Response to reviewer #2

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.

One of my main concerns relates to the lack of detail in some of the sections, but particularly in the section that is supposed to demonstrate the added value from assimilating low-cost sensor data (Sec 6). Given that multiple previous studies have already successfully assimilated regular stations observations using OI (e.g. Tilloy et al. 2013), one of the more novel aspects of this study is the assimilation of low-cost air quality sensors. The manuscript goes through great lengths of building up a dispersion model system with simplified emissions as well as an OI assimilation scheme, but the added value from low-cost sensor networks is covered in just a few lines towards the end without much detailed analysis. I think the manuscript could be a lot stronger and have more impact if a more comprehensive analysis of this were carried out in Section 6.

The work of Tilloy et al. is now referenced in the manuscript. This work is different in the sense that it presents a flexible urban dispersion model, it presents an alternative covariance model, and it studies the added value of low-cost sensors (details below). The section on the added value of low-cost sensor data has been extended with a sensitivity study of different network configurations.

Secondly, Section 2.2.1 on traffic emissions (which are crucially controlling NOx/NO2) left me scratching my head at times. I realize that the system is designed to be portable and thus the necessary input data should be kept to a minimum, but I wonder if some of the simplifications taken here are defensible. In particular spatial interpolation of traffic monitoring sites seems to me a quite crude approximation that introduces significant uncertainties in the modelling.

Spatial interpolation of traffic flow might indeed seem a crude way of solving the lack of traffic data. However, part of the introduced inaccuracy is avoided by distinguishing between two road types, highway and urban roads. Both having different diurnal patterns and vehicle counts, the traffic flow interpolation of one road type does not affect the other road type. The validity of the approach is now assessed in two different manners, 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.

Further, what about distinguishing different types of vehicles? Regular cars versus heavy trucks? Euro 4/5/6 emissions categories? These things can have a very significant impact on the modelling results for NO2 and I wonder if some more care in setting up the modelling would not be beneficial in the long run? This is particularly a concern in the sense that OI should technically only be used when model and observations are unbiased against each other and ignoring certain high polluting vehicle classes could introduce potentially damaging biases.
The reviewer is right. We consider only two emission factors, one for highway traffic and one for urban traffic. The emission factors, however, are estimated from analysis against observations in the calibration phase, which implicitly compensates for i.e. different fleet composition. Further refinement of the traffic model is desirable (e.g. based on the COPERT database), and will definitely improve local model performance. However, introducing a detailed traffic model, including fleet composition, was considered outside the scope of this work.

At the very least, the author should discuss these potential issues and lay out future steps to resolve these problems. In the best case scenario, it would be good to see some sensitivity studies testing the modelled NO2 sensitivity to inclusion of these different classes.
New validation results in the Supplementary Material show that the current method of interpolated traffic flow predicts the traffic volume within a 50% error margin at most locations. Better results are obtained when more counting locations are available, or when they are selected strategically around crossings and access roads. It is now remarked in the Discussion & Conclusion section that improved traffic emission models should take local differences in local fleet composition into account.

Thirdly, I feel that the manuscript could benefit from some more detail on how the error characteristics of the observations were derived. Estimating uncertainties from reference instruments and low-cost sensors on its own is a difficult subject and the paper does not provide the reader with information on how these were estimated or how such uncertainties were then transformed into error characteristics suitable for ingestion in the OI scheme. Such a discussion should be included in the paper and I believe this would strengthen the authors conclusions.
The section on observations has been elaborated hereupon. The accuracy of reference instrumentation is determined following the EN 14211 standard (now referenced in the section on observations), which includes all aspects of the measurements method: uncertainties in calibration gas and zero gas, interfering gases, repeatability of the measurement, derivation of NO_2 from NO_X and NO, and averaging effects. The error in the low-cost sensors is determined by side-by-side comparison against reference instruments. Details can be found in the referenced Mijling et al. (2018) paper.

Finally, given that 90% of the paper deals with modelling and data assimilation, I do find the choice of AMT for this manuscript slightly puzzling and I feel that the paper would probably be better suited for a journal more focused on modelling or general air quality issues. However since the editor has accepted the paper to AMTD I assume that the material is considered suitable for the journal.

AMT was chosen because the study describes an observation-driven framework which can be used for processing air quality measurements of different sources. Therefore, it was felt that it is of interest to the air quality measurement community. As the involved data assimilation sits on top of both observation and model, the current work would neither fully qualify for an air quality modelling journal. To put more emphasis on the importance of observations, Section 2.3 on observations is now put forward in the manuscript as an independent section.

DETAILED COMMENTS

L15: "Retina" - why is it called this? Include the full name if this is an acronym.
The algorithm's name reflects its ambition to produce high-resolution imagery; it is not an acronym. A simple name was preferred above a badly constructed or unpronounceable acronym.

L16/17: how are these percentages to be interpreted? Would be good to mention here how accuracy is defined. Something as simple as "... a typical accuracy (defined here as [...]) of 39%" or similar
Added "(defined here as the ratio between the root means square error and the mean of the observations)"

L23: "enhanced understanding of reference measurements". Please [missing]

L38: "adding value". I suggest you give an example of what you consider as adding value to the measurements or otherwise better write "exploiting the measurements"
Changed to "exploiting the measurements"

L39: In single-author papers it looks quite odd to use plural terms such as "our" and"we". Consider revising.
I replaced the inappropriate use of *pluralis majestatis* by the passive tense.

L43: I would add here that it depends on the mapping resolution and the pollutant. The required sampling density increases with the desired spatial resolution of the map. Furthermore, NO2 with its very sharp spatial gradients will always require a much denser network than for example mapping PM2.5 with its relatively smooth spatial gradients.
Added: "To obtain high-resolution information *of air pollutants with sharp concentration gradients*, (...)". From the first paragraph it is clear that is especially the case for NO2 concentrations, "which can vary considerably from street to street".

L60: The introduction/background section is missing a reference and a discussion of Tilloy et al (2013) (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/jgrd.50233),who have essentially done the same as this paper (OI of point-based observations into an urban-scale AQ model).
The paper of Tilloy is now referenced twice.
In the Introduction: "*Tilloy et al. [2013] use the 3-hourly output of a well-developed implementation of the AMDS Urban dispersion model in Clermont-Ferrand, France, to assimilate in-situ NO2 measurements at 9 reference sites in an optimal interpolation scheme. With a leave-one-out validation they show a strong reduction in root mean square error of the time series after assimilation.*"

In Section 5.1 on the modelling of the error covariance matrix: "*Tilloy et al. [2013] choose to model the covariances depending on the road network. Error correlations are assumed to be high on the same road or on connected roads. For background locations, the correlation decreases fast in the vicinity of a road, while the error correlation between two background locations remains significant across a larger distance. The error covariances are kept constant in time, and taken independent of traffic conditions.*"

L61: Again, the name "Retina" comes a bit out of the blue. You should probably introduce here what the acronym stands for.
Retina is not an acronym, please see the answer above.

L70: I would be a bit careful with the term "calibration" in this context, given that it has a very specific meaning for measurements (both reference as well as sensors). Maybe reword or describe a bit more thoroughly what happens in this step.
I prefer to stick to *calibration* here, as I think that within the context of model calibration it is sufficiently clear that it refers to adjusting model parameters to best match the evaluation criteria.

L98: A reference to AERMOD would be useful here.
AERMOD has already been introduced and referenced shortly above (Cimorelli et al., 2004). This particular version of AERMOD has no specific scientific reference.

L99: "local equidistant coordinate system" - at this point you might as well give the actual projection you used. Presumably something UTM-like?
Clarified to: "*All coordinates are reprojected in a custom oblique stereographic projection (EPSG:9809) around the city center coordinate, such that the coordinate system can be considered equidistant at the urban scale.*"

L100: "road-following grid". This is used as if this a commonly known term, which in my opinion it is not. So first of all you might want to introduce this term a bit more carefully by saying something like "we use a road-following grid, which is essentially....". Secondly, to me it sounds a bit weird to use the term "grid" in this context, when you are basically talking about a spatially irregular and scattered set of receptor points with higher density along road links. I think the term grid should be reserved for a somewhat regular arrangement of cells.
The occurrences of *road-following grid* are replaced by *road-following mesh*.

L148-152: I realize that the goal of this paper is not to build the world's best model so a certain amount of simplification is expected, but interpolating traffic flow using IDW seems to be an incredibly crude method. How can this method possibly work?Between two loop counters there will likely be many road segments that either have much more or much less traffic than at the observation sites, so I fail to understand how simply interpolating here can lead to useful results. I think this section needs more detail on how this is carried out and a robust demonstration that the chosen methods are meaningful.
This concern is shared with Reviewer #1. In the revised article, the traffic model is assessed in 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.

L164: Don't they assimilate UTD data? In that case it wouldn't be a day old but just a few hours (maybe better write "up to a day old" or so).
Clarified to: "*The analysis of the ensemble is based on the assimilation of up-to-date (UTD) air quality observations provided by the European Environment Agency (EEA).*"

L230: This should be Figure 3b? Also, I think this Figure should be discussed a bit more (maybe in the discussion section?) for example with respect to potential reasons for the difference between model and observations, particularly for the highway location.
Reference to Figure 3b now included. Differences between model and observations are now mentioned in the discussion section and mainly attributed to its inability to resolve all small-scale structures provoked by local built-up area, and sketchy traffic emission modelling. The latter is especially true for highway location NL49007, which is very near this strong source.

L235: I recommend to remove the term "geostatistical" here. While OI is mathematically very similar to kriging-based techniques (and it can in fact be shown that it provides identical results to kriging if the same inputs are used) it is not traditionally considered a part of the field of geostatistics. Geostatistics was developed in the mining and earth resources community (Matheron et al.), whereas OI was developed within the meteorological community (Gandin 1965).
Rephrased to "the interpolation technique of choice here is".

L241: Again, I would not use the term "grid" for what is essentially a set of irregular, scattered receptor points.
Grid replaced by *mesh*.

L265: I think it would be good to mention here that Statistical Interpolation/OI is essentially the same assimilation scheme (just a different mathematical framework) asprevious kriging-based approaches. The main advantage of OI over geostatistics (but also an added complexity) is that one has detailed manual control over the Pb covariance matrix, which allows for a more comprehensive specification of the area of influence for each contributing observation.
This valuable remark has been included in the beginning of the section, where OI is first introduced.

L287: extend -> extent (or maybe magnitude?); also reflect -> reflects
Corrected

L335-340: I think this section should be either left out entirely or expanded upon significantly. As it is currently it does not represent a robust demonstration that low-cost sensors add value to the system, since the effect has only been shown at a single site and not been analysed in detail. Demonstrating that the information from low-cost sensors can improve urban-scale air quality modelling is clearly a very worthwhile goal but this short section reads unfortunately more like an afterthought than a proper analysis.

This section has been expanded with new material: an additional analysis and discussion for 5 different assimilation scenarios. The different assimilation scenarios show that low-cost sensor data assimilation can improve the results locally, even in absence of reference data.

L354: I think it would also be worthwhile noting here that, while CAMS is definitely useful for providing background conditions and initial conditions, for NO2 the CAMS forecasts can be very misleading when interpreted at the local scale. The predicted diurnal cycle can often deviate substantially from that observed at urban AQ stations.
This valuable remark has been added in the discussion.

L356: Agreed. And in addition the higher resolution from a dispersion model is also much more appropriate than CAMS for applications such as exposure estimates etc.

L374: "Traffic data tend to be harder to obtain". That is very true (and maybe even an understatement) and is one of the most limiting factors in running local-scale dispersion models at random locations. Given that traffic is typically the most important source for NOx I have the suspicion that even the comparatively portable Retina methodology is likely to fail when no such traffic data is available at all. I think it would be worthwhile discussing here that at some point, if nearly all the crucial input data to Aermod is either of low quality or entirely missing, the resulting forecasted concentration fields will be so bad that any type of sophisticated data assimilation of observations is no longer very meaningful.
A traffic degradation study is now included in the Supplemental Material. From this can be seen that the calibration is partly capable of compensating for inferior traffic data by relating traffic emissions more to population density, at the expense of higher RMSE and lower correlation. In general, degraded input data and imperfections in the dispersion modelling will deteriorate the system's capability to resolve local structures; it will lower the effective spatial resolution of the simulations. In its extreme it will only describe the blurry urban background pollution contribution added to the rural background. Oppositely, with improved input data and atmospheric modelling, the effective resolution will improve. This insight has been added to the discussion.

L396: It might be more detailed, but is it really much better? This section is too qualitative to draw much of a conclusion. As I said above, I think the manuscript would benefit from a more robust analysis along these lines.
This is now better addressed in the section on the added value of low-cost sensors.

Figure 2: The caption should indicate more clearly that the thin lines represent the traffic at individual stations.
The caption is rewritten.

Figure 3a: These maps would benefit from some basic cartographic elements, e.g. a background map from OpenStreetmap similarly to Figure 7/8, scale bar, coordinates etc.
A scale bar and North arrow have been added. A cartographic background has not been added as the domain can now be inspected by comparison with new Figure 1.

Figure 5: "Units are in meters" - not all of them. I recommend to either label all axes properly or to have a more thorough caption describing the various elements of this busy Figure in

more detail. It would also be helpful to have labels for each subplot (a,b, c..) so that they can be better referred to in the caption.
The representation has been improved by removing distance units on the axes by scale bars, by better formulating plot titles, and adding grid lines. The figure caption has been reformulated.

Figure 6: I think it is a bit confusing that the time series is only for 8 days, whereas the scatter plots show an entire month of data. Why not show the time series also for the entire period? If it is a visualization issue, it could be plotted over multiple rows. Similar to my earlier comments I also think that the analysis here would benefit from looking at more than just a single station.
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.

Figure 8: "IJ-tunnel" Should be marked on the map since non-locals will not be familiar with this.
A new panel has been added showing the location of reference stations and low-cost sensors in the central area, and the location of the IJ-tunnel.