Anonymous Referee #2

General comment: This paper presented a work analyzing the contribution from Europe to China's atmospheric particle concentrations and haze events, with intensive chemistry transport modeling. The authors made great efforts on incorporating multiple transport models to understand the difference between models and to reduce the uncertainty of simulation. They also evaluated the impacts of emission inventory on the simulation, as the accuracy of emission inventory for anthropogenic pollutants is always a big concern on the air quality research community. Before it can be accepted as a final atmospheric chemistry physics paper, however, the following issues need to be further discussed or stressed.

Response: We thank the referee for the encouraging comment and providing insightful suggestions.

1. The significance of the paper needs to be reconsidered and relevant statement should be revised. In current format, the authors stated that there were limited studies conducted on regional transport to China and it might be important as China is controlling its emissions. The results, however, show that the impacts of Europe was very few, and the studying period was before 2010, during which the emissions in East Asia were expected to still increase. It seems that the current work did not fully answer the question they raised. The most serious haze events after 2010 were not included in this study?

Response: We have added some references documenting the long-range transport into China (Lee et al., 2007; Kong et al., 2010; Akimoto 2003; Fu et al., 2012) and a short introduction of their findings. In our study, Table 2 summarized the annual average long-range transport contribution from EUR and RBU regions to EAS region in year 2000 and 2008-2010. Table 3 summarized the long-range transport contributions during the haze episode. We raised the research question in "Introduction" section that "the background concentrations of PM and the contributions from outside China import of air pollutants to the haze problem, is poorly documented." So the question is answered by Tables 2-3 and the related discussions. The severe haze event in 2010 is included in this study but not specifically highlighted. Some places in China has more than 300 days with haze identified with NCDC observation. We analyzed the annual total haze events and reported the contribution of long-range transport to these events, as shown with Figures 9-10. An overview of the haze events for full year 2010 is provided in supplementary material Figure S1.

2. Lines 9-10, Page 2: this results is quite old, there are recently more studies on health impacts of China's air pollution.

Response: We have added several most recent studies that reported the premature deaths attributable to $PM_{2.5}$ pollution in China from 2013 to 2017, these references include: Huang et al., 2018; Cao et al., 2017; Li et al., 2018.

3. Lines 13-16, Page 6: Please define how the MNB was calculated. Is there any criterion indicating the acceptable range of MNB?

Response: The MNB is calculated as:

$$MNB = \frac{1}{S}\sum_{i=1}^{S}\sum_{j=1}^{T}\frac{sim_{ij} - obs_{ij}}{obs_{ij}}$$

Where S is the number of observation stations, T is the total number of month, obs_{ij} is the observed value at the station i and month j, and sim_{ij} is the corresponding simulation value at the closet grid point to the station. There is no well documented threshold for an acceptable MNB, but the AERCOM research program have been frequently cited by the research community, so we used their MNB values to demonstrate our participating models' performance.

4. There are limited $PM_{2.5}$ observations used in model evaluation for China. I understand that the official data were not available until 2013. However, could the data published in previous studies be available and could the evaluation be improved?

Response: This is a very insightful comment and we agree that there many some publications reporting the measured $PM_{2.5}$ concentrations in China (Zhang and Cao, 2016; Lowsen and Conway, 2016), but the diversities in instrument, measuring method, and sampling period make it difficult to develop a consistent observation database from the literatures. In addition to the potential uncertainties within each individual measurement literature, these measurements are usually presented with charts or figures so we would have to roughly read the values from the figures, which may introduce more uncertainty and is not proper for the HTAP program as it requires applying official downloadable data by all participating groups so all experiments and analysis could be reproduced. In contrast, the EANET data used in this study provides measurement collected with the same type of instrument and method. Considering the limited number of EANET sites, we also included AERONET and MODIS AOD which are all public accessible for model evaluation with better spatial and temporal coverage. In addition, since examining long-range transport of surface $PM_{2.5}$ into China is the main objective of this study, evaluating models performances with literature review collected data would require intensive efforts and make the manuscript lengthy.

5. Figures 5 and 6 illustrated the surface aerosol response under EURALL and RBUALL. Can you explain why the seasonal variations were different between models? In Fig 6, for example, larger response was found in summer for CAM-chem and SPRINTARS, while smaller was found in summer in EMEP and GEOSCHEMADJOINT. By the way, caption of Figure 6 should be revised (Figure XXX?).

Response: We apologize for the typo in the figure caption, it is corrected in the revised manuscript. We agree with the referee that prominent different seasonality was found between CAM-chem and other models. Despite applying the same emission inputs, several aspects of the participating models lead to the different seasonality of PM2.5 response. One of the other two referees also pointed out this issue, and we briefly probe into these aspects. These aspects including the meteorology, aerosol mechanisms, and convection mechanisms. We first examined the meteorology differences by comparing the model simulated air temperature the following figure shows domain average monthly mean surface air temperature from CAM-chem, CHASER, GEOS5, and SPRINTARS. In EUR region, CAM-chem and SPRINTARS simulated surface air temperatures are systematically higher than other models by ~2.5K in winter. A higher temperature in the emission source region may facilitate the PM precursors' chemistry and subsequently allow less precursors enter long-range transport. In RBU region, SPRINTARS simulated temperature is ~2K lower than other models in summer, which may lead to more precursors transport into EAS and subsequently induce larger PM response. But on the other hand, temperature is apparently not the only influencing factor as CAM-chem showed highest temperature in summer over EUR region yet largest PM response in summer too. Wind speed and PBL height may play a more important role as indicated by Im et al. (2018) but unfortunately only one of the participating model provided wind and PBL data.

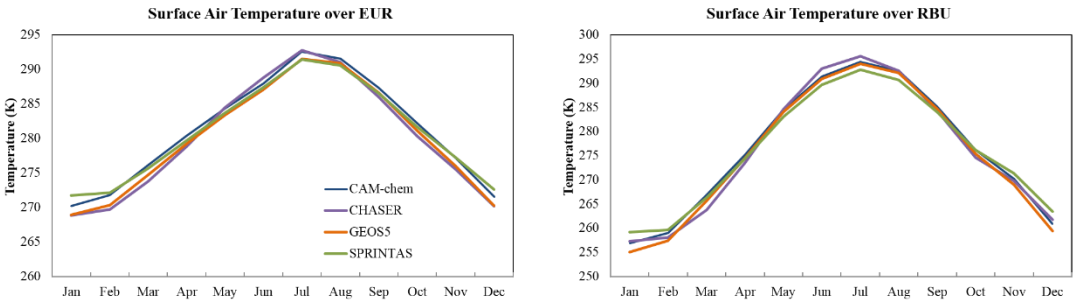

Figure. Monthly surface air temperature from CAM-chem, CHASER, and GEOS5

We then examined the convection schemes among models. CAM-chem applied the modified Zhang-McFarlane approach (Zhang and McFarlane, 1995) with shallow convection follows Hack et al. (2006). GEOS5 applied the modified scheme by Moorthi and Suarez (1992), which is a relaxed Arakawa-Schubert algorithm. These schemes are functionalized and parameterized substantially different and will subsequently lead to differences of aerosol vertical distribution, lifetime, transport, and total suspended aerosol concentration in the atmosphere (Stjern et al., 2016). Aerosol parameterization also lead to different PM2.5 formula. CAM-chem simulates secondary organic aerosol (SOA) with the 2-product approach using laboratory-determined yields from photooxidation of monoterpenes, isoprene and aromatics, while GEOS5 has no SOA. The differences of aerosol scheme, heterogeneous chemistry, treatment of OC, OA, and SOA lead to additional inter-model variability. In addition, grid resolutions diversity is also responsible as Molod et al (2015) demonstrated that different grid resolutions will result in different scavenging aerosol even with the same model. In short summary, the abovementioned aspects may all contribute to the different seasonality of PM2.5 response, and more a set of more specifically designed model experiments is necessary to explicitly identify their influences, yet this is beyond the current scope of HTAP program. We have added a short discussion of the seasonality in the revised manuscript to point out this issue with the clues mentioned here.

References:

Hack, J. J., Caron, J. M., Yeager, S. G., Oleson, K. W., Holland, M. M., Truesdale, J. E., and Rasch, P. J.: Simulation of the Global Hydrological Cycle in the CCSM Community Atmosphere Model Version 3 (CAM3): Mean Features, J. Climate, 19, 2199–2221, doi:10.1175/JCLI3755.1, 2006.

Molod, A., Takacs, L., Suarez, M., and Bacmeister, J.: Development of the GEOS-5 atmospheric general circulation model: evolution from MERRA to MERRA2, Geosci. Model Dev., 8, 1339–1356, doi:10.5194/gmd-8-1339-2015, 2015.

Moorthi, S. and Suarez, M. J.: Relaxed Arakawa-Schubert. A Parameterization of Moist Convection for General Circulation Models, Mon. Weather Rev., 120, 978–1002, doi:10.1175/1520-0493(1992)120<0978:RASAPO>2.0.CO;2, 1992.

Zhang, G. J. and McFarlane, N. A.: Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian climate center general circulation model, Atmos.-Ocean, 33, 407–446, 1995.

6. Lines 23-27, Page 10. I am not persuaded by the authors by the linearity assumption. They estimated the full impact by scaling the PM responses under the 20% emissions perturbation conditions by a factor of 5. If this is the case, why not directly estimate the PM response by removing 100% emissions for given region? Nonlinearity of $PM_{2.5}$ concentrations to precursor emissions was strong, and the uncertainty of the assumption should be carefully analyzed and quantified.

Response: We agree with the referee that $PM_{2.5}$ concentrations to precursor emissions are strong, and this is the merit of applying atmospheric models to simulate the "real" aerosol response instead of simply estimating it with a certain emission change ratio. The 20% emission perturbation is the first priority of model experiment designed by the HTAP Phase2 program because it is a relatively reasonable and applicable control rate for air quality management. The impact of long-range transport however, indicates the overall contribution of the total emission in the source regions, so 100% emission reduction would be a stronger but unrealistic experiment. While the 100% emission perturbation simulation is not available, the "full impact" calculated from 20% emission perturbation is the best estimates we can derive. This method was applied by several investigations for estimating inter-continental transport of O_3 (Fiore et al., 2009; Lin et al., 2010; West et al., 2009; Zhang et al., 2009), which has an even more significant nonlinear response to the precursors. But as the referee mentioned, this method may introduce uncertainty due to the nonlinear response, and we also noticed this issue while analyzing the modeling data. We have applied the Response Surface Method (RSM) with Hemispheric-CMAQ model to quantify the source-

receptor relationship with more detailed simulation design than HTAP. We are analyzing the data now and expect to report the findings later.