Dear Dr. Ehret,

Thank you very much for your positive assessment and constructive suggestions to our study. After carefully reading your comments, we do agree that an in-depth analysis should be made on the rationality of our choice of the modeling and evaluation domains. Therefore, we have repeated the analysis on the D03 scale and compared the results with those derived from the D02 scale. Before we come to our conclusions, some information related to your concerns is added and presented first.

• The main objective of this study is to evaluate the effect of WRF setups on the precipitation simulations during the sub-daily heavy rainfall event on July 21st, 2012 in Beijing, China (sub-daily, regional scale).

• Beijing is the area of interest where the convective processes (convective-scale) happened, while the synoptic features (larger-scale) triggered the event, e.g., the eastward-moving vortex and the source of water vapor, were more spread outside this range. Therefore, 3 nested domains (D01, D02, D03) were set up, with D01 being the largest in the coarsest resolution, which covers the leading synoptic features, and D03 the smallest, which covers the area of interest (i.e., Beijing region).

• The initial conditions of the 3 domains were all provided by the ERA-Interim reanalysis and the regional geographical dataset. As for the lateral boundary conditions, D01 was forced by the ERA-Interim reanalysis, while D02 by D01, and D03 by D02.

• For the analysis/evaluation, the rainfall fields from D03 were mapped back onto the grid of the D02 domain. The analysis and evaluation were then done with the hybrid dataset D02+partly D03, on the D02 domain, against the ground rainfall observations and the ERA-interim reanalysis.

• The reasons for doing so are that (1) the spatial resolution of the reference truth (an hourly gridded dataset publicly available with the spatial resolution of 0.1 degrees) is commensurate with the D02 resolution, (2) the effect of some WRF model configurations (e.g., the domain size) on simulating this heavy rainfall event could not be well presented if it is just evaluated on the D03 scale.

In order to verify whether the choice of the evaluation domain is reasonable, we have recalculated all the metrics on the D03 scale by using another 3-hourly gridded dataset with a finer resolution of 0.05 degrees (Huang et al., 2013). A detailed comparison was then made based on the results derived from the two different scales (the D02 and D03 scale).

Huang, C., Zheng, X., Tait, A., Dai, Y., Yang, C., Chen, Z., Li, T., and Wang Z.: On using smoothing spline and residual correction to fuse rain gauge observations and remote sensing data, J. Hydrol., 508, 410–417, 2013.

By comparison, we have noticed that most of the evaluated results on the D03 scale were similar to those got on the D02 scale (See **Fig. 1** to **Fig. 8**). This has indicated that the experiment performs well on the larger scale could also have a good performance on the finer scale. It makes sense as the one with higher similarity to the larger-scale synoptic features tends to provide more accurate boundary conditions for the modeling in the inner domain. This, in other words, means that the experiment with good performance in the inner domain should also perform well in the larger-scale domain, which could be useful in evaluating the regional weather forecasts.

[Figure 1 to Figure 8]

Taking the domain size scenario as example, when it was evaluated on the D03 scale (See **Fig.1**), Case 0 with the smallest domain size performed better than the other two experiments in terms of the accuracy of rainfall during the first 18 hours. But from the D02 scale (See **Fig.2** and **Fig. 9**), it could be noticed that either the moving speed of the rain-belt or the magnitude of the maximum precipitation simulated by Case 0 was much different from the reference truth.

[Figure 9]

However, we agree that since our goal is to evaluate the effect of the WRF configurations on the heavy rainfall process in Beijing region. The choice of the hybrid domain for evaluation could lead to the possible ambiguity on distinguishing the source of the effect. Therefore, we would like to adopt the first option as suggested by you: repeating the evaluation on the D03 scale. As we have already recalculated the metrics for all the experiments on this scale, we are confident that 2-3 months are sufficient for us to revise this paper.

We appreciate your help in improving this manuscript, and we hope that our replies have addressed your remaining concerns.

Kind Regards!

The Authors

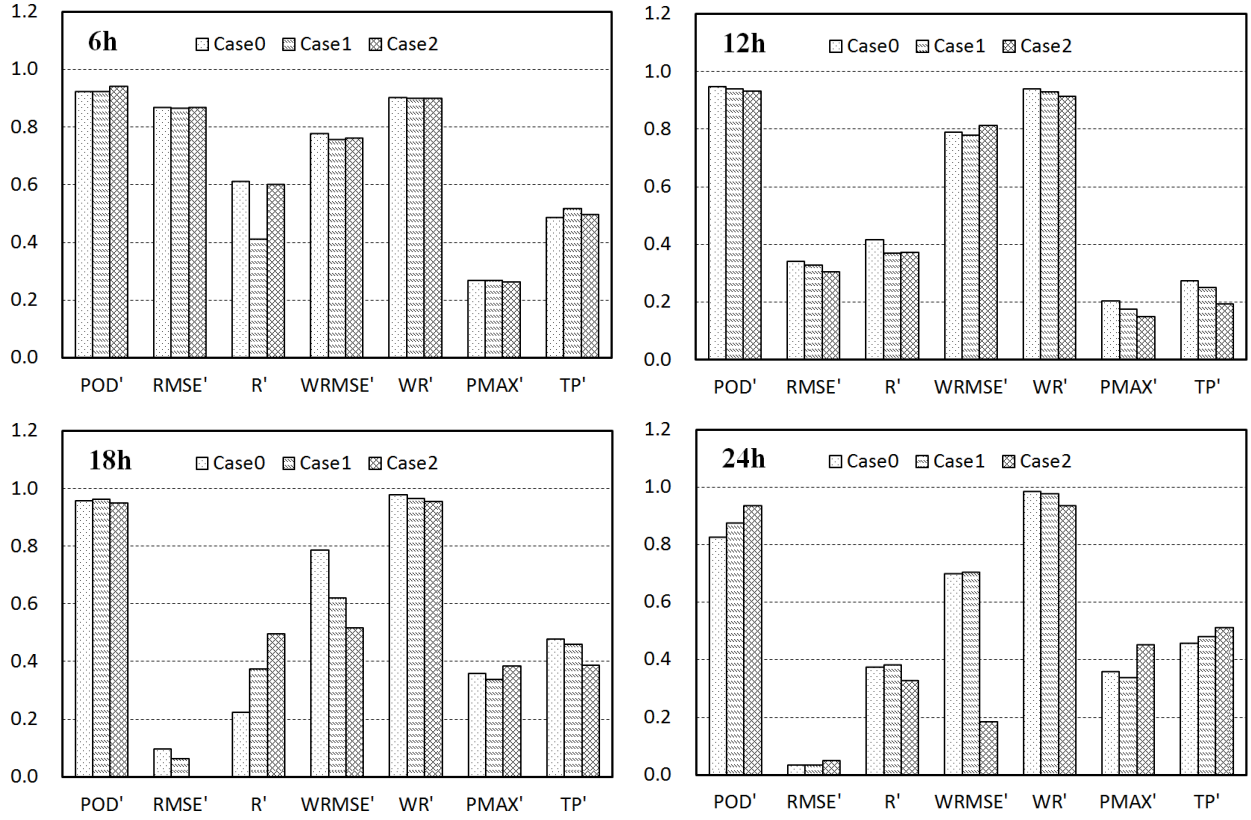

Figure 1: Spatial values of the verification metrics calculated on the D03 scale for the experiments with different domain size (Case 0 has the smallest domain size covering Northern-Central China; Case 1 has the intermediate domain size covering Northern China and a part of the Mongol Country; Case 2 has the largest domain size covering the northeastern hemisphere; the statistic time durations are 6 h, 12 h, 18 h, and 24 h, respectively, counting from 12 am 21 July 2012).

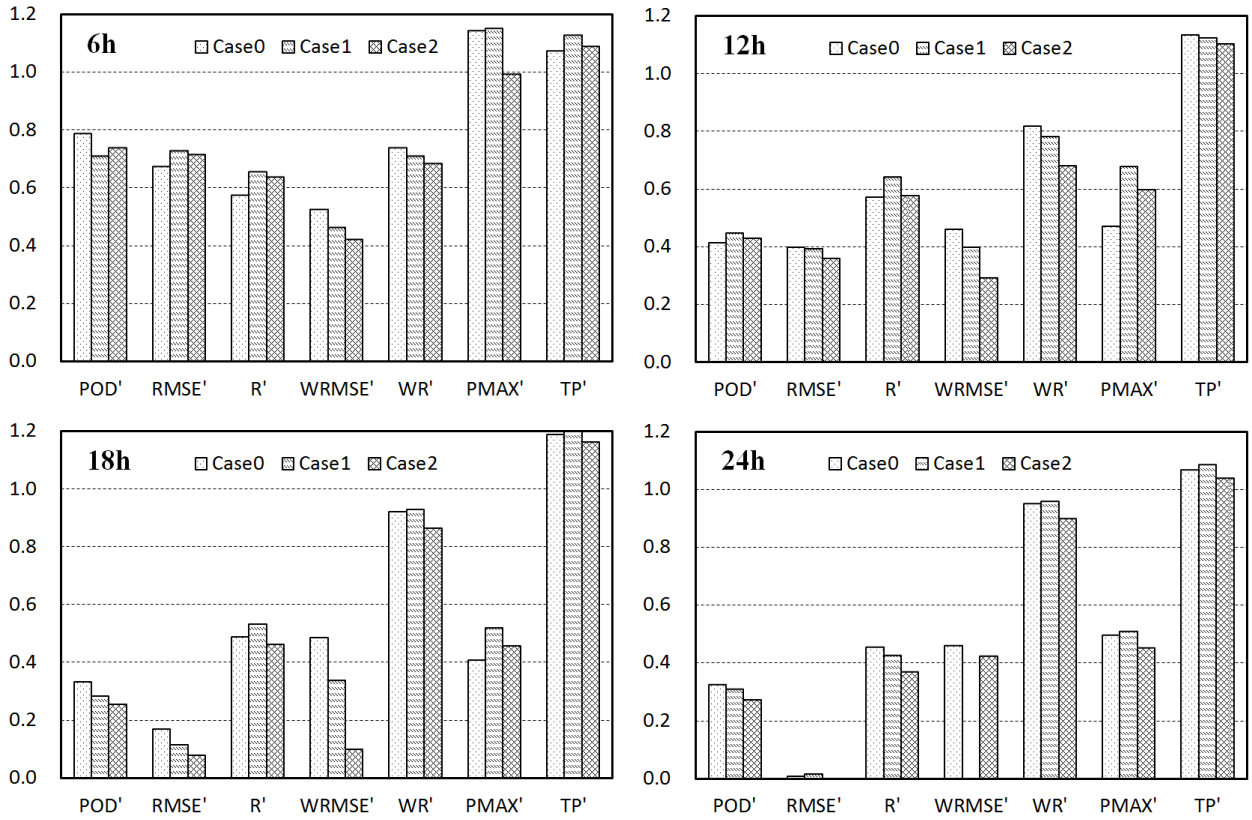

Figure 2: As in Fig. 1, but the metrics were calculated on the D02 scale.

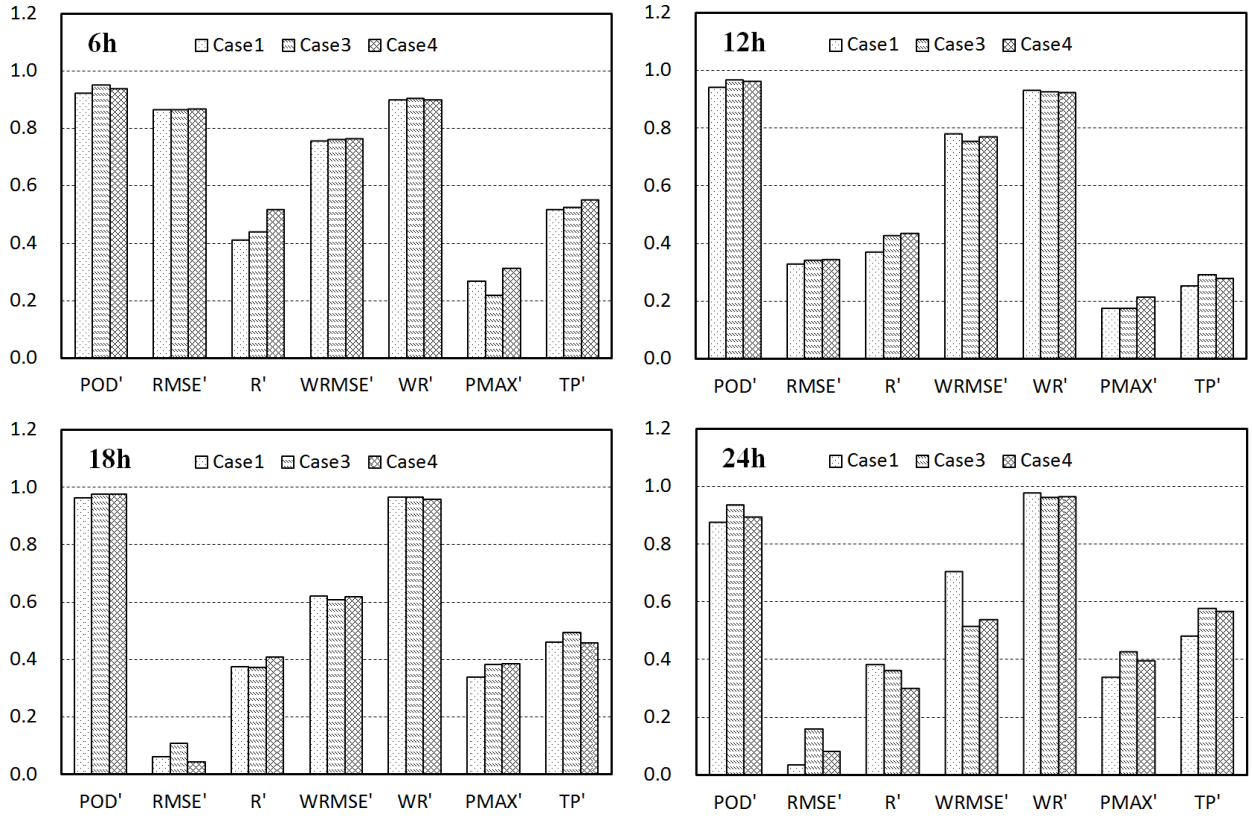

Figure 3: Spatial values of the verification metrics calculated on the D03 scale for the experiments with different vertical resolution (Case 1 has 29 vertical levels equal to that of the ERA-Interim reanalysis, Case 3 and Case 4 has doubled and tripled vertical levels, respectively).

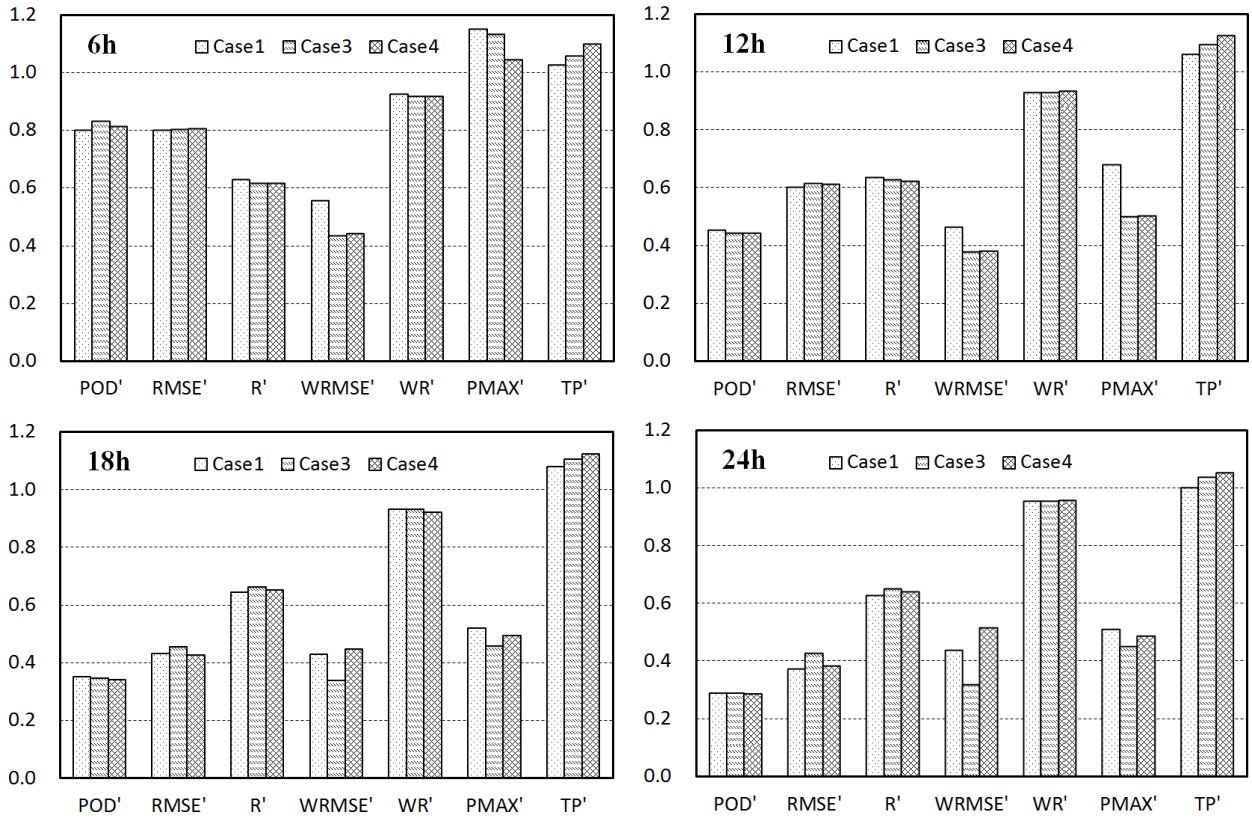

Figure 4: As in Fig. 3, but the metrics were calculated on the D02 scale.

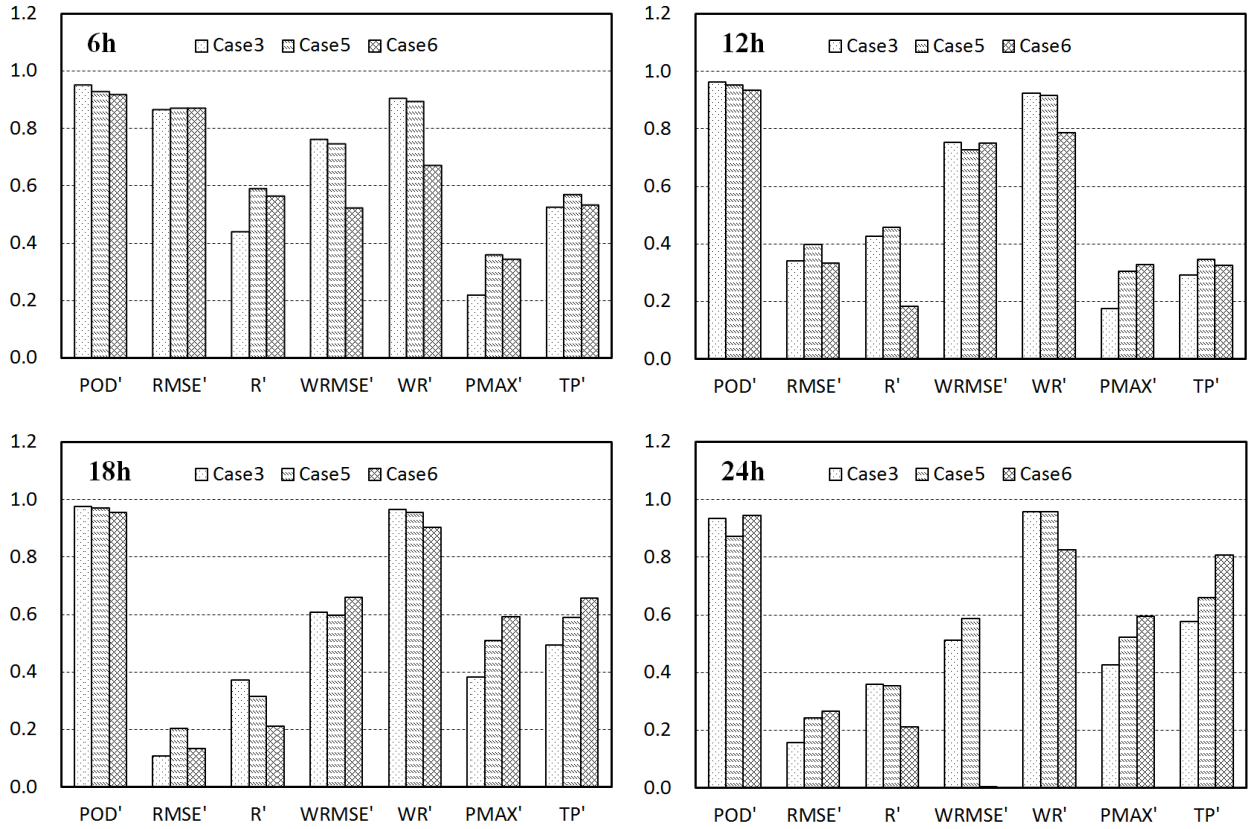

Figure 5: Spatial values of the verification metrics calculated on the D03 scale for the experiments with different horizontal resolution (Case 3 has the initial downscaling ratio of 1:3:3 with the outermost grid size of 40.5 km; Case 5 and Case 6 have the same outermost grid size with 1:5:5 and 1:7:7 downscaling ratio, respectively).

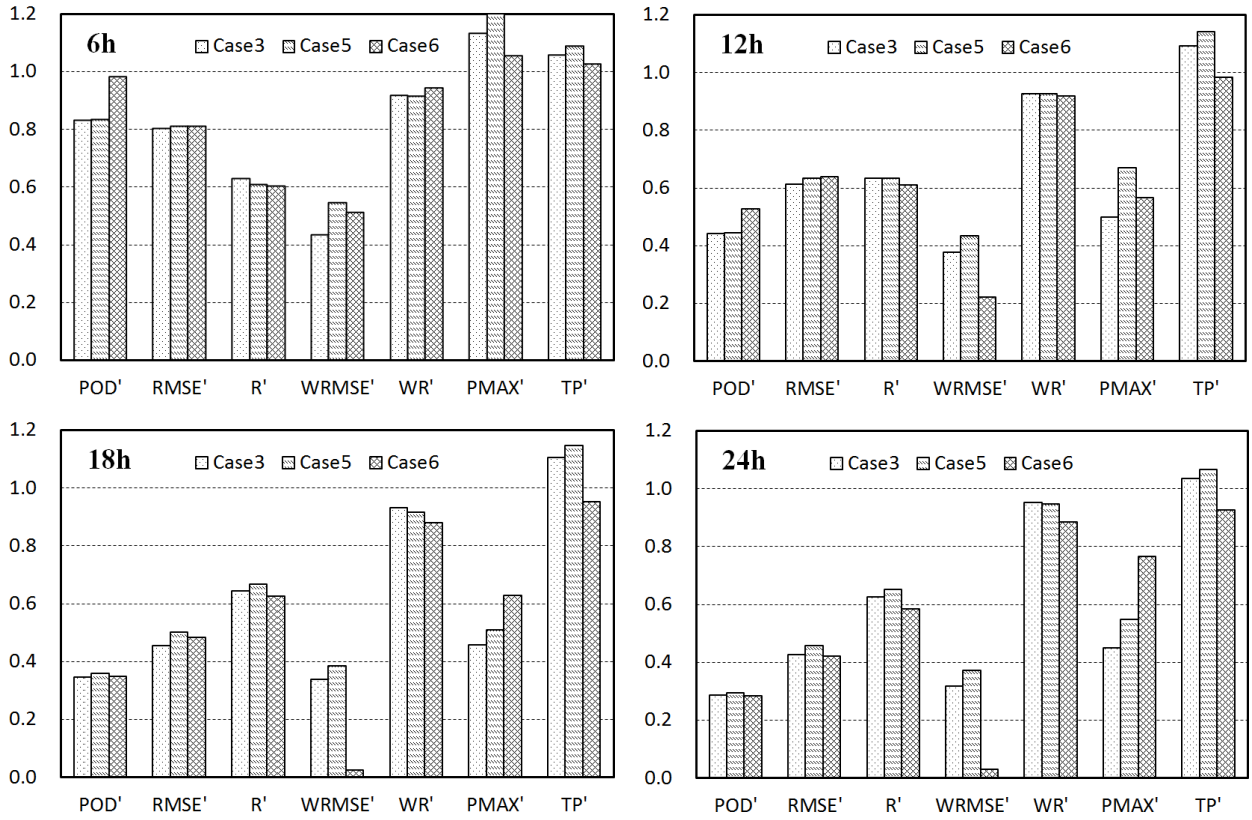

Figure 6: As in Fig. 5, but the metrics were calculated on the D02 scale.

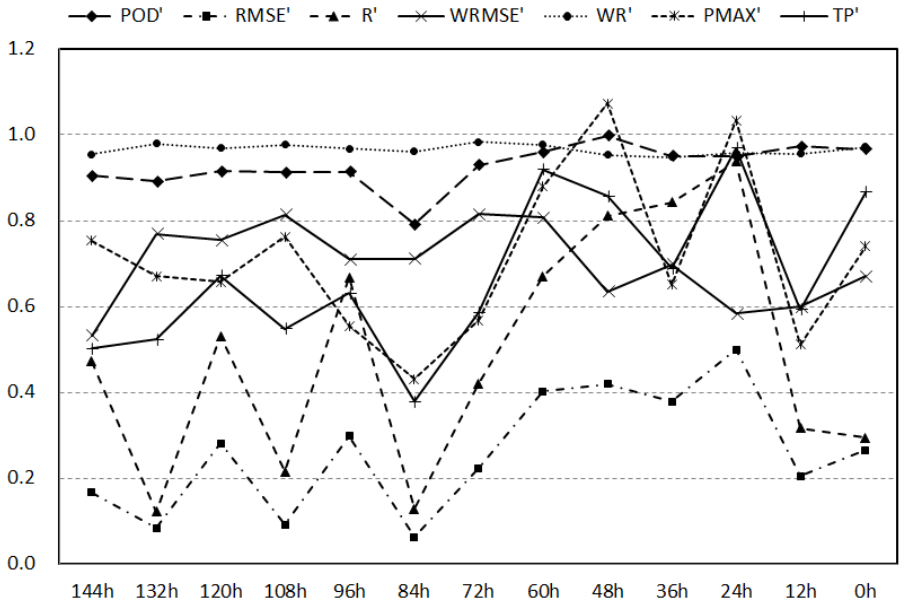

Figure 7: Spatial values of the verification metrics calculated within 18 h time periods on the D03 scale for the WRF spin-up experiments (Case 5 has the initial spin-up time of 12 h; Case 7 is designed with 0 h spin-up time; From Case 8 to Case 18, the spin-up time is increased by 12 h from 24 h).

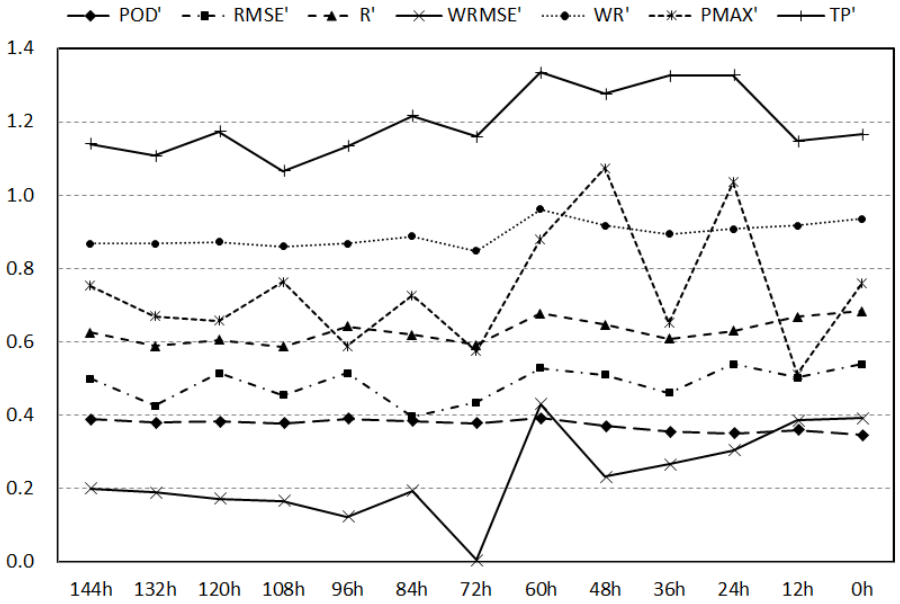

Figure 8: As in Fig. 7, but the metrics were calculated on the D02 scale.

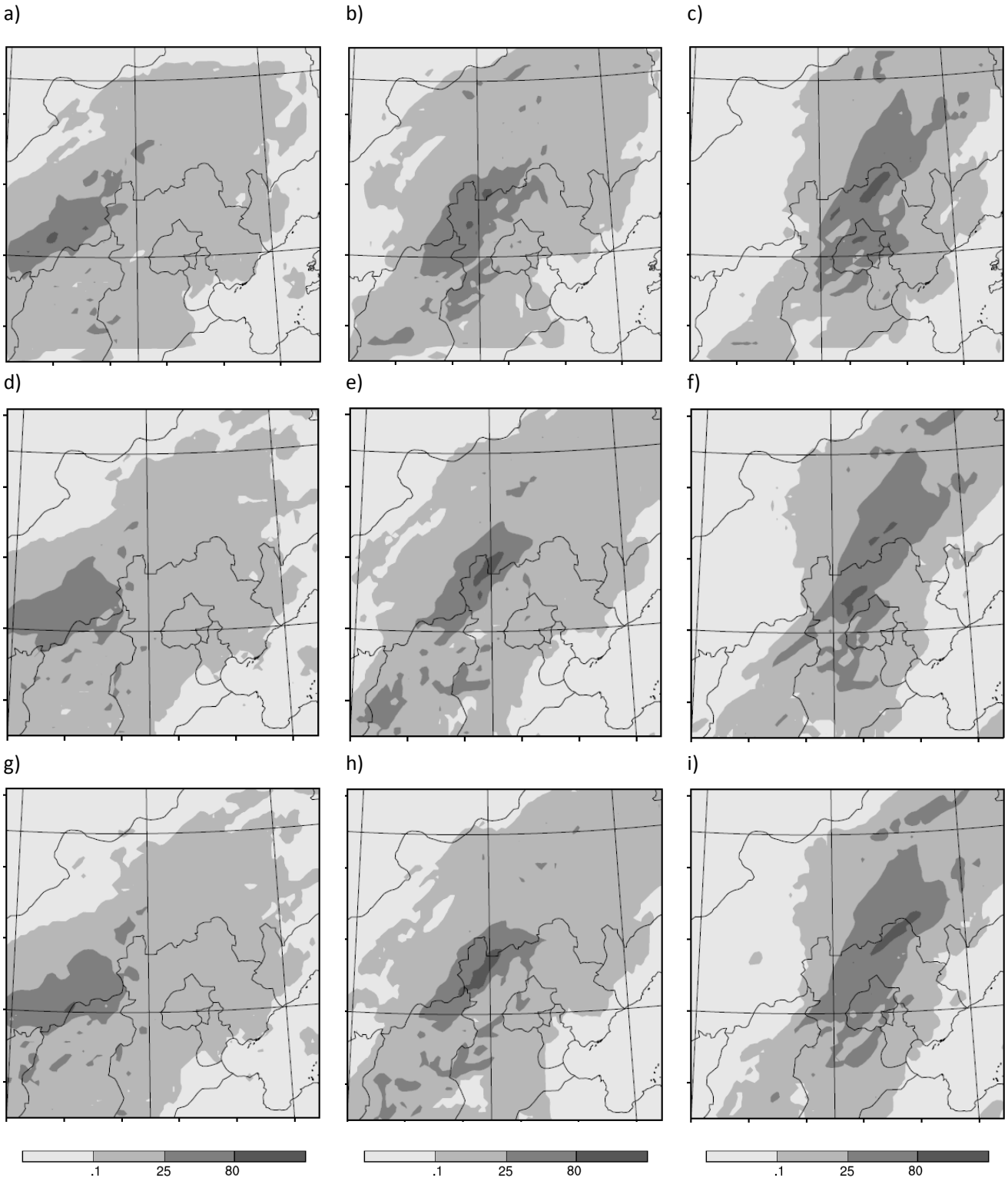

Figure 9: Spatial distribution of six hour accumulated precipitation for the domain size experiments within the D02 domain during the heavy rainfall event from 12am, July 21, 2012 (a) precipitation of Case 0 (with the smallest domain size) in the first 6h; (b) precipitation of Case 0 in the second 6h; (c) precipitation of Case 0 in the third 6h; (d) precipitation of Case 1 (with the medium domain size) in the first 6h; (e) precipitation of Case 1 in the second 6h, (f) precipitation of Case 1 in the third 6h; g) precipitation of Case 2 (with the largest domain size) in the first 6h; h) precipitation of Case 2 in the second 6h; and i) precipitation of Case 2 in the third 6h.