Response RC2 to Referee # 4

Dear reviewer,

Thank you for the comments to help improve the quality of the paper. We have revised the manuscript to address your comments and a detailed response to each comment is provided in this file. The comments are in regular font and the responses are in red.

RC2, Anonymous Referee #4

General Comment

The authors showed an inclusive validation about the ability of CMAQ model to simulate the air pollutants (O3 and PM2.5) in China with using four different EI data in recent year (2013). They used the widely-used statistical indices for the validation and observations which covered wide areas in China. An ensemble method to obtain better prediction of air pollutants in China was proposed which is the main part of this paper. This paper is well within the scope of this journal, however, I noticed several issues in this paper which cannot be passed over to be published. I suggested that the authors should consider the following comments: two major and several specific comments.

Major Comment 1:

My biggest concern is the lack of carefulness in the manuscript. Several typos, mistakes in table and figure, and the insufficient explanations can be found which make the manuscript difficult to read and greatly damage the value of this paper. I pointed out some of those points in the specific comment below, and I strongly suggest that the authors consider those comments and should carefully and thoroughly check the manuscript again before revised submission.

Response: Thanks for pointing out the typos and mistakes. We have checked the manuscript carefully and made correction to the typos and mistakes in the revised manuscript. The changes can be found in the manuscript with changes marked.

Major Comment 2:

The authors set a goal of this paper on proposing a method for using the model simulation to health impact study and so the authors put "for health effect study" in the title. However, it was not clear which part of the manuscript was particularly dedicated for the health effect study. I concerned if the indices of air pollutants used in the manuscript: daily, monthly, and annual means, 1hourly and 8hourly O3, are appropriate for this purpose. I think more sentences is necessary to discuss the validity of those indices to be used for the health impact research, if they want to claim it as, at least, a part of heath effect study.

Response: This study is part of a project to investigate the long-term health impacts of the severe outdoor air pollution in China. This is the first part of the series study aiming to provide more accurate air pollution exposure assessment for health analysis. The predicted air pollution fields then will be used in a number of epidemiology studies. Actually, the first such analysis using the annual ensemble $PM_{2.5}$ predictions to investigate the premature mortality attributable to various sources of $PM_{2.5}$ in China and the responses of premature mortality to the $PM_{2.5}$ reduction objectives in different regions of China was recently accepted for publication in *Environmental Science & Technology* (Hu et al., 2017). A few studies are undergoing to analyze the correlations between air pollutants and certain health outcomes in China using the ensemble predictions of gaseous pollutants, PM mass and compositions.

A few epidemiology groups expressed their interest of using the ensemble predictions of $PM_{2.5}$ and O_3 from this study for short-term health effect studies in China. Therefore, we also evaluated the performance of daily and monthly ensemble predictions for both $PM_{2.5}$ and 1h- and 8h- O_3 in this manuscript so that it can provide a validation for future applications for such dataset.

We added a brief discussion on the current and future applications of our dataset for health effect studies in China at the end of Section 3.3.

Specific Comments:

- Model description: There was no descriptions about the model domain. Figure S1 can be moved from the supplement to the manuscript since the abbreviation for the different regions in China were frequently used in the manuscript.

Response: We moved Figure S1 from the supplement to the manuscript.

- E1-E4: How did you treat the observation from 422 sites? Are these data once averaged out to form the city average for each of 60 cities, and then calculate the statistical indices (MNB, MNE, MFB, MFE)? Please make it clearly described in the manuscript.

Response: Yes, the city averages were firstly calculated by averaging the observations in all the sites located in that city, and then the statistical indices were calculated based on the city averages. We added above information in the revised manuscript.

- L249-251: It is better to briefly describe the reason why different statistical indices are used for O3 and PM2.5.

Response: In air quality modeling studies, it has been common to use MNB and MNE to evaluate the model performance for O_3, and use MFB and MFE to evaluate the model performance for $PM_{2.5}$. And accordingly the MNB and MNE criteria and goals have been set for O_3, and MFB and MFE criteria and goals for $PM_{2.5}$. We added above information in the revised manuscript.

- E6: A brief explanation of the method to minimize the function Q is necessary.

Response: The linear least square solver 'lsqlin' in matlab was used to minimize the function Q. This information was added in the revised manuscript.

- Table 1: Are these statistical indices calculated using annual mean? not clearly described.

Response: The original statistics in Table 1 were calculated using hourly average concentrations. We clearly added this information in the table caption.

- L286-288: The description here is inconsistent with Figure1. Is this sentence correct?

Response: The description here is about the 'overall' performance, i.e., the average indices over the entire modeling period and over the entire regions of China. Figure 1 shows the performance in different months and regions. Therefore, there seems some difference, but we double checked the numbers, they are correct.

- L295: Why were January and February omitted?

Response: The national air quality monitoring network started publishing ambient air quality observations since March 2013. Therefore, no observations were available for January and February in 2013.

- L300-301: I couldn't understand the meaning of this sentence. Are there any typo or mistake?

Response: We corrected the sentence to "O_3 predicted using MEIC, EDGAR, and REAS2 meets the performance criteria in most regions except for the YRD by MEIC and the PRD by EDGAR."

- L302-304: It is difficult to see what this sentence said from in Figure 1.

Response: We modified and expanded the sentence to be clearer: "CO and NO_2 are under-predicted in all regions, with the largest under-predictions in NW and Other. This pattern is similar among the results with all inventories. SO_2 is generally under-predicted in all regions, but over-predicted in the Sichuan Basin (SCB) by all inventories. SO_2 is also over-predicted by EDGAR in the PRD region. SO_2 in Northeast (NE) is substantially under-predicted by MEIC and REAS2. In general, model performance in the more developed regions such as YRD, NCP, and PRD are relatively better, compared to NW and Other regions."

- Figure 2: The explanation to properly see this figure is highly insufficient. What does the x-axis stand for? Is it the absolute concentration of observation or simulation? Furthermore, "goal" and "criteria" in Figure 2 should be explained somewhere in the manuscript. Otherwise the readers cannot take the messages properly from this

figure.
Response: The x-axis shows the observed $PM_{2.5}$ and PM_{10} concentrations. We added definitions of "goal" and "criteria" in the figure caption of Figure 3 in the revised manuscript (Figure 2 in the original manuscript):
"The model performance goals represent the level of accuracy that is considered to be close to the best a model can be expected to achieve, and the model performance criteria represent the level of accuracy that is considered to be acceptable for modeling applications."

- L311: typo?, a period -> comma?
Response: corrected.

- Figure 3: Are these indices (O3-1h, -8h) maximum 1h- or 8h- mean concentration in a day (=daily maximum 1h or 8h-mean O3)? If so, should be more clearly stated.
Response: We added the definitions for O3-1h (daily maximum 1h O3) and O3-8h (daily maximum 8h mean O3) in the figure caption.

- L327-328: I don't think so. There were large differences between SOE and MEIC over the oceanic area east of China.
Response: The O_3 difference between SOE and MEIC is generally less than 1ppb over the oceanic area east of China, indicated by the 'green' color (the color scheme is shown in the bottom of the figure). To be more accurate, we added the "(the difference is generally less than 1ppb)" in the sentence.

- L344: typo?, South Asia -> Southeast Asia
Response: We corrected it to Southeast Asia.

- L354 & L361: What is NCY?
Response: We corrected it to NCP.

- L362 typo?, YRD -> PRD?
Response: We corrected it to PRD.

- Table2: This is too detailed information. It can be moved to supplement.
Response: We moved Table 2 to the supplemental materials as Table S2

- L410-412: Why are the values referred here as the MFB of individual simulation (-0.25 – -0.16) different from those appeared in Table 1 (-0.32 – -0.21)? If the definitions are different for both, it should be clearly written in the manuscript. I really confused here.
Response: Following the discussion of annual average concentration in Table 2, the values in L410-412 refer to the MFB and MFE calculated using the annual averages. The MFB and MFE values in Table 1 were calculated using the hourly averages. We clearly clarify the calculation of the values in the paragraph.

- L412-413 Something wrong with English.
Response: We corrected "and" to "any" in the sentence.

- L413-415: Same as the two comments above, why are the values of MNB of individual simulation (0.06 – 0.19) different from those appeared in Table 5?
Response: Again, the values in L413-415 were calculated using annual averages, while the values in Table 5 were calculated using the daily averages. We clearly clarify the calculation of the values in the paragraph.

- Table 3: The authors showed that the weighting factor of each EI can vary for different averaging time. in general, EDGAR and REAS have large weight for daily and monthly, and the other two Chinese EI were weighted large for annual time scale. I encouraged the authors to discuss more on the interpretation of it.
Response: The weighting factors in different averaging times were determined by the model performance. The model performance in different averaging times was affected by the total emission rates, temporal profiles (which

assigned the annual total emission rates into different months/days). The results probably indicate the annual total emission rates of MEIC and SOE were accurate but the temporal profiles were not as good as the ones in EDGAR and REAS.

We added above discussion in the revised manuscript.

- Table 4: This is also too detailed information. If you only want to say how many cities out of 60 can improve their prediction with ensemble and do not intend to describe its regional differences, this table can be moved to supplement and it is enough to briefly describe the result in the manuscript.
Response: We moved Table 4 to the supplemental materials as Table S3 and only brief description was kept in the manuscript.

- Table 5: This table showed that the weighting factor can vary large depending on the region. Table 3 demonstrated the factor also change for different averaging time scale. And the factor may be different for the different year. The purpose of this study is proposing an ensemble method for obtaining the better air pollutants concentration data for health effect estimation, from this point of view, how do the authors think the best way to calculate the weighting factor in China? Need some more sentences on it.
Response: Even though the weighted factors vary depending on the regions, averaging times and different years, the ensemble method that we proposed in this study is to minimize the difference between predictions and observations and can be applied in different regions with different averaging time scales, and for any years. The ensemble analysis is a post-process method to improve the agreement between predictions and observations in any averaging time scales, as shown in the manuscript. The way to calculate the weighting factors depends on the objectives of specific studies. But in general, more observation data used in the calculation, more accurate the ensemble prediction would be.

We added above discussion in the revised manuscript.

Reference:

Hu, J., Huang, L., Chen, M., Liao, H., Zhang, H., Wang, S., Zhang, Q., and Ying, Q.: Premature Mortality Attributable to Particulate Matter in China: Source Contributions and Responses to Reductions, Environ Sci Technol, 10.1021/acs.est.7b03193, 2017.