Response to reviewer #1

I want to thank the reviewer for the useful comments. It resulted in a revised manuscript which puts more emphasis on the added value of low-cost sensors by including results for different assimilation configurations. Also, a sensitivity study on traffic emissions have been added, and the accuracy and limitation of the method are better discussed. To put more emphasis on actual measurements governing the spatial interpolation, the subsection on observations is now put forward as an independent section. New figures have been added to better support the content. Results which are considered of importance but distracting from the main argument have been moved to Supplemental Material. The Discussion section and Conclusion section have been joined together and rewritten.

Below is my response. In blue the original comments, and in black the answers.

I favor the publication of the manuscript in AMT after the following two main issues and the specific comments have been carefully addressed and the manuscript has been substantially revised.

1. Due to the modularity of the modelling framework it is necessary that the accuracy and limitations of the individual components are determined and described in more detail and a little bit more critical. For example, the used dispersion model is computationally fast but it seems that it has clear limitations when applied in an urban environment as e.g. building geometries are not an input. Similarly, the emissions used as input for the dispersion model are derived in a simple way that similar data is probably available also for many other cities but the accuracy of the derived emission inventory seems by far not to be optimal. I think that the gain in accuracy of a spatially variable pollutant field by assimilating measurement data strongly depends on the model's capability to resolve small scale structures. It should be made more clear in the manuscript if measurements adjust local deviations in emission source activity or only general model deficiencies. In the latter case the assimilation of measurements does not necessarily lead to throughout improved results in a local environment around the sensor.

The revised manuscript elaborates on the comments above. The assimilation corrects both general model deficiencies and local deviations. The leave-one-out validation results show that in most cases the assimilation improves the accuracy, by reducing the local modelling bias, and improving the precision. This is better shown by including histograms of Observation minus Forecast (OmF) and Observations minus Analysis (OmA) for all validation locations. The potential shortcomings of the method are better discussed. Further reduction of the bias will be the subject of future studies, concentrating on (1) improving the modelling of traffic emissions, (2) including street canyon effects, (3) improving NO_X chemistry (ratio NO_2/NO_X and lifetime), (4) improving the modelling of the model covariance.

2. The validation part should be extended. The model uses a proxy for residential emissions as input data. Residential emissions probably have a distinct seasonality. Therefore, I strongly recommend to use an additional winter period to validate the modelling framework and to analyse the resulting alpha_pop. Actually, the simulation of a whole year would be best. No low-cost sensor data are required for this analysis.

I have added an analysis of the seasonality of the emission proxy calibration in Section 4.1. It compares a summer period (July 2016) with a winter period (January 2017). The results show that the average emissions do not agree well with the expected emissions from a bottom-up inventory. The regression analysis of the dynamic calibration finds the best linear combination of traffic proxies and residential proxies which explains the observations. Unlike traffic, the diurnal cycle for the residential contribution is shaped by the regression analysis. The seasonal analysis shows that the fitted diurnal cycle for the residential sector not only describes the cycle of the residential emissions, but also compensates for changing NO_2/NO_X ratios over the day due to changing photochemistry and temperature. Also, due to collinearity, part of the traffic emissions can be explained by population density. Therefore, the found emission factors (and the corresponding sectoral emissions) should be considered as "effective" rather than real, i.e. factors which best describe the observations under the given model assumptions.

Specific comments

Introduction (section 1) or Setting up an urban air quality model (section 2): For completeness, a short discussion/list of other dispersion models that could be used as an alternative to AERMOD should be included.

Added to the end of the introduction in Section 3: "Note that any other dispersion model can be used in the Retina methodology, as long as it is capable of simulating concentrations from individual emission sectors on an arbitrary receptor mesh."

Traffic emissions (section 2.2.1): The author writes that traffic emissions are the dominant source of NO2 in the Amsterdam domain. Hence, traffic emissions are an important input factor for the simulation of the NO2 concentration. The interpolation of the vehicle counts for arbitrary locations based on the counting sites using IDW is practical but, at least for urban roads, possibly limitedly accurate as network characteristics are neglected. When I think of parallel roads in close distance, IDW would assign them similar vehicle counts, but in reality the true counts can be very different. An analysis/description/discussion of the accuracy of the resulting traffic input data should be added.

Indeed, the question remained how well this IDW interpolation describes the traffic flow differences found for nearby roads of the same road type. This is now assessed by two different approaches (presented in the Supplementary Material, and mentioned at the end of Section 3.2.1): a leave-one-out validation to study the error in local traffic flow estimations, and a concentration validation study of dispersion simulations done under different traffic scenarios. The results show that for this counting network IDW predicts the traffic volume within a 50% error margin at most locations. The model simulations show that using inferior traffic data is partly compensated by the calibration dynamics, at the expense of less pronounced concentration gradients.

Population data (section 2.2.2): The indication of the magnitude of different contributions (heating, cooking, others) to the total residential emissions in Amsterdam would be helpful. For the all-season applicability of the model: does the population database also include the spatial distribution of employees to account for heating emissions of office buildings? And for the selected period: Are heating emissions substantial in this summer period?

This is now assessed in newly added Section 4.1, where the diurnal cycles of sectoral emissions are analysed for summer and winter. There is no clear answer to the reviewer's questions, as the NOx emissions used by the model after calibration should be regarded as “effective” rather than real, i.e. best describing the spatial concentration patterns under given model assumptions. The effective residential emissions contain an unquantified contribution which compensates for the simplified NOx chemistry assumptions.

Calibrating the model (section 3): First, lines 217 to 219 are unclear for me. After reading these sentences I was confused if cj(t) in Eq. 6 is the measured NOX concentration. But it is not, correct?

The confusion arises from the fact that *P* represents proxies for NO_x emissions, while *c* represents the observed NO_2 concentrations. The conversion from NO_x to NO_2 is implicitly done by dispersion *f*, using an NO_2/NO_x ratio from the Ozone Limiting Method.

In Gaussian dispersion modelling there is a linear relation between emission strength and concentration, but this linearity breaks down when conversion from NO_x to NO_2 is included. This is now better formulated by eliminating lines 217 to 219 and changing the description of Eq. (6) to: “(...) with f_{ij} describing the dispersion of a unit emission from *i* to *j*, including the conversion from NO_x to NO_2 from the OLM. Eq. (6) is assumed to describe a linear relation between emission and concentration, although strictly speaking the variable NO_2/NO_x ratio introduces a weak nonlinearity.”

Second, can you comment on how worse the model is performing when residential emissions are omitted? I guess that in the selected summer period heating emissions are nearly zero. Are the estimated two-hourly alpha_pop plausible and can you show that the temporal pattern of the values are related e.g. to cooking emissions? In Figure 3b, the contribution of residential emissions to the NO2 is surprisingly large given the fact that residential emissions are only 1/3 of the traffic emissions (stated in section 2).

This is now assessed in Section 4.1, where also the collinearity between the traffic and residential proxies is mentioned.

Assimilation of observations (section 4): The described algorithm is applied by using the pollutant concentrations transformed into the log-space. Here, one has to be aware that the distribution of the pollutant concentration at a particular location is not equal to the measurement error that is required for the algorithm in this section. The measurement error is described in the manuscript as being dependent on the concentration (section 2.3). So, the reasoning of this transformation is not correct. However, I suppose that the transformation of the measurements into the log-space has a positive effect on the stability of the results as the modeling framework becomes less sensitive to (less frequent) measurements of high concentrations by reducing their impact. The transformation into the log-space is fine, but the respective paragraph should be reformulated.

The reviewer is right. Changed “*The analysis is therefore done in log-space ($z_j = \ln c_j$), which converts lognormal distributions to Gaussian, for which the Bayesian assumptions behind Equation 8-10 are valid*.” to “*The analysis is done in log-space ($z_j = \ln c_j$), stabilizing the results by reducing the impact of less frequent measurements of high concentrations.*”

Modelling the model error covariance matrix (section 4.1): The interpolation of the model error by IDW might result in an error field that is too smooth in the urban environment given that the model is limitedly capable to represent small-scale structures (e.g. buildings). At least a comment should be included in the manuscript that points out this issue.

It is impossible to assess the model error at all locations and under all conditions when the "ground truth" is only available at 15 locations. In my opinion, interpolation of the model error gives a good first impression of the local model performance away from validation locations. However, the reviewer is right that small-scale structures provoked by the local built-up area might not be well represented in the model. Inclusion of the street canyon effect, together with refinements in covariance modelling, is subject of further study. It will reduce local bias in the modelling, but also suppress the introduction of bias by the assimilation. This is now pointed out in the discussion section.

Validation (section 5): The first paragraph is, as I understand, only an example where the modelling framework works well. It can be removed. Start with the second paragraph ("overall assessment"). Figure 6 can be presented directly after the overall assessment by discussing sites where the model performs well (NL49019) and where it performs less optimal (e.g. NL49002, NL49014). Figure 6 should include examples for both types. The time period of the used data in this Figure should correspond to table 3. Omit in the Figure the performance analysis of low-cost sensors but extend chapter 6. Moreover, add a file to the manuscript with supplementary materials where the scatter plots of the remaining, in the manuscript not presented air quality monitoring sites are shown analogue to Fig. 6 in order to provide the reader a clear picture of the model performance.
The reviewer's suggestions have been implemented. The first paragraph has been deleted. Figure 6 has been replaced by two validation examples, for a well performing location (NL49012) and a worse performing location (NL49014). The time series plots have been removed, regarded as redundant as the performance can also be read from the scatter plots. Bar plots with error distributions have been added to better illustrate the effect of the assimilation. The validation plots for all reference locations are included in the Supplementary Material.

Added value of low-cost sensors (section 6): The material presented in this chapter is only qualitative. The two average concentration maps presented are not validated and so the accuracy of their differences is unclear. The single example of the "Oude Schans" site is not sufficient to show that low-cost sensors add value. I have some questions here: What is the reasoning of largely adjusting the results of the dispersion model by low-cost sensors when there is also the option to improve the input data for the dispersion model?
To start with the last question: better input data is not always available. In this particular case there was no detailed information available about traffic flow and changing traffic patterns. To better assess the added value of low-cost sensors, additional results have been added from different assimilation configurations. The different assimilation scenarios show that low-cost sensor data assimilation improves the results locally, even in absence of reference data. Generally, the best results are obtained when both reference data and low-cost data are included. This is the configuration used to obtain the map showing the NO_2 reduction during the holiday period (Figure 11).

Is it possible to generate traffic input data for the dispersion model for each month based on the traffic counts you have access to?
Yes, it would be possible to generate monthly traffic data. This is done for instance in traffic scenario TS3, which can be found in the Supplementary Material. However, as both the holiday period and the closure of the tunnel started at half July and ended half August, neither traffic data for July or August would be representative for the whole month.

Anticipating the application for other cities where such detailed traffic data might not be available, it was decided to evaluate the system for a yearly averaged traffic “climatology”.

Moreover, NO2 concentrations also depend on meteorology. What fraction of the differences in the monthly aggregated concentration fields is related to different weather conditions?
The influence of meteorological variability is now included in the discussion of the figure. It can be estimated from the NO_2 reduction found at rural stations NL49565 and NL49703. Average values drop from 16.60 and 12.57 ug/m^3 during 15 June - 15 July, to 15.53 and 11.55 ug/m^3 during 16 July - 15 August. Added to this section: “*Based on averaged NO_2 measurements at rural stations NL49565 and NL49703, the NO_2 reduction due to meteorological variability is estimated to be 7%*.”

As I understand the main benefit of the low-cost sensors for the modelling framework is the increased spatial resolution of the measurements. Here, I miss some sensitivity analyses or similar material regarding measurement network design. What options exist when using this modelling framework in reducing traditional air quality monitoring sites and adding low-cost sensors? The accuracy of low-cost sensors is reported to be about 30%. Is this enough for adding substantial information?
These questions are now implicitly answered by the results of the different assimilation scenarios. Validation results are comparable when only observations of 14 low-cost sensors are assimilated (AS3), instead of observations at 3 reference sites (AS2). Even a notable improvement is visible in bias and RMSE at location NL49019, where the low-cost sensors are relatively nearby.

Technical comments
page 4, lines 102-103. How are the parallel distances of 75 and 125 m related to the grid? Maybe reformulate this sentence to make it clearer.
Changed to: “Receptor locations are chosen at every 75 m along the parallel curves with 25 m distance to the road, and at every 125 m along the parallel curves with 50 m distance to the road.”

page 5, lines 143-144. Refers traffic "climatology" to counting sites?
Changed “For each location” to “For each counting site”.

page 7, line 197. Mijling (2018) instead of Mijling (2017)?
Changed to Mijling (2018)

page 7, lines 204-205. Pik, Pki: keep consistent.
Adapted

page 7, lines 213-214. I would not say that b(t) is observed. It is rather the output to another modelling system.
Agreed, changed to “*Note that both background concentrations b(t) and local concentrations $c_j(t)$ are taken from external data*”

page 9, line 266. Section 3 instead of Section 2.3?
c refers to the measurements of the reference network, described (according to the new numbering) in Section 2.

page 10, line 285-286. "Isotropic" is the wrong word here as it is not isotropic.
True. Changed to "*The correlation of model errors between different locations is parametrized with a downwind correlation length* L_{dw} *and a crosswind correlation length* L_{cw}*.*"

page 11, line 336. Change to "lower accuracy".
Changed accordingly

Figure 1. Add units to the x and y axes.
Units are mentioned in the figure caption.

Figure 2. What do the depicted lines show? Sample week, yearly aggregation? Add more information.
Caption changed to "*Weekly cycle of highways and urban roads at counting locations, aggregated from hourly data from 2016. The morning and evening rush hours on working days are clearly visible for highways. Urban traffic has, apart from lower volume, less distinct peaks. The thick lines show the median of traffic flow for both road types.*"

Figure 3a. Add north arrow and scale in one of the four Figures. The meaning of the three dots should be explained already in the caption of Figure 3a.
Scale bar and arrow added. The figure caption now explains the three dots.

Figure 3b. Indicate in the Figure in an appropriate manner the weekday the dates refer to. 2016-07-07 –> Thu, 2016-07-07. In Figure 2 the distinct traffic pattern is shown. It is interesting if there is a clear relation between traffic and NO2 in the modelling results.
Labels on x-axis of Figure 3b changed.

Figure 5. Add units for x and y axes and scale in all Figures. Remove the first "and" in the second sentence of the caption.
Figure updated, scale bars added.

Figure 7. Add scale in both Figures. Moreover, the visibility of the points could be better in all the presented maps.
Figure 8. Add location of IJ-tunnel and of the historic center.
A new panel has been added showing the location of reference stations and low-cost sensors, the location of the IJ-tunnel, and the scale.

Table 3. Indicate more precisely the date period the analysed measurements refer to.
Table caption changed to "*Validation results at reference locations, June 1 - August 31, 2016*"