Answer to Referee #2

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

General comments: Overall this is an interesting paper comparing methods for estimating spatial concentrations of PM2.5 using crowd sourced low-cost sensor measurements. I think it will be highly valuable for many researchers in the field interested in spatial variation. However, I think there is a lack of discussion of the limitations of low-cost optical particle sensors especially with the limited performance evaluation presented in this manuscript. I suggest major revisions for this paper. There are a number of places where the text is unclear and the authors should take care to thoroughly edit the next draft of this paper.

Specific comments:
Abstract: It's not clear what the different periods are referring to, morning versus afternoon? Line 19: I don't think that a range is the best statistic to show that 2 sets of numbers are "clearly different". Line 48: What do you mean by: "and a promising access to the prevention of exposure risks for individuals in their daily life."

Response: the statistic of range was replaced by Mean±SD. Meanwhile, we rewrote these confusing sentences and replaced lines 15 –27 on Page 1 by:

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.

Page 3 line 24-25: What does “data consistency” mean? Can you please elaborate. Also, where do you get the resolution data from? The manufacturer? Lab studies? Please cite.

Response: we rewrote these confusing sentences and more details about this monitor were added.

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 3 Line 30: Why would you only select 30 monitors to collocate? Without the collocation data from the other monitors you have no idea what the bias is of the other measurements.

Response: In fact, the 86 portable laser air quality monitors we used in the sampling were selected from 115 monitors through preliminary indoor and outdoor experiments. The relative errors between each other were no larger than 5%, which guaranteed the reliability of sampling data of the other measurements to a certain extent. Sentences about this were added in **2.1.1 Measurement instrument**. Under the circumstance that the national monitoring stations do not have enough room for more laser monitors to conduct the comparison experiments, we only randomly selected 30 monitors.

Section 2.2.1: Can you mention if these monitors or internal sensors are commercially available or have been evaluated in any other studies, etc. Oh, I see in the supplement they are SDL307 but I think this may be important to add to the text.

Response: added.

Page 3 Lines 28-29: This is confusing to me. I don't see K factors anywhere when I look at the figure. Please clarify this sentence and/or move the figure reference to a more appropriate location.

Response: made the changes as the reviewer suggested in the revised manuscript.

Page 3 Line 30-Page 4 Line 3: I think the performance needs more discussion. How do the monitors compare to each other? If you are looking at spatial variability, bias/error between different monitors will be important. Were all monitors at the reference site for the same period? Is this 1-hr data shown in the plot or some other averaging time? Knowing the bias of individual monitors is very important because it will help determine at what threshold you can say there is likely spatial variation versus just bias in the sensor measurements. In addition, RH is known to significantly influence optical PM measurements. RH should be reported throughout. If RH is >75% during one of the periods (1,2, or comparison) this may be an issue. In addition, you have no data above ~100 ug/m3 but during your second period the concentrations are in the 170-180 range. I think it is important to know how the sensors perform at these high concentrations if you are going to try to draw conclusions. Has any previous work evaluated these sensors at high concentrations? You cannot assume that just because they work well from the 40-100 range they will work the same below and above that.

Response: 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 (including RH) 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.

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 associated with monitors and RH 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.

Replaced lines 30 on Page 3 and lines 1 –3 on Page 4 by:

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}.

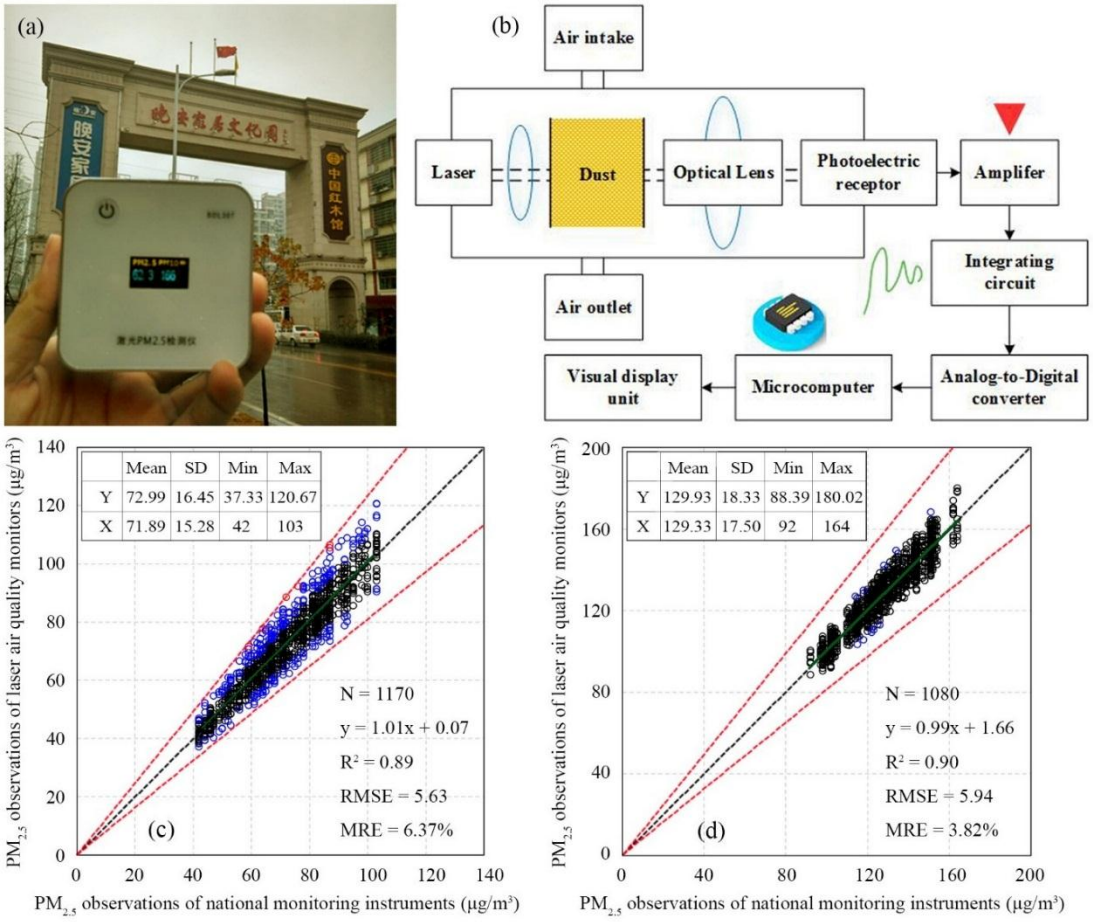

Figure 1: Principle and accuracy of measurement instrument. Y and X are laser air quality monitors and national monitoring instruments, respectively. The black dots, blue dots and red dots indicate $PM_{2.5}$ observations with relative error of <10%, 10%–20%, and >20%, respectively, between two instruments.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

Replaced lines 4 –13 on Page 9 by:

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 $PM_{2.5}$ 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.

Page 5 lines 17-18: Meteorological data with a spatial resolution of roughly 0.4 sites per 100 km2 (wind speed, atmospheric pressure, relative humidity, temperature) that-I think it might be clearer to just list the number of stations you had in total over your sampling area.

Response: changed the sentence as the reviewer suggested:

Meteorological data including wind speed, atmospheric pressure, relative humidity, and temperature of 107 sites in and around the sampling area, which may affect the dispersion of $PM_{2.5}$, were also obtained.

Section 2.1.2: I'm not clear how this data is crowdsourced can you please include more

information about how each monitor got to each monitoring point.

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 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.

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.

Page 6 lines 19-21: Is this the highest and lowest one-hour average from a single site and single monitor? Why are these and the times they occurred important?
Line 20: These what? Averages?
Line 20-21: I don't know what the numbers in parenthesis are please clarify
Lines 25 and 26: Is there more traffic at noon than at morning rush hour? Also is the average concentration at the different hours significantly different?

Response: this is the highest and lowest one-hour average for all crowdsourced sampling sites, we tended to present the rather large range of crowdsourced $PM_{2.5}$ observations. We rewrote the confusing sentence.
"These" are the maximum and minimum values of crowdsourced $PM_{2.5}$ concentrations. Rewrote the confusing sentence.
"numbers in parenthesis" are the mean values and SD of the $PM_{2.5}$ concentrations and the maximum and minimum values of national monitoring $PM_{2.5}$ concentrations. Rewrote the confusing sentence.
There may be more traffic at morning than at noon, the higher $PM_{2.5}$ concentrations at noon than at morning may relate to the peaked cooking emission of stir-fry at noon. The average concentration at the different hours in the same period is rather close. Sentence about these were added or replaced.

Replaced lines 16 –26 on Page 6 by:

Table 3 shows the descriptive statistics of hourly $PM_{2.5}$ concentrations for the crowdsourced sampling sites and the national monitoring stations. Generally, the statistics differed. For Period 1, the mean values and SD of the $PM_{2.5}$ concentrations for the crowdsourced sampling sites ranged from (69.67±18.81) to (76.45±14.55) μg m^{-3}. These values were substantially higher than those for the national monitoring stations (i.e., (36.9±10.97) – (41.2±8.68) μg m^{-3}). The maximum and minimum values of crowdsourced $PM_{2.5}$ concentrations were higher than the national values. However, the mean values and SD of $PM_{2.5}$ concentrations of the crowdsourced sites are lower than those of the national stations in period 2. The former values ranged from (162.72±15.96) μg m^{-3} to (171.89±21.5) μg m^{-3}, while the latter values ranged from (177.8±16.91) μg m^{-3} to (188.3±22.4) μg m^{-3}. Although the minimum values of crowdsourced $PM_{2.5}$ concentrations were also lower than those of the national stations, the maximum values were higher. The average $PM_{2.5}$ concentrations of Period 2 were substantially higher than those of Period 1, and the highest values occurred when traffic and emissions from cooking had peaked (i.e., 12:00 and 18:00) for both periods.

Figure 3. Does each of these points represent a single monitor? Why are they fewer monitors during period 2?

Response: each point represents a single sampling site with no less than three observations for each hour. Because some volunteers quit after the sampling of the first period, sampling sits in period 2 were mainly concentrated in the central study area and thus fewer than in period 1. Sentences about this were added in section **2.1.3 Sampling and data processing** as mentioned before.

Line 13: I don't understand what you are comparing that increased. What is the first set of numbers versus the seconded set of numbers?
Line 14: What do you mean significant and steady decrease? Decreased by hour by the same amount?

Response: the first set of numbers are the average validation R^2 of OK with the smallest number of training sites for each hour; the seconded set of numbers are the average validation R^2 of OK with the largest number of training sites for each hour; rewrote these confusing sentences.
We thank the reviewer for pointing out the fault in line 14; rewrote the confusing sentence.

Replaced lines 7 –29 on Page 7 by:

The box plots of Fig. 4 show the variation in the hold-out validation R^2 for the three mapping approaches in relation to the number of training sites. The average and standard deviation of the RMSE and MRE between the observed concentration and predicted

concentration of $PM_{2.5}$ in the hold-out validation were presented in the Supporting Information (Table S3–S4). The average values and variability ranges of R^2 for OK, LUR and RK were positively associated with an increase in the number of training sites. RK performed best in Period 2 and at 8:00 and 12:00 of Period 1 with training sites less than ~100. The LUR demonstrated the poorest performance for both periods of the models tested.

For Period 1, the $PM_{2.5}$ estimating accuracy was generally highest at 9:00 and lowest at 12:00. The average validation R^2 ranges for different number training sites of OK at 8:00, 9:00, 10:00, 11:00 and 12:00 were 0.58–0.72, 0.56–0.78, 0.51–0.82, 0.47–0.71, and 0.24–0.48, respectively. Compared with OK, the accuracy of LUR was substantially lower. The ranges were 0.26–0.55, 0.29–0.54, 0.16–0.40, 0.16–0.36, and 0.24–0.34. The average R^2 for RK were weakly smaller than OK at 9:00, 10:00, and 11:00 with ranges of 0. 59–0. 69, 0. 50–0.66, and 0. 48–0.60, respectively. The average R^2 of RK at 8:00 and 12:00 were higher than OK when less than ~100 sampling sites were divided into training datasets (8:00: 0.65–0.69 vs. 0.58–0.68; 12:00: 0.40–0.44 vs. 0.24–0.41). For Period 2, the validation R^2 from high to low followed the sequence RK > OK > LUR. The average validation R^2 for a different number of training sites of OK were considerably lower in Period 1. The ranges at 14:00, 15:00, 16:00, 17:00 and 18:00 were 0.25–0.49, 0.34–0.50, 0.40–0.59, 0.27–0.39, and 0.18–0.27, respectively. The average R^2 of LUR were even lower; the lowest values were 0.08, 0.07, 0.15, 0.06, and 0.04, and the highest values were 0.22, 0.25, 0.42, 0.22, and 0.16, respectively. Combining OK and LUR, the performance of RK improved with an average R2 that ranged from 0.43, 0.44, 0.43, 0.36, and 0.32 to 0.60, 0.68, 0.52, 0.54, and 0.57.

Page 7 Line 30: Since readers can see the individual R2 on the figure it may be easier to digest if you just include an average or range instead of so many lists of numbers.

Page 8 Line 5: I read this paragraph a couple times and I'm still a bit confused which method performs the best. Can you add a summary sentence at the end just stating the conclusion? Or reorganize more clearly.

Response: we thank the reviewer for the suggestion. These confusing sentences were rewritten.

Replaced lines 30 –33 on Page 7 and lines 1 –5 on Page 8 by:

Fig. 5 shows scatterplots of holdout-validation results with 90% training sites. For Period 1, the lowest total R2 of OK and the highest total R2 of OK were 0.46 for 12:00 and 0.82 for 10:00 (Fig. 5a), respectively, while R2 of RK were lower with the range of 0.44–0.68 (Fig. 5c); they were both higher than the LUR (0.29–0.53, Fig. 5b). Correspondingly, the RMSE and MRE from low to high were OK (5.95–10.36; 6.80%–9.91%) < RK (8.23–10.92; 9.80%–11.91%) < LUR (10.68–13.16; 12.91%–14.97%). For Period 2, however, the RK presented the highest accuracy with a R2 that ranged from 0.45 (17:00) to 0.66 (14:00) (Fig. 5f). The OK ranked second (R^2: 0.27–0.54, Fig. 5d), while the LUR achieved the poorest performance (R^2: 0.06–0.36, Fig. 5e).

Section 3.3: Can you clarify: did you use 90% training sites for only the sensor measurements and then only 90% of the reference stations? As far as I could tell previously you only used withholding from the sensor data and didn't evaluate the models using the reference data?

Response: the method that performed best with 90% training sites was chosen as the mapping method. Using this method, the spatial distributions of the $PM_{2.5}$ concentration for each hour were estimated with all samples. Spatial distributions of $PM_{2.5}$ concentration for each hour with measurements of 10 national monitoring stations were estimated using the same method for comparison.
We rewrote these confusing sentences in section **2.3 $PM_{2.5}$ concentration mapping** and corresponding results.

Replaced lines 10 –13 on Page 6 by:

The method that performed best with 90% training sites was chosen as the mapping method. Using this method, the spatial distributions of the $PM_{2.5}$ concentration for each hour were estimated with all samples. In this study, nearest neighbour distances between two sampling sites ranged from 15 to 60 metres for Period 1 and 54 to 98 metres for Period 2. Considering the resolutions of the potential predictors, 100 metres was used as the mapping grid size. The spatial distributions of the $PM_{2.5}$ concentration for each hour with measurements of 10 national monitoring stations were estimated using the same method for comparison.

Page 8 line 9: "Significant difference can be found between two sources," what do you mean?
Page 8 Line 13: What do you mean three-step growth?

Response: It means that the hourly $PM_{2.5}$ concentrations for the crowdsourced sampling sites and the national monitoring stations were rater different.
It means gradual growth.
We rewrote these confusing sentences and replaced lines 7 –20 on Page 8 by:

Fig. 6a and Fig. 6b reveal the spatial distributions of OK interpreted $PM_{2.5}$ concentrations for Period 1 from the crowdsourced sampling sites and the national monitoring stations, respectively. Fig. 6c and Fig. 6d demonstrate the spatial distributions of the RK estimated $PM_{2.5}$ concentrations for Period 2. The crowdsourced hourly $PM_{2.5}$ concentration maps demonstrate more detailed intra-urban variations than the national monitoring maps, especially for Period 1.
For Period 1, crowdsourced $PM_{2.5}$ concentrations generally increased from south-east to north-west with multiple hot spots. In the central and south regions of the study area, areas with a larger number of factories that experience a relatively higher $PM_{2.5}$ concentration than other areas. The national monitoring $PM_{2.5}$ concentrations, however,

were less than 55 μg m^{-3} with limited spatial variation. For Period 2, with the exception of 14:00, the national monitoring $PM_{2.5}$ concentration maps showed high-east and low-west patterns. $PM_{2.5}$ concentrations of central Yuelu district were rather low (<175 μg m^{-3}). Crowdsourced $PM_{2.5}$ concentrations demonstrate extensive cold spots of $PM_{2.5}$ concentrations in southern Changsha County and the southern Kaifu district, while southern Yuelu and western Tianxin with a high-density of factories and roads were hot spots of $PM_{2.5}$ concentration.

Page 8 Line 15: I don't understand based on the figure it seems like there are almost no factories and roads in the top left corner but that is where most of the pollution is.

Response: 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. Sentences addressing this were added in the section of Discussion.

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. This finding suggests that optimising the distribution of land use may improve the air quality to some extent and strengthening the control of local emission may be the primary way to reduce pollution in the light-polluted period. As the urban air quality grade has an important effect on the spatial distribution of samples (spatial autocorrelation, and heterogeneity), which may also be affected by sample size, the mechanism for this influence is somewhat equivocal and needs further research.

Page 8 Line 30-Page 9 Line 3: I think you need to mention though the limitations of low-cost monitors and the inaccuracies in these measurements compared to federal methods.
Page 9 Line 10: It seems likely the low-cost sensors may have been saturated at the high concentrations and this may have led to the difference between the sensors and the reference methods.

Response: we thank the reviewer for the suggestion. Sentences about this were added.

Replaced lines 4 –13 on Page 9 by:

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.

Technical corrections:

Suggest rewording the title for clarity, possibly: Strategies of method selection for Fine Scale PM2.5 mapping in an intra-urban area using crowdsourced monitoring

Response: changed as the reviewer suggested.

Fine particulate matter (particulate matter singular remove s)

Response: corrected.

Line 9: to “the” public – there are a number of grammatical errors throughout the text and I have not had a chance to identify them all in this review. Please review for grammar.
Page 6 Line 20 ug/m3 formatting
Page 7 Line 4: Remove “had” assuming you are talking about this work where the sites experienced extreme PM

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.