Dear Associate Editor/Reviewers,

We would like to thank the editor and the referees for the helpful comments and thorough suggestions that surely help us improve the manuscript. We have made corrections according to these comments and marked the amended portion in blue in the annotated version of the manuscript. Below, we provide our point-to-point response to the comments and presented the main corrections in the revised manuscript.

Response to the reviewer's comments

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 the issues and added the missing information to increase the traceability of the results.**

General Comments

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 an urban flooding study. As you suggested, **we recalculated the metrics for all the experiments over domain three (D03)** by comparing the simulations with 3 hourly 0.05-degree gridded dataset, produced by fusing rain gauge observations and the CMORPH data (Huang et al., 2013). The evaluation was then repeated on the scale of D03.

Detailed results can be seen in Section 5. Besides, the evaluation made over the domain two was also used, but as an auxiliary method for subjective verification. This is based on the assumption that an experiment with good performance in the inner domain should also capture the large-scale features in the outer domain, as the appropriate representation of these large-scale features will result in more accurate boundary conditions.

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.

2. - 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 have removed this confusing sentence in the revised manuscript.**

3. - The authors decided to use ERA-Interim reanalysis data to initialize their model simulations. As mentioned on page four, the resolution is 0.75°. I am not sure if such a coarse 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 may not 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.

4. - Although you mention that a small domain may benefit from the lateral boundary conditions, I doubt that such a small outer domain of effectively 30*30 (40×40) grid cells (due to boundaries of at least 5 cells in each direction) is sufficient here. This is also mentioned on page five in your manuscript. If you carefully checked the WRF webpage, you may have noticed that at least 100*100 cells are recommended 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 this 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 were set between the outermost grid and the interim grid, and eleven cells were set between the interim grid and the innermost grid. We believe that these distances are sufficient for relaxation.

As for 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 the most 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.

5. - 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 is set to 50 hPa.

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 is also verified in our study where the experiment with 57 vertical levels (C3) shows better performance than the one with 29 vertical levels (C1). 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 the convective processes mainly happen (See **Table 1**). Second, the NWP models used by the Chinese Meteorological Centre mainly employ 31 vertical levels in 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

Here, 1000pha-800pha corresponds to the lower troposphere; 800pha-250pha corresponds to the middle and upper troposphere. The approximation of the height was provided by a ground study in Beijing during the summer. 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 conditions in other regions. **To illustrate this issue, one experiment forced with ERA-Interim 38 pressure level data (C5) has been added in S2 for comparison.** The results showed that C5 shows either better or worse performance than C1 in each period but produces less accurate rainfall simulations than C3 over most of the evaluated durations.

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]

6. - 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 on the same latitude (42.25° N) and longitude (114.0° E).

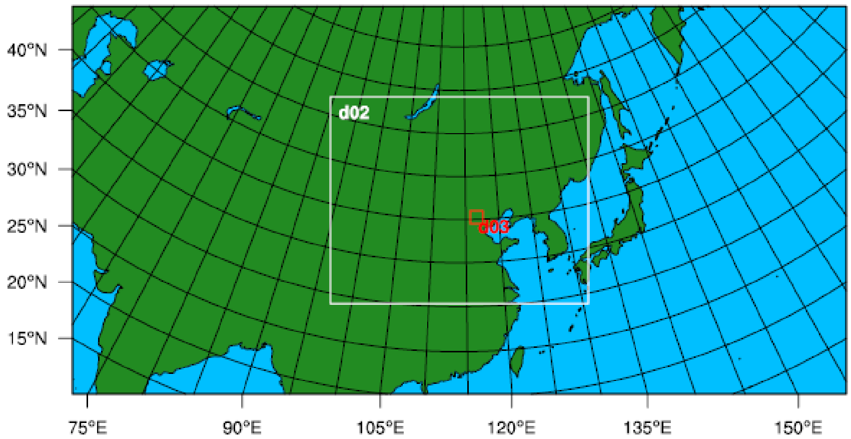

Figure 1. The relative location of the nested domains adopted in C2 that with the largest outer domain of all the experiments.

Figure 1 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 information has been added to the revised manuscript. Please see Page 8, Line 22 – 25.**

7. - Choosing an adaptive time step may save computation time but is not the best way for scientific experiments (see WRF webpage).

Response: Thank you for raising this problem. 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 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 was adopted, where the maximum time step was up to 6×DX with CFL value set to 1.2. **We have removed this confusing sentence in the revised manuscript.**

8. - The rescaling of the error measures may lead to a misinterpretation. In case of POD, how did 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 the reference in which the same method was adopted to allow a convenient and multidimensional assessment of precipitation simulation quality for various WRF configurations. Please see Table 2 in Sikder and Hossain (2016).

In the case of POD, the factor of 0.115 was 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 has been added to the revised paper. **Please see Page 11, Line 10 - 11.**

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.

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

Response: Thank you for raising this problem. 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. **The related information has been added into the revised manuscript. Please see Page 10, Line 23 – 27.**

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.*, **17**, 563-579, doi:10.5194/nhess-17-563-2017, 2017.

10. - It is also hard to believe that the POD remains constant, independent whether you start one week or 12 hours before the event?

Response: Thank you for raising this problem. In our study, as the POD was averaged among the large domain with at least 72×72 grid. 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.

11. - 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. **We have added the information into the revised version to make it clear. Please see Page 10, Line 16 – 17.**

12. - 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 WRF simulations to the reference grid. In this method, four nearest points for each WRF grid were searched to accomplish the bi-linear interpolation process. In WRF-ARW, the bi-linear interpolation method is used as the default choice to interpolate the initial meteorological fields. **The related illustration has been added in the revised manuscript in Page 10, Line 14 - 15.**

13. - 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. **The related illustration has been added in the revised manuscript. Please see Page 10, Line 22 - 23.**

14. - 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 have been added to the revised manuscript to make the results with more clarity. Please see Page 12, Line 29 – 30 to Page 13, Line 1 - 11.**

"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 has been thoroughly proof-read by a native English speaker.**

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

1. - 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. Besides, 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).

2. - 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 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 have removed this confusing statement in the revised manuscript.**

3. - 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. **As you suggested, we have recalculated the metrics for all the experiments over**

domain three (D03) and repeated the evaluation on the scale of D03. The analyzed results can be seen in Section 5 of the revised manuscript.

4. - 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. **We have added the information into the revised manuscript to avoid the confusion. Please see Page 8, Line 23 - 25.**

5. - 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 is also verified in this study where the experiment with 57 vertical levels (C3) shows better performance than the one with 29 vertical levels (C1). 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 the 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 (C5) has been added in S2 for comparison.** The results shows that C5 has either better or worse performance than C1 in each analyzed period, but is verified with less accuracy in the rainfall simulations than C3 throughout most of the evaluated time durations.

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]

6. - 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 have checked all the abbreviations in the text and figure captions to make sure their full names are given before they are used. Please see the figure captions in Page 24.**

7. - 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 have been assigned with a new set of symbols after the statistics are rescaled (see **Table 2**). **The explanation of the performance measures has been added in the revised manuscript as suggested. Please see Page 11, Line 7 - 14.**

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

8. - 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 has been added in the discussion section as suggested. Please see Page 3, Line 1 - 3, Line 10 – 11 and Page 17, Line 20 – 23.**

Specific remarks

1. - 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 have rephrased this sentence in the manuscript to avoid the confusion. Please see Page 2, Line 27 – 30.**

2. - 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 have amended this statement. Please see Page 5, Line 1.**

3. **-** 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 have edited this sentence** to: "The radiation processes are represented by the RRTMG shortwave radiation and the RRTMG longwave radiation schemes (Iacono et al., 2008)." **Please see Page 8, Line 10 - 11.**

4. - 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 have followed this advice, as noted above. Please see Page 14, Line 23 – 24 and Page 33, Table 1.**

5. - 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 have highlighted this finding and discussed it as suggested. Please see Page 15, Line 27 – 30.**

6. - 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 (C5) has been added in Scenario two for comparison. Besides, the possible effect of the quality of the forcing data on the forecasts has been briefly discussed in the revised manuscript. **Please see Page 17, Line 23 – 25.**

Typographical corrections

1. **-** Page 4, line 10: "coaster-scale" -> "coarser-scale"

2. **-** Page 4, line 21: "Earth-system system"

3. **-** Page 4, line 29: "WRFV3.7.2" or "WRFV3.7.1"

4. **-** Page 5, line 14: "They two together" -> "The two together"”

5. **-** Page 5, line 6: the correct reference should be Skamarock et al. (2008).

6. **-** Page 13, line 11: "less sensititvity" -> "less sensitive"

Response: Thanks for reading our manuscript so meticulously – **these have been corrected**.

Grammar corrections

1. **-** 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.

2. **-** 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 have carefully checked through the manuscript to correct the grammatical mistakes with more precise descriptions**.

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

Response to the editor’s comments:

Associate Editor

Specific Comment

"- I have read the referees' comments and your replies. While for most points, your replies are satisfactory, I need some more information from you about a major point both referees raised: The choice of modeling and evaluation domains (comment 2 (General Comment **1**) of referee #1, comment 4 (General Remarks **3**) of referee #2). In the following, I will explain my current understanding of what you were doing, and my related conclusions. Before I take my decision about how to proceed, I would like to verify that I understood things correctly:"

Response: Thank you very much for your positive assessment and constructive suggestions to our study.

" •You model an extreme rainfall event using WRF set up on 3 nested domains (D01, D02, D03) with D01 being the largest domain in coarsest resolution, D03 the smallest, which covers the area of interest, Bejing region.

- When running the models, D01 is forced by ERA-Interim reanalysis, D02 by D01, D03 by D02
- For analysis/evaluation, you map rainfall from D03 back into the results and onto the grid of the D02 domain.
- Analysis and evaluation is the done with this hybrid data set D02+partlyD03, on the D02 domain, against ground rainfall observations and ERA-interim reanalysis.
- The reasons for doing so is that a) the reference truth is available only in a resolution comparable to the D02 resolution, and b) differences among the models are less obvious in the D03 domain than in the D02 domain."

Response: Thank you for reading our manuscript so meticulously**.** We agree that most of your understanding is correct. Beijing is the area of interest where the convective processes (convective-scale) happened. Three 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 Beijing region. When running the model, D01 is forced by ERA-Interim reanalysis, D02 by D01, D03 by D02. In the original version of the manuscript, the analysis/evaluation is conducted on the D02 scale by comparing the hybrid dataset D02+partly D03 against ground rainfall observation. 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.

"So if this is correct, I have two main concerns:

- If your goal is to evaluate different WRF setups with respect to regional (here: D03 or Bejing-scale) heavy precipitation, then a) the evaluation should be done exactly on this scale and b) they should be compared to reference data with adequate resolution for that scale.

- If your goal is to evaluate different WRF setups with respect to larger-scale (here: D02-scale) precipitation patterns, then the D03 run on its small scale is unnecessary. You could then run WRF on the D02 domain with different configurations (the D02 and the D03 settings with respect to parameterization, grid size etc.) and compare these.

So if my concerns are based on correct understanding, you will either have to

- gather better reference data and repeat the evaluation on the D03 scale, or
- do new model runs on the D02 scale and repeat evaluation on the D02 scale."

Response: To verify whether our choice of the evaluation domain is reasonable, we 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). The results were then compared with the results derived from the D02 scale.

By comparison, we have noticed that most of the evaluated results on the D03 scale were similar to those got on the D02 scale. This indicates 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.

Taking the domain size scenario as example, when it was evaluated on the D03 scale (See **Fig. 2**), 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.3** and **Fig. 4**), 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 2 to Figure 4]

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 have adopted the first option as suggested by you: repeating the evaluation on the D03 scale. Detailed results can be seen in Section 5. Meanwhile, the evaluation made on the D02 scale has also been adopted, but as an auxiliary method for subjective verification. Related results can be seen in Section 5.1 and Section 5.4.**

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

Kind Regards,

The Authors

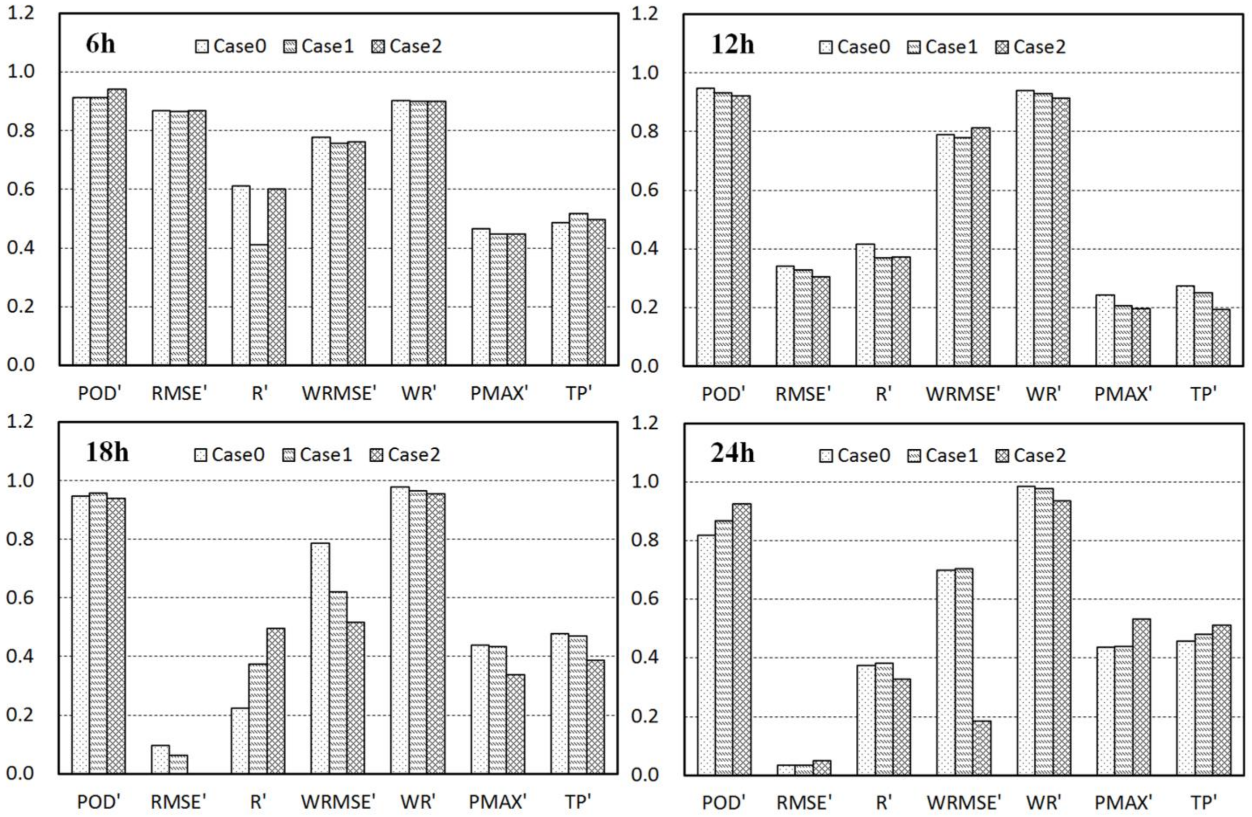

Figure 2: Spatial values of the verification metrics for the WRF domain size experiments, calculated over different temporal durations and over domain three. Case 0 incorporates the smallest domain, which covers north-central China; Case 1 incorporates a domain of intermediate size that covers northern China and part of Mongolia; and Case 2 incorporates the largest domain, which covers the Northeastern Hemisphere. The metrics are calculated over time periods of 6 h, 12 h, 18 h, and 24 h that begin at 12 am on 21 July 2012.

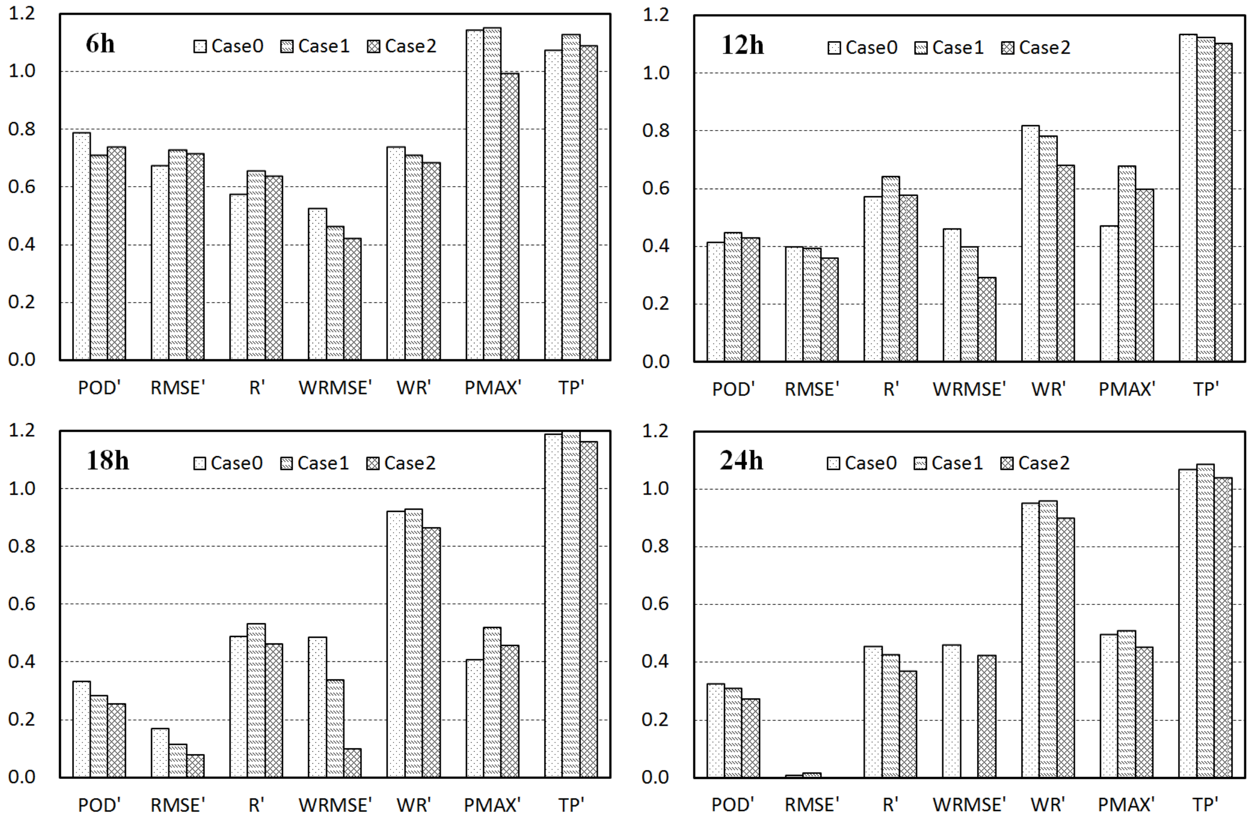

Figure 3: As in Fig. 1, but the metrics were calculated on the D02 scale.

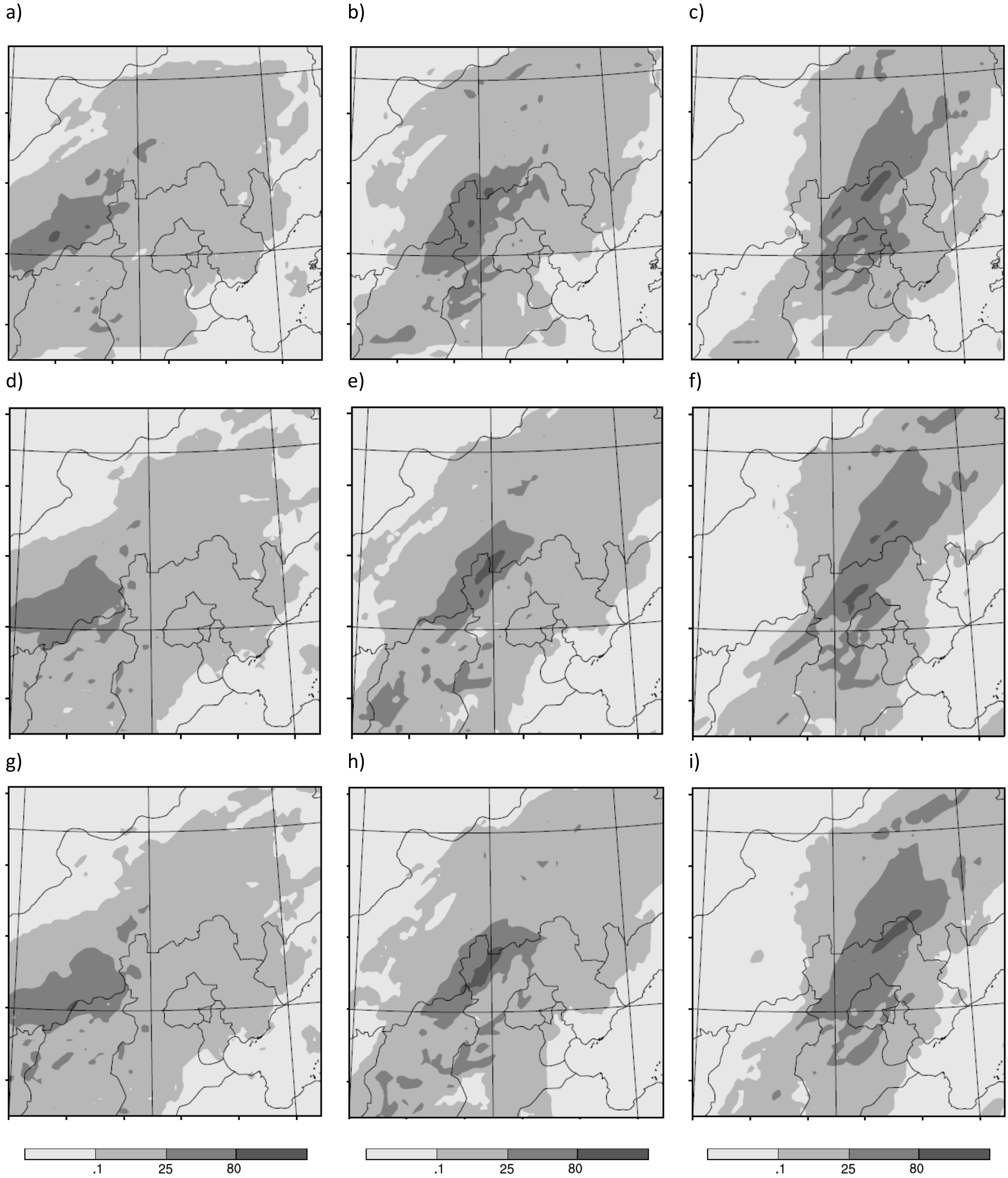

Figure 4: Spatial distribution of 6-h accumulated precipitation for the domain size experiments within domain two of Case 0 during the heavy rainfall event beginning at 12 am on July 21, 2012. (a) Precipitation in Case 0 (with the smallest domain size) during the first 6-h period; (b) precipitation in Case 0 during the second 6-h period; (c) precipitation in Case 0 during the third 6-h period; (d) precipitation in Case 1 (with the medium-sized domain) during the first 6-h period; (e) precipitation in Case 1 during the second 6-h period; (f) precipitation in Case 1 during the third 6-h period; g) precipitation in Case 2 (with the largest domain) during the first 6-h period; h) precipitation in Case 2 during the second 6-h period; and i) precipitation in Case 2 during the third 6-h period.

List of all relevant changes

New Title "Evaluation of the ability of the Weather Research and Forecasting model to reproduce a sub-daily extreme rainfall event in Beijing, China using different domain configurations and spin-up times".

Page 1, Line 4 and Line 6 one affiliation of the co-authors was added.

Page 1, after Line 18 the original statements were replaced by "A comparison of the optimal run and the initial run performed using the most common settings reveals clear improvements in the verification metrics. Specifically, R' increases from 0.226 to 0.67; the relative error of the maximum precipitation at a point increases from 0.44 to 0.883; and the cumulative spatial error decreases by 33.65 %."

Page 2, after Line 12 we added:

"Recently developed statistically-based rainfall generation methods and remote sensing data have been shown to enable the extension of the lead time to 24 hours. However, this lead time is still insufficient to provide effective flood mitigation for medium or large urban areas with very short hydrologic response times (Shih et al., 2014; Li et al., 2017)."

The correction was made based on the General Remark 8 of Reviewer #2.

Page 2, after Line 27 the original statements were replaced by:

"Increasing numbers of meteorological operational centres and research groups are adopting these new NWP models to carry out simulations of heavy rainfall events or real-time forecasting. The resolutions of the rainfall products have improved from tens of kilometres to less than a kilometre, and the lead times have increased from less than a day to more than a week (WMO, 2013)."

The correction was made based on the Specific Remark 1 of Reviewer #2.

Page 2, after Line 30 we added:

"Meanwhile, case studies have been carried out using regional convective-resolving models to evaluate the local rainfall predictions generated by sophisticated regional nesting techniques or the global smooth grid transition approach on unstructured grids (Hong and Lee, 2009; Soares et al., 2012; Sikder and Hossain, 2016; Heinzeller et al. 2016)."

The correction was made based on the General Remark 8 of Reviewer #2.

Page 3, after Line 10 we added:

"These uncertainties are expected to be further magnified by downscaling or the use of mesh transition procedures, so re-evaluation and calibration of the related model configurations are commonly required."

The correction was made based on the General Remark 8 of Reviewer #2.

Page 4, Line 15 “a coaster-scale reanalysis” was replaced by “a coarser-resolution reanalysis”.

Page 4, Line 26 “an Earth-system system” was replaced by “an Earth system”.

Page 4, after Line 30 the original statement was replaced by:

“The final reanalysis product, ERA-Interim, is a global gridded dataset that is available at a spectral resolution of T255 and at both the 60 levels used in the model and 38 interpolated pressure levels for all dates beginning on 1 January 1979 (Berrisford et al., 2009; Dee et al., 2011).”

The correction was based on the Specific Remark 2 of Reviewer #2.

Page 5, after Line 2 we added:

“Here, the ERA-Interim pressure-level data are selected as the initial forcing. One reason is that, as is necessary, the vertical grid spacing between the adjacent pressure layers is less than 1 km in the free troposphere, where the convective processes mainly occurred during the Beijing SDHR event. In addition, the NWP models used by the Chinese Meteorological Centre mainly employ 31 vertical levels in regional forecasting (WMO, 2013).”

The correction was made based on the General Comment 5 of Reviewer #1 and General Remark 5 of Reviewer #2.

Page 5, Line 7 “WRF 3.7.2” was replaced by “The advanced WRF (ARW-WRF) model, version 3.7.1”.

Page 5, Line 14 the reference was amended to “Skamarock et al. (2008)”.

Page 5, Line 20 “they two together” was replaced by “Together”.

Page 8, after Line 10 the original statement was replaced by:

“The radiation processes are represented by the RRTMG shortwave radiation and the RRTMG longwave radiation schemes (Iacono et al., 2008).”

The correction was made based on the Specific Remark 3 of Reviewer #2.

Page 8, after Line 22 we added:

“The initial fields and the model physics are the same for all of the domains throughout the entire comparative procedure.”

The correction was made based on the General Remark 4 of Reviewer #2.

Page 8, after Line 23 we added:

“Because the area of interest is located in the middle latitudes, the Lambert conformal projection centred on the same latitude (42.25 °N) and longitude (114.0 °E) is employed in all of the experiments.”

The correction was made based on the General Comment 6 of Reviewer #1.

Page 9, after Line 21 we added:

"In the Beijing SDHR event, the pressure-level data meet the requirement of a grid spacing of less than 1 km in the troposphere; however, this condition is not necessarily satisfied in other regions. Thus, an experiment forced by the ERA-Interim model-level data with 38 vertical levels (C5) is also designed for comparison."

The content was added according to the General Remark 5 of Reviewer #2.

Page 9, after Line 24 the original statement was replaced by:

"The three experiments (OS2, C6, and C7) in scenario three (S3) differ in terms of their horizontal resolutions and nesting ratios; the increased nesting ratios are 1:3:3 (4.5 km grid spacing in D03), 1:5:5 (1.62 km in D03) and 1:7:7 (0.826 km in D03)."

The correction was made based on the Specific Remark 4 of Reviewer #2.

Page 10, Line 1 the original statement was replaced by:

"Both objective and subjective verification methods are applied to the innermost domain (D03) at a sub-daily scale.**"**

The correction was made based on the Editor's comments.

Page 10, after Line 1 the original statements were replaced by:

"Both objective and subjective verification methods are applied to the innermost domain (D03) at a sub-daily scale. D03 is selected because it covers the area of interest, Beijing, and the convective processes in this domain can be explicitly resolved in all of the experiments. The rainfall data used for comparison in D03 are 3-hourly 0.05-degree data that were produced by fusing rain gauge observations and the CMORPH data (Huang et al., 2013)."

The correction was made based on the Editor's comments.

Page 10, after Line 9 we added:

"The comparison on the scale of D02 is used only as an auxiliary method for subjective verification, based on the assumption that an experiment with good performance in the inner domain should also capture the large-scale features in the outer domain, as the appropriate representation of these large-scale features will result in more accurate boundary conditions."

Page 10, after Line 14 we added:

"Five are rainfall-related and compared by bilinear interpolation of the output of the simulations to the grid of the ground truth data."

The correction was made based on the General Comment 12 of Reviewer #1.

Page 10, after Line 16 we added:

“The percentage of correct rainfall hits is measured using the probability of detection (POD) with a threshold of 0.1 mm.”

The correction was made based on the General Comment 11 of Reviewer #1.

Page 10, after Line 22 we added:

“For comparison, the PW fields of the reanalysis are remapped to the grids of the model outputs using the WRF Preprocessing System (WPS).”

The correction was made based on the General Comment 13 of Reviewer #1.

Page 10, after Line 23 we added:

“In this study, all of the metrics are calculated between the simulations and the reference data on the same grid at each time step (3 h in D03). The values of these metrics are then averaged over four different sub-daily time periods (6 h, 12 h, 18 h, and 24 h) counted from 12 am on 21 July 2012. Different time periods are selected with the purpose of determining whether the performance of WRF differs when the evaluation is conducted using different durations.”

The correction was made based on the General Comment 9 of Reviewer #1.

Page 11, after Line 7 the original statements were replaced by:

“To facilitate evaluation, the metrics are further adjusted to ensure that the ideal value of all of the metrics is 1. In this study, only $RMSE$ and $WRMSE$ must be rescaled. They are first divided by a rescaling factor to fall into the range of 0-1 and then subtracted from 1 to provide an indication of good performance. The rescaled metrics, $RMSE'$ and $WRMSE'$, have the value 1 representing the lowest accumulated error (highest accuracy). The factor used for rescaling is determined by the largest values of the error metrics in all of the experiments (Sikder and Hossain, 2016). The other metrics are not rescaled because they already have ideal values of 1, but they are assigned a new set of symbols to distinguish them from the original metrics used before rescaling. For example, RE_{PMAX} is replaced with $PMAX'$, and RE_{TP} is replaced with TP'. **Table 2** shows the correlations between the original metrics and the rescaled metrics.”

The correction was made based on General Remark 7 of Reviewer #2.

Page 11, after Line 10 we added:

“The factor used for rescaling is determined by the largest values of the error metrics in all of the experiments (Sikder and Hossain, 2016).”

The content was added based on the General Comment 8 of Reviewer #1.

Page 11, after Line 20 the Results and Discussions (Section 5 and Section 6) were rewritten based on reevaluation of the performance of WRF model on the D03 scale by using a higher resolution reference dataset.

The corrections were mainly made based on the General Comment 1 of Reviewer #1 and the General Remark 3 of Reviewer #2.

Page 12, after Line 29 we added:

"In this scenario, if the experiments are merely evaluated on the scale of D03, the conclusion that C0 displays the best performance during most of the evaluated time periods may be reached. However, at the scale of D02, clear differences between C0 and the ground truth in both the spatial characteristics of the rainfall and the magnitude of the maximum precipitation are detected. Fig. 4 shows the spatial distribution of the accumulated six-hour precipitation over domain D02 in C0. Note that the speed of movement of the belt of heavy rain simulated in C0 is a few kilometres per hour faster than those in C1 and C2, leading to an early end of the heavy rainfall event. This difference may explain why the modelling skill of C0 declines significantly as the end of the rainfall event approaches. The belt of heavy rain in C0 displays an orientation that is shifted nearly ten degrees northward from those simulated in C1 and C2 during the first six hours, and the storm centre in C0 displays the smallest range; it is nearly half of the area in C2. The results indicate that the domain size of C0 is not broad enough to allow the model physics to fully develop the small-scale features that favour heavy rainfall. The spatial characteristics of precipitation are relatively similar in the other two experiments, but C1 outperforms C2 in both the rainfall-related and the PW-related features on the scale of D02. It may be that C2 does not yield better performance than C1 because of its inefficient use of boundary conditions to adjust the false perturbations generated by the local model run."

The content was added mainly based on the General Comment 14 of Reviewer #1.

Page 13, Line 28 "less sensitivity" was replaced by "less sensitive".

Page 14, after Line 15 we added:

"As shown in Fig. 5, C5 shows either better or worse performance than C1 in each period but produces less accurate rainfall simulations than C3 over most of the evaluated durations."

The content was added according to the General Comment 5 of Reviewer #1 and General Remark 5 of Reviewer #2.

Page 14, after Line 23 the original statements were replaced by:

"Over most of the evaluated time periods, C6, which has a grid spacing of 1.62 km, displays better performance than C3 and C7 having grid spacings of 4.5 km and 0.826 km, respectively."

The correction was made based on the Specific Remark 4 of Reviewer #2.

Page 15, after Line 27 the original statements were replaced by:

“Positive biases are detected in $PMAX'$ in C9 (which is run 24 hours ahead) and C11, in which the largest positive biases are detected in the simulated amount of water vapour across the analysed periods and earlier (during the initialization period). This result agrees with intuition because the atmospheric water vapour content determines the maximum possible rainfall amount.”

The correction was made based on the Specific Remark 5 of Reviewer #2.

Page 17, after Line 20 we added:

“Given that the uncertainties in the regional NWP studies result mainly from the inaccurate boundary conditions associated with grid nesting techniques, methods that can serve as alternate schemes to reduce these uncertainties are also worth studying. Examples include the mesh transitions approach used on irregular grids.”

The content was added according to the General Remark 8 of Reviewer #2.

Page 17, after Line 23 we added:

“In addition, more accurate simulations are expected when the model is driven with forcing data with higher temporal or spatial resolutions than those of the ERA-Interim reanalysis because the uncertainties and errors introduced by the input data are then further reduced.”

The content was added according to the Specific Remark 6 of Reviewer #2.

Additional references:

Heinzeller, D., Duda, M. G. and Kunstmann, H.: Towards convection-resolving, global atmospheric simulations with the Model for Prediction Across Scales (MPAS) v3. 1: an extreme scaling experiment. *Geosci. Model Dev.*, **9(1)**, 77, 2016.

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.

Sikder, S., and Hossain, F.: Assessment of the weather research and forecasting model generalized parameterization schemes for advancement of precipitation forecasting in monsoon-driven river basins. *J. Adv. Modeling Earth Syst.*, **8(3)**, 1210-1228, 2016.

Swinbank, R. and James Purser, R.: Fibonacci grids: A novel approach to global modeling. *Q. J. R. Meteorol. Soc.*, **132(619)**, 1769-1793, 2006.

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.*, **17(4)**, 563-579, 2017.

Figure and table captions:

All the abbreviations in the figure and table captions were checked to make sure their full names are given before they are used. Besides, the captions were rewritten to avoid rewording.

This correction was made based on the General Remark 6 of Reviewer #2.

Figures were amended:

Figure 3, 5, 6, 7 were replaced by the subfigures which showed the spatial values of the verification metrics for the WRF experiments calculated over domain three.

New figure was added:

Figure 4 was added, showing the spatial distribution of 6-h accumulated precipitation for the domain size experiments within domain two of Case 0 during the heavy rainfall event beginning at 12 am on July 21, 2012.

Data in the tables were amended:

The results in the tables were updated after we repeated the evaluation on the D03 scale.

The marked-up manuscript version

Evaluation of the ability of the Weather Research and Forecasting model to reproduce a sub-daily extreme rainfall event in Beijing, China using different domain configurations and spin-up times

Qi Chu[1,2,3], Zongxue Xu[1,2], Yiheng Chen[3], and Dawei Han[3]

[1] College of Water Sciences, Beijing Normal University, Beijing, 100085, China

[2] Beijing Key Laboratory of Urban Hydrological Cycle and Sponge City, Beijing 100875, China

[3] Department of Civil Engineering, University of Bristol, Bristol, BS8 1TR, UK

Abstract. The rainfall outputs from the latest convection-scale Weather Research and Forecasting (WRF) model are shown to provide an effective means of extending prediction lead times in flood forecasting. In this study, the performance of the WRF model in simulating a regional sub-daily extreme rainfall event centred over Beijing, China is evaluated at high temporal (sub-daily) and spatial (convective-resolving) scales using different domain configurations and spin-up times. Seven objective verification metrics that are calculated against the gridded ground observations and the ERA-Interim reanalysis are analysed jointly using subjective verification methods to identify the likely best WRF configurations. The rainfall simulations are found to be highly sensitive to the choice of domain size and spin-up time at the convective scale. A model run covering northern China with a 1:5:5 horizontal downscaling ratio (1.6 km), 57 vertical layers (0.5 km), and a 60-hour spin-up time exhibits the best performance in terms of the accuracy of rainfall intensity and the spatial correlation coefficient (*R'*). A comparison of the optimal run and the initial run performed using the most common settings reveals clear improvements in the verification metrics. Specifically, *R'* increases from 0.226 to 0.67; the relative error of the maximum precipitation at a point increases from 0.44 to 0.883; and the cumulative spatial error decreases by 33.65 %. In summary, re-evaluation of the domain configuration options and spin-up times used in WRF is crucial in improving the accuracy and reliability of rainfall outputs used in regional sub-daily heavy rainfall (SDHR)-related applications.

1 Introduction

The possibility that sub-daily heavy rainfall (SDHR) will increase with climate change is of significant societal concern. SDHR-driven flash floods (FFs) are among the most destructive natural hazards that threaten many urban areas in northern and central China and many other parts of the world. In these regions, SDHR is triggered mainly by regional mesoscale circulation systems (MCSs) and occurs with increased intensity and frequency in warm seasons (Yu et al., 2007; Chen et al., 2013). Records from the Emergency Events Database (EM-DAT) indicate that the damages and losses caused by FF events in China have increased significantly over the past several decades. The risks are expected to continue to grow, given the increase in the magnitude of SDHR predicted by most general circulation models (Chen et al., 2012; Willems et al., 2012; Westra et al., 2014). The accelerating pace of urbanization also contributes to the increase in risk; urbanization has already changed the hydrologic characteristics of the land surface considerably, resulting in higher peak flows and shorter flow concentration times (Xu and Zhao, 2016; Gao et al., 2017). In such cases, very short-term (< 6-h) rainfall predictions are not sufficient to provide adequate warning and mobilize emergency response activities. Recently developed statistically-based rainfall generation methods and remote sensing data have been shown to enable the extension of the lead time to 24 hours. However, this lead time is still insufficient to provide effective flood mitigation for medium or large urban areas with very short hydrologic response times (Shih et al., 2014; Li et al., 2017). Therefore, numerical weather prediction (NWP), which represents a means of forecasting heavy rainfall with lead times exceeding 24 h, has come into wide use in flood-related studies and applications (Cuo et al., 2011).

Precipitation uncertainty accounts for a large proportion of the uncertainty in flood forecasts. Hence, given the large uncertainties of NWP, its use in flood forecasting has long been questioned (Castelli, 1995; Bartholmes and Todini, 2005). The ice wasn’t broken until substantial improvements in the predictive skill of NWP were made that resulted from increases in computational power and storage capacity, which enable parallel processing of high-resolution forcing data and the resolution of convective-scale physical processes (Done et al., 2004; Clark et al., 2016). NWP models developed during this period can perform regional and convective-scale modelling and display good performance in simulating heavy rainfall. Experimental studies have shown that NWP models of this kind, such as the Weather Research and Forecasting (WRF) model, tend to capture greater numbers of small-scale processes and the triggers of convective storms (Klemp, 2006; Prein et al., 2015). Increasing numbers of meteorological operational centres and research groups are adopting these new NWP models to carry out simulations of heavy rainfall events or real-time forecasting. The resolutions of the rainfall products have improved from tens of kilometres to less than a kilometre, and the lead times have increased from less than a day to more than a week (WMO, 2013). Meanwhile, case studies have been carried out using regional convective-resolving models to evaluate the local rainfall predictions generated by sophisticated regional nesting techniques or the global smooth grid transition approach on unstructured grids (Hong and Lee, 2009; Soares et al., 2012; Sikder and Hossain, 2016; Heinzeller et al. 2016). The results of these studies demonstrate that, over relatively short periods of time, regional modelling is often superior to large-scale modelling because it better resolves surface heterogeneities, topography and small-scale features in air flow, such as growing instabilities (Miguez et al., 2004; En-Tao et al., 2010; Prein et al., 2015; Brommel et al. 2015).

Despite the great potential of NWP models to predict heavy rainfall, a number of uncertainties remain that must be considered. The errors induced by the initial and boundary conditions represent one source of these uncertainties; others stem from cognitive errors and the scale effect in the solution of physical models, both of which may be exacerbated by the chaotic nature of NWP. In regional simulations, these uncertainties are expected to be further magnified by downscaling or the use of mesh transition procedures, so re-evaluation and calibration of the related model configurations are commonly required (Warner, 2011; Vrac et al., 2012; Liu et al., 2012). As an example, running the WRF model at convective scales means that convective processes are more likely to be resolved by explicit physical schemes than when sub-grid parameterizations are used, which may incorporate new structural uncertainties related to the model physics (Done et al., 2004; Ruiz et al., 2010; Crétat et al., 2012). In addition to model physics, several other aspects of model configuration, such as the spatial resolution and the spin-up time, may also have a substantial impact on the uncertainty of rainfall forecasts through their effects on the initial and boundary conditions (Aligo et al., 2009; Fierro, 2009; Cuo et al., 2011). However, these aspects of model configuration have received relatively little attention in regional case studies because of the relative insignificance of their effects on rainfall forecasts in coarse-resolution and long-term model simulations when compared to the physics of the WRF model. These model configuration aspects are commonly left at the settings recommended by the official website of the WRF model and by some experimental regional heavy rainfall studies.

Precipitation is one of the most sensitive variables to NWP model uncertainties. In this study, a re-evaluation of WRF is performed to determine whether the recommended configuration of WRF represents the best choice in reproducing a regional SDHR event. The WRF model is assessed here because of its superior scalability and computational efficiency; these traits are valued in interdisciplinary studies (Klemp, 2006; Foley et al., 2012; Coen et al., 2013; Yucel et al., 2015). As the latest NWP community model, WRF incorporates up-to-date developments in physics, numerical methods and data assimilation and is thus widely used in theoretical studies and practical applications (Powers et al., 2017). The selected regional SDHR event occurred on July 21st, 2012 and was centred over Beijing, China. Beijing is among the most vulnerable cities to SDHR-induced floods in central China (Yu et al., 2007). The precipitation in this area is caused mainly by monsoon weather systems and enhanced by local orographic effects, and 60 % - 80 % of the total annual precipitation occurs during just a few SDHR events (Xu and Chu, 2015). The SDHR event that occurred on 21 July 2012 caused the most disastrous urban flood in Beijing since 1950. The national operational NWP system failed to predict this event, which resulted in 79 deaths and more than 1.6 billion dollars in damage (Wang et al., 2013; Zhou et al., 2013). Thus, several convective-scale studies have been carried out to re-evaluate the optimal combination of the physics options used in the WRF model (Di et al., 2015; Wang et al., 2015), and these studies represent background information that stimulates this research.

The second question we attempt to explore is to what extent rainfall simulations could be improved through the use of the likely best set of settings if the recommended model configurations are not the best choices. The aspects of the model configuration that are evaluated in this study are the domain size, vertical resolution, horizontal resolution and spin-up time. These options have been found to have substantial impacts on daily-scale extreme rainfall outputs (Leduc and Laprise, 2009; Aligo et al., 2009; Goswami et al., 2012). A comparative test with four scenarios is designed. Each scenario evaluates one model configuration option to ensure that the simulated disparities can be attributed solely to a single factor each time. In addition, the test is conceived as a progressive process: the optimal setting identified in each scenario will be adopted as the primary choice for the next scenario to help quantify the overall improvement in the accuracy of rainfall outputs. The 'ground truth' datasets are gridded observations obtained from Beijing Normal University and the China Meteorological Centre. A coarser-resolution reanalysis called ERA-Interim is also employed in identifying departures of the WRF simulations from the driving weather fields as the model setup is varied. Seven objective verification metrics that reflect different features of the model performance are adopted and considered jointly as part of a subjective verification process because no single verification approach has been shown to provide comprehensive information about the quality of rainfall simulations (Sikder and Hossain, 2016). Most of the metrics adopted here are those used to assess the performance of WRF over daily or longer time periods (Liu et al., 2012; Tian et al., 2016). In this research, these metrics are calculated on an hourly basis and averaged over different sub-daily time spans to evaluate the performance of the WRF model using different configurations from a sub-daily and convective-scale perspective.

2 Numerical Models Used to Forecast Heavy Rainfall

The downscaling is performed using a global atmospheric reanalysis dataset called ERA-Interim, which is produced by an integrated forecasting system (IFS) used by the European Centre for Medium-Range Weather Forecasts (ECMWF). The IFS is an Earth system model that incorporates a data assimilation system and an atmospheric model that is fully coupled with land-surface and oceanic processes. The atmospheric model provides output every 30 min at a spectral resolution of T255 (approximately 81 km over Beijing). This output is then employed as prior information and combined with available observations twice a day to produce the reanalysis output using the four-dimensional variation (4D-Var) assimilation system. The final reanalysis product, ERA-Interim, is a global gridded dataset that is available at a spectral resolution of T255 and at both the 60 levels used in the model and 38 interpolated pressure levels for all dates beginning on 1 January 1979 (Berrisford et al., 2009; Dee et al., 2011). Here, the ERA-Interim pressure-level data are selected as the initial forcing. One reason is that, as is necessary, the vertical grid spacing between the adjacent pressure layers is less than 1 km in the free troposphere, where the convective processes mainly occurred during the Beijing SDHR event. In addition, the NWP models used by the Chinese Meteorological Centre mainly employ 31 vertical levels in regional forecasting (WMO, 2013).

The advanced WRF (ARW-WRF) model, version 3.7.1, is utilized as the dynamical downscaling tool. ARW-WRF is a compressible non-hydrostatic and convection-permitting regional NWP model that employs the conservative form of the dynamic Euler equations. As the latest regional NWP community system, WRF is composed of two dynamic cores, a data assimilation system and a platform that facilitates parallel computation and function portability. Observations, model output or assimilated reanalysis output can be used to initialize WRF. In terms of discretization, WRF uses a third-order Runge-Kutta method for temporal separation and an Arakawa C-grid staggering scheme for spatial discretization. The model is capable of conducting either one-way or two-way nested runs for regional downscaling. A detailed introduction to the physics and numerical properties of ARW-WRF can be found in Skamarock et al. (2008). Given its emphasis on efficiency, portability and updates to reflect the state of the art, WRF has been employed in settings ranging from research to applications and has been incorporated into various operational systems, such as the Hurricane-WRF system for hurricane forecasting and the WRF-Hydro system for hydrologic prediction.

In WRF, the domain size implicitly determines the large-scale dynamics and terrain effects, whereas the vertical and horizontal grid spacings determine the smallest resolvable scale (Goswami et al., 2012). Together, these domain configuration options affect the spectrum of the resolved scales and the nature of scale interactions in the model dynamics (Leduc and Laprise, 2009). Thus, they are responsible for the generation and distribution of precipitation. In regional simulations, small domain sizes are commonly preferred for computational efficiency. Seth and Rojas (2003) demonstrated that simulations with small domain sizes are more likely to benefit from the lateral boundary conditions (LBCs) by dampening the feedback from local perturbations on the large-scale general circulation. However, insufficiently large domains have been shown to prevent the full development of small-scale features over areas of interest. To solve this issue, the official website of WRF provides general guidance (Warner, 2011). This guidance recommends that the ranges of domains should include the major features of the leading MCSs and local surface perturbations, and more than five grid points should exist between adjacent nested domains to allow sufficient relaxation.

As for grid spacing, it appears plausible that WRF model runs performed with relatively small grid spacings would provide more accurate outputs because such runs would resolve more small-scale phenomena of interest that are not present in the

LBCs. This statement is generally accepted as true when a relatively coarse-resolution run (>10 km horizontally or >1 km vertically) is compared with a relatively finely resolved run at the convective scale (1 - 5 km horizontally or <1 km vertically) in representing a convective storm. However, this conclusion is controversial when the comparison is conducted among convective-scale model runs. Taking the horizontal resolution as an example, although there is evidence to show that WRF runs performed at relatively high resolution capture more convective-scale features, the accuracy of rainfall outputs either shows considerable or no statistical improvement (Roberts and Lean, 2008; Kain et al., 2008; Schwartz et al., 2009). In one study, Fierro (2009) suggested that some features detected in convective-scale runs with excessive horizontal grid spacings tend to weaken the kinetic structures that favour torrential rainfall. A similar conclusion was drawn by Aligo et al. (2009) in evaluating the impact of the vertical grid spacing on simulations of summer rainfall performed using WRF. Thus, horizontal and vertical grid spacings of approximately 4 km and 1 km, respectively, have been employed as a reasonable compromise between accuracy and computational efficiency in several regional studies.

In regional modelling, a spin-up period is often required to balance the inconsistencies between the results simulated by the model physics and the initial and boundary conditions provided by the forcing data (Luna et al., 2013). The proper spin-up time depends on the time needed for initialization, which can be affected by the size of the domain and the local boundary perturbations (Warner, 1995; Kleczek et al., 2014). Moreover, the presence of chaotic behaviour, which causes reductions in the predictive skill of models over time, imposes an upper bound on the spin-up time. Therefore, in cases where short spin-up times are expected, e.g., real-time rainfall forecasting, the spin-up time is mainly determined by the domain size and the regional initial and boundary conditions. However, in cases where long spin-up times are needed, e.g., warnings of extreme rainfall, the effects of chaotic behaviour should be relatively evident. In practice, this issue is commonly addressed by regularly updating the lateral boundary information derived from the latest forecasts or analyses to maintain consistency between the regional model solutions and the atmospheric forcing conditions. In such cases, the best-fit performance may occur for model runs with long spin-up times. Based on most previous studies, a spin-up time of 12 hours is recommended to obtain an initial state; however, this spin-up time has been regarded as the most suitable choice in many regional case studies without further verification.

3 Studied Event and Experimental Design

As mentioned above, one aim of this study is to re-evaluate whether the recommended WRF domain configuration options and spin-up time represent the optimal model configuration for reproducing a regional SDHR event when evaluated at a sub-daily scale. Here, the SDHR event that occurred on 21 July 2012 and was centred on Beijing, China is selected as a case study. The reasons why this event is selected, the synoptic and physical features that drove this event, and the model physics adopted in this study are presented before the entire procedure of the experimental design is introduced.

3.1 Study Event Selection and WRF Configurations

Beijing is selected as the study area because it is one of the most vulnerable cities to SDHR-induced FF hazards in China. Beijing is located in central China. It has an area of 16 411 km^2, and its weather is mainly affected by the semi-humid warm continental monsoon climate. The flows of air that favour local precipitation are cold, dry flows of air from high-latitude areas to the north and hot, wet flows of air from the ocean to the south. The interactions between these two flows of air lead to clear divergence in the temporal distribution of rainfall amount; 60 % **-** 80 % of the annual precipitation occurs during just a few heavy rainfall events during the warm season. Of all of the heavy rainfall events, the intensity and frequency of SDHR events have been shown to display the greatest increasing tendencies over the past several decades. Meanwhile, Beijing, as the capital of China, has experienced a significant expansion of its urban area and rapid increases in its population and economic development. The negative effects of this expansion, such as losses of natural water bodies, increases in land cover with low permeability and increases in urban drainage pipe networks, have led to continuous decreases in the hydrologic response time. In addition, most of the population lives in the southwestern plain area. This region is downstream of mountainous areas with steep terrain that varies in elevation from 2 300 m to 60 m (**Fig. 1**). All of these factors contribute to the continuing increase in the exposure of this city to the high risks of flooding and waterlogging caused by SDHR events (Xu and Chu, 2015).

[Figure 1]

The case study examines the largest heavy rainfall event that has occurred in Beijing in the past 65 years. The rainfall event lasted for 16 hours (from 2 am to 6 pm) on 21 July 2012 (UTC), and the highest hourly rainfall intensity (100 mm/h) was experienced in the southwestern part of the plain area. The associated FF hazard led to 79 deaths and damages totalling 1.6 billion US dollars, and more than 1.6 million people were affected. In addition to Beijing, the adjacent provinces, including Hubei and Liaoning, were all significantly affected by this event and experienced severe FF hazards. The synoptic features that triggered the rainfall were an eastward-moving vortex in the middle to high troposphere, a northward-moving zone of subtropical high pressure and sharp vertical wind shear (Sun et al., 2013). The rainfall event as a whole can be divided into two phases. From 2 am to 2 pm, the convective rain was dominated and enhanced by the orographic effect. The frontal rain was then followed by the arrival of a cold front moving from the northwest until 6 pm (Guo et al., 2015). The rainfall intensity in the second phase was relatively low compared to that in the first phase, due to the lack of strong kinetic forcing to maintain the occurrence of precipitation.

The ERA-Interim reanalysis and 30-second static geographical data are employed to initialize the surface and meteorological fields of the WRF. As shown in **Fig. 2**, ERA-Interim captures the vortex and the subtropical high pressure well that occurred at the beginning of the rainfall event. In addition, the patterns of the leading MCSs and the primary synoptic features shown in this figure also correspond well to those described in previous studies (Zhou et al., 2013). The setup of the model physics is based mainly on the results of sensitive, high-resolution studies on the physics of the WRF model in simulating the same event (Wang et al., 2015; Di et al., 2015). The 'resolved rain' is driven by the single-moment 6-class microphysics scheme (Hong and Lim., 2006), whereas the 'convective rain' is resolved using the Grell-Devenyi cumulus parameterization scheme (Grell and Devenyi, 2002). The Noah land-surface model (Chen and Dudhia, 2001) is used and coupled with the surface layer model once utilized in MM5 (Ek et al., 2003). The radiation processes are represented by the RRTMG shortwave radiation and the RRTMG longwave radiation schemes (Iacono et al., 2008). For the planetary boundary layer scheme, the Yonsei University method (Hong et al., 2006) is adopted.

[Figure 2]

3.2 Experimental Design: Domain Configuration Options and Spin-Up Time

The comparative test is designed as a progressive process to help quantify the overall improvement in the performance of WRF after re-evaluating the WRF experiments performed using different domain configuration options and spin-up times. The test is classified into four successive scenarios. The first three scenarios investigate the domain configuration options, including the domain size, vertical resolution, and horizontal resolution; the fourth scenario concerns the spin-up time. During the entire procedure, the optimum configuration identified in each scenario is then adopted as the primary choice for the corresponding configuration in the following scenario. The initial fields and the model physics are the same for all of the domains throughout the entire comparative procedure. Because the area of interest is located in the middle latitudes, the Lambert conformal projection centred on the same latitude (42.25 ° N) and longitude (114.0 ° E) is employed in all of the experiments. Moreover, sigma vertical coordinates with a top level of 50 hPa are used in all of the experiments.

Initially, the WRF domain configuration options and the spin-up time are set to the recommended values described in Section 2. Three levels of nested domains are adopted so that the horizontal resolution in the smallest domain is sufficiently high to explicitly resolve convective-scale processes (**Fig. 1**). An odd downscaling ratio (1:3:3) is selected to reduce the initial error introduced by interpolating the initial fields to the assigned domains. For the same reason, the boundaries of each domain are set along specific grid lines of the ERA-Interim dataset. Of the three nested domains, the outermost domain (D01) has the largest horizontal grid spacing of 40.5 km over north-central China, where the main perturbed synoptic features occur. The innermost domain (D03) has the smallest horizontal grid spacing of nearly 4.5 km over the area of interest, Beijing. The second domain (D02) is the child of D01 and the parent of D03 and has a horizontal grid spacing of 13.5 km. The distance between D01 and D02 is similar to that between D02 and D03, both of which exceed five grid points. The eta values utilized in the initial run are set based on the pressure values at the 29 vertical layers of the ERA-Interim pressure-level data. A spin-up time of twelve hours (12 h) is selected; the outputs are saved every three hours in D03 and every hour in D02. The LBCs are updated every six hours using ERA-Interim.

As shown in **Table 1,** the first experiment (C0) adopts the model configuration options mentioned above. To determine whether the domain configuration options and the spin-up time used in C0 are the likely best set, four scenarios are designed. The first scenario (S1) focuses on evaluating the effect of the WRF domain size. For computational efficiency, the leading MCS systems that drive the perturbed synoptic features are not completely contained within the outermost domain of C0, the information of which is compensated by the updated LBCs from ERA-Interim. Two comparative experiments, C1 and C2, are devised to verify that the domain size assigned to C0 is large enough to enable the full development of small-scale features. Of the three experiments, C2 has the largest outermost domain size, which incorporates the leading MCS systems over the entire Northeastern Hemisphere. The intermediate domain, which is centred between the outermost and innermost domains, is then adopted as the outermost domain of C1. The purpose of scenario two (S2) is to evaluate whether the use of a higher vertical resolution in a WRF model run results in better performance. In this scenario, the starting experiment is the optimal experiment identified in S1 (OS1), forced by the ERA-Interim pressure-level data with 29 vertical levels. This starting experiment is then followed by two experiments, C3 and C4, which incorporate one and two times more vertical levels than OS1 (57 and 85 vertical levels), respectively. In the Beijing SDHR event, the pressure-level data meet the requirement of a grid spacing of less than 1 km in the troposphere; however, this condition is not necessarily satisfied in other regions. Thus, an experiment forced by the ERA-Interim model-level data with 38 vertical levels (C5) is also designed for comparison. The three experiments (OS2, C6, and C7) in scenario three (S3) differ in terms of their horizontal resolutions and nesting ratios; the increased nesting ratios are 1:3:3 (4.5 km grid spacing in D03), 1:5:5 (1.62 km in D03) and 1:7:7 (0.826 km in D03). The last scenario (S4) is designed to identify a reasonable optimal model run with the maximum spin-up time after minimizing the uncertainties introduced by inappropriate domain configuration options. It contains one starting experiment (OS3) and twelve comparative experiments (C8-C19). Except for C8, which includes no spin-up time, the remaining experiments (C9-C19) include spin-up times that increase from 24 hours to 144 hours by every twelve hours.

[Table 1]

4 Verification Schemes

Both objective and subjective verification methods are applied to the innermost domain (D03) at a sub-daily scale. D03 is selected because it covers the area of interest, Beijing, and the convective processes in this domain can be explicitly resolved in all of the experiments. The rainfall data used for comparison in D03 are 3-hourly 0.05-degree data that were produced by fusing rain gauge observations and the CMORPH data (Huang et al., 2013). The ERA-Interim reanalysis is utilized as well to monitor the possible departures of the model simulations from the driving fields. Because the sub-daily rainfall is not available from the reanalysis, the atmospheric precipitable water vapour (PW), which determines the possible maximum precipitation, is instead compared with the model outputs every six hours. In addition, the model outputs that cover a larger domain (D02) are compared with an hourly 0.1-degree gridded dataset obtained from the China Meteorological Centre. The comparison on the scale of D02 is used only as an auxiliary method for subjective verification, based on the assumption that an experiment with good performance in the inner domain should also capture the large-scale features in the outer domain, as the appropriate representation of these large-scale features will result in more accurate boundary conditions.

Seven error metrics that describe different features of precipitation are selected for use as objective verification metrics. Five are rainfall-related and compared by bilinear interpolation of the output of the simulations to the grid of the ground truth data. The accumulated areal rainfall is assessed using the relative error of the total precipitation (RE_{TP}). The percentage of correct rainfall hits is measured using the probability of detection (POD) with a threshold of 0.1 mm. The root mean squared error ($RMSE$) represents the amount of continuous error in the predicted precipitation. Detailed illustrations of these three metrics can be found in Liu et al. (2012) and Tian et al. (2016). The other two rainfall-related metrics are the relative error of the maximum grid precipitation (RE_{PMAX}) and the Pearson correlation coefficient (R), which describe the spatial association between the simulations and the ground truth data (**Eq. (1) and Eq. (2)**). The two metrics selected for the verification of PW are the root mean squared error ($WRMSE$) and the Pearson correlation coefficient (WR). For comparison, the PW fields of the reanalysis are remapped to the grids of the model outputs using the WRF Preprocessing System (WPS). In this study, all of the metrics are calculated between the simulations and the reference data on the same grid at each time step (3 h in D03). The values of these metrics are then averaged over four different sub-daily time periods (6 h, 12 h, 18 h, and 24 h) counted from 12 am on 21 July 2012. Different time periods are selected with the purpose of determining whether the performance of WRF differs when the evaluation is conducted using different durations.

$$RE = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{f-r}{r}\times 100\ \%\right] \tag{1}$$

$$R = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sum_{j=1}^{M}(f_j-\bar{f})(r_j-\bar{r})}{\sqrt{\sum_{j=1}^{M}(f_j-\bar{f})^2\sum_{j=1}^{M}(r_j-\bar{r})^2}}\right) \tag{2}$$

Here, R is the empirical spatial correlation coefficient; M is the total number of grid points within the evaluated domain of the starting experiment; f_j is the value of the j th grid point in the tested field at time step i; r_j is the value of the reference field; N is the total number of time steps, which is 6, 12, 18, or 24, depending on the time period considered; and RE is the relative error. For the maximum precipitation, f is the tested value of the maximum gridded precipitation over the area of interest, and r is the reference value of the maximum gridded precipitation over the same area.

To facilitate evaluation, the metrics are further adjusted to ensure that the ideal value of all of the metrics is 1. In this study, only $RMSE$ and $WRMSE$ must be rescaled. They are first divided by a rescaling factor to fall into the range of 0-1 and then subtracted from 1 to provide an indication of good performance. The rescaled metrics, $RMSE'$ and $WRMSE'$, have the value 1 representing the lowest accumulated error (highest accuracy). The factor used for rescaling is determined by the largest values of the error metrics in all of the experiments (Sikder and Hossain, 2016). The other metrics are not rescaled because they already have ideal values of 1, but they are assigned a new set of symbols to distinguish them from the original metrics used before rescaling. For example, RE_{PMAX} is replaced with $PMAX'$, and RE_{TP} is replaced with TP'. **Table 2** shows the correlations between the original metrics and the rescaled metrics. Given that the metrics describe different features of the rainfall simulations, the values of these metrics are checked and considered together in subjective verification to determine the likely best set of domain configuration options and to search for the longest reasonable spin-up time.

[Table 2]

5 Results and Analyses

In the first three scenarios, all of the experiments are run for 36 h, from 12 pm on 20 July 2012 to 12 am on 22 July 2012. The time periods selected for evaluation are 6 h, 12 h, 18 h and 24 h, all of which begin at the same time at 12 am on 21 July 2012. In each scenario, the metrics are compared among the experiments that consider different durations and cover the same domain (D03). The results are presented in four sub-graphs; each sub-graph shows the values of the metrics calculated for individual evaluated time periods. The spatial distribution of rainfall is also presented over a slightly larger area when evident discrepancies are noted in the results obtained for the inner domain (D03) and the outer domain (D02). **Table 1** shows the categories of the scenarios and the model configurations adopted in each experiment. In the following section, the domain size scenario (S1) is evaluated first, followed by the vertical resolution scenario (S2) and the horizontal resolution scenario (S3).

5.1 Results of the Domain Size Scenario

Fig. 3 shows the spatial values of the verification metrics for the WRF domain size experiments. The performance of the experiments clearly worsens as the evaluated temporal duration increases from 6 h to 24 h. The most evident deteriorations are detected in the point-to-point accuracy of the rainfall; the reversed root mean squared error ($RMSE'$) decreases by 0.8, which represents a six-fold increase in the cumulative spatial error. The spatial association between the simulations and the gridded observations also declines; the correlation coefficient (R') decreases by 0.3 on average. Although a slight increase is observed in the percentage of correct hits (POD') during the first 18 hours, this increase is followed by a rapid decrease of nearly 14 % during the last stage of the rainfall event. The relative bias in the accumulated areal rainfall (TP') indicates that the total rainfall amount is underestimated throughout the entire evaluated temporal period. The maximum gridded precipitation ($PMAX'$) is also underestimated; the largest negative bias occurs during the heavy convective rainfall stage. For PW, a slight decrease is found in the reversed accumulated error ($WRMSE'$), whereas an increase of 5 % - 9 % is detected in the spatial correlation coefficient (WR'). Such variations may be attributable to the role of the updated boundary conditions in adjusting the local model solutions to approach the large-scale atmospheric circulation conditions.

[Figure 3]

Comparison of the four sub-graphs shows that the values of the metrics do not point to a single perfect experiment in a given period, and their ranked predictive skills determined using a given metric differ when evaluated over different time periods. During the early stage of the rainfall event (6 h), C0 yields better performance than C1 and C2 in terms of $RMSE'$, R' and $PMAX'$; it simultaneously displays the lowest value of POD' and the largest bias in estimating the total precipitation. Although the superiority of C0 is more evident in the second period, a sharp deterioration is then observed in capturing the point-to-point accuracy of precipitation for the 18-h duration, where the lowest R' is obtained. Meanwhile, C1, which employs a domain of moderate size, displays greater skill than C0 in capturing the correct hits and the spatial pattern of the simulated rainfall. C2 employs the largest domain. Although it shows the best fit to the rainfall observations on the daily scale (24 h), it displays the worst performance over the three shorter time periods. For the PW fields, the highest similarity with the ERA-Interim reanalysis is found for C0, whereas the lowest similarity is found for C2. These results demonstrate indirectly that small domains are more likely to be influenced by updated boundary conditions.

In this scenario, if the experiments are merely evaluated on the scale of D03, the conclusion that C0 displays the best performance during most of the evaluated time periods may be reached. However, at the scale of D02, clear differences between C0 and the ground truth in both the spatial characteristics of the rainfall and the magnitude of the maximum precipitation are detected. **Fig. 4** shows the spatial distribution of the accumulated six-hour precipitation over domain D02 in C0. Note that the speed of movement of the belt of heavy rain simulated in C0 is a few kilometres per hour faster than those in C1 and C2, leading to an early end of the heavy rainfall event. This difference may explain why the modelling skill of C0 declines significantly as the end of the rainfall event approaches. The belt of heavy rain in C0 displays an orientation that is shifted nearly ten degrees northward from those simulated in C1 and C2 during the first six hours, and the storm centre in C0 displays the smallest range; it is nearly half of the area in C2. The results indicate that the domain size of C0 is not broad enough to allow the model physics to fully develop the small-scale features that favour heavy rainfall. The spatial characteristics of precipitation are relatively similar in the other two experiments, but C1 outperforms C2 in both the rainfall-related and the PW-related features on the scale of D02. It may be that C2 does not yield better performance than C1 because of its inefficient use of boundary conditions to adjust the false perturbations generated by the local model run. Therefore, C1 is verified as reasonable from both statistical and physical perspectives and is chosen as the optimal experiment in the domain size scenario (OS1).

[Figure 4]

5.2 Results of the Vertical Resolution Scenario

Based on the analysed results, C1 is selected as the starting experiment in the vertical resolution scenario. As mentioned above, C1 is forced with the ERA-Interim pressure-level data with 29 vertical levels. C3 and C4 are forced with the same pressure-level data with 57 and 85 vertical levels, respectively, whereas C5 is forced with the model-level data with 38 vertical levels. As shown in **Fig. 5**, a decline in model performance is also obtained for all of the vertical resolution experiments as the evaluated time period increases in length. Moreover, the largest deterioration in $RMSE'$ is also observed; it decreases by 0.82 on average. The values of TP' and $PMAX'$ derived from the simulations are slightly higher than those predicted in S1 but are still less than those calculated for the actual precipitation over the entire rainfall event. POD' displays an evident decrease during the end stage of the rainfall event, and its magnitude decreases 50 % less relative to that shown in C1. The most obvious difference from the domain size scenario is that the values of R' calculated between the simulations and the ground truth vary slightly and remain almost the same between the different time periods. In addition, the performance of the vertical resolution experiments seems to be less sensitive to the boundary conditions because they result in relatively small variations in $WRMSE'$ and WR'.

[Figure 5]

Unlike the apparent discrepancies noted in the metrics obtained for the domain size experiments, the differences in the rainfall-related metrics among the experiments with different numbers of vertical levels are not evident, especially during the less rainy period (6 h) and the period when convective rainfall dominates (12 h). During the first 12 hours, C4 displays better agreement with the gridded observations than the other three experiments in terms of the accuracy and spatial correlation of the rainfall amount. However, over the longer time periods, C3 displays the greatest skill, according to most of the verification metrics. Comparing C3 and C1 shows that increases in the vertical resolution may increase WRF's ability to explicitly resolve small-scale physical processes and improve the accuracy of the amount and distribution of the simulated rainfall. Comparing C3 and C4 shows that, although C4 include further refinement of the vertical resolution, its performance is worse than that of C3 when the evaluated time period increases to more than 12 hours. This result may occur because progressive reductions in the vertical resolution magnify the propagation of surface perturbations through the vertical grid columns, potentially weakening the kinetic energy that favours precipitation. Examining the values of $WRMSE'$ and WR' shows that the differences between the simulations and the reanalysis are more distinct in C4. This discrepancy may occur due to the exaggeration of the initial errors introduced by the interpolation process and the incorporation of false surface perturbations introduced by the limited accuracy and resolution of the initial forcing data. C5 shows either better or worse performance than C1 in each period but produces less accurate rainfall simulations than C3 over most of the evaluated durations. As such, C3 is identified as yielding the best performance in the vertical resolution scenario.

5.3 Results of the Horizontal Resolution Scenario

Based on the results obtained for scenario S2, C3 is selected as the starting experiment in the horizontal resolution scenario. The modelling skill of the S3 experiments shows similar temporal trends as that of the S2 experiments (**Fig. 5** and **Fig. 6**). However, the sensitivity of the metrics to the variation of the horizontal resolution is more evident than that with different vertical resolutions. Over most of the evaluated time periods, C6, which has a grid spacing of 1.62 km, displays better performance than C3 and C7 having grid spacings of 4.5 km and 0.826 km, respectively. Comparison of C3 and C6 shows that C6 tends to produce more accurate spatial patterns of rainfall throughout the heavy rainfall event in Beijing. The $WRMSE'$ values indicate that model runs with higher horizontal resolutions are more likely to benefit from updated LBCs. Higher values of $PMAX'$ and TP' are also detected in C6 when compared to C3. This result stems in part from the explicit resolution of the convective processes by the WRF microphysics scheme, which may explain why the $PMAX'$ of C7 is higher than C6 over most of the tested durations. Note that the modelling skill of C7 deteriorates rapidly after the heavy rain begins (12 h); the lowest POD' and R' values of the three experiments are obtained for this simulation and time period. Analysis of the $WRMSE'$ values suggests that simulation C7 displays significant departures from the coarser-scale PW fields that are used to force the model. Thus, model simulations with excessively high horizontal resolutions may also display poor performance. Theoretically, this deterioration may be attributed to the accumulated errors introduced by the imperfect model physics or biases in the initial and boundary conditions, which can be exaggerated by the chaotic nature of NWP systems. According to the above analysis, C6 yields the best agreement with the ground truth data among the horizontal resolution experiments.

[Figure 6]

5.4 Searching for the Likely Ideal Spin-up Time

To limit the effects of the chaotic nature of NWP on the model simulations and extend the lead time, the scenario in which the spin-up time used in WRF is varied is placed at the end of the experimental design, after the possible errors introduced by inappropriate domain configuration options have been reduced. In S4, C6 is adopted as the starting experiment (OS3). Unlike the previous scenarios, the ranks of the spin-up time experiments, as sorted by the metrics, are nearly the same across the different time periods. Hence, **Fig. 7** presents only the modelling skill of the spin-up time experiments over the time period of 18 h. The model performance of WRF in simulating heavy rainfall clearly varies with the spin-up time. For most of the metrics, an obvious diurnal tendency is found from 0 h to 60 h, followed by a short-term decrease until 72 h; random fluctuations occur after 72 h. Before 72 h, the variations in the rainfall and PW metrics are almost consistent; thus, the good fits of the simulations produced by the model runs with longer spin-up times are also physically reasonable within this period. The discrepancies among these experiments may be due to differences in the initial conditions (e.g., the water vapour amounts and the times of day when the simulations begin).

[Figure 7]

From TP', it is found that all of the spin-up time experiments underestimate the total rainfall amount during the heavy rainfall event. Of all of the rainfall-related metrics, POD' is found to display the least sensitivity to the spin-up time; however, it displays similar variations over time as $PMAX'$, R', and $RMSE'$ before 72 h, with the highest values shown in the experiment with a spin-up time of 48 h (C11). Positive biases are detected in $PMAX'$ in C9 (which is run 24 hours ahead) and C11, in which the largest positive biases are detected in the simulated amount of water vapour across the analysed periods and earlier (during the initialization period). This result agrees with intuition because the atmospheric water vapour content determines the maximum possible rainfall amount. C12, which includes a spin-up time of 60 h, is ranked third in terms of $PMAX'$, whereas it displays better performance than C9 and C11 in terms of TP', WR', and $WRMSE'$. As seen in

Fig. 8, C9, C11 and C12 also rank in the top three, based on the values of the rainfall-related metrics on the scale of D02. However, larger departures from the forcing PW fields are seen in C9 and C11 than in C12. The difference is that C12 shows the best agreement with the ground truth data in terms of both the rainfall- and PW-related fields. Overall, C12 is regarded as the experiment that best reproduces the Beijing SDHR event with the optimal set of domain configuration options and the longest spin-up time.

[Figure 8]

6 Discussion

The results reveal that the initial experiment with the most commonly employed WRF domain settings does not yield the best performance in reproducing the temporal and spatial characteristics of SDHR on the convective scale. In S1, the assigned domain size of C0 is not sufficiently broad to allow the model physics to fully develop local small-scale features, resulting in obvious reductions in modelling skill as the evaluated time duration increases from 12 h to 24 h. Further refinement of the grid spacing of C0 in S2 and S3 is shown to enable more explicit resolution of convective processes, leading to more accurate rainfall simulations. The comparison made in S4 suggests that the proper spin-up time is determined by both the time needed for model initialization and the accuracy of the initial conditions fed into the model run. Moreover, experiments with excessively large domains, excessively high spatial resolutions, or excessively long spin-up times also yield poor performance in rainfall simulations. Therefore, the reasonableness of these WRF settings should be checked before the model is utilized in regional NWP systems for flood forecasting or as a reference for the design of flood mitigation strategies.

In addition to exploring whether the recommended WRF domain configuration options and spin-up time are optimal for application in SDHR-prone urban areas, the performance of the model is quantified, and its total improvement is evaluated by comparing the values of the verification metrics yielded by the experiments. **Table 3** compares the values of the verification metrics obtained for the optimal experiments in each scenario with the values obtained for the initial experiment. Here, the 18 h time duration is selected for evaluation because it covers most of the heavy rainfall event, and the metrics calculated over this period display a greater range and thus greater ability in identifying the simulation with the best performance. One exception is the domain size scenario, in which C0 presents the most obvious reduction in performance during the last stage of the rainfall event (24 h). Therefore, the improvement in C1 relative to C0 is mainly represented by R' and POD' across D03 over the 18-h time period. The improvement produced by refining the vertical resolution is indicated by all of the rainfall-related metrics but is accompanied by a decrease in $WRMSE'$ that stems in part from the reduction in kinetic energy, which promotes rainfall. C6 yields higher values of POD', $RMSE'$, R', and $PMAX'$ when compared with C3, indicating that appropriate increases in the horizontal resolution can increase the accuracy of rainfall simulations. The largest differences in the metrics between C6 and C12 occurs for $PMAX'$, which may relate to the different initial weather conditions at the different starting times of the model runs.

[Table 3]

Overall, although the magnitudes of the increases in the rainfall metrics differ, they all reflect an increase in model skill after the re-evaluation process has been conducted. Specifically, R' increases from 0.226 in C0 to 0.67 in C12; $RMSE'$ increases from 0.098 to 0.402; and $PMAX'$ increases from 0.44 to 0.883. As the complete assessment is based on objective verification metrics and checked by subjective verification methods, it can be concluded that the domain configuration options and the spin-up time have significant effects on regional simulations of SDHR. Therefore, re-evaluating the values of those settings used in high-resolution regional studies is certainly worthwhile, and the accuracy of predictions of heavy rain clearly benefit from these analyses. For the evaluated metrics, evaluations based on a single type of metric or a single time period may clearly result in partially accurate conclusions. The use of datasets from multiple sources in verification can help increase the comprehensiveness of the analyses, such as the use of $WRMSE'$ and WR' in this study. The use of different time periods helps to better determine physically reasonable optimal configurations, such as the selection of the proper domain size. In addition, the verification results may also depend on the fields and temporal-spatial scales of interest. To further understand the effects of WRF model configuration options on regional simulations of sub-daily heavy rainfall, more objective verification metrics for SDHR should be developed, and more case studies of SDHR events are also needed. Given that the uncertainties in the regional NWP studies result mainly from the inaccurate boundary conditions associated with grid nesting techniques, methods that can serve as alternate schemes to reduce these uncertainties are also worth studying. Examples include the mesh transitions approach used on irregular grids. In addition, more accurate simulations are expected when the model is driven with forcing data with higher temporal or spatial resolutions than those of the ERA-Interim reanalysis because the uncertainties and errors introduced by the input data are then further reduced.

7 Conclusions

In this study, a comparative test is designed to evaluate the effects of WRF domain configuration options and the spin-up time on simulations of the precipitation during the SDHR event that occurred on July 21st, 2012 in Beijing, China. Three nested domains are established; D01 is the largest, has the coarsest resolution, and covers the leading synoptic features, and D03 is the smallest and covers the area of interest, Beijing. The initial conditions of the three domains are provided by the

ERA-Interim reanalysis and regional geographical datasets. For the LBCs, D01 is forced by the ERA-Interim reanalysis, whereas D02 is forced by D01, and D03 is forced by D02. The reference ground truth data used for verification is 3-hourly 0.05 gridded rainfall observations and the coarser-scale ERA-Interim reanalysis. Five rainfall-related error metrics and two PW-related indices that monitor the departure of the model simulations from the driving fields are calculated at the convective-resolving scale over different sub-daily time spans. These metrics are then checked and considered together as part of a subjective verification process that is intended to pinpoint the likely best combination of the domain configuration options and spin-up time and to help quantify the possible improvements in the model performance of WRF in reproducing severe SDHR events after carrying out the entire re-evaluation process.

Precipitation simulations are sensitive to changes in domain size, vertical resolution, horizontal resolution and spin-up time. Of all of the configurations, the most obvious variations are found when adjusting the domain size and the spin-up time. This analysis shows that domains that cover only the area of interest may be insufficiently broad to permit full development of small-scale features, resulting in poor performance in capturing the spatial pattern of heavy rainfall, especially in the early stages of rainfall events. Despite the dominant role of chaotic processes, it is still possible that model runs with longer spin-up times may result in better rainfall simulations, given favourable initial weather conditions. The effects of the vertical and horizontal resolutions are smaller, but the accuracy of the rainfall amount and the correct hits exhibit evident increases in runs with slightly higher spatial resolutions. A comparison of C12, which uses the evaluated optimum configurations, and C0, which uses the recommended settings, shows that the metrics clearly increase. Specifically; R' increases from 0.226 to 0.67; $PMAX'$ rises from 0.44 to 0.883; and the cumulative spatial error decreases by 33.65 %. Thus, substantial benefits may result from re-evaluating the WRF domain configuration options and spin-up times used in regional studies of SDHR.

Given the intensification of SDHR and the increased risks posed by SDHR-induced hazards, the demands of the operational flood management community for more accurate rainfall predictions with longer lead times, especially over highly affected areas with very short hydrologic response times, are increasing. One method that has now been proven to be effective is to dynamically downscale freely available global NWP products to areas of interest using high-resolution regional NWP models (e.g., WRF). Therefore, the uncertainties associated with the downscaling process, such as errors in boundary conditions and the issues associated with grid nesting, should be carefully evaluated to ensure that the rainfall simulations produced are both statistically accurate and physically reasonable before they are employed in flood forecasting systems. This study illustrates the importance of re-evaluating the domain configuration options and spin-up times used in WRF in improving regional rainfall simulations. Comparisons of the metrics indicate that evaluations based on just one category of metrics or values of metrics calculated over only one time period (e.g., 24 h) do not result in comprehensive comparisons and may lead to partially accurate conclusions. The use of PW fields calculated against reanalysis output is verified to be helpful in determining the optimal set of model configurations when analyses of rainfall-related metrics do not yield uniform conclusions. In addition, evaluations conducted over larger-scale domains are demonstrated to have utility in establishing the reasonableness of the evaluated results. Overall, the evaluation process is partly subjective. To simplify the assessment process, verification methods that can replace this subjective verification procedure should be developed. More regional case studies are also needed to further investigate the effects of configuration options in simulations of regional SDHR and to explore methods of reducing the uncertainties in regional NWP modelling associated with the scale-variation procedures. In addition, the use of more accurate forcing data with higher temporal and spatial resolutions is also expected to reduce the errors in the initial and boundary conditions and could thus be helpful in further improving the accuracy of rainfall simulations and extending the lead times of forecasts.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgement

This study is supported by the key research projects “Sponge city construction and urban flooding/waterlogging disaster in the sub-center of Beijing City” (Z171100002217080), Beijing Municipal Science and Technology Commission, and “Urban flood/waterlogging hazard and disaster reduction strategies in Beijing” (8141003), of Beijing Natural Science Foundation. Support is also received from the Resilient Economy and Society by Integrated Systems modeling (RESIST), Newton Fund via Natural Environment Research Council (NERC) and Economic and Social Research Council (ESRC) (NE/N012143/1), and the National Natural Science Foundation of China (No：4151101234). The China Scholarship Council supports the first author for her academic visit to the University of Bristol, UK.

References

Aligo, E. A., Gallus Jr., W. A., and Segal, M.: On the impact of WRF model vertical grid resolution on Midwest summer rainfall forecasts. *Weather and Forecasting*, **24**, 575-594, 2009.

Bartholmes, J., and Todini, E.: Coupling meteorological and hydrological models for flood forecasting. *Hydrol. Earth Syst. Sci.: Discussions*, **9(4)**, 333-346, 2005.

Berrisford, P., Dee, D. P., Fielding, K., Fuentes, M., Kallberg, P., Kobayashi, S., and Uppala, S. M.: The ERA-Interim Archive. *ERA Report Series*, **1**, 1-16, 2009.

Castelli, F.: Atmosphere modeling and hydrologic-prediction uncertainty. U.S. - Italy Research Workshop on the Hydrometeorology, impacts and management of extreme floods, Perugia, 1995.

Chen, F., and Dudhia, J.: Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Mon. Weather Rev.*, **129(4)**, 569-585, 2001.

Chen, H., Sun, J., Chen, X., and Zhou, W.: CGCM projections of heavy rainfall events in China. *Int. J. Climatol.*, **32(3)**, 441-450, 2012.

Clark, P., Roberts, N., Lean, H., Ballard, S. P., and Charlton-Perez, C.: Convection-permitting models: a step-change in rainfall forecasting. *Meteor. Appl.*, **23(2)**, 165-181, 2016.

Coen, J. L., Cameron, M., Michalakes, J., Patton, E. G., Riggan, P. J., and Yedinak, K. M.: WRF-Fire: coupled weather–wildland fire modeling with the weather research and forecasting model. *J. Appl. Meteor. Climatol.*, **52(1)**, 16-38, 2013.

Cr état, J., Pohl, B., Richard, Y., and Drobinski, P.: Uncertainties in simulating regional climate of Southern Africa: sensitivity to physical parameterizations using WRF. *Clim. Dyn.*, **38(3-4)**, 613-634, 2012.

Cuo, L., Pagano, T. C., and Wang, Q. J.: A review of quantitative precipitation forecasts and their use in short-to medium-range streamflow forecasting. *J. Hydrometeor.*, **12(5)**, 713-728, 2011.

Dee, D. P., and Coauthors: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.*, **137(656)**, 553-597, 2011.

Di, Z. H., and Coauthors: Assessing WRF model parameter sensitivity: A case study with five-day summer precipitation forecasting in the Greater Beijing Area. *Geophys. Res. Lett.*, **42**, 579-587, 2015.

Done, J., Davis, C. A., and Weisman, M.: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) model. *Atmos. Sci. Lett.*, **5(6)**, 110-117, 2004.

En-Tao, Y. U., Hui-Jun, W. A. N. G., and Jian-Qi, S. U. N.: A quick report on a dynamical downscaling simulation over China using the nested model. *Atmos. Oceanic Sci. Lett.*, **3(6)**, 325-329, 2010.

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., and Tarpley, J. D.: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.: Atmos.*, **108(D22)**, 2003.

Fierro, A. O., Rogers, R. F., Marks, F. D., and Nolan, D. S.: The impact of horizontal grid spacing on the microphysical and kinematic structures of strong tropical cyclones simulated with the WRF-ARW model. *Mon. Weather Rev.*, **137(11)**, 3717-3743, 2009.

Foley, A. M., Leahy, P. G., Marvuglia, A., and McKeogh, E. J.: Current methods and advances in forecasting of wind power generation. *Renewable Energy*, **37(1)**, 1-8, 2012.

Gao, Y., Yuan, Y., Wang, H., Schmidt, A. R., Wang, K., and Ye, L.: Examining the effects of urban agglomeration polders on flood events in Qinhuai River basin, China with HEC-HMS model. *Water Sci. Technol.*, **75(9)**, 2130-2138, 2017.

Goswami, P., Shivappa, H., and Goud, S.: Comparative analysis of the role of domain size, horizontal resolution and initial conditions in the simulation of tropical heavy rainfall events. *Meteor. Appl.*, **19(2)**, 170-178, 2012.

Grell, G. A., and Dévényi, D.: A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Lett.*, **29(14)**, 38-31, 2002.

Guo, C., Xiao, H., Yang, H., and Tang, Q.: Observation and modeling analyses of the macro-and microphysical characteristics of a heavy rain storm in Beijing. *Atmospheric Research*, **156**, 125-141, 2015.

Heinzeller, D., Duda, M. G. and Kunstmann, H.: Towards convection-resolving, global atmospheric simulations with the Model for Prediction Across Scales (MPAS) v3. 1: an extreme scaling experiment. *Geosci. Model Dev.*, **9(1)**, 77, 2016.

Hong, S. Y., and Lee, J. W.: Assessment of the WRF model in reproducing a flash-flood heavy rainfall event over Korea. *Atmos. Res.*, **93(4)**, 818-831, 2009.

Hong, S. Y., and Lim, J. O. J.: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42(2)**, 129-151, 2006.

Hong, S. Y., Noh, Y., and Dudhia, J.: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Weather Rev.*, **134(9)**, 2318-2341, 2006.

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.

Kain, J. S., and Coauthors: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Weather and Forecasting*, **23(5)**, 931-952, 2008.

Kleczek, M. A., Steeneveld, G. J., and Holtslag, A. A.: Evaluation of the weather research and forecasting mesoscale model for GABLS3: impact of boundary-layer schemes, boundary conditions and spin-up. *Boundary-layer meteor.*, **152(2)**, 213-243, 2014.

Klemp, J. B.: Advances in the WRF model for convection-resolving forecasting. *Adv. Geosci.*, **7**, 25-29, 2006.

Leduc, M., and Laprise, R.: Regional climate model sensitivity to domain size. *Clim. Dyn.*, **32(6)**, 833-854, 2009.

Liu, J., Bray, M., and Han, D.: Sensitivity of the Weather Research and Forecasting (WRF) model to downscaling ratios and storm types in rainfall simulation. *Hydrol. Processes*, 26(20), 3012-3031, 2012.

Li, J., Chen, Y., Wang, H., Qin, J., Li, J., and Chiao, S.: Extending flood forecasting lead time in a large watershed by coupling WRF QPF with a distributed hydrological model. *Hydrol. Earth Syst. Sci.*, **21(2)**, 1279, 2017.

Luna, T., Castanheira, M., and Rocha, A.: Assessment of WRF-ARW forecasts using warm initializations. 2013. [Available online at http://climetua.fis.ua.pt/publicacoes/APMG_extended_abstract_2013_Luna_et_al.pdf]

Miguez-Macho, G., Stenchikov, G. L., and Robock, A.: Spectral nudging to eliminate the effects of domain position and geometry in regional climate model simulations. *J. Geophys. Res.: Atmos.*, **109(D13)**, 2004.

Mlawer, E. J., and Clough, S. A.: Shortwave and longwave enhancements in the rapid radiative transfer model. *Proceedings of the 7th Atmospheric Radiation Measurement (ARM) Science Team Meeting*, 499-504, 1998.

Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., and Clough, S. A.: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.: Atmos.*, **102(D14)**, 16663-16682, 1997.

Prein, A. F., and Coauthors: A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Rev. Geophys.*, **53(2)**, 323-361, 2015.

Powers, J. G., and Coauthors: The Weather Research and Forecasting (WRF) Model: Overview, System Efforts, and Future Directions. *Bull. Amer. Meteor. Soc.*, 2017.

Roberts, N. M., and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Weather Rev.*, **136(1)**, 78-97, 2008.

Ruiz, J. J., Saulo, C., and Nogués-Paegle, J.: WRF model sensitivity to choice of parameterization over South America: validation against surface variables. *Mon. Weather Rev.*, **138(8)**, 3342-3355, 2010.

Schwartz, C. S., and Coauthors: Next-day convection-allowing WRF model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Weather Rev.*, **137(10)**, 3351-3372, 2009.

Seth, A., and Rojas, M.: Simulation and sensitivity in a nested modeling system for South America. Part I: Reanalyses boundary forcing. *J. Clim.*, **16(15)**, 2437-2453, 2003.

Shih, D. S., Chen, C. H., and Yeh, G. T.: Improving our understanding of flood forecasting using earlier hydro-meteorological intelligence. *J. Hydrol.*, **512**, 470-481, 2014.

Sikder, S., and Hossain, F.: Assessment of the weather research and forecasting model generalized parameterization schemes for advancement of precipitation forecasting in monsoon-driven river basins. *J. Adv. Modeling Earth Syst.*, **8(3)**, 1210-1228, 2016.

Skamarock, W. C., and Coauthors: A description of the advanced research WRF Ver. 30, NCAR Technical Note. NCAR/TN-475, 2008.

Soares, P. M., Cardoso, R. M., Miranda, P. M., de Medeiros, J., Belo-Pereira, M., and Espirito-Santo, F.: WRF high resolution dynamical downscaling of ERA-Interim for Portugal. *Clim. Dyn.*, **39(9-10)**, 2497-2522, 2012.

Sun M. S., Yang L. Q., Yin Q., Niu Z. Y., and Gao L. M.: Analysis of the cause of a torrential rain occurring in Beijing on 21 July 2012(Ⅱ). *Torrential Rain and Disasters (in Chinese)*, **32(3)**, 218-223, 2013.

Swinbank, R. and James Purser, R.: Fibonacci grids: A novel approach to global modeling. *Q. J. R. Meteorol. Soc.*, **132(619)**, 1769-1793, 2006.

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.*, **17(4)**, 563-579, 2017.

Vrac, M., Drobinski, P., Merlo, A., Herrmann, M., Lavaysse, C., Li, L., and Somot, S.: Dynamical and statistical downscaling of the French Mediterranean climate: uncertainty assessment. *Nat. Hazards Earth Syst. Sci.*, **12(9)**, 2769, 2012.

Wang, K., Wang, L., Wei, Y. M., and Ye, M.: Beijing storm of July 21, 2012: observations and reflections. *Nat. hazards*, **67(2)**, 969-974, 2013.

Wang S. L., Kang H. W., Gu X. Q., and Ni Y. Q.: Numerical Simulation of Mesoscale Convective System in the Warm Sector of Beijing '7.21' Severe Rainstorm. *Meteor. Mon.*, **41(5)**, 544-553, 2015.

Warner, T. T., Peterson, R. A., and Treadon, R. E.: A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bull. Amer. Meteor. Soc.*, **78(11)**, 2599, 1997.

Warner, T. T.: Quality assurance in atmospheric modeling. *Bull. Amer. Meteor. Soc.*, **92(12)**, 1601-1610, 2011.

Westra, S., and Coauthors: Future changes to the intensity and frequency of short-duration extreme rainfall. *Rev. Geophys.*, **52(3)**, 522-555, 2014.

Willems, P., and Coauthors: Climate change impact assessment on urban rainfall extremes and urban drainage: methods and shortcomings. *Atmos. Res.*, **103**, 106-118, 2012.

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]

Xu, Z.X., and Chu, Q.: Climatological features and trends of extreme precipitation during 1979–2012 in Beijing, China. *Proceedings of the International Association of Hydrological Sciences*, **369**, 97-102, 2015.

Xu, Z. X., and Zhao, G.: Impact of urbanization on rainfall-runoff processes: case study in the Liangshui River Basin in Beijing, China. *Proceedings of the International Association of Hydrological Sciences*, **373**, 7-12, 2016.

Yu, R., Xu, Y., Zhou, T., and Li, J.: Relation between rainfall duration and diurnal variation in the warm season precipitation over central eastern China. *Geophys. Res. Lett.*, **34(13)**, 2007.

Yu, W., Nakakita, E., Kim, S., and Yamaguchi, K.: Impact Assessment of Uncertainty Propagation of Ensemble NWP Rainfall to Flood Forecasting with Catchment Scale. *Adv. Meteor.*, 2016.

Yucel, I., Onen, A., Yilmaz, K. K., and Gochis, D. J.: Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation and satellite-based rainfall. *J. Hydrol.*, **523**, 49-66, 2015.

Zhou Y. S., Liu L., Zhu K. F., and Li J. T.: Simulation and evolution characteristics of mesoscale systems occurring in Beijing on 21 July 2012. *Chinese J. Atmos. Sci. (in Chinese)*, **38 (5)**, 885-896, 2014.

Figure captions

Figure 1: Relative location of the study area.

Figure 2: Initial wind field and geopotential height field at 12 pm on 20 July 2012 over the Northeastern Hemisphere obtained from the ERA-Interim reanalysis.

Figure 3: Spatial values of the verification metrics for the WRF domain size experiments, calculated over different temporal durations and over domain three.

Figure 4: Spatial distribution of 6-h accumulated precipitation for the domain size experiments within domain two of Case 0 during the heavy rainfall event beginning at 12 am on July 21, 2012.

Figure 5: As in Fig. 3, but for the experiments in scenario two with different vertical resolutions.

Figure 6: As in Fig. 3, but for the experiments in scenario three with different horizontal resolutions.

Figure 7: Spatial values of the verification metrics for the WRF spin-up experiments, calculated over 18-h periods and over domain three.

Figure 8: As in Fig. 7, but the metrics are calculated over 18-h periods and over domain two in Case 6.

Table captions

Table 1: Categories of experiments with different domain sizes, vertical resolutions, horizontal resolutions and spin-up times.

Table 2: Correlations between the original and rescaled objective verification metrics.

Table 3: Comparison of the values of the error metrics in the initial experiment and the optimum experiments identified for each scenario.

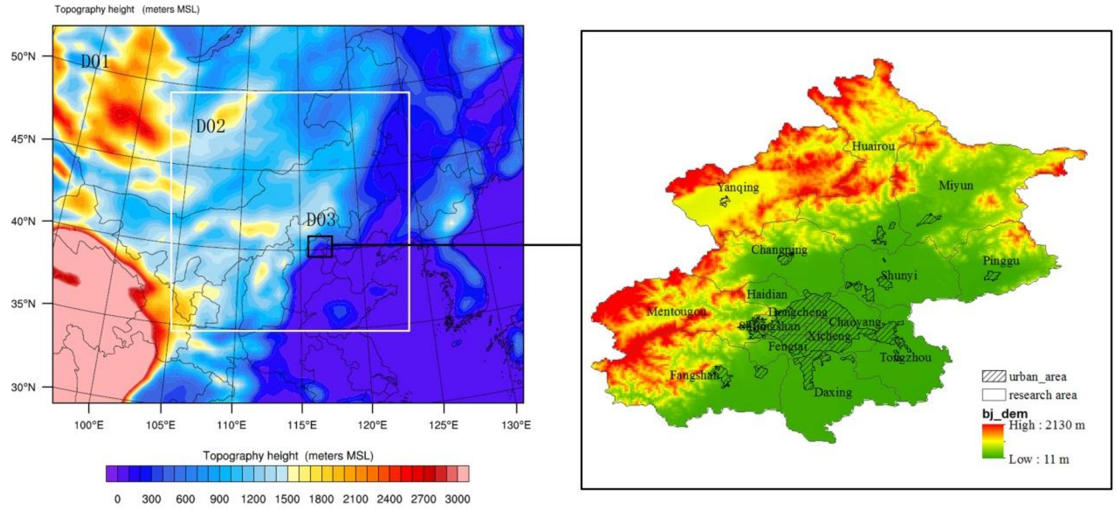

Figure 1: **Relative location of the study area. The left panel shows the three nested domains adopted in most of the experiments, of which domain three (D03) covers the entire Beijing area; the right panel depicts the geographic features of the Beijing area.**

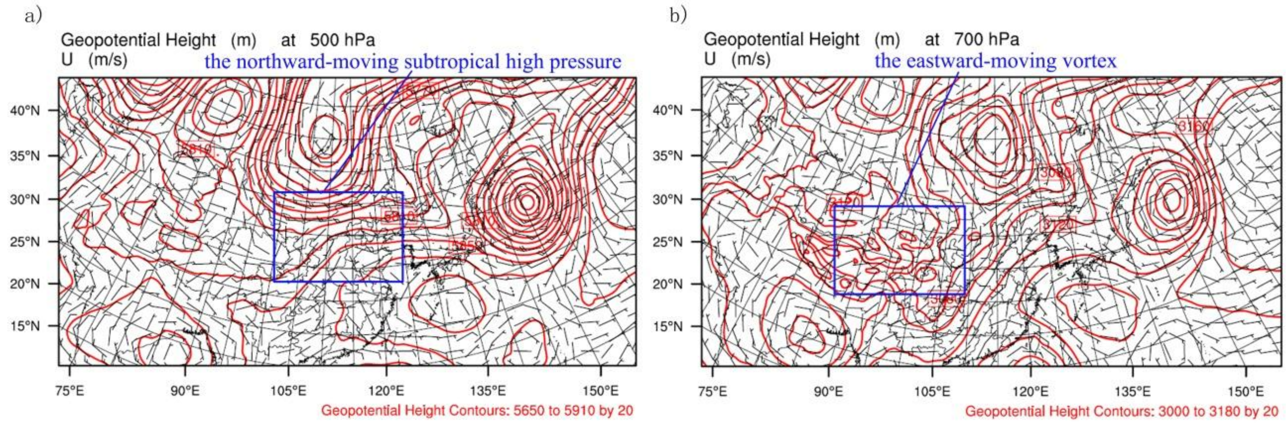
a)
Geopotential Height (m) at 500 hPa
U (m/s)
the northward-moving subtropical high pressure
40°N
35°N
30°N
25°N
20°N
15°N
75°E
90°E
105°E
120°E
135°E
150°E
Geopotential Height Contours: 5650 to 5910 by 20
b)
Geopotential Height (m) at 700 hPa
U (m/s)
the eastward-moving vortex
40°N
35°N
30°N
25°N
20°N
15°N
75°E
90°E
105°E
120°E
135°E
150°E
Geopotential Height Contours: 3000 to 3180 by 20

Figure 2: Initial wind field and geopotential height field at 12 pm on 20 July 2012 over the Northeastern Hemisphere obtained from the ERA-Interim reanalysis. (a) The fields at 500 hPa; (b) the fields at 700 hPa.

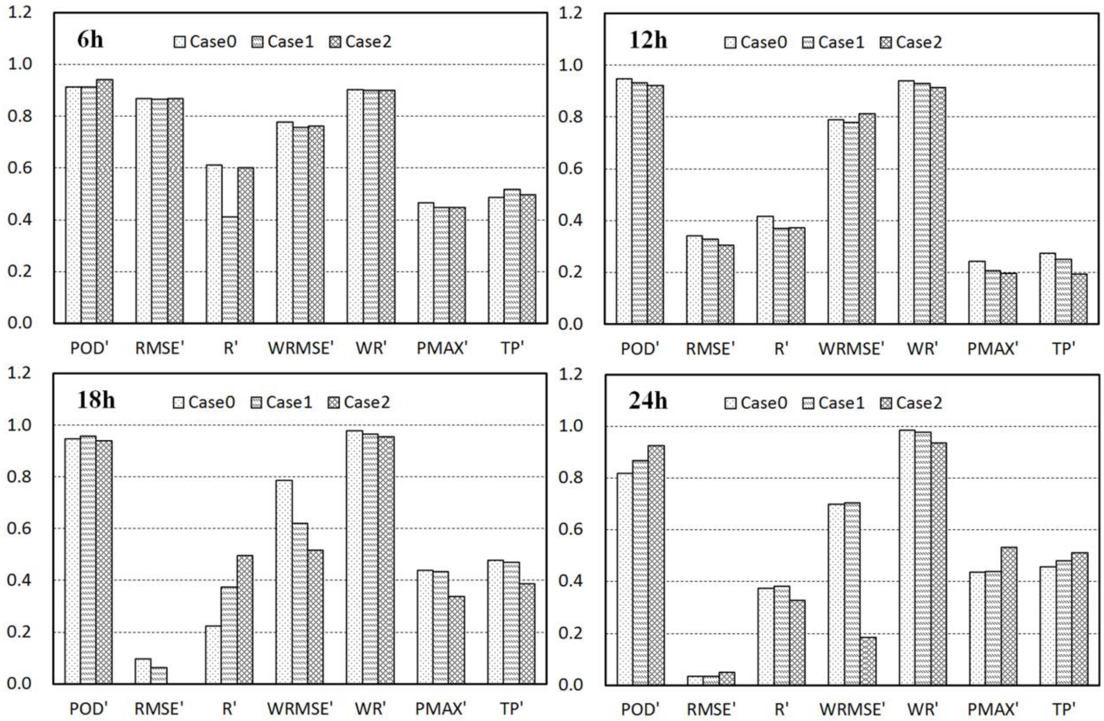

Figure 3: Spatial values of the verification metrics for the WRF domain size experiments, calculated over different temporal durations and over domain three. Case 0 incorporates the smallest domain, which covers north-central China; Case 1 incorporates a domain of intermediate size that covers northern China and part of Mongolia; and Case 2 incorporates the largest domain, which covers the Northeastern Hemisphere. The metrics are calculated over time periods of 6 h, 12 h, 18 h, and 24 h that begin at 12 am on 21 July 2012.

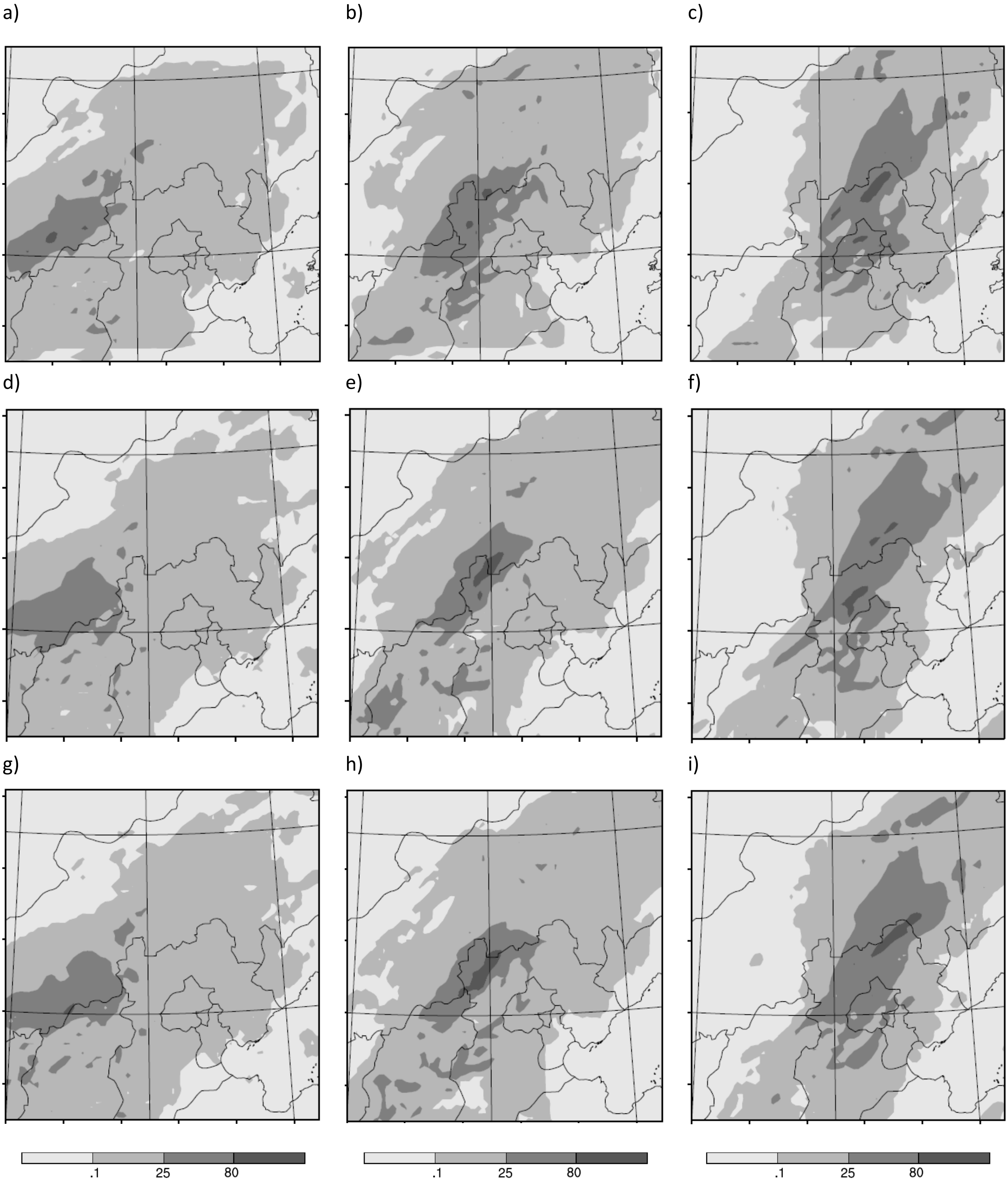

Figure 4: Spatial distribution of 6-h accumulated precipitation for the domain size experiments within domain two of Case 0 during the heavy rainfall event beginning at 12 am on July 21, 2012. (a) Precipitation in Case 0 (with the smallest domain size) during the first 6-h period; (b) precipitation in Case 0 during the second 6-h period; (c) precipitation in Case 0 during the third 6-h period; (d) precipitation in Case 1 (with the medium-sized domain) during the first 6-h period; (e) precipitation in Case 1 during the second 6-h period; (f) precipitation in Case 1 during the third 6-h period; g) precipitation in Case 2 (with the largest domain) during the first 6-h period; h) precipitation in Case 2 during the second 6-h period; and i) precipitation in Case 2 during the third 6-h period.

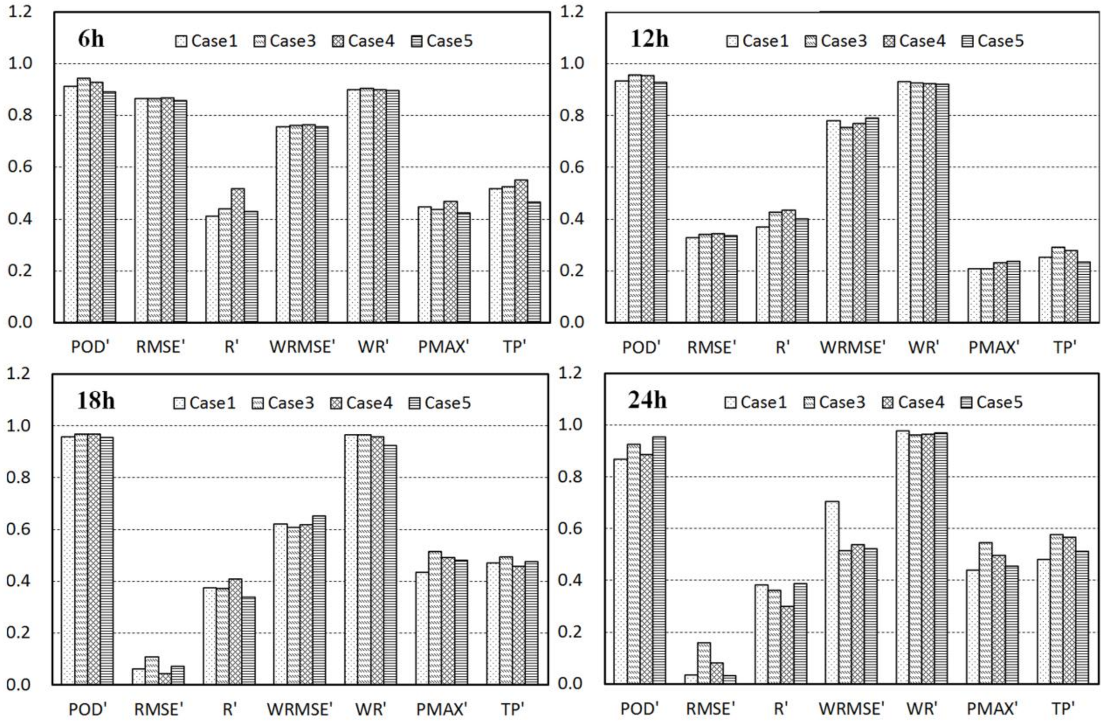

Figure 5: As in Fig. 3, but for the experiments in scenario two with different vertical resolutions. Case 1 is forced by the ERA-Interim pressure-level data with 29 vertical levels; Cases 3 and 4 are forced by the same data but include double and triple the number of vertical levels, respectively; Case 5 is forced by the ERA-Interim model-level data with 38 vertical levels.

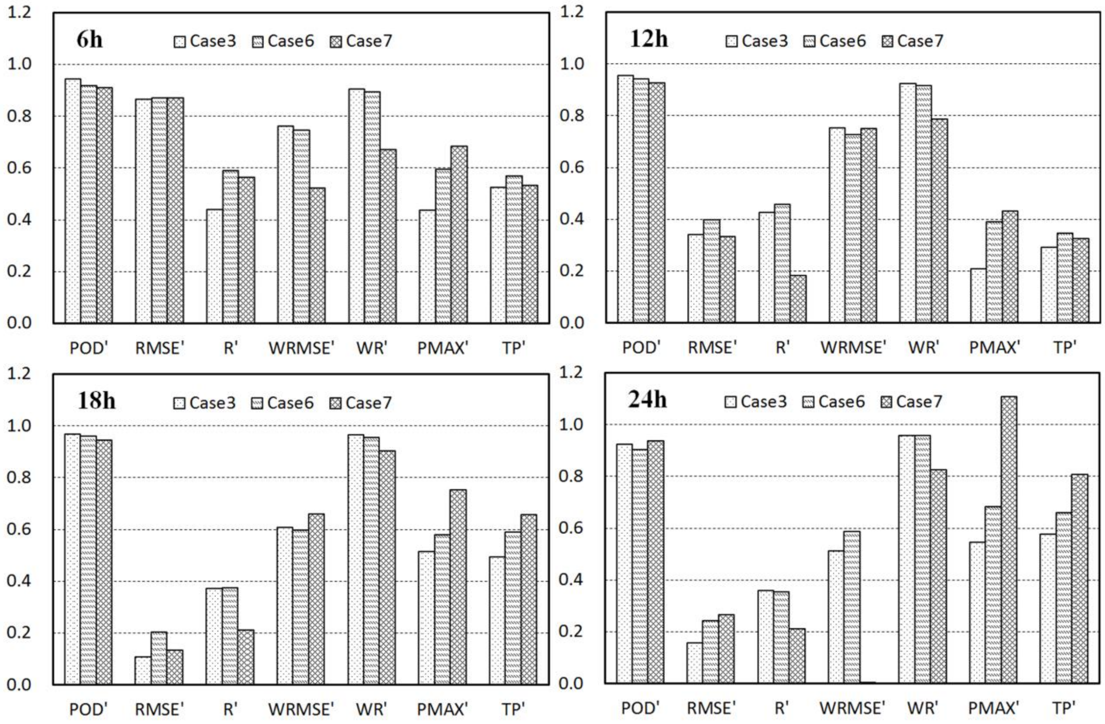

Figure 6: As in Fig. 3, but for the experiments in scenario three with different horizontal resolutions. Case 3 has an initial downscaling ratio of 1:3:3 with horizontal grid spacings of 40.5 km, 13.5 km and 4.5 km, whereas Cases 6 and 7 have the same large horizontal grid spacing with nesting ratios of 1:5:5 and 1:7:7, respectively. The innermost grid spacing is 1.62 km in Case 6 and 0.827 km in Case 7.

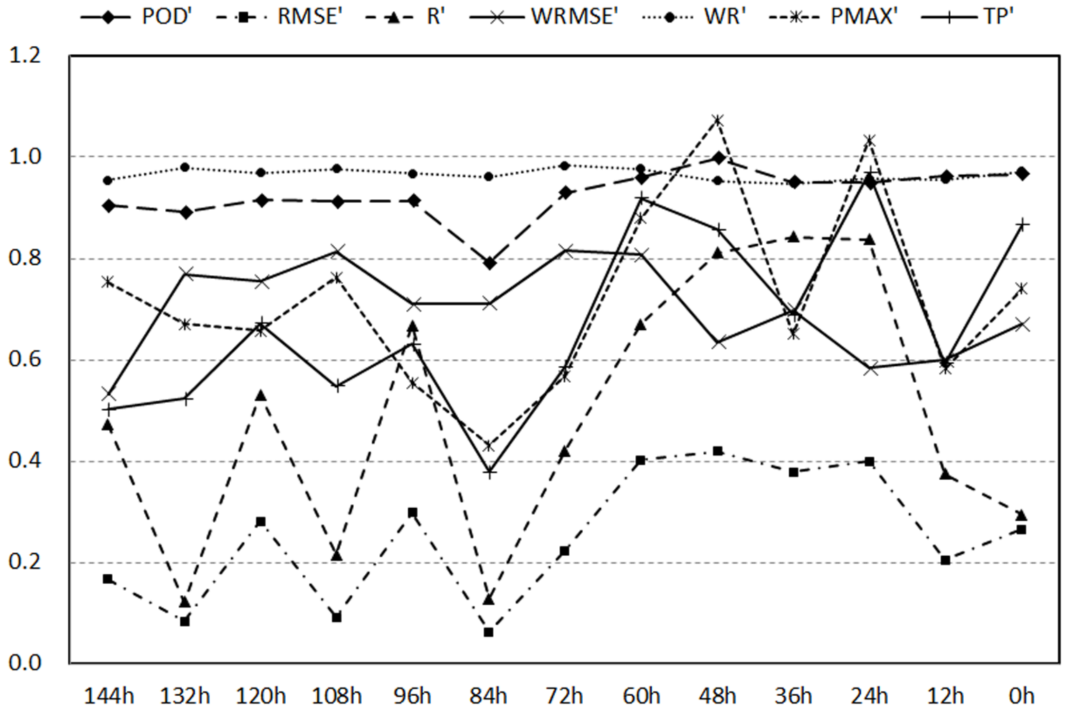

Figure 7: Spatial values of the verification metrics for the WRF spin-up experiments, calculated over 18-h periods and over domain three. Case 6 employs an initial spin-up time of 12 h; Case 8 employs a spin-up time of 0 h; and from Case 9 to Case 19, the spin-up time is increased from 24 h to 144 h by every twelve hours.

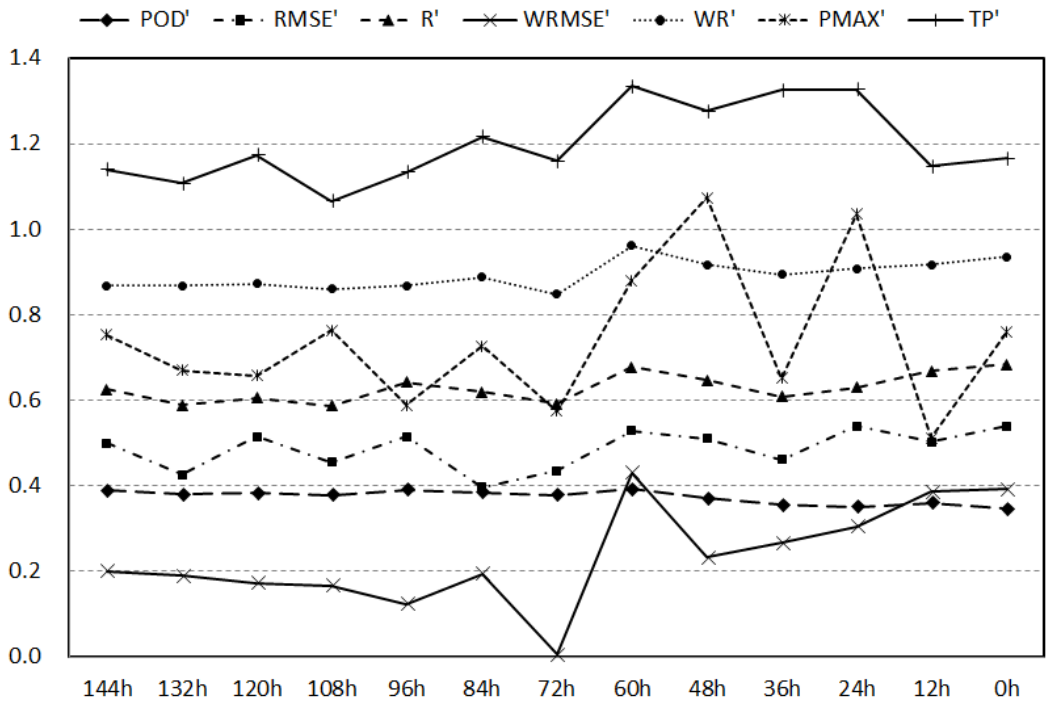

Figure 8: As in Fig. 7, but the metrics are calculated over 18-h periods and over domain two in Case 6.

Table 1: Categories of experiments with different domain sizes, vertical resolutions, horizontal resolutions and spin-up times.

Scenario	Experiment Number	Domain Size (grid numbers)	Vertical Levels	Horizontal Resolution (nesting ratio)	Spin-up Time
Domain Size (S1)	Case 0 (C0)	D01 40×40 D02 72×72	29 (pressure level)	D01 40.5km; D02 13.5km; D03 4.5km 1:3:3	12 h
	Case 1 (C1)	D01 80×64 D02 120×120	As C0	As C0	As C0
	Case 2 (C2)	D01 160×128 D02 240×192	As C0	As C0	As C0
Vertical Resolution (S2)	Optimal Case in S1 (OS1)	As OS1	29	As C0	As C0
	Case 3 (C3)	As OS1	57	As C0	As C0
	Case 4 (C4)	As OS1	85	As C0	As C0
	Case 5 (C5)	As OS1	38 (model level)	As C0	As C0
Horizontal Resolution (S3)	Optimal Case in S2 (OS2)	As OS1	As OS2	1:3:3	As C0
	Case 6 (C6)	As OS1	As OS2	D01 40.5km; D02 8.1km; D03 1.62km 1:5:5	As C0
	Case 7 (C7)	As OS1	As OS2	D01 40.5km; D02 5.785km; D03 0.826km 1:7:7	As C0
Spin-up Time (S4)	Optimal Case in S3 (OS3)	As OS1	As OS2	As OS3	12 h
	Case 8 (C8)	As OS1	As OS2	As OS3	0 h
	Case 9-Case 19 (C9 - C19)	As OS1	As OS2	As OS3	24 h – 144 h per 12 h

Table 2: Correlations between the original and rescaled objective verification metrics.

Original Metrics	Representative Meaning	Rescaled Metrics	Threshold Value
POD	Probability of Detection	$POD' = POD$	N/A
$RMSE$	Root Mean Squared Error	$RMSE' = 1 - RMSE/RMSE_{max}$	+ 62.5 max
R	Pearson Correlation Coefficients	$R' = R$	N/A
$WRMSE$	RMSE of the Precipitable Water	$WRMSE' = 1 - WRMSE/WRMSE_{max}$	+ 8.3 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

Table 3: Comparison of the values of the error metrics in the initial experiment and the optimum experiments identified for each scenario.

Experiment Number	POD'	$RMSE'$	R'	$WRMSE'$	WR'	$PMAX'$	TP'
Case 0 (C0)	0.950	0.098	0.226	0.789	0.980	0.440	0.478
Case 1 (C1)	0.960	0.064	0.376	0.622	0.967	0.436	0.471
Case 3 (C3)	0.969	0.110	0.373	0.610	0.967	0.515	0.496
Case 6 (C6)	0.963	0.205	0.375	0.600	0.956	0.582	0.592
Case 12 (C12)	0.959	0.402	0.670	0.807	0.977	0.883	0.920