The revised paper has not been uploaded because we are waiting for the comments from other reviewers and the current version was sent to an English native speaker for checking now. But we want to respond to the comments received.

We would like to thank the referee for the helpful comments and thorough suggestions that will surely help us improve the manuscript. We will take all of them into consideration while revising the paper. Below, we provide our point-to-point response to the reviewer's comments and describe the modifications to address those issues.

Reviewer #1

"The authors present a study to investigate the impact of different domain sizes, vertical resolution, nesting ratios and spin-up time on a heavy precipitation event over Beijing. The simulations were forced by ERA-Interim reanalysis data available on 0.75 ° resolution in six hourly intervals. The different experiments were performed using three domains with a two-way nesting approach and the innermost domain centered on Beijing. Sub-daily precipitation of the second domain was verified against gridded precipitation observations from the China Meteorological Center. In addition, the precipitable water content of the WRF simulations was validated with ERA-Interim reanalysis data as a proxy for the maximum possible precipitation.

In a general sense, this type of experiments is of great relevance for flash flood forecasting and early warning systems. However, in the current experimental setup, I see several critical points preventing the traceability of the results."

**Response**: Thank you for pointing this out. We really appreciate your time and effort invested in the review. We have carefully checked all issues and added the missing information to increase the traceability of the results.

"1) The authors apply 2-way nesting from the outer to the inner domains. This means that precipitation patterns of the 3rd domain (which is not analyzed in your study) are reflected in the second domain. This actually means that you verify the precipitation from domain three mixed with terrain and land use data from domain two. Why did you verify domain two instead of domain three in your study? When looking at Fig. 1, Beijing has complex terrain which is not accurately represented at 13.5 km resolution."

**Response**: Thank you for pointing this out. Indeed, domain three covers the main convective process and should be the best choice for urban flooding study. We did attempt to verify domain three when we started this research. But then we noticed that the influence of some WRF model configurations, such as domain size and spin-up time, on simulating this heavy rainfall event was not well presented by this innermost domain.

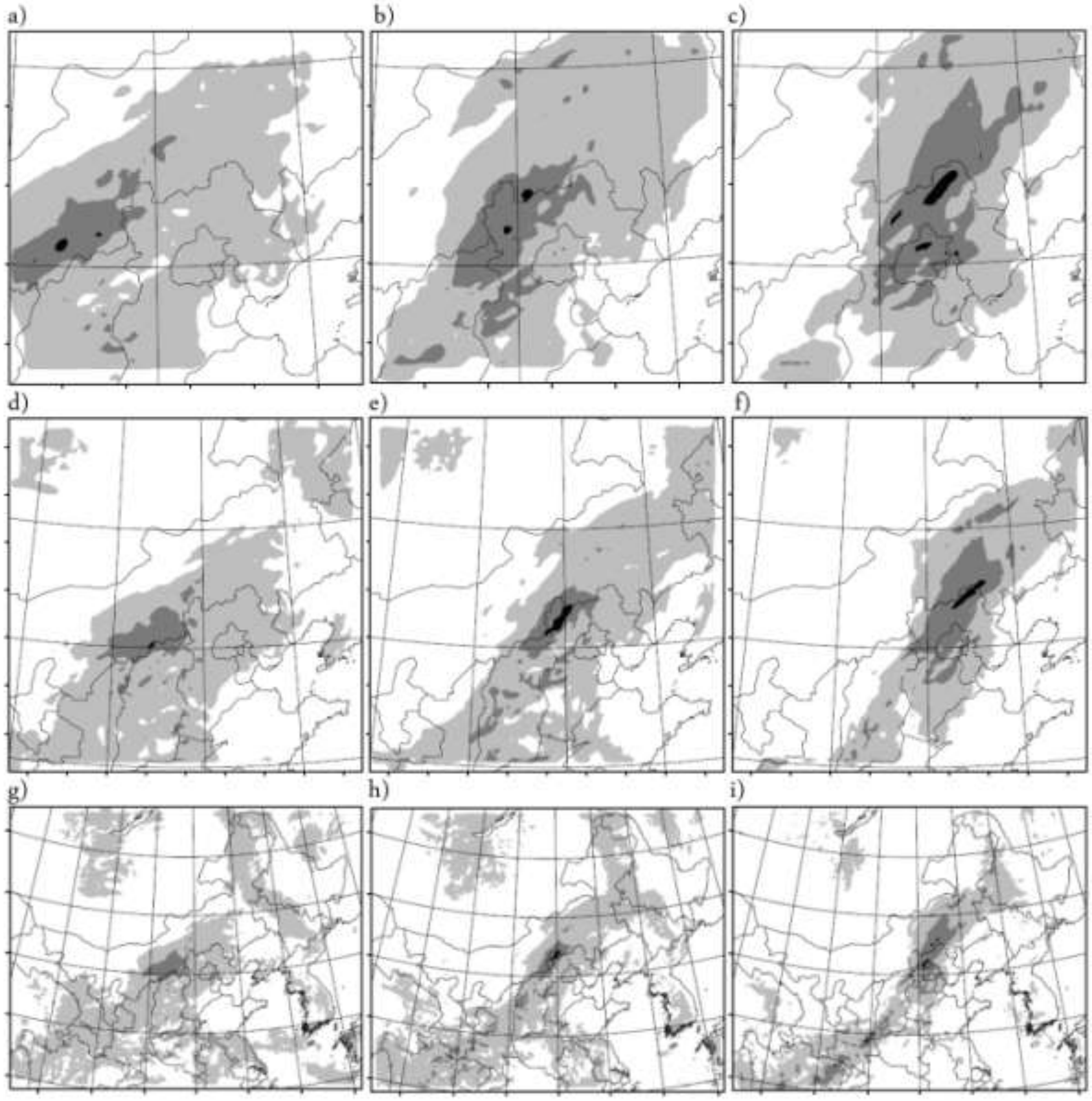


**Figure 1. Spatial Distribution of 6 hours Accumulated Precipitation with different domain size within D02 during the heavy rainfall event from 12 pm, July 21, 2012 (a) precipitation of C0 with the smallest domain size in the first 6h (b) precipitation of C0 in the second 6h (c) precipitation ofC0 in the third 6h (d) precipitation of C1 with the medium domain size in the first 6h (e) precipitation of C1 in the second 6h (f) precipitation ofC1 in the third 6h (g) precipitation of C2 with the largest domain size in the first 6h (h) precipitation of C2 in the second 6h (i) precipitation of C2 in the third 6h.**

Taking the domain size for example (**Fig.1**), it can be seen that the differences shown in domain three are less obvious among these experiments, while the spatial pattern significantly differs when looking at domain two. Besides, the spatial resolution of the CMC ground observation dataset, the one that is publicly available with the highest spatial resolution of 0.1 degrees, approximates to the resolution of domain two. Therefore, we compromised to extend the analyzed range to domain two for comparison in the submitted version to illustrate the possible influence on the whole pattern of the rain-belt during this rain storm. We are sorry for not making it clear in our previous manuscript. The reason for choosing domain two for verification will be added to the revised version. Besides, the representative results of domain three were provided in the discussion section to illustrate this issue.

"Also with a 2-way nesting approach, you do not balance any kind of model physics with respect to the lateral boundary conditions. In a 2-way nesting approach, the fine grid resolution replaces the coarser scale resolution over the area of domain three."

**Response**: Thank you for pointing it out. We are sorry for not carefully checking this statement. Indeed, in a two-way nesting approach, the parent domain has its independent run, but it serves the child domain's boundary condition at each time step. And then after the child domain's dynamic modeling run, the child domain result (including the rainfall field) is mapped onto the target parent domain grids. We will remove this confusing sentence in the revised manuscript.

"2) The authors decided to use ERA-Interim reanalysis data to initialize their modelsimulations. As mentioned on page four, the resolution is 0.75°. I am not sure if such acoarse resolution is able to provide reasonable initial conditions, especially when focusing on sub-daily rainfall."

**Response**: Thank you for raising this question. We agree that this dataset maynot be the best choice if other sources of input fields with higher resolution were available. However, the ERA-Interim reanalysis dataset has been widely used in downscaling studies with acceptable results, and it is the best source of the reanalysis data accessible in the study area.

"Although you mention that a small domain may benefitfrom the lateral boundary conditions, I doubt that such a small outer domain ofeffectively 30*30 (40×40) grid cells (due to boundaries of at least 5 cells in each direction) issufficient here. This is also mentioned on page five in your manuscript. If you carefullychecked the WRF webpage, you may have noticed that at least 100*100 cells arerecommended for every domain."

**Response**: Thank you for pointing it out. In WRF-ARW, at least five cells along the boundaries of each domain are indeed required to mitigate sharp gradients (i.e., short wavelength features) that may exist along the lateral boundaries where the specified lateral boundary conditions differ from their values on the interior of the limited-area domain. As you pointed out, we carefully rechecked our settings of all the experiments to make sure that we followed this recommendation. However, in our smallest domain size experiment (C0), eight cells are set between the outermost grid and the interim grid, 11 cells are set between the interim grid and the innermost grid, and 18 cells are set in the innermost grid in both x- and y-direction. We believe that these distances are sufficient for relaxation.

As to the domain size of at least 100×100, this should be related to the weather system in the study area. As long as the domain size covers the key weather features, it should be fine with a smaller domain than 100×100. Examples can be seen in Figure 3c of Bukovsky and Karoly (2009) where 45×70 cells were adopted in the outer domain covering the U.S. and in Figure 1a of Dasari et al. (2014) where 98×55 cells were used in the outer domain covering most of Europe. In our initial case (C0), the main feature of the weather systems that led to this storm was included within the outermost domain. Besides, to estimate the potential influence of lateral boundary conditions to the rainfall outputs, two other cases with larger domain size were designed for comparison.

Bukovsky, M. S., and Karoly, D. J.: Precipitation simulations using WRF as a nested regional climate model. *J. Appl. Meteorol. Climatol.*, **48(10)**, 2152-2159, doi:10.1175/2009JAMC2186.1, 2009.

Dasari, H. P., Salgado, R., Perdigao, J., and Challa, V. S.: A regional climate simulation study using WRF-ARW model over Europe and evaluation for extreme temperature weather events. *Intl. J. Atmospheric Sci.*, **9**, 2101-2122, doi:10.1155/2014/704079, 2014.

"3) It is also not clear how the WRF model levels are distributed in your simulation. From table 1, I only see that you used 29 levels up to 50 hPa. If a constant grid spacing of 1 km is applied, your model simulations will fail because processes in the PBL are not at all resolved. If you are in the middle troposphere, this spacing can be sufficient. Also the WRF tutorial and website suggest a vertical grid spacing of less than 1km. If you look at the user guide (e.g. http://www2.mmm.ucar.edu/wrf/users/docs/user guide V3.8/users guide chap5.htm#examples) you will see that at least ~40 levels are recommended when the model top isset to 50 hPa."

**Response**: Thank you for pointing it out. Indeed, vertical grid spacing of less than 1km was suggested, especially in the lower and middle troposphere layer where convective process mainly happens. In this study, nonuniform grid spacing was adopted, and the eta value for WRF was calculated by using the vertical eta levels provided by the ERA-Interim. The detailed configuration can be seen in **Table 1**.

**Table 1. Vertical levels set in the initial experiment and the corresponding height in Beijing, China.**

| Eta Level | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pressure (hpa) | 1000 | 975 | 950 | 925 | 900 | 875 | 850 | 825 | **800** | 775 | 750 | 700 | 650 | 600 | 550 |
| Eta value (0-1) | 1 | 0.973 | 0.947 | 0.921 | 0.894 | 0.868 | 0.842 | 0.815 | 0.789 | 0.763 | 0.736 | 0.684 | 0.631 | 0.578 | 0.526 |
| Height (km) | 0 | | | | 0.988 | | 1.457 | | 1.949 | | | 3.012 | | 4.206 | |

| Eta Level | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pressure (hpa) | 500 | 450 | 400 | 350 | 300 | **250** | 225 | 200 | 175 | 150 | 125 | 100 | 70 | 50 |
| Eta Value (0-1) | | | | | | | | | | | | | | |
| Height (km) | 5.574 | | 7.185 | | 9.164 | 10.363 | | 11.784 | | | | 16.18 | 18.442 | 20.576 |

Here, 1000pha-800pha corresponds to the lower troposphere; 800pha-250pha corresponds to the middle and upper troposphere. The approximation of the height was also provided by referring the summer value provided by a ground study in Beijing. By comparing the values, we can see that the setting of vertical levels meets the requirements. We are sorry for not making it clear in the manuscript. The detailed illustration will be added to the revised manuscript.

“As your outer domain gets enlarged towards the pole, how did you deal with the map factors?”

**Response**: Thank you for pointing it out. In this study, our main focus is on the simulation in domain two and domain three which are located in the medium latitude of the northern hemisphere, so the Lambert conformal projection was selected and centered at the same latitude (42.25° N) and longitude (114.0° E).

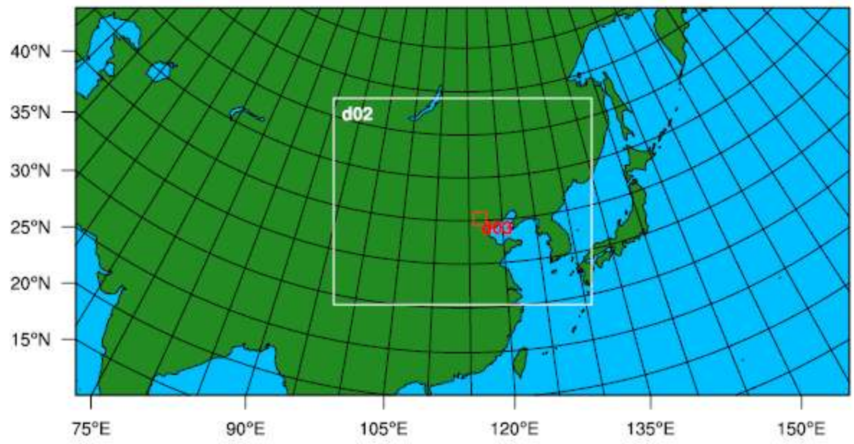


**Figure 2. The relative location of the nested domains adopted in C2 that with the largest outer domain of all the experiments.**

**Figure 2** shows the domain configuration with the largest outer domain. In this figure, it can be seen that the boundary of the outer domain is still far away from the pole and centered at the medium latitude area. The above-mentioned info will be added to the revised manuscript.

“Choosing an adaptive time step may save computation time but is not the best way forscientific experiments (see WRF webpage).”

**Response**: Thank you for raising this question. As you mentioned, choosing an adaptive time step may not be the best way for scientific experiments. We are sorry for not carefully checking this statement and giving the impression that all the experiments in this study adopt the adaptive time step. In fact, C4 with the finest vertical resolution has used this setting, as well as C6 with the finest horizontal resolution. When running these two cases, the recommended minimum time step was set (about 3×DX seconds for the outermost domain), but instability was encountered, and the model ran much slower than we expected and stopped before the end time. To deal with this problem, an adaptive time step is adopted, where the maximum time step was up to 6×DX with CFL value set to 1.2.

“4) The rescaling of the error measures may lead to a misinterpretation. In case of POD, howdid you choose the factor 0.115? This is not clear from the manuscript.”

**Response**: Thank you for pointing it out. It is our negligence to forget adding the related reference to illustrate how the scale for each verification parameter was selected. Here, we add two references in which the same method was adopted to allow a convenient andmultidimensional assessment of precipitationsimulation quality for various WRF configurations. Please see Table 2 in Sikder and Hossain (2016).

In the case of the POD, the factor of 0.115 is determined by the largest POD value calculated from all the experiments. Then all the POD values were divided by this factor to ensure they fell within the range of 0-1. The detailed illustration and reference will be added to the revised paper.

Sikder, S., and Hossain, F.: Assessment of the weather research and forecasting model generalized parameterization schemes for the advancement of precipitation forecasting in monsoon-driven river basins, *J. Adv. Model. Earth Syst.*, **8**, 1210–1228, doi:10.1002/2016MS000678, 2016.

"Is the maximum RMSE used for each individual time step or is it calculated from an average over all thetime steps?"

**Response**: Thank you for raising this question. The RMSE was calculated through the same process as POD. The procedure has been illustrated by Tian et al. (2016) in comparing the simulations to the observations from the ground meteorological stations. As the temporal duration was shorter and the spatial calculation was at grid scale, we merely mentioned the difference when we adopted the method. In this study, all the metrics were firstly computed between the observations and simulations of the same grid at each time step and then averaged within different time durations (6h, 12h, 18h, 24h) for the final analysis.

Tian, J. Y., Liu, J., Li, C. Z., and Yu, F. L.: Numerical rainfall simulation with different spatial and temporal evenness by using WRF multi-physics ensembles. *Nat. Hazards Earth Syst. Sci. Discuss.*, **17**, 563-579, doi:10.5194/nhess-17-563-2017, 2017.

"It is also hard to believe that the POD remains constant, independentwhether you start one week or 12 hours before the event?"

**Response**: Thank you for raising this question. In our study, as the POD was averaged among the large domain with at least 72×72 grids. The differences presented could be less obvious than expected. Besides, when shown in the same graph with the other parameters, the differences were much less obvious than others. But it doesn't mean it is constant.

"Also, what is the precipitation threshold used to calculate POD? Is it 0.1mm? Usually, POD is applied for different thresholds."

**Response**: Thank you for raising this idea. In this study, 0.1 mm is used to calculate POD. POD with different thresholds may be useful to investigate the accurate hit of the heavy rainfall area further. Considering that RMSE and R could also reflect this feature, we only choose POD with 0.1mm as one of the verification parameters in this study.

"How did you match both grids together? Did you use CDO, NCO, or NCL for this? It seems that you applied a $1/R^2$ approach to remap the CMC precipitation observations to the WRF grid. What is the radius of influence in this case? This can strongly determine the resulting field, especially in case of heavy and localized precipitation."

**Response**: Thank you for pointing it out. We are sorry to miss the detailed illustration of the interpolation process. Here, for the rainfall values, the bi-linear interpolation method was adopted to remap the CMC precipitation observations (about 10 km×10 km) to the WRF grid (including 13.5 km×13.5km, 8.1 km×8.1 km and one with 5.785km×5.785km). In this method, four nearest points for each WRF grid were searched to accomplish the bi-linear interpolation process. And the bi-linear interpolation method is now the default option used by WRF-ARW to interpolate the initial meteorological fields.

"Have the integrated water vapor fields been handled in the same way?"

**Response**: Thank you for raising this question. The initial water vapor field was extracted from WPS outputs by adding TCWV (Total column water vapor) in the Vtable files. This means that it has the same location as the field calculated from the WRF outputs. Therefore, there is no need to remap one field to another.

"It would be very useful, if the authors provide horizontal plots of the precipitation patterns to substantiate the results. The applied scores do not necessarily tell if the precipitation is simulated spatially correctly."

**Response**: Thank you for your suggestion. The plots of the spatial precipitation patterns indeed would be helpful to substantiate the results, such as the examples of domain size experiments. Some representative plots will be added to the revised manuscriptto make the results with more clarity.

"In my opinion, a lot of important information is missing here and I also see deficiencies in the experimental setup. I strongly suggest that a native English speaker reads through the manuscript."

**Response**: We hope our replies have addressed your concerns. The revised manuscript will be thoroughly proof-read by a native English speaker.