Anonymous Referee #1

General comment: The manuscript submitted by Dong et al. reports a basic statistical analysis of 6 simulations from HTAP2 global modelling exercise, aimed at assessing the simulated impact of long-range transport of pollutants from Europe and Russia on China's haze events. The scope of the work is well defined, I think there is some gap that may be filled in terms of link with the existing literature, and there is generally no attempt by the authors in explaining the reasons for inter-model differences. The manuscript is basically a description, sometimes lengthy, of the materials presented in the figures and the tables. Considering the relevance of the topic, I think the manuscript could be published on ACP, after considering some suggestions given below, and after careful English editing.
Response: We appreciate the referee for the overall positive comment and providing the helpful detailed suggestions. We have rewrote unnecessary long sentences and read through the whole manuscript for English editing.

Specific comments:
1. In the introduction the authors very briefly review the literature regarding existing studies on haze in China. It is mentioned that long-range transport contribution to haze episodes is poorly documented (indeed they do not insert any reference). However, the literature on long-range transport to China is not null, and part of the phenomenology and underlying mechanisms might be in common with period of haze episodes. From a very quick literature search I identified, as potential references:
- Lee et al., Heavy metals and Pb isotopic composition of aerosols in urban and suburban areas of Hong Kong and Guangzhou, South China. Evidence of the long-range transport of air contaminants, Atmospheric Environment, Volume 41, Issue 2, January 2007, Pages 432-447
- Kong et al., Receptor modeling of $PM_{2.5}$, PM_{10} and TSP in different seasons and long-range transport analysis at a coastal site of Tianjin, China, Science of The Total Environment, Volume 408, Issue 20, 15 September 2010, Pages 4681-4694
- Akimoto, Global Air Quality and Pollution, Science 05 Dec 2003: Vol. 302, Issue 5651, pp. 1716-1719 (and references therein)
I suggest to deepen the review of the literature on long-range transport from Europe to East Asia and put it into the fourth paragraph of the introduction. The same material might be subsequently used in the interpretation of some of the results illustrate afterwards (e.g. in section 3.1 and 3.2).
Response: The references listed above are closely related to our study thus they have been properly cited in the revised manuscript. The objective of this study is to evaluate the contribution of long-range transport to $PM_{2.5}$, with special focus on the haze episode. Previous studies about long-range transport mainly focused on the exported air pollutants from China to other areas or the transport of O3. Although some studies (e.g., the Akimoto 2003 publication, and the HTAP Phase1 report) pointed out that mitigating global air pollution requires international participations of multiple countries or continents, the contribution of long-range transport to $PM_{2.5}$ in China remains poorly documented. The references suggested by the referee are very helpful. We also added a more detailed description of the research status about long-range transport of air pollutants to China.

2. page 4, lines 1-5: I think these very general statements, without any specific reference, on physical processes should be avoided in the manuscript. Please add proper reference and try to be more specific on the region and the situation you are referring to.
Response: Thanks for the reminder, we have added proper references (Eckhardt et al., 2003; Sthol et al., 2002) and related descriptions.

3. section 2.2: all the data versions and source of data are missing. Please add the exact product names of the data used, the web source used, and the version of the algorithms. This is necessary for the reproducibility of the work.

Response: Thanks for the comment, we have added all the products names, versions, and web sources in the revised manuscript. These details are also summarized in the following table (API, EANET, and EBAS has no version updates information, the data is downloaded from the web source).

Table. Version details and web sources of the data used for model evaluation

Data used	Web source
AERONET (Level2.0, version2)	http://aeronet.gsfc.nasa.gov
API	http://datacenter.mep.gov.cn
EANET	http://www.eanet.asia/
EBAS	http://ebas.nilu.no
MODIS (MOD08, MYD08)	https://modis.gsfc.nasa.gov

4. Figures 2 and 3 and related comments: there are some apparent inconsistency between the results presented in these figures. For example, $PM_{2.5}$ is overestimated by GEOSCHEMADJOINT and underestimated by CHASER, but then AOD at AERONET sites has the opposite bias for these models. Why is that? Perhaps it could be useful to include a comparison only for some specific station for which all the datasets are available, or at least within the same model grid. From Figure 1 it seems to be possible for some stations.

Response: This is a very interesting point and we thank the referee for mention it. Figure 2 shows PM2.5 was overestimated by GEOSCHEMADJOINT by 7.5 μg/m^3 (63%) in EAS (EANET and API stations), by 8.6 μg/m^3 (66%) in EUR (EBAS stations). Figure 3 suggests that GEOSCHEMADJOINT underestimates AERONET-AOD by -0.08 (-23%) in EAS and overestimates AOD by 0.004(4%) in EUR. As suggested by the referee, we selected EANET-Oki (36.28ºN, 133.18ºE) as the PM2.5 site and AERONET-Osaka (34.65ºN, 135.59ºE) as the AOD site in EAS region, and selected EBAS-Revin (49.90ºN, 4.63ºE) as the PM2.5 site and AERONET-Brussels (50.78ºN, 4.35ºE) as the AOD site in EUR region. These are the closest nearby sites in each of the domain. Simulation bias of GEOSCHEMADJOINT at these sites are shown in the following figure.

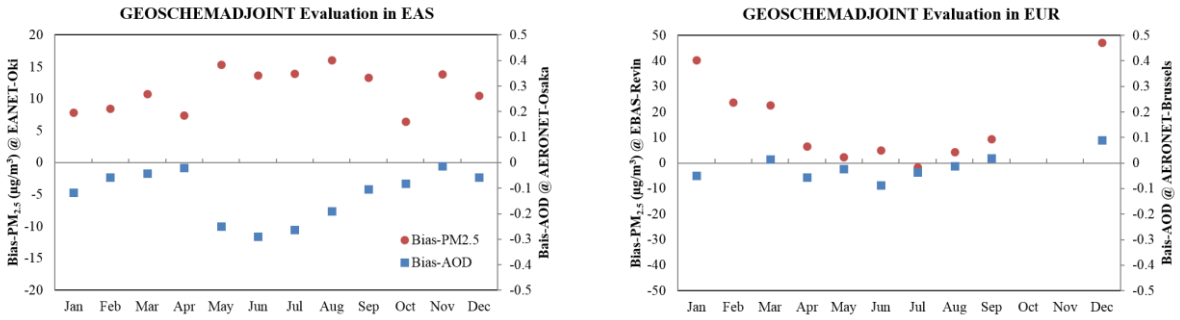

Figure. Simulation bias of GEOSCHEMADJOINT for $PM_{2.5}$ (solid red circles) and AOD (solid blue squares)

As shown in the figure, GEOSCHEMADJOINT overestimated $PM_{2.5}$ but underestimate AOD throughout the full year 2010 in EAS region. We examined this issue and found out there are two reasons: first, there are relatively less $PM_{2.5}$ observation sites (2 in EAS, 5 in EUR) compared to large number of AOD observation sites (15 in EAS, 73 in EUR). The EANET-Oki station was located on a small island ~50 miles from west coast of Japan thus represents the background concentration, while the AERONET-Osaka site is located in the downtown area of Osaka City. Evaluation in EUR region has the similar condition, the EBAS-Revin site is in a national park, while the AERONET-Brussels site is close to downtown. Although some AERONET sites are also located in remote areas, it generally has a more comprehensive representation of different surroundings including both rural and urban, but the $PM_{2.5}$ data

used in this study are most located in rural area. Second, GEOSCHEMADJOINT are reported as tend to overestimate the surface layer $PM_{2.5}$ concentration in Asia (Figure 2 in Gu and Liao, 2016; Figure 2 in Xu et al., 2015) and underestimate the column density AOD (Figure 4 in Choi et al,. 2009) in East Asia, although the explicit reason for this inconsistency hasn't been well documented. So generally the performance and evaluation results of this HTAP Phase 2 modeling effort is consistent with those previous studies.

References:

Yi-Xuan GU & Hong LIAO (2016) Response of fine particulate matter to reductions in anthropogenic emissions in Beijing during the 2014 Asia–Pacific Economic Cooperation summit, Atmospheric and Oceanic Science Letters, 9:6, 411-419, DOI: 10.1080/16742834.2016.1230465

Xu, J.-W., Martin, R. V., van Donkelaar, A., Kim, J., Choi, M., Zhang, Q., Geng, G., Liu, Y., Ma, Z., Huang, L., Wang, Y., Chen, H., Che, H., Lin, P., and Lin, N.: Estimating ground-level PM2.5 in eastern China using aerosol optical depth determined from the GOCI satellite instrument, Atmos. Chem. Phys., 15, 13133-13144, https://doi.org/10.5194/acp-15-13133-2015, 2015.

5. Figure 4 and related comments: the modelled AOD over China and elsewhere in the domain differ among models by more than a factor of two. As for previous results on point measurements, there is no attempt to explain the differences. For example, considering the same anthropogenic emissions, the difference over China CHASER and SPRINTARS is quite remarkable.

Response: We also notice the large difference between model performances with the same emission inputs. Explicitly clarify the causes of the difference would require deep detailed investigation of the model schemes, algorithms, and parameterization, which is not within the scope of this study. But the other HTAP Phase2 related studies (Im et al., 2018; Palacios-Peña et al., 2018; Astitha et al., 2018) do present a few investigations into the multi-model comparison between the models used in this study, and the different model performance are attributed to meteorology (in particular wind speed and PBL height), different aerosol mechanisms, treatment of wind-blown dust emission, and biomass burning emission injection heights. Previous multi-modeling efforts such as the AEROCOM also pointed out that these aspects can lead to modeled AOD and surface PM concentration differ by a factor of 2 and 10 respectively, although the some AEROCOM participating models are different from HTAP. We agree with the referee that briefly explain the difference is necessary as our discussion is based on multi-model simulations, so we have added a short discussion in the revised manuscript. The above-mentioned references are also added into the revised manuscript.

References:

Im, U., Christensen, J. H., Geels, C., Hansen, K. M., Brandt, J., Solazzo, E., Alyuz, U., Balzarini, A., Baro, R., Bellasio, R., Bianconi, R., Bieser, J., Colette, A., Curci, G., Farrow, A., Flemming, J., Fraser, A., Jimenez-Guerrero, P., Kitwiroon, N., Liu, P., Nopmongcol, U., Palacios-Peña, L., Pirovano, G., Pozzoli, L., Prank, M., Rose, R., Sokhi, R., Tuccella, P., Unal, A., Vivanco, M. G., Yarwood, G., Hogrefe, C., and Galmarini, S.: Influence of anthropogenic emissions and boundary conditions on multi-model simulations of major air pollutants over Europe and North America in the framework of AQMEII3, Atmos. Chem. Phys., 18, 8929-8952, https://doi.org/10.5194/acp-18-8929-2018, 2018.

Palacios-Peña, L., Jiménez-Guerrero, P., Baró, R., Balzarini, A., Bianconi, R., Curci, G., Landi, T. C., Pirovano, G., Prank, M., Riccio, A., Tuccella, P., and Galmarini, S.: Aerosol optical properties over Europe: an evaluation of the AQMEII Phase 3 simulations against satellite observations, Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2017-1119, in review, 2018.

Astitha, M., Kioutsoukis, I., Fisseha, G. A., Bianconi, R., Bieser, J., Christensen, J. H., Cooper, O., Galmarini, S., Hogrefe, C., Im, U., Johnson, B., Liu, P., Nopmongcol, U., Petropavlovskikh, I., Solazzo, E., Tarasick, D. W., and Yarwood, G.: Seasonal ozone vertical profiles over North America using the AQMEII group of air quality models: model inter-comparison and stratospheric intrusions, Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-98, in review, 2018.

6. Figures 5-6 and related comments. The figures are interesting because they nicely illustrate the model diversity. For example, the seasonal cycle of contributions from some models is opposite to that of others (e.g. CAM-Chem peaks in summer, CHASER in winter, and GEOS5 in spring). It would be useful to have some inspection of these difference. I suspect that differences in the meteorological fields used in these models are responsible for the large variability.

Response: CAM-chem showed the largest PM response in summer under EUR emission reduction scenario, and SPRINTARS showed the largest PM response in summer under RBU emission reduction scenario, while the other models all showed larger PM responses in winter or spring. We agree with the referee that meteorology difference might be one of the reasons for simulation diversity. We examined the surface air temperature used by the participating models. Domain averages of monthly temperature over EUR and RBU are shown in the following figure.

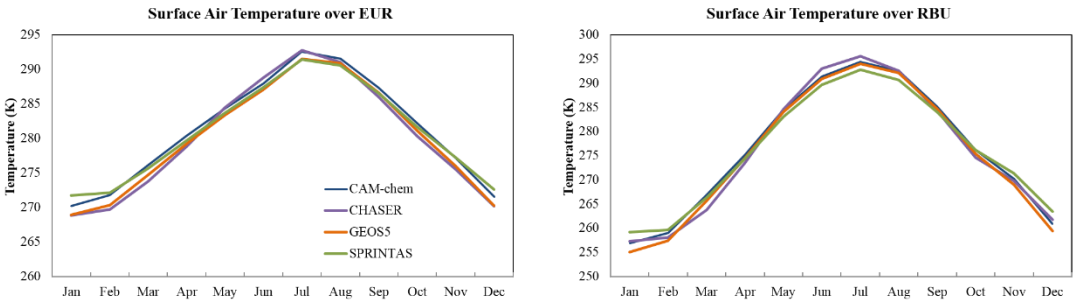

Figure. Monthly surface air temperature from CAM-chem, CHASER, and GEOS5

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. Explicitly identify the impact of meteorology on modeled PM response would require a set of more detailed experiments, and this is beyond the scope of HTAP program. 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 $PM_{2.5}$ 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 fact, not only the $PM_{2.5}$ responses but also the baseline $PM_{2.5}$ concentrations show prominent different seasonality among the models in both the HTAP Phase1 (Dentener et al., 2010) and Phase2 program, and this is also why multi-model mean is used to estimate the source-receptor relationship. We have added the abovementioned discussion and references in the revised manuscript.

7. Figures 9-10: some panels look patchy, for example EMEP, SPRINTARS and all in

Figure 10. Why is that?
Response: Figures 9-10 are designed to demonstrate the full impact of long-range transport during the haze episode, so the NCDC surface observation data is used to identify where and when haze exist. For those with finer grid resolution such as EMEP (0.5×0.5º) and SPRINTARS (1.1×1.1º), there are some model grids having no NCDC observation site, and these grids are filled with missing value, and this makes the figures look patchy. Although haze (visibility) can also be estimated with aerosol extinction coefficient, using the direct measurements from NCDC is apparently a more solid method to identify haze, and only SPRINTARS provides the aerosol extinction coefficient data. We have added a short sentence in the figure caption to explain the patchy panels to avoid misleading.

8. I recommend English editing of the manuscript. The use of language is imaginative and makes understanding difficult. A few random examples:
- p. 3, l. 40-41: "These datasets are essential to estimate surface PM response compare the aerosol transport in different atmosphere layers". What is "response compare"? "atmosphere" –> "atmospheric"
Response: We agree that English editing is necessary, it's also pointed out by another referee. This sentence is removed because it is not necessary.

- p. 5. l. 6-7: "the models all tend to underestimate the high peaks in spring (Mar.-Apr.) and low bottoms in summer". Not clear what "low bottoms" means.
Response: The sentence is changed to "Temporal variation of O_3 is also simulated well in EAS, although the models all tend to underestimate the high values in spring (Mar.-Apr.) and low concentrations in summer (Jul.-Sep.)"

- note 2 on caption of Figure 2: "PM2.5 observations in EUR and EAS region have no standard because there are no sites with valid measurements fall into the same model ensemble mean grid". Very difficult to understand: why a standard deviation cannot be calculated even if stations are not in the same model cell?
Response: The standard deviation is calculated between the observations from different sites in the same model grid, we have mentioned in the caption of Figure 2 that "vertical error bars depict the standard deviation across the sites in the same ensemble grid."