **discharge** simulation 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, several of the best ensemble members only could not be selected from the NSE of the inflow discharge and NSE of the catchment averaged rainfall. Herein lies the problem that, 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. Likewise, discharge simulations with similar NSEs until X hours before the onset of forecasting produce different future forecasts after the Xth hour. In other word, we cannot select the best discharge simulation from the NSE only until X hours. Thus, in this sense the **dynamical** selection of the best rainfall field from rainfall simulations is required by comparing the simulated rainfall field with observed Radar fields, etc. using some

methods, such as SOM, SVM, pattern recognition, machine learning, etc., 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.jmbasc.or.jp/>. The rain gauge and discharge data were provided by MLIT, Niigata Prefecture and JMA.

References

- Apip, Sayama, T., Tachikawa, Y. and Takara, K.: Spatial lumping of a distributed rainfall-sediment-runoff model and its effective lumping scale. *Hydrol. Proc.*, 26(6), 855-871, doi: 10.1002/hyp.8300, 2009.
- 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.
- Duc, L. & Saito, K.: A 4DEnVAR data assimilation system without vertical localization using the K computer. Japan Geoscience Union meeting, Chiba, Japan, 2017.
- 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.
- 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).

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

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), 2013a (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., Nagatani, G. and Takara, K. (2004). Development of stage-discharge relationship equation incorporating saturated/unsaturated flow mechanism (in Japanese), *Ann. Hydraul. Eng., Jpn. Soc. Civil Engrs.*, 48, 7-12. doi: 10.2208/prohe.48.7, 2004.

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.

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.

List of Table

Table 1. The equivalent roughness coefficient of the forest, the Manning coefficient of the river, and identified soil-related parameters.

Forest [$m^{-1/3}/s$]	River [$m^{-1/3}/s$]	D [m]	Ks [ms^{-1}]
0.170	0.00536	0.234	0.00084

5

10

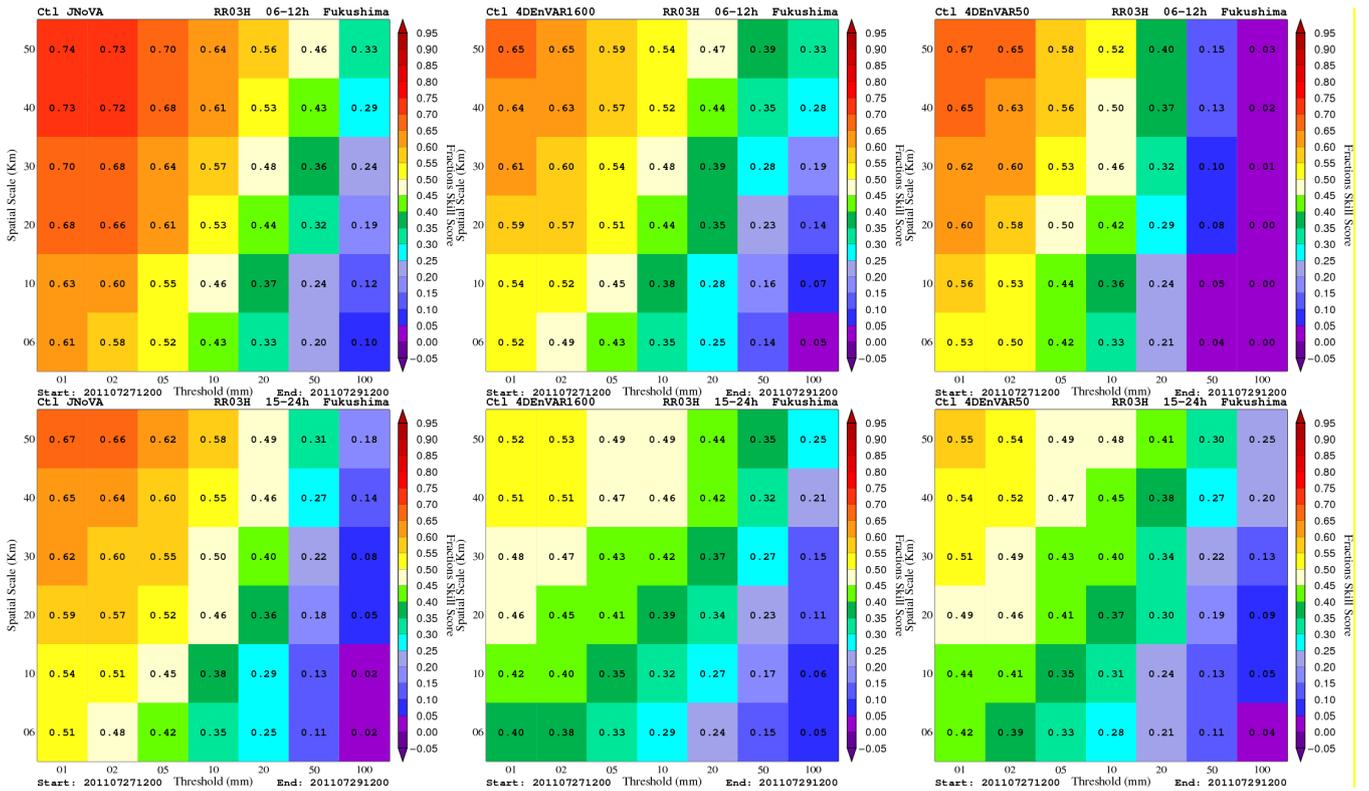
15

20

25

30

List of figures



5 **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.**

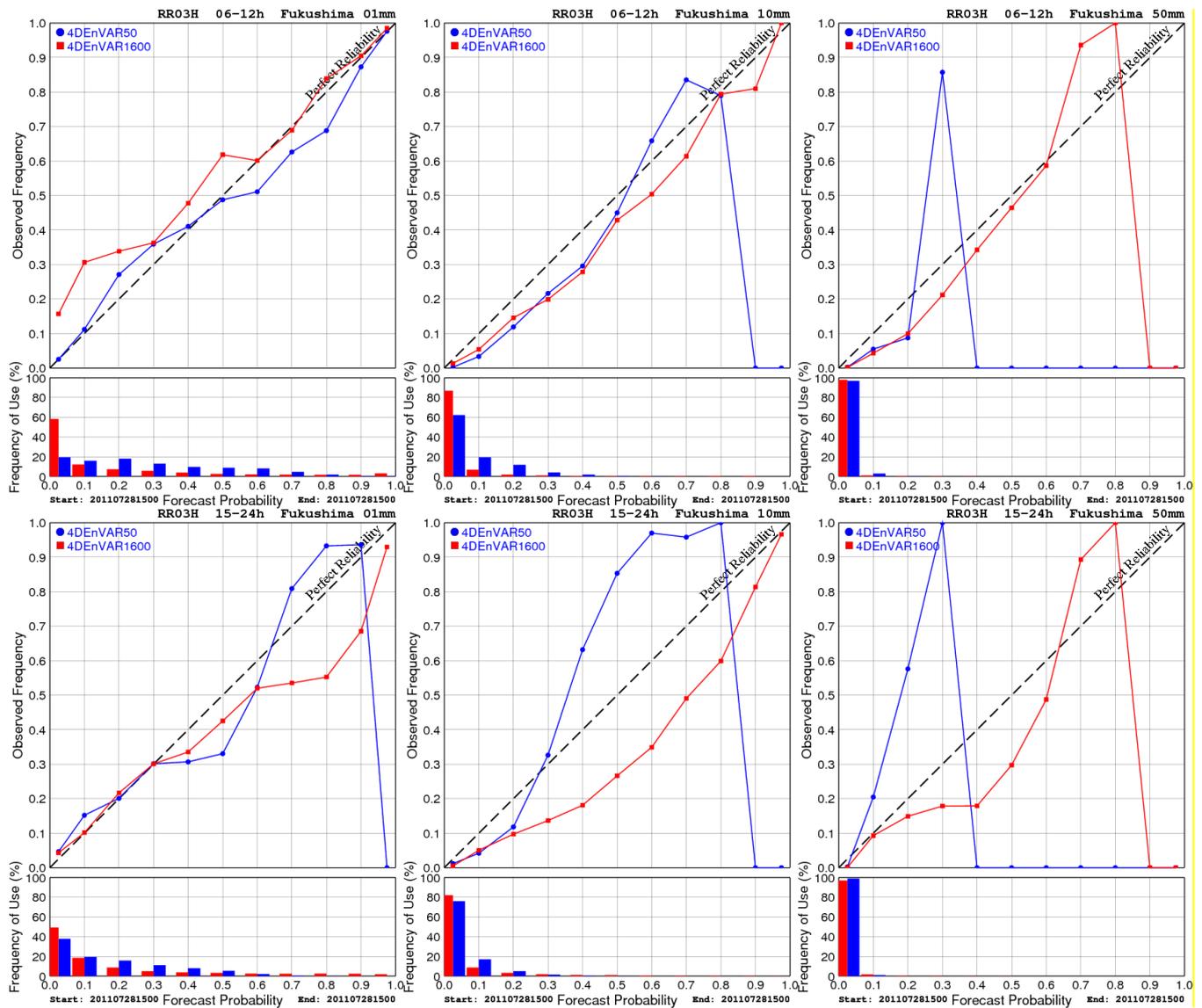


Figure 2. As Figure 1 but for reliability diagrams of 3-hour precipitation from ensemble forecasts initialized by analysis ensembles of 4D-EnVAR-NHM using 1600 and 50 members. Three precipitation thresholds of 01 mm (left), 10 mm (center), and 50 mm (right) are chosen. Note that the ensemble forecasts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

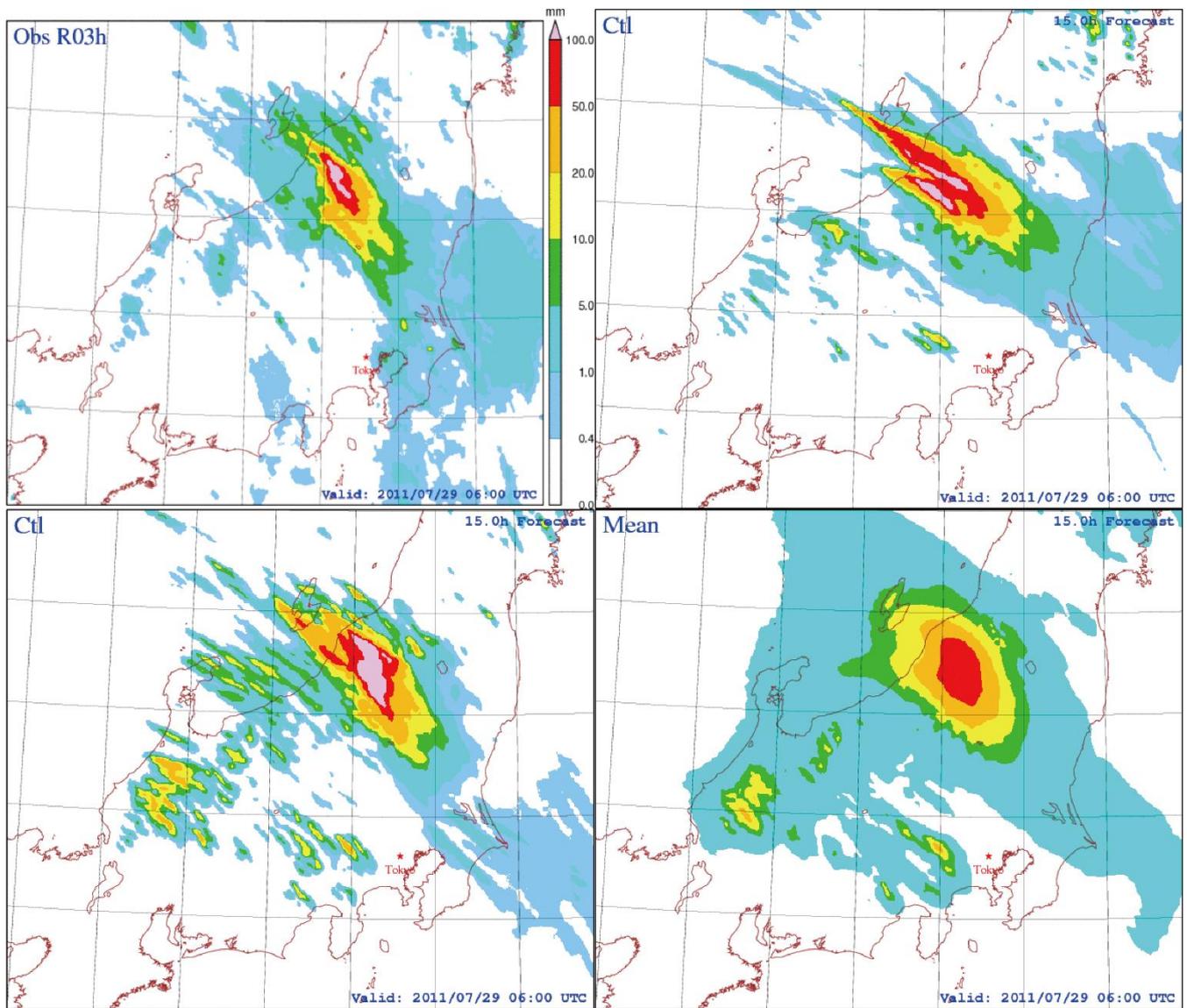


Figure 3. Three-hour accumulated precipitation for 1200-1500 JST July 29th, 2011 at Fukushima-Niigata as observed by Radar-AMeDAS (R/A; top left), forecasted by NHM initialized by the analysis of JNoVA (top right), forecasted by NHM initialized by the analysis of 4D-EnVAR-NHM (bottom left), and the ensemble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

5

Hourly Rainfall Forecast

Exp: Japan02km, Ensemble members: 1600

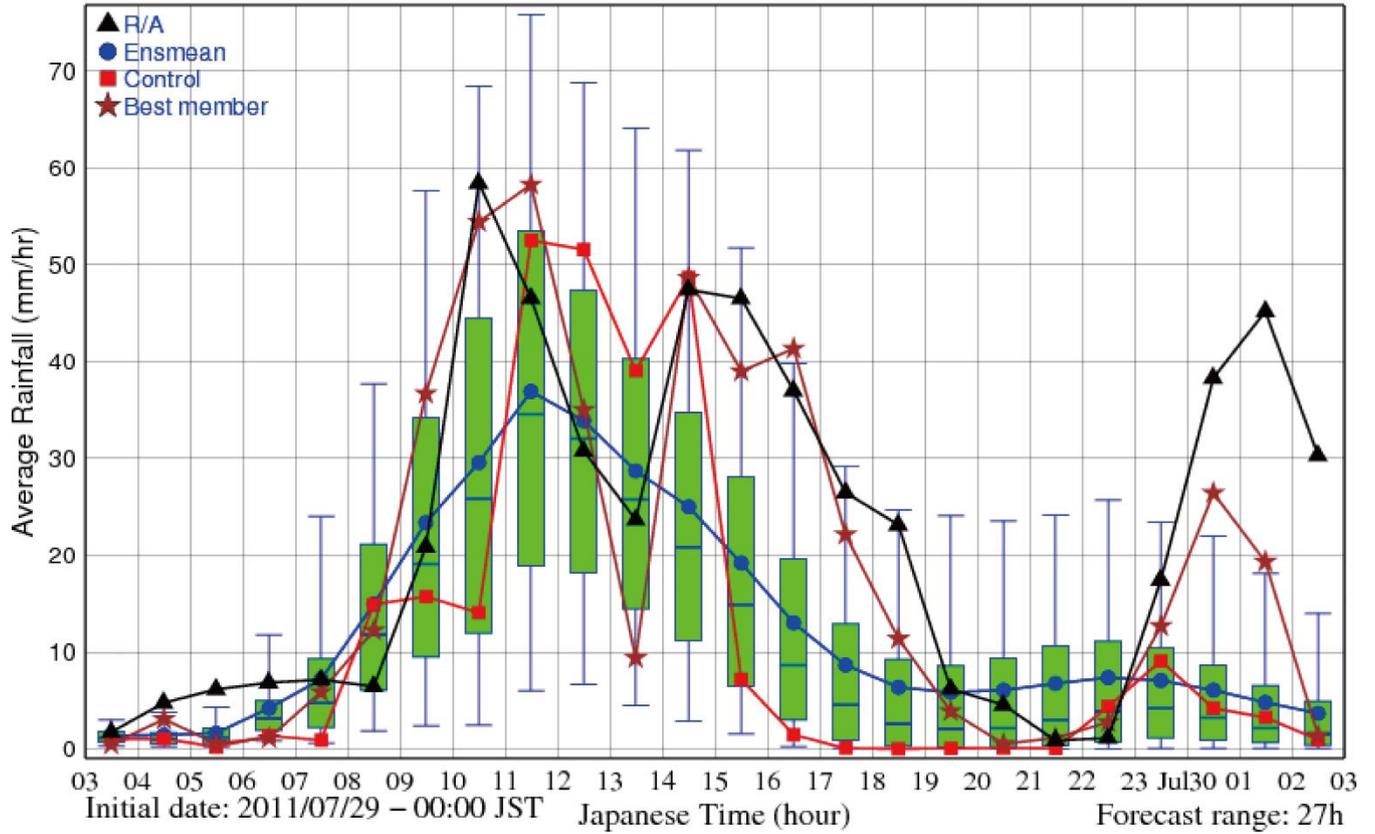


Figure 4. Time series of one-hour accumulated rainfall over the catchment as forecasted by all ensemble members. The two whiskers in each box-and-whisker diagram show the inter-quartile and 5th and 95th percentile of forecasted precipitation. The observation, control forecast, ensemble mean forecast, 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.

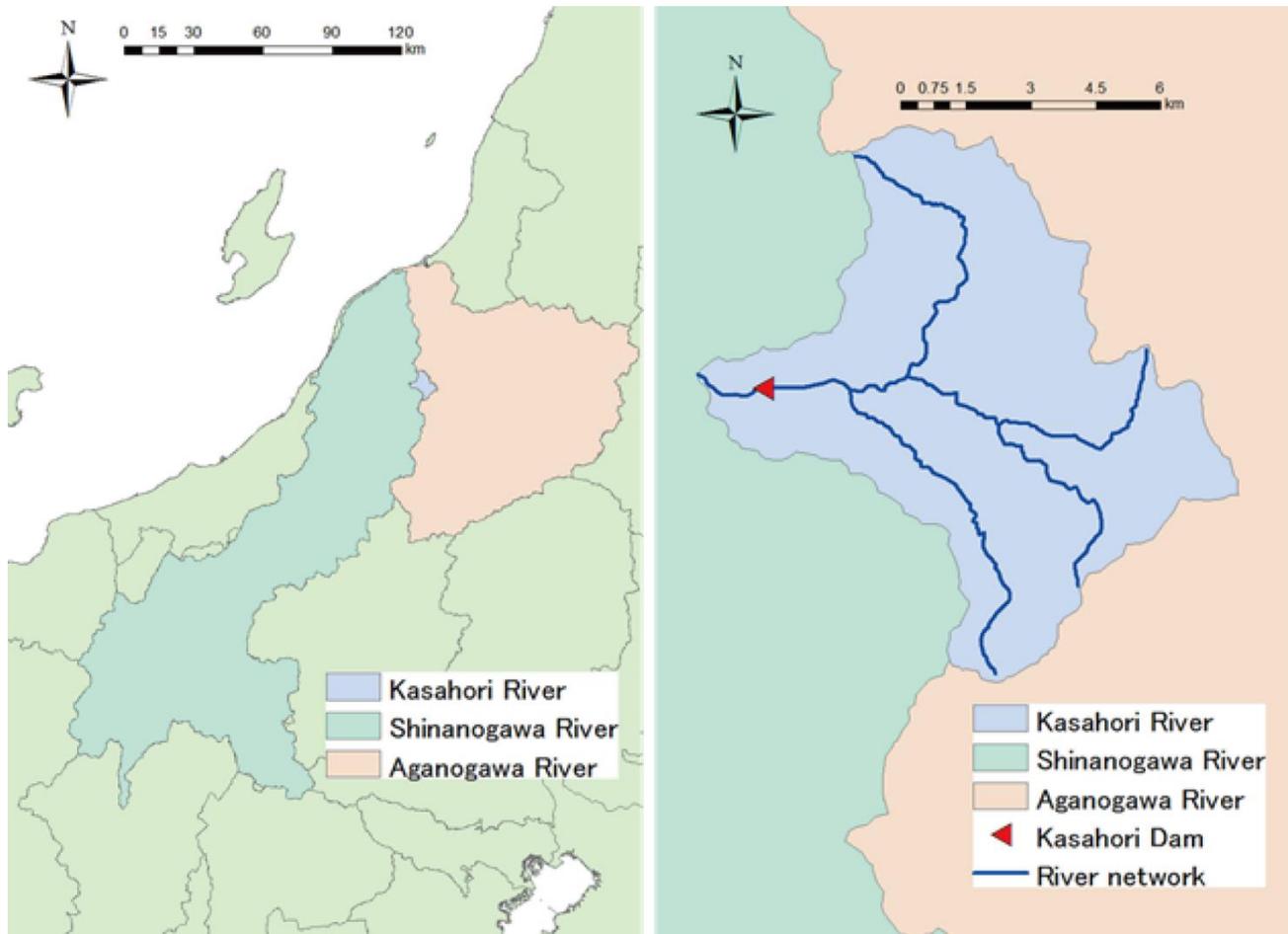


Figure 5. Map of the Kasahori, Shinanogawa, and Aganogawa river catchments in Niigata, Fukushima, and Nagano prefectures, Japan (left), and detailed view of Kasahori River catchments (right).

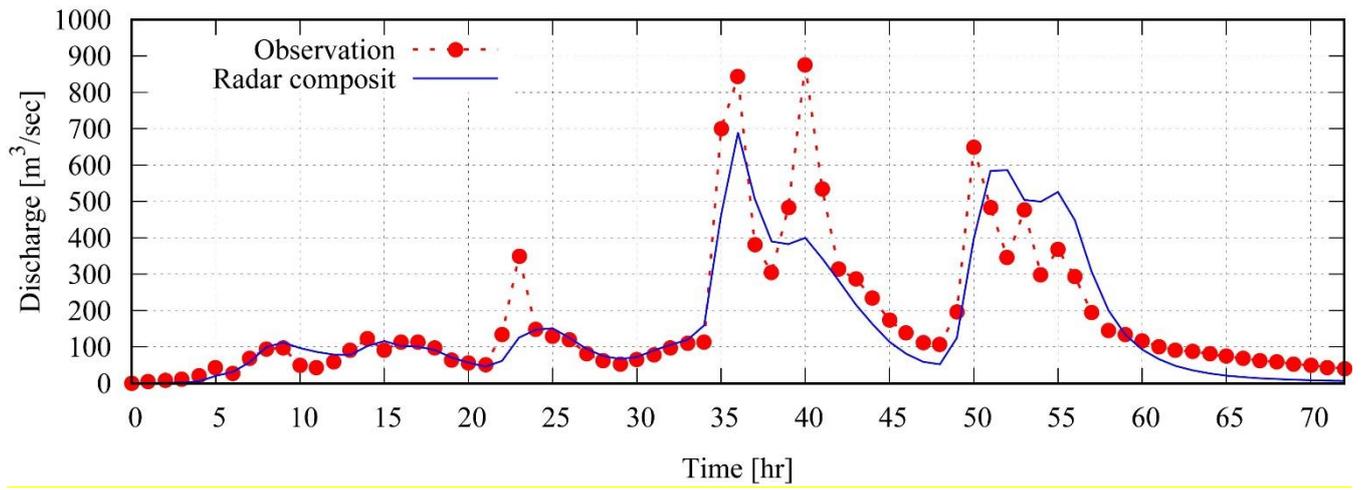


Figure 6. The observed inflow and simulated dam inflow using Radar-Composite.

5

10

15

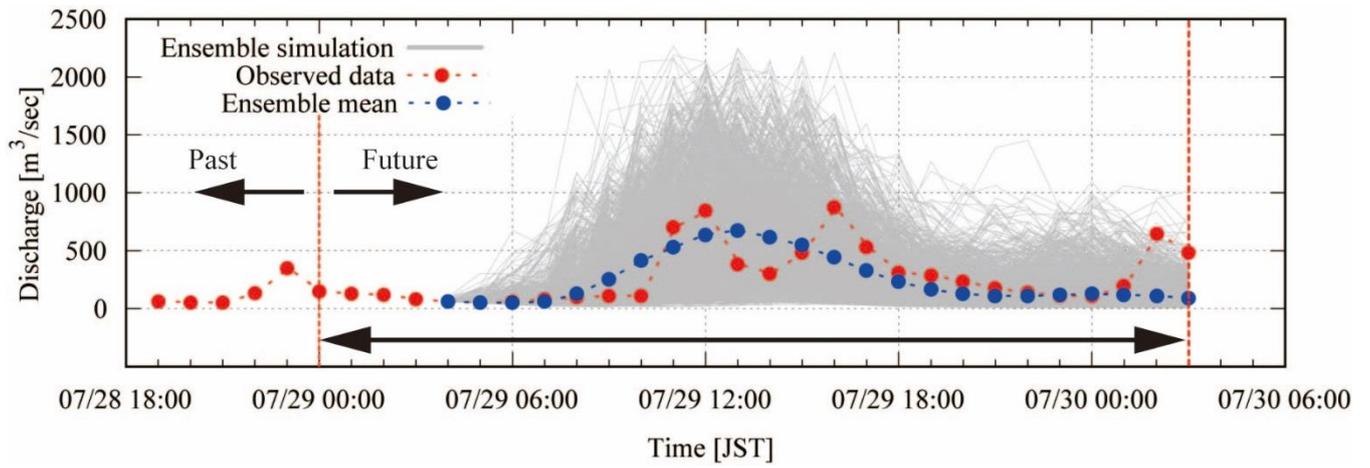


Figure 7. Total of 1600 ensemble inflow simulations to the Kasahori Dam, as well as mean and the observation.

5

10

15

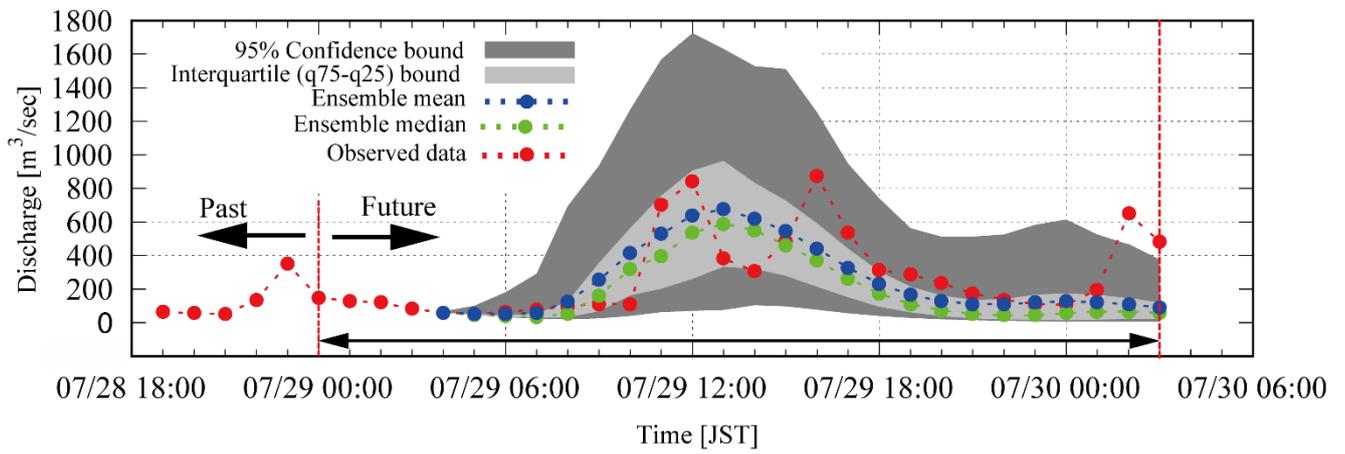


Figure 8. The 95% confidence limits and inter-quartile limits of the 1600 ensemble members.

Hourly Discharge Forecast Probability

Critical discharge: 140 m³/s

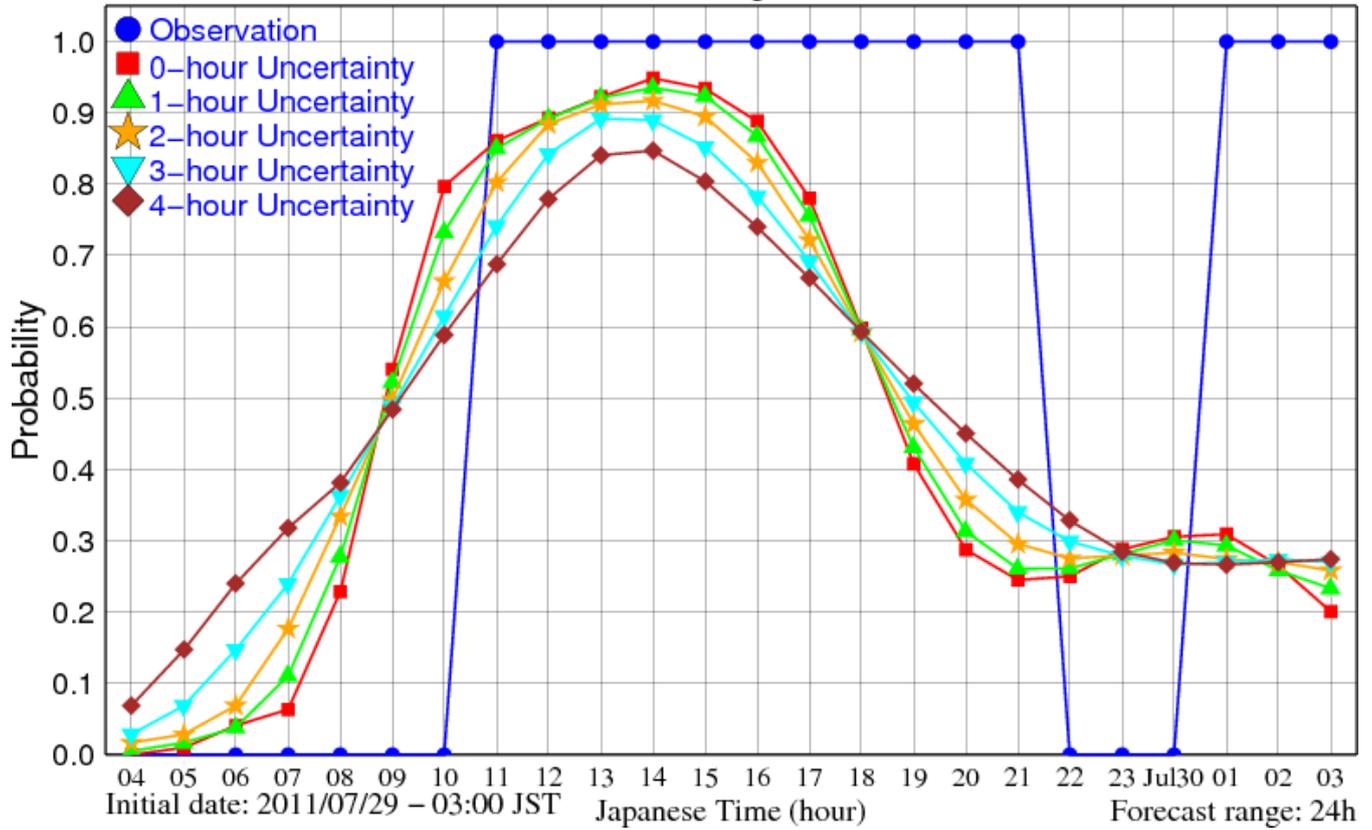


Figure 9. Probability that the simulated inflow is beyond 140 m³/s considering temporal uncertainty.

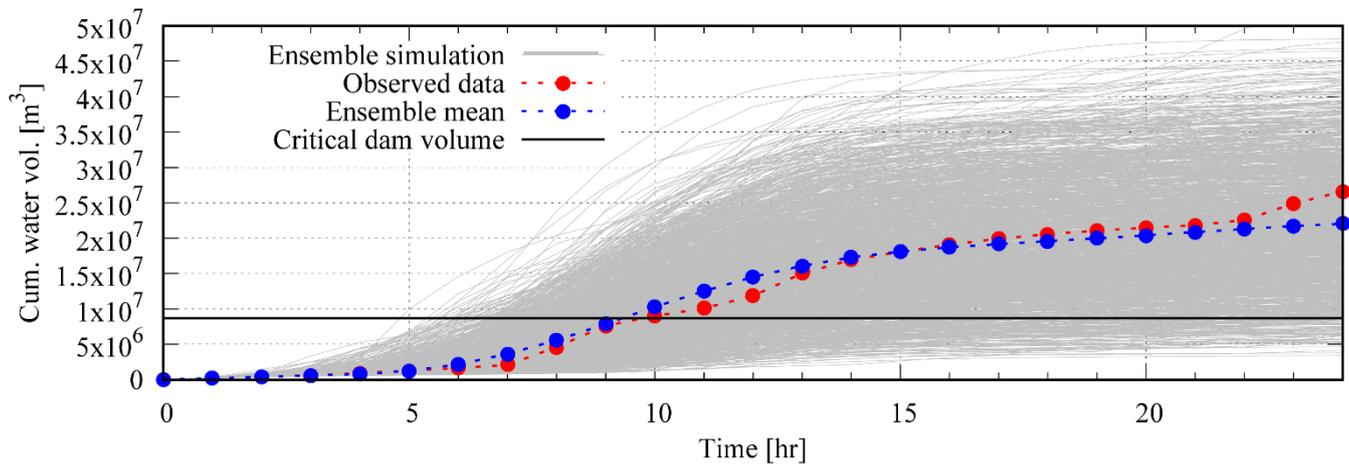


Figure 10. Cumulative dam inflow by the ensemble simulations, mean of simulation and observations, as well as critical dam volume.

5

10

15

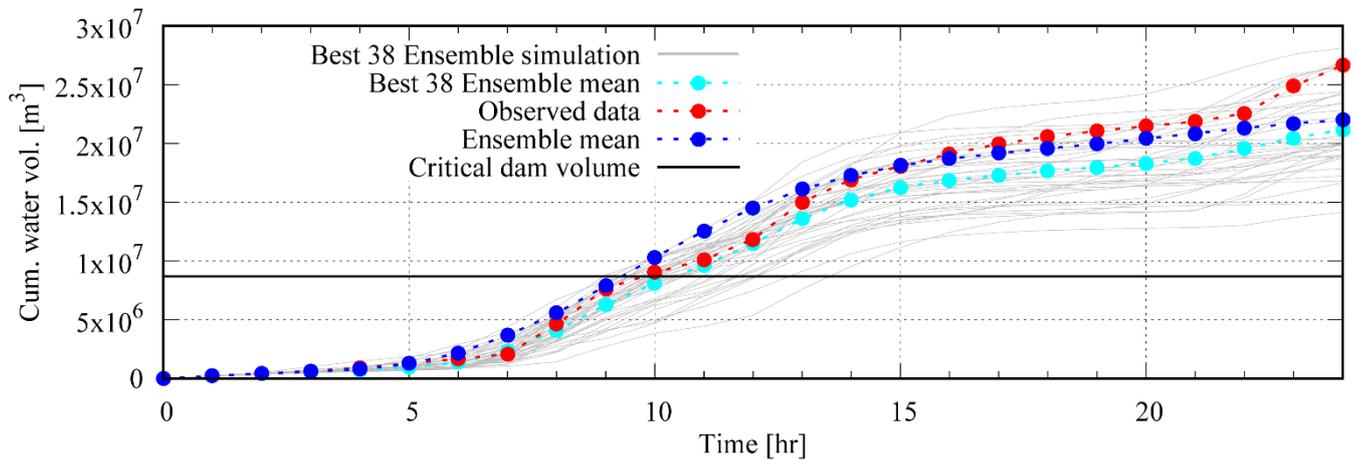


Figure 11. Cumulative dam inflow by the best 38 ensemble members, mean of the best 38 ensemble members, mean of all ensemble members, and observations, as well as critical dam volume.

5

Accumulated Volume Forecast Probability

Critical volume: 8700000 m³

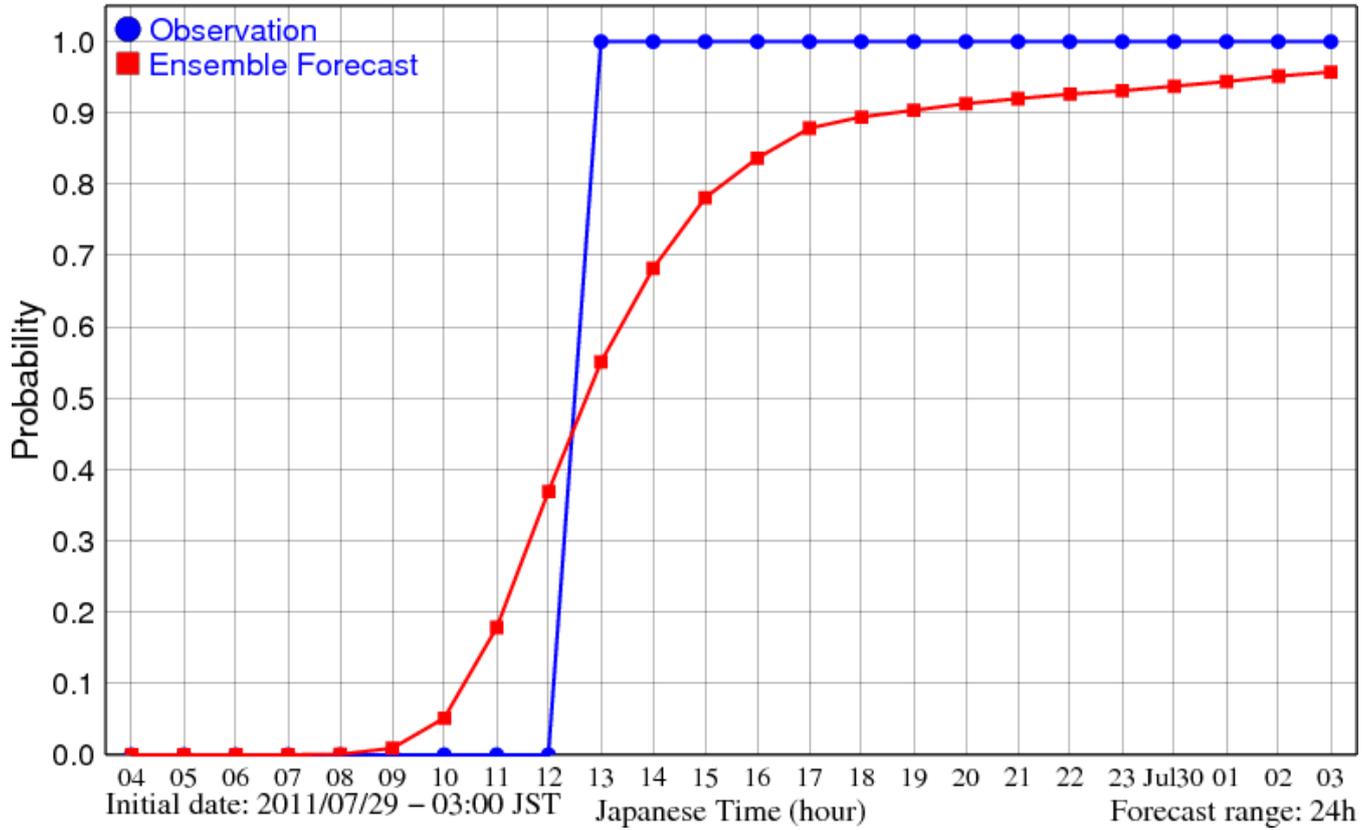


Figure 12. Probability that the dam needs emergency operation.

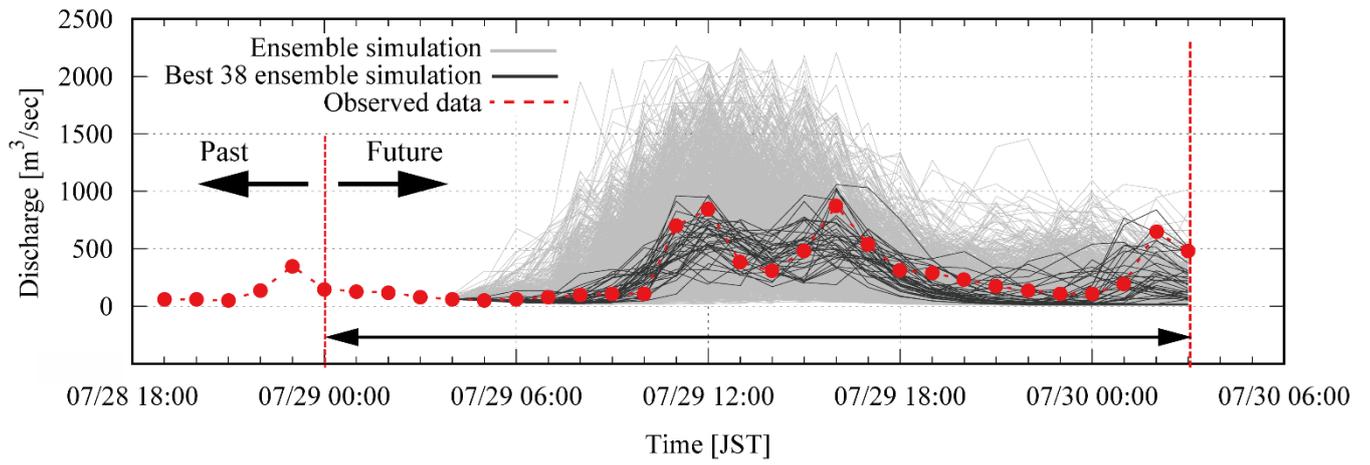


Figure 13. Hydrographs of all 1600 ensemble members, the 38 best ensemble members, and observations.

5

10

15

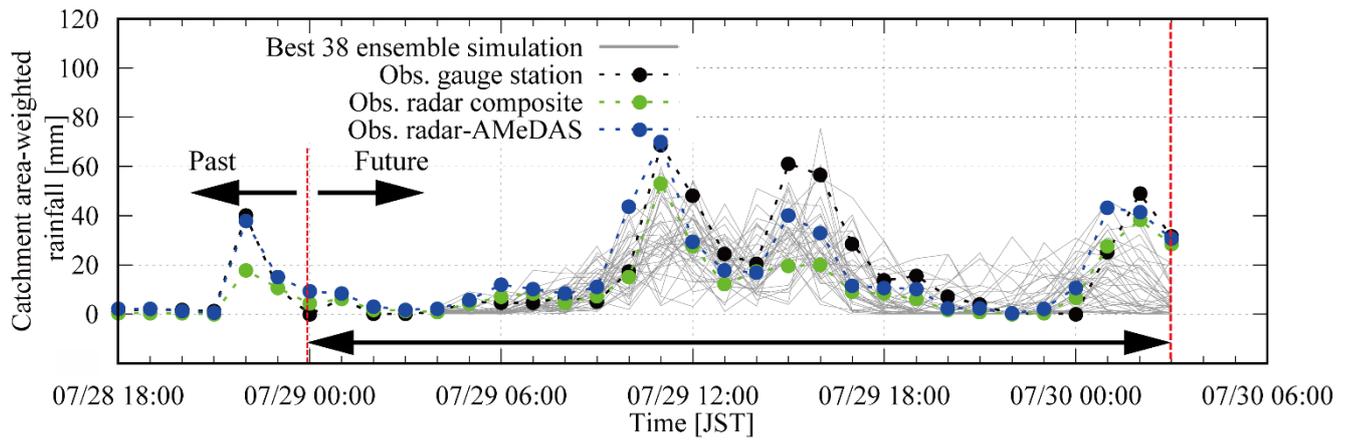


Figure 14. Rainfall intensity of the 38 best ensemble inflow simulation members, of Radar AMeDAS, of Radar-Composite, and ground observations.

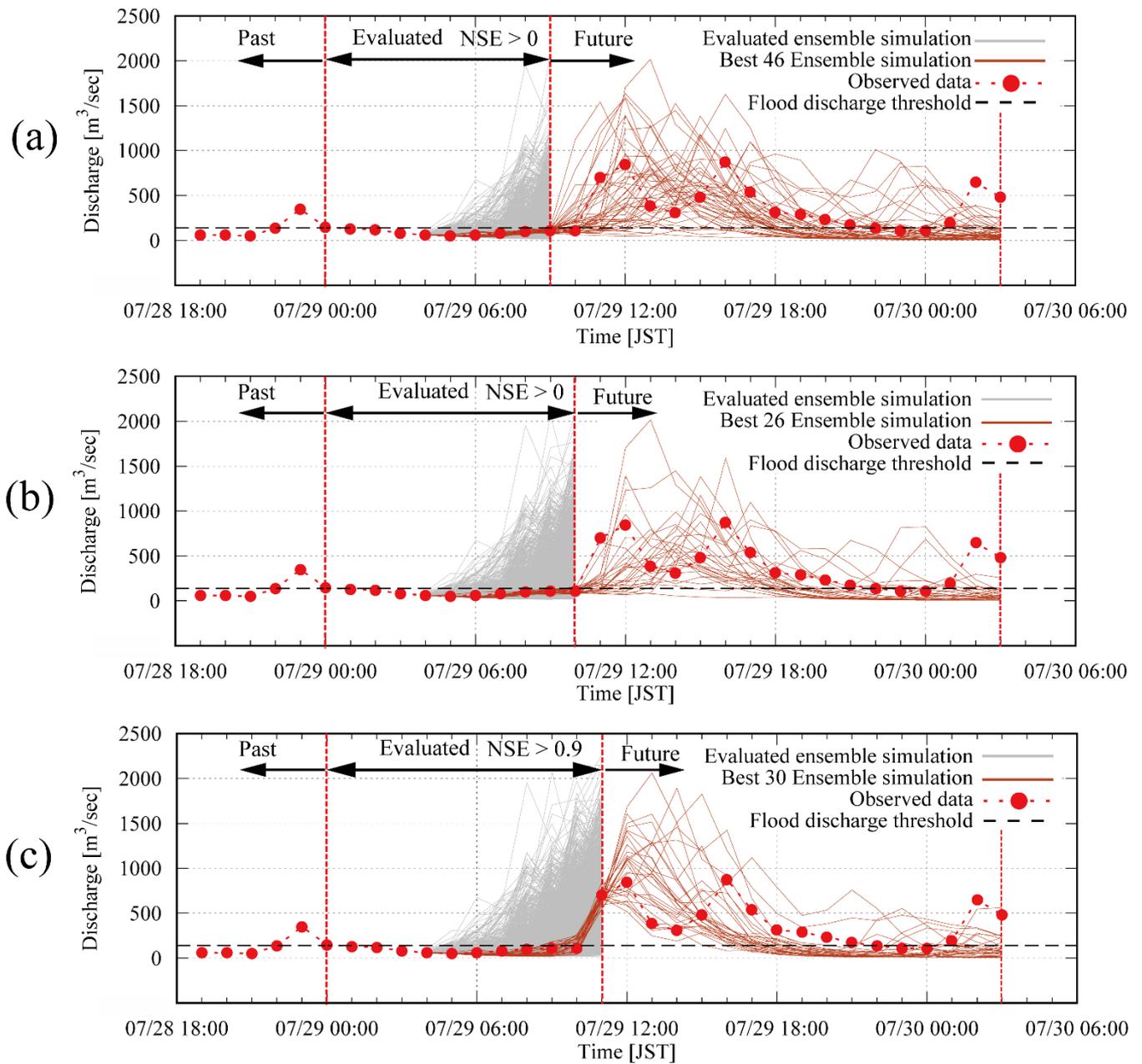


Figure 15. (a) 46 ensemble members ($NSE > 0.0$) selected from first 9-hour forecast, (b) 26 ensemble members ($NSE > 0.0$) selected from first 10-hour forecast, and (c) 30 ensemble members ($NSE > 0.9$) selected from first 11-hour forecast.

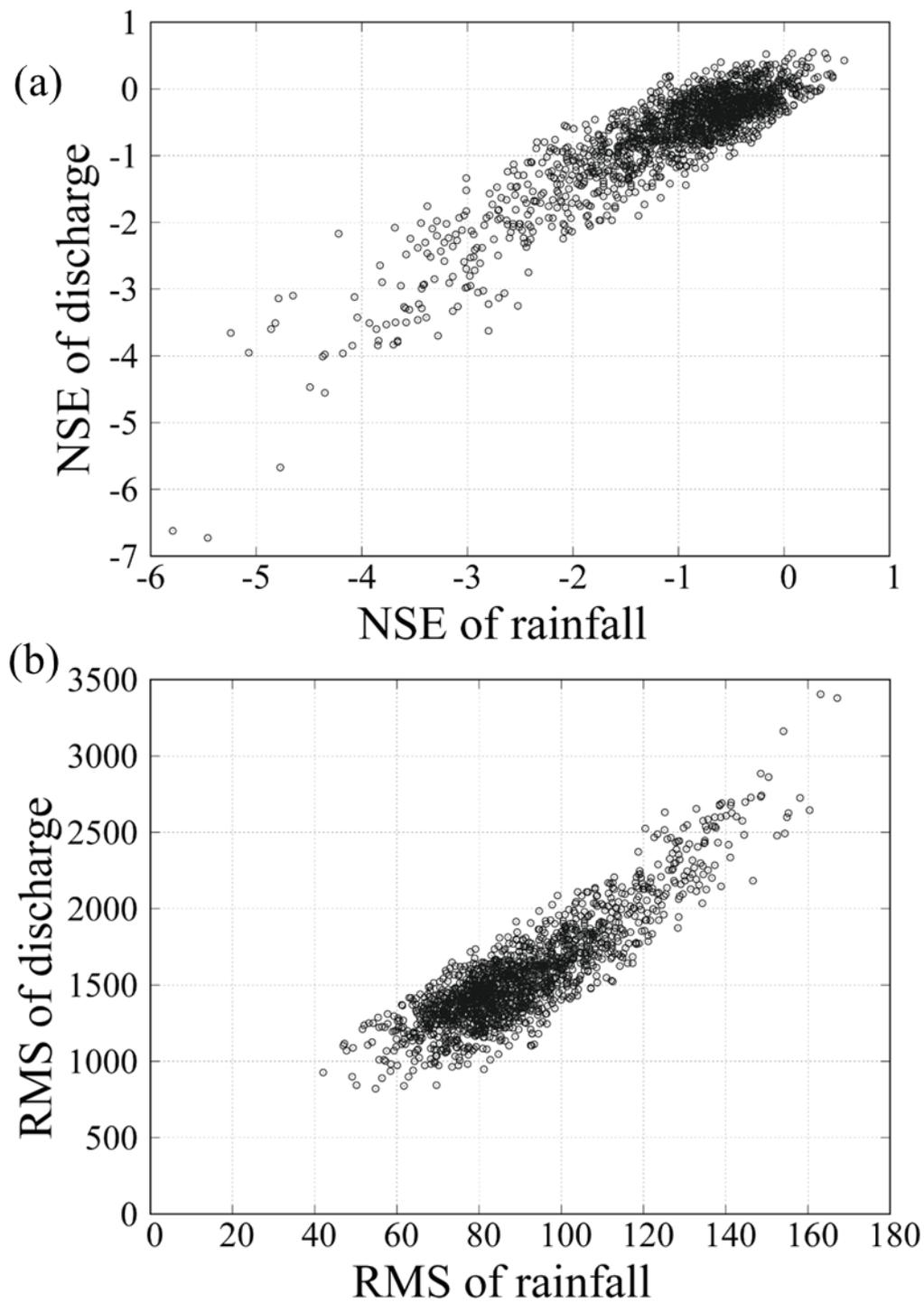


Figure 16. (a) NSE of rainfall vs NSE of discharge, (b) RMS of rainfall vs RMS of discharge