General Comments

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 nesting approach while the innermost domain is centered on Beijing. Sub-daily precipitation of the third domain was verified against a merged precipitation product in 3 hourly intervals.

Generally, this type of study is of importance as domain size, vertical and horizontal resolution can either have a positive or detrimental impact on the forecast quality.

However, in the present manuscript, it is not really clear how the impact of the previously mentioned can be isolated. For me it is also still a major concern to investigate a sub-daily heavy rainfall event and drive the model with coarse resolution ERA-Interim data. I highly recommend to repeat at least one of the simulation by applying the high-resolution ECMWF operational analysis or ERA5 reanalysis data.

I also feel that the rescaled scores causes more confusion to the reader, rather than improving the readability.

Compared to the first draft, the English spelling and grammar is considerably improved.

The manuscript can be accepted after major revisions. Please the reasons below.

Specific Comments:

From the manuscript it is not clear to me, whether you applied 1-way or 2-way nesting. This information is missing for the reader.

The size of your innermost domain is never mentioned but this can play an important role. In case it's just 100*100 cells, you have to subtract 10 cells in both direction due to boundary effects so you will have an effective grid size of 90*90 levels and I have a feeling that this is not sufficient at this particular high resolution.

For a better readability, I suggest the following section ordering:

2 – Meteorological situation: The paragraph from section 3.1 on page 7 can be moved here.

3 Experimental design

3.1 WRF Model: The paragraph on page 5, lines 7-17 can be moved here.

3.2. Model configuration

- Forcing: The paragraph on page 4, starting line 24 can be moved here.

- Model setup end experiments: The text between page 8, lines 5 tp page 9, line 29 can be moved here.

Abstract: What do you mean with “cumulative spatial error”? I do not find any explanation in the manuscript.

Page 2, line 24: Which period?

Page 2, line 25: Please use Skamarock et al. (2008) as reference here.

Page 3, line 18: “relatively” repeated twice. I think “relative” is not appropriate here.

Page 3, line 21: I do not agree here. For sure there are studies about this topic.

Page 3, line 25: I guess you are talking about Beijing here.

Page 4, line 4: Did you use the setting from the two publications in your study? If this is the case, please mention this.

Page 4, line 6: What is the first question here?

Page 4, line 15: Reference to ERA-Interim? This paragraph can be splitted into two parts and the second half may be moved to section 4.

Page 5, line 1: I think Dee et al. (2011) is sufficient here.

Page 5, lines 4-5: What is the CMC model saying in this case? Is it better than WRF? I am also not confident that the CMC model configuration is the justification to apply coarse resolution input and boundary conditions.

Page 5, lines 19-29: Please check if the full paragraph is necessary here. The focus in your study is not evaluating LBCs.

Page 5, line 31: Do you really mean grid spacing here? Or do you mean the number of grid cells?

Page 6, lines 1-2: I think a vertical resolution of > 1 km in the PBL is far not sufficient here. I also do not agree with this statement in general.

Page 6, line 7: What is an excessive grid spacing?

Page 6, line 10: see my comment two lines above. I thinks this is wrong.

Page 6, line 18-20: Do you really mean spin-up times here? Or do you mean sth. like forecast lead times?

Page 6, line 25: Are you sure?

Page 7, line 15: I would rather write 60 m to 2300 m.

Page 7, line 22: You mention 100 mm/h rain rate, but your verification is performed over at least 6 hourly windows. How does this fit together?

Page 8, line 5: What is the reason you also apply the cumulus parametrization in D03? Please explain. I am not sure if the GD is designed to run at very high resolution. If you see the

necessity to use a convection parametrization in D03, why didn't you choose a more recent scheme like G3 or Grell-Freitas?

Page 8, line 10: I think Ek et al. is the wrong reference for the MM5 surface layer scheme.

Page 8, lines 22-25: I think this is not correct as you also change the spin-up period. There is no other choice than using the same map projection in a multi domain simulation.

Page 8, line 27: Please mention the recommended setting you refer to in section 2. What is the time step you applied? This is not mentioned in the manuscript? If an adaptive time step is applied, then the results are not comparable.

Page 8, line 29: The odd ratio is selected due to the applied Arakawa grid and not because of a reduction of the initial error.

Page 9, lines 2-3: This should be clear.

Page 9, line 6: Why did you output the D03 data only in 3 hourly intervals? Is it due to the merged CMORPH data set? Why didn't you use the hourly gridded precipitation data set from the CMC? Especially when you are interested in sub-daily extreme rainfall?

Page 9, lines 11-13: This implies that your simulation C0 is unusable here as it does not capture the situation. What is a perturbed synoptic feature?

Page 9, line 16: This is totally confusing. D02 of the experiment C0 is now the new D01 for C1? I do not see the equivalence here: 72*72 cells at 13.5 km vs. 80*64 cells at 40.5km resolution.

Page 9, line 22: I guess you mean the middle troposphere. Why is this condition not satisfied in other regions?

Page 9, line 25: I think there is no "increased nesting ratio" of 1:3:3.

Page 10, line 6: ERA-Interim provides precipitation on a 6 hourly basis.

Page 10, line 12: What do you mean with "scale of D02"? Is it the area covered by domain 2? In case 2-way nesting is applied, you will see the results of D03 in D02. I also think that you cannot assume that in case the results in D03 are reasonable, it's the same in D02.

Page 10, lines 21-27: I think this should be put behind the paragraph explaining the formulas. 24 h sums are not sub-daily anymore.

Page 11, first paragraph: What is the "tested value"? Is it the forecast model? In your formulas, the total number of time steps should be 2,4,6,8 as you are using 3 hourly precipitation data. Did you consider this in your evaluation?

Page 11, line 11: It is still not clear to me, how the maximum errors are defined. Are they based on 6h, 12h, 18h, or 24 h precipitation differences? Is it the same value for all experiments? It should be reasonable to use a single value for the different time periods and spin-up times. In general, it is very confusing to read and interpret the rescaled scores. Did you choose the scores

only to fit all scores into one figure? Sikder and Hossain (2016) had a different intention behind. What is R' (not explained)?

Page 11, lines 21-22: This is already mentioned before and can be deleted.

Page 11, lines 23-24: Is it the same domain or the same area? This makes a difference.

Page 12, line 29: Are you interested in D02 or D03? Are you evaluating on the area of D03, even if you are talking about D02? "scale of D03" is strongly misleading here.

Page 13, line 10: What is a "PW-related" feature? This is unclear.

Page 13, line 26: I do not think that there is a really obvious variation in R' here.

Page 13, line 28-29: How can this be justified? Please explain.

Page 14, line 13: This is hardly visible for me in case of WR'.

Page 14, lines 13-15: What do you mean by surface perturbations? Of course, the coarse resolution of the initial conditions severely limits the meaningfulness.

Page 14, line 25: I do not see this in your plots.

Page 15, line 16: What is meant by a "diurnal tendency"?

Page 15, lines 27-29: Bias is commonly used when showing absolute differences.

Page 15, line 30: Not necessarily true.

Page 16, first paragraph: The data sets you applied for verification of precipitation are different in Fig. 7 and Fig. 8 (CMORPH merged vs. CMC).

Page 17, line 17: Avoid "excessively". This really gives a negative touch to the application of large domains and very high resolution.

Page 17, line 9: It's quite confusing for the reader that an increasing RMSE is good.

Page 17, line 14: What do you mean here? Please be more precise.

Page 17, line 21: This a major point. In my opinion you cannot expect that you are able to reproduce a single extreme event in case you apply very coarse initial and boundary conditions. There is a high chance, that applying data assimilation is essential here. See e.g. Sun et al. (2013) MWR. I do not think that on a short time period, the boundary conditions take the leading role.

Page 18, line 1: What kind of regional geographical data sets did you apply? Did you provide your own landuse and/or soil texture data set? Did you try this data set: https://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=WRF_NOAH_HWSD_world_TOP_SOILTYP ? I am not sure if the default coarse FAO data set is sufficient here.

Page 18, line 19: Where does the cumulative spatial error come from? How is it calculated? Please mention this in the manuscript.

Figures: Please use panel subscripts when applicable.

Figure 1: Please rework on the terrain plot. The default NCL color bar is not appropriate here. The image is either saturated in blue or red.

Figure 2: This image is overloaded. What is the meaning of the blue box?

Figure 4: The label bar in its current stage is useless. Why is such a coarse interval chosen? The reader does not see any major differences here. I would also appreciate a color plot here. For readers who are not familiar with China, please indicate where Beijing is located. Please rewrite the figure caption. This is very hard to understand.

Figure 6: In the 24 h diagram, WRMSE' of C7 appears to be missing.

Tables

Table 1: Is the threshold value the same for all time periods (see my comment above)?

Table 2: Are the results based on the area of domain 3?