Answer to Referee #4

We thank the referee for his/her very careful review, and his/her constructive suggestions. In the following, we answer his/her specific questions. In order to facilitate the reference to the questions and proposed changes, we use the following color coding:

Color coding:
reviewer comment
our answer
proposed change in manuscript

This study used in situ PM2.5 measured by portable laser sir quality monitors to replace traditional PM2.5 data collected by ground monitoring stations or derived from remote sensing images and developed a new hybrid (land use regression plus geostatistical) method to map PM2.5 concentrations in an urban area. Generally, this manuscript is well organized and clearly written, even though a few of sentences need to be rephrased and more details need to be supplemented. I recommend the editor to accept this manuscript after a minor or moderate revision.
The authors developed a hybrid model in which the deterministic component of the PM2.5 concentration was fitted by LUR and the stochastic component (i.e. residual) was interpolated by kriging. Thus this is a typical LUR based REGRESSION kriging but not universal kriging. Please see Liu et al. (2018). Incorrectly naming the method is my biggest concern for the manuscript.
Liu, Y. et al., 2018. Improve ground-level PM2.5 concentration mapping using a random forests-based geostatistical approach. Environmental Pollution, 235, 272-282.

Response: the naming of this method followed Mercer et al. (Atmospheric Environment 2011). They proposed a 2-step approach in which simple kriging is applied to the residuals from LUR. This approach is similar but not identical to UK. Thus, we agree with the reviewer that the Regression Kriging is more appropriate and thank him/her for the suggestion. We implemented the changes in the revised manuscript.

Mercer, L. D., Szpiro, A. A., Sheppard, L., Lindström, J., Adar, S. D., Allen, R. W., Avol, EL., Oron, A. P., Larson, T., Liu, L. J., and Kaufman, J. D.: Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NOx) for the multi-ethnic study of atherosclerosis and air pollution (MESA Air), Atmos Environ, 45, 4412–4420, doi:10.1016/j.atmosenv.2011.05.043, 2011.

I am afraid that the Abstract from line 16 to 27 is not clear for a new reader especially who has not read the Method section. What do the "Period 1" and "Period 2" represent?

Response: "Period 1" and "Period 2" represent the light-polluted period and heavy-polluted period. We rewrote these confusing sentences and replaced lines 15 –27 on Page 1 by:

During this process, $PM_{2.5}$ concentrations were measured by laser air quality monitors and uploaded by a group of volunteers via their smart phone applications during two periods. Three extensively employed modelling methods (ordinary kriging (OK), land use regression (LUR), and regression kriging (RK) were adopted to evaluate the performance. An interesting finding is that $PM_{2.5}$ concentrations in micro-environments significantly varied in the intra-urban area. These local $PM_{2.5}$ variations can be effectively identified by crowdsourced sampling rather than national air quality monitoring stations (light-polluted period: (69.67±18.81) – (76.45±14.55) μg m^{-3} vs. (36.9±10.97) – (41.2±8.68) μg m^{-3}; heavy-polluted period: (162.72±15.96) – (171.89±21.5) μg m^{-3} vs. (177.8±16.91) – (188.3±22.4) μg m^{-3}).. The selection of models for fine scale $PM_{2.5}$ concentration mapping should be adjusted according to the changing sampling and pollution circumstances. Generally, OK interpolation performs best in conditions with non-peak traffic situations during a light-polluted period (hold-out validation R^2: 0.47–0.82), while the RK modelling can perform better during the heavy-polluted period (0.32–0.68) and in conditions with peak traffic and relatively few sampling sites (less than ~100) during the light-polluted period (0.40–0.69). Additionally, the LUR model demonstrates limited ability in estimating $PM_{2.5}$ concentrations on very fine spatial and temporal scales in this study (0.04–0.55), which challenges the traditional point about the good performance of the LUR model for air pollution mapping. This method selection strategy provides empirical evidence for the best method selection for $PM_{2.5}$ mapping using crowdsourced monitoring, and this provides a promising way to reduce the exposure risks for individuals in their daily life.

(Page 2, line 19) The authors should cite Liu et al. (2018) that is a typical study combining two technologies to estimate PM2.5 concentrations.

Response: Liu et al. (2018) adopted a random forests-based regression kriging approach which integrates recent advancements of machine learning with conventional kriging methods in geostatistics. We thank the reviewer for the suggestion and cited this article in the revised manuscript.

In the Measurement Instrument section, the authors may add more details for their portable air quality monitors, e.g. the company producing the equipment and other practical uses of the portable monitor.

Response: more details were added as the reviewer asked. Replaced lines 23 –29 on Page 3 by:

The portable laser air quality monitor SDL307 (produced by NOVA FITNESS Co., Ltd.) is employed to perform sampling. The monitor manual can be downloaded from http://www.inovafitness.com/index.html. This monitor can be conveniently carried with a total size of 25×34×14 cm (Fig. 1a). According to the test report provided by the Center for Building Environment Test at Tsinghua University, the maximum relative error of this monitor is ±20% compared with a regulatory monitor in the 20–1000 μg m^{-3} range and has a resolution of 0.1 μg m^{-3}. The concentration of particulate matter is measured using the light-scattering method (Fig. 1b). The monitor contains a special laser module, and the signals are recorded by a photoelectric receptor when particulate matter passes through laser light. The count and size of particulate matter are then analysed by a microcomputer after the signals are amplified and converted. Their mass concentrations are calculated based on the conversion factor between the light-scattering method and the tapered element oscillating microbalance technology.

(Page 4, lines 13-20). The sentences here are unclear and the authors may need to rewrite them. "Sampling was carried out in two time periods in the winter of 2015…" I am wondering whether the authors can provide a specific time periods (e.g. from November 1 to December 31) to replace "the winter". "The second period was between 14:00 and 18:00, when Orange warning signals of haze were released by Changsha Meteorology Bureau…" I guess Orange warning signal was not released every day, but from your last sentence "The first period was between 8:00 and 12:00, representing a light-polluted period" it seems the Orange warning signal is released every afternoon. So please make it clear whether you measured PM2.5 concentrations during the two time slots all days or only Orange days. Additional, I suggest using "time slots" to replace "time periods". The "period" may be used for the days when you collected the PM2.5 concentration samples.

Response: In fact, due to the difficulties in implementing the campaign (e.g. the financial burdens of volunteers' recruitment and the extensive investment of time and efforts for technology part and procedures to ensuring data quality), we only carried out this sampling between 8:00 and 12:00 on December 24 and 14:00 and 18:00 on December 25. In the first period, the official air pollution levels were "Good" and "Moderate", in the second period, the Changsha Meteorology Bureau released an Orange warning signal of haze (i.e. the official air pollution level was "Heavily Polluted"). We rewrote these confusing sentences and replaced lines 16 – 25 on Page 4 by:

Sampling was performed in two time periods in the winter of 2015 to examine the effect of air quality grades on the mapping results. The first period fell between

8:00 and 12:00 on December 24. In this period, the official air pollution levels were "Good" and "Moderate" (i.e., Period 1, light-polluted period). The weather was overcast with occasional rain or drizzle, and the relative humidity (RH) ranged from 95% to 98%. The second period extended between 14:00 and 18:00 on December 25, when an orange warning signal of haze (i.e., official air pollution level was "Heavily Polluted") was released by the Changsha Meteorology Bureau (i.e., Period 2, heavy-polluted period). The weather was cloudy with some sunshine, and the RH ranged from 39%–43%.
Before sampling started, every volunteer received one monitor and went to the corresponding area. At each potential monitoring site, the volunteer lifted the monitor (~2 metres above the ground) and held it for at least 60 seconds to measure the PM2.5 concentration. The observations were uploaded twice to four times hourly using a smart phone application (App) that we developed. The geographic coordinates of the sampling sites were also uploaded. For each hour, we eliminated the sampling sites with less than three observations. The valid observations were then averaged at each site. As some volunteers quit after the sampling of the first period, the sampling sites in period 2 were concentrated in the central study area. A total of 179-208 samples were successfully collected at each hour in Period 1, and 105-118 samples were successfully collected in Period 2. The official observations at 10 national monitoring stations in the study area were also obtained (China Environmental Monitoring Center, CEMC: http://106.37.208.233:20035/) and averaged for comparison purposes.

(Page 4, line 20). "The official observations at 10 national monitoring sites stations."

Response: corrected.

(Page 6, lines 21-22) "Clearly, the average PM2.5 concentrations of Period 2 were two times higher than those of Period 1…" I wonder why the authors emphasized "two times" higher here. It gave me a deep impression that "two times" implied something, but I have not seen any explanation for the "two times" in the following text. I would simply say: the average PM2.5 concentrations of Period 2 were much higher than …

Response: we thank the reviewer for the suggestion and rewrote this sentence.

(Page 9, lines 1-10) I cannot accept the authors' discussion in this paragraph whatsoever. Compared with the authors' cheaper potable air pollution monitors, I more trust instruments from national monitoring stations. "This suggests the inconvenient truth (what a strong word! It is just a possible.) that the exposure risk remains relatively high for the public when official air pollution levels are "Good" and "Moderate" and this risk …" I completely understand what the authors intend to express, but if the government intentionally falsified the air quality data, it was more likely to lower the heavy- rather than light-pollution data. I thought of another possibility: the authors' portable monitors were not sensitive for the low PM2.5 concentrations and are proneto be saturated in the heavy-pollution days. In that case, it will also get the result the authors showed in the manuscript. The authors intended to emphasize that the large error (difference) on PM2.5 concentrations over the city is due to the relatively small number of national monitoring stations and thus their method using portable monitors to collect PM2.5 data is useful. However, based on the authors' statement, large differences on PM2.5 concentrations have existed even if concentrations are measured by the instruments of the national monitoring stations and the portable equipments of the authors at the same location.

Response: we agree with the reviewer that the instruments of national monitoring stations are more accurate and reliable than the potable air pollution monitors, and that is the reason why we conducted the comparison experiments between laser air quality monitors and the national monitoring instruments at the same positions and heights before and after the crowdsourcing sampling. The point we intend to make is that the crowdsourced $PM_{2.5}$ measurements demonstrated obvious spatial variation between urban microenvironments, and these variations can hardly be disclosed by sparse national air quality monitoring stations. The difference of hourly $PM_{2.5}$ concentrations between the two types of instruments in sampling campaign is possibly because of the different sampling heights and the change of the major pollution sources in the study area. We thank the reviewer for pointing out this issue. We rewrote these confusing sentences and replaced lines 25 –30 on Page 8 and lines 1 –13 on Page 9 by:

The number of sampling sites were 18 and 10 per 100 km^2 for Period 1 and Period 2, respectively. These data comprise a considerable improvement compared with a density of approximately 0.015 sites per 100 km^2 in the national air quality monitoring network in China. As expected, crowdsourced $PM_{2.5}$ measurements demonstrated detailed spatial variation among urban microenvironments, and these variations can hardly be disclosed by sparse national air quality monitoring stations. This finding suggests that crowdsourced sampling can effectively improve the density of $PM_{2.5}$ monitoring at a rather low monetary cost and can be supportive of the short-term air pollution exposure assessment for epidemiologic studies at a fine scale. To explore the spatial variation in the $PM_{2.5}$ concentration for various urban microenvironments and compare with the national air quality measurements, the crowdsourced monitoring is assumed to cover a certain number of areas. However, persuading the general public in these areas to continuously observe and upload $PM_{2.5}$ concentrations during their activities of daily living through a designed study is difficult. We employed a batch of volunteers to model their behaviours on the general public's behaviour and simultaneously collect data. This approach is a preliminary practice of crowdsourced monitoring and can be further developed and improved in the long-term exposure assessment at the fine scale in the future with the progress in low-cost wearable air quality monitors and automatic processing techniques of crowdsourced data.

The hourly $PM_{2.5}$ concentrations between crowdsourced sampling sites and national monitoring stations were rather different; this difference varied as the official air quality level changed. The crowdsourced $PM_{2.5}$ concentrations were substantially larger than the national concentrations in Period 1 (light-polluted) and slightly lower in Period 2 (heavy-polluted). One possible reason is that the national monitoring stations in the study area were installed on the roofs of mid-rise buildings (i.e., ~15 m) with ventilation and spaciousness, while crowdsourced sampling was conducted on the real ground (i.e., ~2 m). The change in the major pollution sources and meteorological conditions in the study area may contribute to the difference between two periods; the major contribution of local sources, especially the vehicle emission and the very high RH (95%–98%) during the light-polluted period, may cause the accumulation of $PM_{2.5}$ near the ground; and the sources of long-range transport of regional pollution during the heavy-polluted period can increase the concentration of $PM_{2.5}$ on the upper layer. This finding suggests that the air pollution exposure risk may remain relatively high for the public on the ground in some urban microenvironments, even when official air pollution levels are "Good" and "Moderate" and sensitive groups should consider reducing some outdoor activities. The results confirm the necessity of developing real-ground high-density crowdsourced $PM_{2.5}$ monitoring networks. Although the low-cost sensor and the use of optical particle detection of monitors in sampling may cause inaccuracies in measurements, we have attempted to minimise the uncertainty by disusing the relatively inaccurate monitors (MRE>5%) used in preliminary indoor and outdoor experiments. Comparison experiments between laser air quality monitors and the national monitoring instruments were also conducted at the same positions and heights for two time slots; the weather conditions and air quality scenarios of the two time slots were similar to the two sampling periods (i.e., overcast with light rain, RH≥76%: December 20–22 vs. Period 1; cloudy with sunshine, RH≤67%: December 29–31 vs. Period 2). The relatively good agreement between the hourly PM2.5 concentrations of laser monitors and those of national instruments had guaranteed the reliability of sampling data to a certain extent. The relative humidity may have slightly influenced the crowdsourced $PM_{2.5}$ concentrations in the light-polluted period since December 20–22 yielded a slightly lower R^2 and RMSE than those of December 29–31 but a higher MRE than that of December 29–31. However, the relative error of $PM_{2.5}$ observations in preliminary and comparison experiments were generally small and fluctuated without distinct trends and leading factors. During the following procedure of mapping method selection, three methods were performed with the same dataset, which caused a limited influence of uncertainty in measurements on the method comparison results; therefore, we did not correct the measurements in this study. However, more efforts are needed in crowdsourced measurements correction and uncertainty analysis in air pollution concentration mapping at high resolution for accurate exposure assessment in the future.

I suggest the authors cautiously using some very strong adjectives and adverbs, such as clearly, significantly, tremendous, etc. (Abstract, line 25) "This method selection strategy provides solid experimental evidence for method selection of …" I will say "this study provides empirical evidence for …" Although generally clear for me, it is better to further polish the English of this manuscript, especially in the Results and Discussion sections.

Response: we thank the reviewer for the suggestion. Meanwhile, this manuscript was edited for proper English language, grammar, punctuation, spelling, and overall style by one or more of the highly qualified native English speaking editors at American Journal Experts. The certificate may be verified at www.aje.com/certificate with a certificate verification key of E57E-12C6-6B0F-0300-999B.