Dear Editor and Reviewers,

Thank you for the comments to help improve the quality of the paper. We have revised the manuscript to address your comments. A detailed response to each comment is provided in this file with comments from referees in black, author's response in red, and author's changes in manuscript in blue.

Report #1 by anonymous referee #2:

The authors presented their efforts in examining the spatial and temporal distributions of air pollutants over China, with a full year simulation and local observations. Although the WRF/CMAQ model has been applied in many studies over East Asia recently, this full year simulation helps to probe into the modeling discrepancy and stability. The manuscript generally introduced the performance but lacked sufficient discussion or emphasis on the findings as a research paper. More explicit discussion and demonstration of the findings are necessary to illustrate the contribution of this manuscript. Therefore, I recommend the manuscript to be accepted with minor revision if the following questions/issues are being addressed in details by the authors.

Responses: We thank the reviewer for providing important comments that help improve the quality of manuscript. However, the comments are same as the quick review process before its publication on ACPD. We have responded to the comments and here we made additional changes together with other reviewers' comments.

(1) General comment: The manuscript spent a lot efforts describing the results of model evaluation, but lack of detailed discussion about the factors that are/may responsible for the good/bad model performance. For example, section 3.1 describes the statistics summarized in Table 1, but it would be much more informative for the research community if the authors can discuss about why T2 is overestimated in winter. Is it because of physical options in WRF or bias induced by FNL? Section 3.2 describes the performance of O3 and PM2.5 predictions at different sub-regions in China, but there some important issues remain unknown. The research domain is divided into several sub-regions, thus the authors are expected to point out why the performance varies among sub-regions, is it correlated with bias from meteorology, emission, or chem/physical mechanisms applied? These are the real important findings this paper can contribute to the research community rather than demonstrating that this specific model application meets the benchmark.

Responses: We thank the reviewer for pointing out these important issues. We responded specific comments such as meteorology performance, sub-regions etc. below. Here we want to have a general response.

We agree with the reviewer that there are important problems relating simulation of air pollution in China, from meteorology, emissions, and domain setting, to model representations of physical and chemical processes. Even observations have large limitations and China is currently lacking routine long-term measurements of $PM_{2.5}$ chemical components. These issues have to be addressed by contributions from the entire community and it is beyond the scope of study to address them all. Particularly, this paper aims to examine the capability of a state-of-art CMAQ model in reproducing severe air pollution in China by comparing with the recently public available air quality monitoring data for an entire year. By identifying the regions/seasons with good model performance, further studies investigating potential air pollution control strategies and associating human health outcomes with air pollution exposure can be conducted with confidence using the dataset generated from this study. In the meantime, by identifying the regions/seasons with less desirable performance, we can point out which areas need more work in future. This is why we are rigorously comparing with the model application benchmarks suggested by the US EPA. Thus, as the first study to do so, this paper has its own merits.

In fact, we indeed found different model performances for different seasons and sub-regions and we pointed out the possible reasons that can be implied by this study in the discussion part. However, it is beyond the scope of this study to answer very specific questions such as whether physical options of WRF or bias associated with FNL caused the T2 overestimation in winter, or quantifying the uncertainties in the predicted concentrations due to meteorology, emission, or chemical/physical mechanisms. These specific questions are research topics that need to be addressed in separate studies. We are currently evaluating different WRF model configurations to improve our meteorology model performance. Also, we are investigating CMAQ model performance based on different emission inventories. Addressing these questions will be published in follow-up papers and are beyond the scope of this paper.

Changes in manuscript: no changes were made for this comment.

(2) General comment: Some of the discussions in section 4 lack detailed explanation thus are not persuasive to support the conclusions. For example, line#440-448 states the bias from met prediction may affect the chemistry of CMAQ, but line#451 attributes the bias of SO2 and NO2 to anthropogenic emissions. Although CO is a good indicator for estimating bias induced by anthropogenic emission, SO2 and NO2 are often affected by other factors, such as meteorology, biomass burning emission as well. How does the overestimation of T2 affect the chemistry of SO2 and NO2, is it pushing the bias induced by anthropogenic emission towards or away from the benchmarks? Are there any evidences, such as satellite products/surface monitors/field studies, that can help to identify the location and intensity of emission bias? There are some published modeling studies using nested domains with different grid resolutions, are their findings support the statements in line#478-486? As the manuscript mainly focused on model performance, it is necessary to probe deeper into at least some of these issues, to investigate the contributions from different sources of uncertainties, such as meteorology, anthropogenic emission, biomass burning/biogenic/dust emission.

Responses: Model bias is affected by the combination of bias in meteorology prediction, emissions, and air quality model algorithms. In line#451 in the original manuscript, we attribute the bias of CO, NO2 and SO2 in NW to anthropogenic emissions, not to meteorology bias because we compared the biases of air quality and meteorology predictions in NW to other regions. When checking the model performance among difference regions, we noticed larger biases in predicted air pollutant concentrations in NW than other regions, but meteorology performance among different regions was similar. In addition, unlike the NCP, YRD and PRD regions where a significant and continuous effort was undertaken to improve the accuracy of the emission estimations, the northwestern regions have not been the focus of previous investigations and it is likely that a lot of the smaller emission sources were not reported when developing the emission inventories. Biases in predicted surface temperature could directly affect the loss rate of SO2 and NO2 through reactions with OH. However, based on the current understanding of the temperature dependence of the OH reactions, one-degree Celsius bias in ambient temperature only changes the loss of SO2 and NO2 through OH reactions by less than 1%. By ruling out other possible explanations for the significant negative biases, we believe that bias in emissions is more likely the reason for the substantial under prediction of SO2 and NO2 in the NW.

A few studies have examined the major emission sources, such as power plants, using satellite observations (Zhang et al., 2012; Liu et al., 2016) and combined with chemical transport model studies (Wang et al., 2012), building confidence in the emissions from the major point sources. However, emissions from area sources, such as residential sources, are difficult to characterize and therefore have much larger uncertainties. Recent studies indicate that residential emissions are an important but generally unrecognized source of ambient air pollution in China (Hu et al., 2005l Liu et al., 2016).

References:

Hu, J., Wu, L., Zheng, B., Zhang, Q., He, K., Chang, Q., Li, X., Yang, F., Ying, Q., Zhang, H., 2015, Source contributions and regional transport of primary particulate matter in China, Environmental Pollution, 207: 31-42

Liu, J., Mauzerall, D.L., Chen, Q., Zhang, Q., Song, Y., Peng, W., Klimont, Z., Qiu, X., Zhang, S., Hu, M., Smith, K.R., Zhu, T., 2016, Air pollutant emissions from Chinese households: A major and underappreciated ambient pollution source, Proceedings of the National Academy of Sciences of the United States of America, 113(28):7756-7761

Liu, F., Beirle, S., Zhang, Q., Dorner, S., He, K., Wagner, T., 2016, NOx lifetimes and emissions of cities and power plants in polluted background estimated by satellite observations, Atmospheric Chemistry and Physics, 16, 5283-5298.

Wang, S.W., Zhang, Q., Streets, D.G., He, K., Martin, R.V., Lamsal, L.N., Chen, D., Lei, Y., Lu, Z., 2012, Growth in NOx emissions from power plants in China: bottom-up estimates and satellite observations, Atmospheric Chemistry and Physics, 12, 4429-4447.

Zhang, Q., Geng, G., Wang, S., Richter, A., He, K., 2012, Satellite remote sensing of changes in NOx emissions over China during 1996-2010, Chinese Science Bulletin, 57(22):2857-2864.

Regarding the suggestion on nested simulations, we added the following references to support that a finer resolution will help improve model predictions for urban locations (lines 513-515 in the revised manuscript):

Fountoukis, C., Koraj, D.H., Denier van der Gon, H.A.C., Charalampidis, P.E., Pilinis, C., Pandis, S.N., 2013, Impact of grid resolution on the predicted fine PM by a regional 3-D chemical transport model, Atmospheric Environment, 68, 24-32.

Gan, C.M., Hogrefe, C., Mathur, R., Pleim, J., Xing, J., Wong, D., Gilliam, R., Pouliot, G., Wei, C., 2016, Assessment of the effects of horizontal grid resolution on long-term air quality trends using coupled WRF-CMAQ simulations, Atmospheric Environment, 132, 207-216.

Stroud, C.A., et al., 2011, Impact of model grid spacing on regional- and urban- scale air quality predictions of organic aerosol, Atmospheric Chemistry and Physics, 11, 3107-3118.

Regarding the suggestion that we took a deeper look at uncertainties caused by meteorology, anthropogenic emissions, biomass burning, biogenic, and dust emissions, we have included discussions in various relevant parts of the manuscript, such as the discussion of the underestimation of emissions in the NW region. The role of natural emission sources, such as biogenic and dust emissions on model predictions needs further investigation. As inputs to these emission models and as well as the underlying parameterizations (e.g. the impact of soil moisture on dust and biogenic VOC emissions) have large uncertainties, a more reasonable approach is to study the modeling of these emissions and their impact on air quality in China in separate manuscripts rather than including them all into a single manuscript. Results of our study here will provide the readers in various areas in the community an opportunity to think and identify research areas within their expertise for further detailed investigation, such as the excellent suggestions recommended by the reviewer.

Changes in manuscript: In lines 513-515 in the revised manuscript, we added three more references to support that a finer resolution will help improve model predictions for urban locations.

(3) Minor comment: Please briefly describe why sub-regions are defined to evaluation model performance.
Responses: We added following description in the revised manuscript:
"Concentrations of pollutants in different regions of China exhibit large variations due to diverse climates, topography, and emission sources. Aiming to identify the model strength and weakness in different regions of China, model performance was evaluated separately for different regions."
Changes in manuscript: In lines 225-228 in the revised manuscript, we added above description.

(4) Minor comment: Table3, in Mar OBS and PRE for PM2.5 is 81.68 and 66.12 respectively, while the MNB is only 0.04, please double check this statistics as it indicates large bias but strong correlation between observation and prediction.
Responses: We checked the calculation and the numbers are correct. The small MNB but large MNE is due to model under prediction of very high PM2.5 concentrations and over prediction of very low PM2.5 concentrations, which compensates each other in the MNB calculation and leads to a small MNB value.
Changes in manuscript: No changes were needed for this comment.

(5) Minor comment: Please briefly describe how the MEIC emission is temporally/spatially allocated as CMAQ-ready inputs, since these factors have large impact on regional model performance.
Responses: The MEIC emissions are already spatially allocated into 0.25x0.25 degree grid cells before they were given to us. We re-gridded the emissions to our model domain, which uses 36 km horizontal resolution, using the Spatial Allocator program provided by the US EPA. Monthly MEIC emissions were obtained and temporal allocation of MEIC emissions were also conducted based on weekly and diurnal profiles from MEIC developers. We have cited the papers describing these methodsin the revised manuscript.
Changes in manuscript: We have cited the papers describing these methods in lines 175-176 in the revised manuscript.

(6) Minor comment: line#115, "interaction of" should be "interaction between A and B"
Responses: Corrected it.
Changes in manuscript: We corrected it in line 122 in the revised manuscript.

(7) Minor comment: How are the initial/boundary conditions generated for CMAQ ?

Responses: Initial and boundary conditions were based on the default vertical distributions of concentrations that represent clean continental conditions as provided by the CMAQ model. The impact of initial conditions was minimal as the results of the first five days of the simulation were excluded in the analyses.

Changes in manuscript: We added description of the initial/boundary conditions in lines 186-189 in the revised manuscript.

(8) Minor comment: line#175-178. The default inline dust emission module in CMAQ was reported to significantly underestimated the emission of total dust by Fu et al. (2013) and Dong et al. (2015) due to the double-count of soil moisture effect. But dust mainly dominates the coarse mode aerosol so it may not influence the performance for O3 and PM2.5 which are the focus of this study. Still I would suggest the author take a look at their PM10 results especially in spring, since PM10 is also in the criteria air pollutants.

Responses: The most recent version of the CMAQ dust module appears to have already been using the adjusted critical U* values (see DUST_EMIS.F, around line 842, in CMAQv5.0.1). For example, the U* for barren land soil was changed from 0.28 to 0.23, the recommended adjustment as mentioned in Dong et al. (2015). In general, the U* values used in our simulations agree with the ones used by Dong et al (2015). We also observed that dust emissions are likely under-estimated by our updated dust module. We calculated the PM10 results and added the results in Table 3 and Table 4 as the reviewer suggested. It is clearly that PM10 is significantly underestimated. The discussions of PM10 were added in the revised manuscript.

Changes in manuscript: We added the PM10 results in Table 3 and Table 4. We added the discussions of PM10 in line 235, lines 297-299, and line 305 in the revised manuscript.

(9) Minor comment: line#286, “Figure ?”

Responses: It is “Figure 4”.

Changes in manuscript: We have corrected it in line 306 in the revised manuscript.

(10) Minor comment: line#504: “this is the first study”. Zhao B. et al. (2013, ERL) did a full-year simulation with WRF/CMAQ in China, and there are some studies with MM5/CMAQ in China prior to 2013 too.

Responses: Although a full year air quality simulations have been conducted previously by Zhao et al. (2013, ERL) and other studies (Gao et al., 2014, Atmos. Environ.; Wang et al., 2011, Atmos. Environ.), model performance on temporal and spatial variations of air pollutants were mostly evaluated against available surface observation at a limited number of sites. In addition, the surface observations were mostly based on the MEP’s air pollution index (API) numbers which could be used to calculate the concentrations of the major pollutant of SO2, NO2 or PM10. Therefore it is still true that no studies have reported “the detailed model performance of O3 and PM2.5 for an entire year”. We modified the introduction section to include the above facts and to avoid confusion, we revised the expression to “this study reports…”

Changes in manuscript: We modified the introduction section to include the above facts in lines 103-110 in the revised manuscript, and we revised the expression in line 536 in the revised manuscript.

(11) Minor comment: line#205-206: “WRF model has acceptable”. Table 1 indicates many of the

variables failed to meet the benchmark. The Zhao B. et al. (2013, ERL) shows WRF performance all falls in the benchmark, so I would suggest the authors check their configurations of WRF namelist to either improve the WRF performance or specify the reason for relatively large bias in this study.

Responses: Many thanks for providing the reference. We checked the model configuration used in the study of Zhao B. et al. (2013, ERL) and compared to ours (list in the Table R1). We also noticed that Zhao B. et al. (2013, ERL) evaluated the WRF performance at ~380 stations, but in our study, we evaluated WRF performance at ~1200 stations. For this reason, we cannot directly compare the model performance between the two studies. Our WRF model performance is consistent with some other previous studies. In addition to the studies we compared in the manuscript, we found another study by Wang et al. (2014, ACP) reported comparable WRF performance as ours. We added both Zhao B. et al. (ERL) and Wang et al. (2014, ACP) in the revised manuscript and pointed out the different WRF performance in different studies. However, improving WRF model performance needs extended efforts and is beyond the scope of this paper.

Wang, L.T., Wei, Z., Yang, J., Zhang, Y., Zhang, F.F., Su, J., Meng, C.C., Zhang, Q., 2014, The 2013 severe haze over southern Hebei, China: model evaluation, source apportionment, and policy implications, Atmospheric Chemistry and Physics, 14, 3151-3173.

Table R1. WRF model configurations in this study and in Zhao B. et al (2013, ERL)

Physics	**This Study**	**Zhao B. et al. (2013, ERL)**
Microphysics	New Thompson scheme	WSM 3-class scheme scheme
Long wave radiation	RRTM scheme	RRTM scheme
Shortwave radiation	Goddard shortwave	RRTM shortwave
Land surface	MM5 Land Surface Model	NOAH Land Surface Model
Planetary boundary layer	Yonsei University scheme	Mellor-Yamada-Janjic PBL scheme
Cumulus Parameterization	Grell-Devenyi ensemble scheme	Grell-Devenyi ensemble scheme

Changes in manuscript: We added both Zhao B. et al. (ERL) (line 218) and Wang et al. (2013, ACP) (line 215) and pointed out the different WRF performance in different studies (lines 217-220) in the revised manuscript.