Answer to Referee #1

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

Xu et al describe measurements and spatial modeling of PM2.5. Measurements were conducted with hand-held optical particle monitors. The spatial modeling compared multiple methods: ordinary kriging, universal kriging, and land use regression. The paper suffers from several critical flaws and is not publishable in its current form. Below I outline five major problems with the manuscript.

Major Issue #1: I do not know what the authors mean by a "crowdsourced" data collection. The authors seem to define crowdsourcing in lines 27-28 of page 2, but "Crowdsourcing activities based on informal social networks and web 2.0 technologies that allowed citizens themselves to produce geospatial data among others" seems more like corporate jargon than a useful explanation of crowdsourcing.

Response: we rewrote this sentence as:

Crowdsourced monitoring that enables citizens to produce geospatial data is constantly growing and shows considerable potential (Heipke, 2010). Large and diverse groups of people who lack formal training can easily describe their environments with a mobile phone or smart phone and upload data via informal social networks and web technology.

The sampling approach seems to be short-term saturation sampling - many volunteers simultaneously sampled at predetermined locations. This sampling approach does not fit my personal notion of crowdsourcing, which would be a more informal data collection leveraging people's normal movements throughout the day. Sending an army of students to collect data in an organized fashion seems less like "crowdsourcing" and more like a sampling campaign. In that sense, this study has little distinction from the large literature on distributed air quality sampling.
What would be the value or longer-term viability of this or a similar sampling approach? This paper focuses on two short sampling periods of a few hours each, so the data are unlikely representative of long-term spatial patterns. Do the authors expect to deploy an army of distributed samplers on a semi-regular basis in order to build up a dataset capable of reproducing longer-term trends? Or to send out volunteers daily to make daily maps? I

don't see how the "crowdsourced" aspect of this adds value or novelty; instead it seems like crowdsourcing is being used as a buzzword.

Response: In order to explore the spatial variation of $PM_{2.5}$ concentration for various urban microenvironment 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 therefore employed a batch of volunteers to model their behaviours on the general public's behaviour and and simultaneously collect data. 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 for two short sampling periods. We agree with the reviewer that this sampling approach is not a complete crowdsourcing activity, but we believe it is a preliminary practice of crowdsourced monitoring and can be further developed and improved by progress in low-cost wearable air quality monitors and automatic processing techniques of crowdsourced data. We rewrote the **2.1.3 Sampling and data processing,** sentences about this were added. Meanwhile, it has to be claimed that the focus of this study is the 'strategies of methods' (e.g. LUR, OK) selection under crowdsourced monitoring data rather than the discoveries of long-term spatial patterns of $PM_{2.5}$ concentrations. Sentences about this were claimed in the **Introduction**.

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.

Replaced lines 25 –30 on Page 8 and lines 1 –3 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.

Major Issue #2: Data quality. Figure 1 shows one short-term comparison between the handheld PM monitors and the regulatory monitors. While there is generally good agreement, there is a fair amount of scatter among the handheld monitors. This scatter is to be expected given the low cost and the use of optical particle detection. However, the authors do not address how uncertainty in the measurements potentially impacts the mapping. Nor do they seem to account for uncertainty in the measurements or make any efforts to correct the measurements (e.g., based on hygroscopic growth).

Response: Although the low-cost sensor and the use of optical particle detection of monitors used in sampling may cause inaccuracies in measurements, we have tried to minimum the uncertainty by disusing the relatively inaccurate monitors (MRE>5%) through 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. In the previous version of the manuscript, we thought one comparison result is enough to demonstrate the reliability of sampling data to a certain extent, we thank the reviewer for pointing out the inadequacies. Sentences about data quality and Figure 1d were added.

Replaced lines 23 –30 on Page 3 and lines 1 –3 on Page 4 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.

To ensure the data quality of this monitor, we placed 115 laser air quality monitors in the same environment and continuously observed them for one week during each of the four seasons. If the relative error between the observation of one monitor and the average observations of the other monitors exceeded 5%, this monitor fell into disuse. This procedure was conducted both indoors and outdoors. Subsequently, 86 monitors with rather stable performance and a small difference between each observation remained. In addition, we randomly selected 30 portable laser air quality monitors to compare with the national monitoring instruments to further guarantee the reliability of the sampling data. First, for ease of operation, three national air quality monitoring stations were selected. Second, for each station, 10 monitors were observed next to the national monitoring instrument (~15 metres above the ground in the study area) from 8:00 to 20:00 on December 20–22, 2015 and from 8:00 to 20:00 on December 29–31, 2015. The weather on December 20–22 was overcast with patchy drizzle and light rain at times, and the relative humidity (RH) ranged from 77% to 94%, while the weather on December 29–31 was cloudy with some sunshine and a RH that ranged from 38%–67%.

The scatter plots and descriptive statistics of the valid hourly average $PM_{2.5}$ concentrations from the laser air quality monitors and the national monitoring instruments were presented in Fig. 1c and Fig. 1d. The hourly average $PM_{2.5}$ concentrations for two types of instruments generally showed good agreement with a correlation coefficient R2 of 0.89 on December 20–22 and 0.90 on December 29–31. The root-mean-square-errors (RMSE) for the former time period was lower than the RMSE for the latter time period (5.63 μg m^{-3} vs. 5.94 μg m^{-3}), while the mean relative error (MRE) was higher than the MRE for the latter time period (6.37% vs. 3.82%). The latter time period demonstrated a smaller difference in hourly average PM2.5 concentrations between laser air quality monitors and the national monitoring instruments with mean values and standard deviations (SD) of 72.99±16.45 μg m^{-3} vs. 71.89±15.28 μg m^{-3} and 129.93±18.33 μg m^{-3} vs. 129.33±17.50 μg m^{-3}.

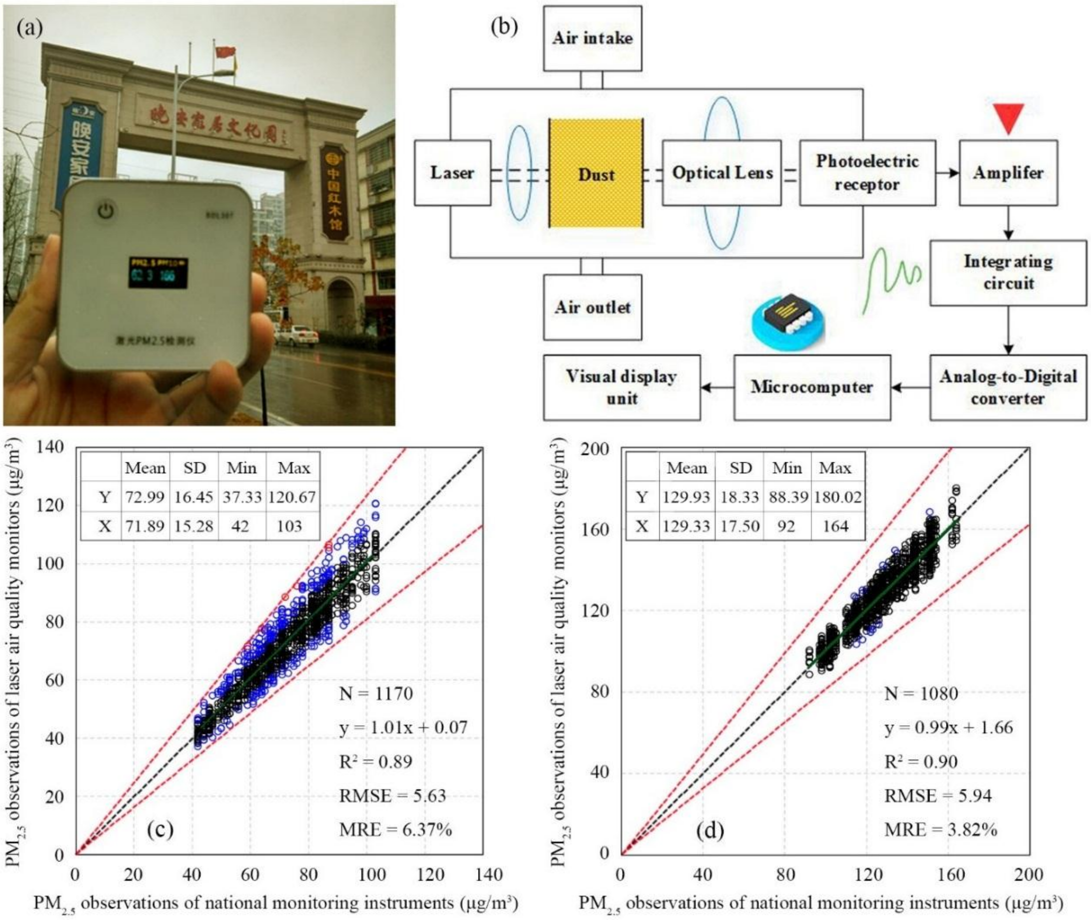

Figure 1: Principle and accuracy of measurement instrument. Y and X are laser air quality monitors and national monitoring instruments, respectively. The black dots, blue dots and red dots indicate $PM_{2.5}$ observations with relative error of <10%, 10%–20%, and >20%, respectively, between two instruments. The black dotted line and red dotted line are the 1:1 line and 1:1.2 line as references.

On the one hand, the relative error of $PM_{2.5}$ observations in preliminary and comparison experiments were generally small and fluctuated without distinct trends and leading factors which make it hard to correct. On the other hand, the main purpose of this study was to propose strategies of method selection for fine scale $PM_{2.5}$ mapping using crowdsourced monitoring, as the three methods we compared were performed with the same sampling dataset, the uncertainty in measurements may cause a limited influence on the method comparison results. We therefore did not correct the measurements in this study. We agree with the reviewer that more efforts are needed in crowdsourced measurements correction and uncertainty analysis in air pollution concentration mapping at high resolution for accurate exposure assessment. Sentences about this were added in the section of Discussion.

Table 3 and section 3.1 - the crowdsourced data read higher PM than the regulatory data. The authors have not convinced me that this is not an artifact of the sensors they have chosen. During some hours there is significant difference between the mean "crowdsourced" PM and the mean regulatory PM. Since the overall spatial extent of the two sampling domains (regulatory and crowdsourced) is roughly similar, I would expect similar mean concentrations from each dataset.
Line 30 on page 8 calls the national monitoring sites "inaccurate." I am not familiar with regulatory measurement policies in China, but if they are anything like the US and Europe, the accuracy standard is high. The spatial pattern derived from these few monitors may be erroneous, but the specific measurements are accurate.

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. In fact, the overall spatial extent of the two sampling domains (regulatory and crowdsourced) is relatively different according to the Figure 3, the color rendering may be the reason why the difference is not so significant. We therefore summarized the statistics of $PM_{2.5}$ concentration. 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.

Major Issue #3: Site selection and sampling strategy. The description of the sampling strategy is insufficient. Were all samplers deployed simultaneously at all sites in Table 1? How were the sampling times defined and chosen? What are significant differences between period 1 and period 2?
Table 1 - A better description of each type of site is needed. For example, Dust surfaces seem to be defined as "dust surfaces," which is not helpful to readers. What qualifies as a dust surface? Some entries in this table have "A" and "U". What do those designations mean?

Response: Table 1 presents the rules to determine the potential $PM_{2.5}$ sampling sites that we would like to monitor. 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 $PM_{2.5}$ concentration. Those observations were uploaded twice to four times hourly using a smart phone application (App) that we developed. So technically, samplers were not deployed simultaneously. For each hour, we eliminated the sampling sites with less than three observations. The valid observations were then averaged at each site. Meanwhile, because some volunteers quit after the sampling of the first period, the final number of samples for each hour were different. Compare with period 1, sampling sits in period 2 were mainly concentrated in the central study area.
Dust surfaces refer to natural and artificial bare surfaces with vegetation cover less than 10% that are easy to produce atmospheric particulate matters. “U” and “A” are subset of the set of potential $PM_{2.5}$ sampling sites and the subset of the union of supporting data. More details for supporting data of site selection were added.

Replaced lines 9-14 on page 4 by:

To ensure that the sampling sites exhibit a relatively even and typical distribution for different urban microenvironments (i.e., residential community, building site, school, and park), a series of rules were designed to determine the potential $PM_{2.5}$ sampling sites based on the distribution of potential emission sources (refer to Table 1). The data that support the sampling design consist of important points of interest (POI), dust surfaces, and main road networks. POI data includes industrial parks, enterprises, factories, depots, hospitals, schools, and parks. Dust surfaces refer to natural and artificial bare surfaces with vegetation that covers less than 10%, which easily produce atmospheric particulate matter, such as construction sites, stacked substance, and natural bare land. These data were collected from the Information Center of Land and Resources of Hunan Province. More than three observations of $PM_{2.5}$ concentrations are required every hour for each potential sampling site to improve the reliability of the sampling data. Given that the number of laser air quality monitors and the distance that a volunteer can walk in one hour are limited, only 2–4 sites can be set in the area in which a monitor can cover during the sampling. Therefore, a total of 208 potential $PM_{2.5}$ sampling sites were selected. The centre of each area covered by a monitor were numbered in sequence (i.e., 1–86). The monitors were also numbered and labelled.

Table 1 was changed as:

Table 1. Rules for potential $PM_{2.5}$ sampling sites selection.

Code	Type	N	Rules
1	Vertex point	5	U1[a] = {X[c] | X∈(Vertex point of the boundary of sampling area ∩ Landmark)}.
2	Industrial park	28	A2[b] = {X | X∈((Industrial park ∪ (Metal & cement & power industrial factories agglomeration)) – High-tech industrial park)}; U2 = {X | X has the largest number of factories within its 100 m buffer zone AND X∈A2}.
3	Dust surface	13	A3 = {X | X∈(POI ∩ Dust surface) AND area of dust surface ranks in the top 4 of each district}; U3 = {X | Distance between X＞200 m AND X∈A3}.
4	Depot	16	U4 = {X | X∈(Coach station ∩ Railway station)}.
5	Scenic area	27	A5 = {X | X∈((Park – Neighbourhood park) ∩ well-known scenic area)}; U5 = {X | Distance between X＞200 m AND X∈A5}.
6	Hospital	11	A6 = {X | X∈(Hospital ranks in the top 3 of each district ∪ Children's hospital ∪ Respiratory special hospital)}; U6 = {X | Distance between X＞200 m AND X∈A6}.
7	Residential area	12	A7 = {X | Distance between X and U1＜200 m OR Distance between X and U3＜200 m, X∈ Residential area }; U7 = {X | Distance between X＞200 m AND X∈A7}.
8	School	15	U8 = {X | Distance between X and U1＜200 m OR Distance between X and U3＜200 m, X∈School, in order of priority: Kindergarten＞Primary＞Secondary＞Universities)}.
9	Commercial area	9	U9 = {X | X is the building with the highest population density, X∈Commercial area}.
10	Other important POI	8	U10 = {X | X∈(Corresponding sampling site of national monitoring station ∪ Background site ∪ Museum)}.
11	Road	56	A11 = {X | X∈(Junction of (Expressway ∪ Main road))}; U11 = {X | X is 50/100 metres away from A11 OR X∈A11}.

12	Supplementary point	3	U12 = {X | X∈ POI where four neighbouring grids have no site}.

[a]U_i (i=1, 2, …): ith subset of the set of potential $PM_{2.5}$ sampling sites.

[b]A_i (i=1, 2, …): ith subset of the union of supporting data.

[c]X: element belongs to the set.

Major Issue #4: Modeling and interpretation. The modeling aspect of this paper is not novel. Since the sampling method seems to be a straightforward saturation sampling campaign, using the resulting data to build spatial models is not a novel contribution. Numerous papers have already done this for PM2.5, as noted by the authors.

One main conclusion seems to be that the modeling approaches work. This is not all that novel - it is more a statistical finding than an atmospheric measurement technique. Numerous papers have shown that LUR and kriging models can be fit to spatially distributed measurements.

Another conclusion is that the models work better when provided with more training sites. Again, this seems like an obvious outcome, especially for the kriging approaches.

Response: we agree with the reviewer that the modeling methods themselves were not novel, but the main purpose of this study was to propose strategies of method selection for fine scale $PM_{2.5}$ mapping using crowdsourced monitoring, not to develop a new model. As we mentioned in the manuscript, the optimal mapping method for conventional sparse fixed measurements may not be suitable for this new high-density monitoring way, and the results were rather different from previous studies as we expected. OK interpolation performs best under conditions with non-peak traffic situation in light-polluted period, while the RK modelling can perform better in heavy-polluted period and for conditions with the peak traffic and relatively few sampling sites (less than ~100) in light-polluted period. Additionally, this study for the first time found and pointed out that the LUR model demonstrates limited ability in estimating $PM_{2.5}$ concentrations at very fine spatial and temporal scale which challenges the traditional point on LUR model's good performance in air pollution mapping.

We would not call "modeling approaches work "and "the models work better when provided with more training sites" the main conclusions of this study, but we do admit that some sentences of the abstract in previous version of the manuscript may make this wrong impression on the readers. We thank the reviewer for pointing out this issue.

We changed the abstract as:

Fine particulate matter ($PM_{2.5}$) is of great concern to the public due to its significant risk to human health. Numerous methods have been developed to estimate spatial $PM_{2.5}$ concentrations in unobserved locations due to the sparse number of fixed monitoring stations. Due to an increase in low-cost sensing for air pollution monitoring, crowdsourced monitoring of fine exposure control has been gradually introduced into cities. However, the optimal mapping method for conventional sparse fixed measurements may not be suitable for this new high-density monitoring approach. This study presents a crowdsourced sampling campaign and strategies of method selection for hundred metre-scale level $PM_{2.5}$ mapping in an intra-urban area of China. 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.

A more relevant analysis would be to evaluate if the models (and measurements) make physical dense. In Figure 5 there is a PM hotspot in the northwestern part of the domain on Day 1 and in the center of the domain on Day 2. Do these hotspots make sense given the distribution of sources and the climatology?

Response: we thank the reviewer for the suggestion and sentences addressing this were added in the section of Discussion. The PM hotspot in the northwestern part of the domain on Day 1 may be attributed to the dust deposition from construction activities promoted by a high RH in this newly developed zone, while the PM hotspot in the center of the domain on Day 2 may relate to the larger number of factories and high-density of roads.

As the crowdsourced $PM_{2.5}$ concentrations maps revealed, areas with a larger number of factories and high-density of roads experienced relatively higher $PM_{2.5}$ concentrations, while areas with high levels of green vegetation cover had lower $PM_{2.5}$ concentrations. The relatively high concentration in the northwest corner of the study area with few factories in Period 1 may be attributed to the dust deposition from construction activities promoted by a high RH in this newly developed zone.

Major Issue #5: The paper needs a thorough review and edit for English grammar.
There are many grammar errors (too many to count or enumerate here), and in other places the language is hard to follow.

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.