The revised paper has not been uploaded because we are waiting for the comments from other reviewers and the current manuscript was under checking by a native English speaker presently. But we want to respond to the comments received.

We would like to thank the referee for the helpful comments. Our point-to-point response to the reviewer's comments is described as follows.

Reviewer #2

"In this contribution, the authors evaluate the performance of the WRF model in different configurations for a single heavy rainfall event centered over Beijing, China. The evaluation differs from other studies in that field in the sense that no physics parameterization evaluation is attempted. Instead, the model setup (domain configuration, number of vertical levels = vertical resolution, nesting ratio = horizontal resolution, forecasting lead time) are explored. In the order of the above, the best configuration is chosen in each step to perform several experiments in the next step. Several verification measures are employed for precipitation and precipitable water (PW).

The design of these experiments is convincing despite a few weak points listed below. The use of English language, however, needs improvement. Grammatical mistakes and strange wordings render some parts of the text unclear. I did not make any attempt to correct this but highlight a few common issues below. With improvements to the language and several changes to the contents, the contribution may be suitable for publication."

Response: Thanks very much for the encouraging feedback.

General remarks

- Dependence of parameters varied: Although discussed in the introduction, the dependence of optimal lead forecasting time on domain extent (and vice versa) is not considered in the study. Instead, based on a standard lead time of 12h, the "best" domain configuration is derived as C1, based on which an optimal lead time of 60h is inferred later on as C11. In my understanding, these two lead times should match if one really found the "best" combination of these two parameters."

Response: We agree that there is a dependence relationship between domain extent and forecasting lead time and this is particularly true in the limited-area modeling cases where data assimilation is not conducted. For instance, a model run with a larger domain size may need longer lead time to spin-up physical processes of interest, such as clouds, precipitation, local ageostrophic circulations, and lateral boundary conditions. Besides, the choice of lead time and domain size at the same time determine the moment at which the initial and lateral boundary conditions are derived and the range of the corresponding synoptic features and water vapor conditions involved at that moment. This, however, means that for a given domain size, the results may also be affected by the degree of similarity between the forcing data and the real conditions at the initialization time. In addition, the choice of updating lateral boundary conditions at a given interval could

affect the time needed to spin-up the physical processes. It is noteworthy that although C1 is detected with the best performance, C2 with nearly doubled domain extent of C1 only differs obviously in the PW fields but with less diversity regarding distribution characteristics of the rain belt. Combining all the aforementioned factors, we believe that it makes sense that C11 (with 60h lead time) is evaluated with better performance when compared with C5 (with 12h lead time).

- The authors choose an adaptive time step to conduct their modeling experiments. This introduces another free parameter, since the actual time step adopted in each simulation may vary and as such influence the results."

Response: As you mentioned, choosing an adaptive time step may introduce another free parameter and as such influence the results. We are sorry for not carefully checking the statement and giving the impression that all experiments adopted the adaptive time step. In fact, only C4 with the finest vertical resolution, and C6 with the finest horizontal resolution have used this setting. 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 was adopted, where the maximum time step was up to 6×DX with CFL value set to 1.2. We will edit this confusing statement and discuss it in the revised manuscript.

- The performance evaluation of the WRF model is performed over the intermediate domain D02 and not over the highest-resolution domain D03. It is argued that the two-way nesting approach does inform D02 about the results in the innermost domain D03, but interpolation to the coarser D02 grid and the (possible) difference in setup of the model physics (see next point) may influence the conclusions drawn. I would like to encourage the authors to conduct simulations without D03 for their best setup at least.

Response: Thank you for raising this question. Indeed, domain D03 covers the main convective processes and should be the best choice for evaluation. However, when we conducted the evaluation over domain D03, 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. 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 D02. Therefore, we compromised to extend the analyzed range to domain D02 for comparison in the submitted version to illustrate the possible influence on the whole pattern of the rain-belt during this rain storm.

Meanwhile, we agree that the two-way nesting may alter the result in D02 (See **Fig. 1**) and it would be useful to explore the one-way nesting for comparison. We are sorry for not making this clear in our previous manuscript. The reason for choosing domain D02 for verification will be added to the revised manuscript.

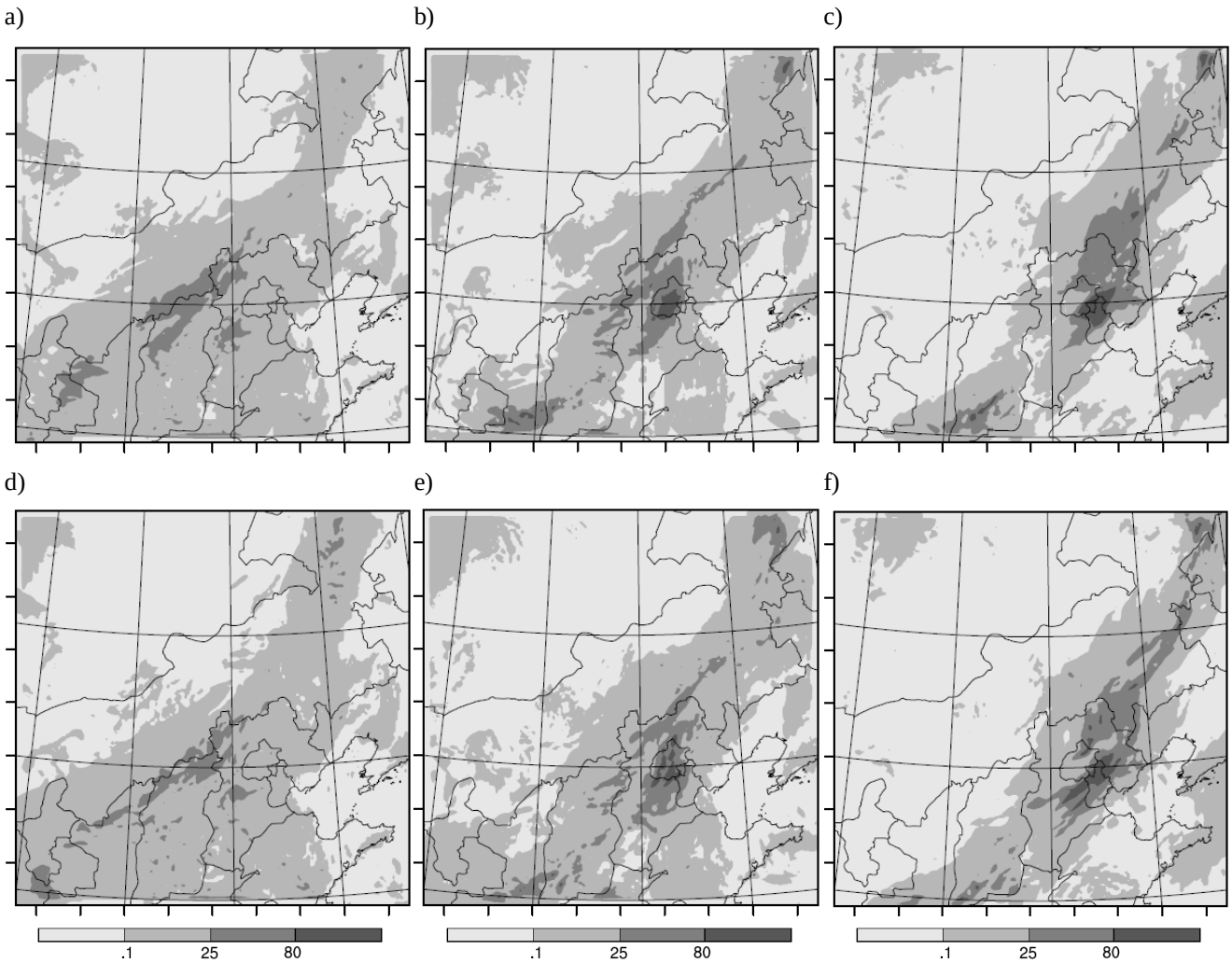

Figure 1. Spatial distribution of six hour accumulated precipitation for C11 (the experiment with the best setup) with different nesting schemes within D02 during the heavy rainfall event from 12 pm, July 21, 2012 (a) precipitation of C11 with two-way nesting in the first 6h; (b) precipitation of C11 with two-way nesting in the second 6h; (c) precipitation of C11 with two-way nesting in the third 6h; (d) precipitation of C11 with one-way nesting in the first 6h; (e) precipitation of C11 with one-way nesting in the second 6h, and (f) precipitation of C11 with one-way nesting in the third 6h.

- Model physics: It is unclear to me whether the GD cumulus parameterization is also employed in D03 at convection-permitting resolution (<5-10km).

Response: Thank you for raising this question. In this study, GD cumulus parameterization was turned on for each domain, including D03 at the convection-permitting resolution between 1km and 5km, to represent the effects of sub-grid scale convective processes which were also detected in the rainfall processes of this heavy rainfall event.

- The forcing data is obtained from ERA-Interim on pressure levels (28 (29) levels). This is not ideal, in particular since the authors are trying to assess the added value of a higher vertical resolution and since ERA-Interim is also available on 38 model levels. I would like to encourage the authors to repeat experiments with their optimal setup C11 using ERA-Interim model-level data and varying the vertical resolution as in S2, for example.

Response: Thank you for your suggestion. We agree that the experiment with slightly higher vertical resolutions could get better performance than the one with lower vertical resolution. This has also been verified in this study where the experiment with 57 vertical levels shows better performance than the one with 29 vertical levels. In this study, the ERA-Interim 29 pressure level data was selected as the initial forcing for two reasons. First, it meets the requirements of less than 1 km distance between each vertical level in the free troposphere where convective processes mainly happen (See **Table 1**). Second, the NWP models used by Chinese Meteorological Center are mainly operated with 31 vertical levels for regional forecasting (See Table 11-2.1 in WMO, 2013).

Table 1. The 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

However, although the ERA-Interim 29 pressure level data meets the requirement when simulating the heavy rainfall event in Beijing, it doesn't mean that it satisfies the condition in other regions. To illustrate this issue, one experiment forced with ERA-Interim 38 pressure level data will be added in S2 for comparison. Besides, the possible effect of the selection of initial forcing data on the forecasts will be discussed in the revised manuscript.

WMO: Anticipated advances in numerical weather prediction, and the growing technology gap in weather forecast. 2013.[Available online at https://www.wmo.int/pages/prog/www/swfdp/Meetings/documents/Advances_NWP.pdf]

- Several of the abbreviations in the text or the figure captions are not introduced before they are used (or not at all), please check and correct

Response: Thank you for pointing it out. We will check all the abbreviations in the text and figure captions and make sure their full names will be added before the abbreviations are used.

- The statistical measures used here have different directions of "good". (i.e. RMSE is good if low, R is good if high). This is not mentioned anywhere in the text and in the figures, which makes the interpretation confusing, also because the statistics are rescaled. I would encourage to state explicitly what value implies a better model performance for which statistical measure, and possibly encode this (in color or differently) in the figures and the tables

Response: Thank you for raising this question. To make it clear, the verification metrics will be assigned with a new set of symbols after the statistics are rescaled (see **Table 2**). The explanation of the performance measures will be added in the revised manuscript as suggested.

Table 2. Correlations between original and rescaled verification metrics.

Original Metrics	Representative Meaning	Rescaled Metrics	Threshold Value
POD	Probability of Detection	$POD' = POD/POD_{max}$	+ 0.115 max
$RMSE$	Root Mean Square Error	$RMSE' = 1 - RMSE/RMSE_{max}$	+ 41 max
R	Pearson Correlation Coefficients	$R' = R$	N/A
$WRMSE$	RMSE of the Precipitable Water	$WRMSE' = 1 - WRMSE/WRMSE_{max}$	+ 7.3 max
WR	R of the Precipitable Water	$WR' = WR$	N/A
RE_{PMAX}	Relative Error of the Maximum Precipitation	$PMAX' = RE_{PMAX}$	N/A
RE_{TP}	Relative Error of the Total Precipitation	$TP' = RE_{TP}$	N/A

- Beneath dynamical downscaling explored here, also statistical downscaling methods and new global modeling approaches on irregular grids (e.g. MPAS) have been used and show promising results to forecast extreme precipitation events. This should be discussed briefly in the introduction or discussion section"

Response: Thank you for your suggestion. We agree that some statistical downscaling methods, along with data assimilation methods, could provide more reliable forecasts of extreme precipitation events, especially for short-term forecasting with 6h-24h lead time. The related content will be added in the discussion section as suggested.

Specific remarks

- Page 2, lines 24-26: WRF being used at resolutions >10km only is not true. Many leading operational NWP centers are employing WRF at convection-resolving resolution (NCEP: HRRR, 3km over CONUS; Meteo Group: 3km over Europe; New Zealand Met Service: <4km over New Zealand) operationally.

Response: Thank you for the clarification. We will rephrase this sentence in the manuscript to avoid the confusion.

- Page 4, line 27: Isn't ERA-Interim available from 1979 (not 1989)?

Response: Thank you for pointing this out. Indeed, ERA-Interim is available from 1979 (originally, ERA-Interim ran from 1989, but the 10-year extension for 1979-1988 was added in 2011). We will amend this statement.

- Page 7, line 31: RRTMG schemes (not RRTM)? Or is it "RRTM" for LW and "Dhudia" for SW?

Response: Thank you for pointing this out. The radiation schemes adopted in our study were the RRTMG schemes. We will edit this sentence to: "The radiation processes were represented by the RRTMG short-wave radiation and the RRTMG long-wave radiation schemes (Iacono et al., 2008), respectively."

- Page 14, lines 2-15: the discussion is confusing for the reader as he/she is expected to translate nesting ratios into effective horizontal resolution. In this paragraph, as well as in Table 1, the effective horizontal resolution should be stated explicitly, alongside with the nesting ratios"

Response: Thank you for the suggestion. We will follow the advice, as noted above.

- Page 14, line 31 to page 15, line 1: the authors state that the positive bias in precipw (PMAX) depends on the initialization time, with largest biases for initialization times with highest amounts of precipw. This, in my opinion, is an important finding and should be highlighted and possibly discussed further.

Response: Thanks for your suggestion. We agree that this is an interesting finding that should be highlighted and further discussed as suggested.

- Page 17, lines 8-11: The authors briefly discuss the dependence on the quality of the forcing data. This highlights the importance to conduct additional experiments with ERA-Interim model level data as described above, and at least discuss (if not evaluate) potential effects when using ECMWF high-res forecasts on 137 model levels and approx. 9km horizontal resolution.

Response: Thank you for pointing this out. As mentioned above, one experiment forced with ERA-Interim 38 model level data will be added in Scenario 2 for comparison. Besides, the possible effect of the quality of the forcing data on the forecasts will be further discussed in the revised manuscript.

Typographical corrections

- Page 4, line 10: "coaster-scale" -> "coarser-scale"
- Page 4, line 21: "Earth-system system"
- Page 4, line 29: "WRFV3.7.2" or "WRFV3.7.1"
- Page 5, line 14: "They two together" -> "The two together"
- Page 5, line 6: the correct reference should be Skamarock et al. (2008).
- Page 13, line 11: "less sensititvity" -> "less sensitive"

Response: Thanks for reading our manuscript so meticulously – these will be corrected.

Grammar corrections

- Language-specific: - the word "occurred" is often missing a leading "that" or "which" -the expression "demonstrated true" seems odd to me - several times in the text, "grid" is used whereas "grid point" should be - singular and plural, as well as the use of articles need to be checked carefully.

- Example for a necessary rewording: caption of figure 1: "Location and topography of the study area. Left panel: three levels of nested domains adopted in most experiments, with D03 covering the Beijing area; right panel: zoom-in on the topography of the Beijing area"

Response: Thank you for pointing this out. We will carefully check through the manuscript to correct the grammatical mistakes with more precise descriptions.

We hope our replies have addressed your concerns, and the revised manuscript will be thoroughly proof-read by a native English speaker.