Suggestions for revision or reasons for rejection

Reviewer 2 (RC2).

AC. Firstly, our thanks once again for your time in preparing these detailed thoughts comments.

1. I find the terminology used for "corrected" sensor data confusing and would prefer to have a terminology that is consistently used throughout the paper (would make reading easier). In Figure 5, the different stages of the applied baseline correction lead to "corrected sensor data". In the following RF modelling step (in particular Fig 6) the baseline corrected data is now considered as "uncorrected data" and the RF step again leads to "corrected data". The authors realise this issue in Figs. 9-11 where the terminology "uncorrected-baseline-normalised" and "fully corrected" is used. Please resolve this and used something like "baseline corrected" and "final (fully) corrected".

AC. Agreed, we have revised the manuscript using the terms below

Full and final corrected (lines 20, 404, 704, 709),
fully corrected (lines 323, 335, 350, 355, 376, 383, 385, 400, 401, 449, 454, 465,468,469, 472, 668, 670, 676, 679, 687, 688, 694, 696, 702, 708),
uncorrected (line 355, 653,656, 660, 670, 676, 679, 687. 694, 697),
part-corrected (lines 298, 309),
baseline corrected (line 667, 687),

AC. In updating figures to accommodate the vocabulary above we have noticed discrepancies in some of the statistics presented. We have updated the manuscript throughout in the interest of transparency & reproducibility. Core conclusions are unchanged.

2. Section 3.2.4 Sensor performance vs. European air quality data objective. I'm still convinced that the validation data set cannot and should not be used for determination of measurement uncertainty. The validation set is an integral part of the RF model building process. It is a random sample of the data used for model training and should only be used for deciding on the RF model parameters and for comparison of different models or modelling approaches. Performance assessments based on the validation should not be considered as representative for the performance that can be achieved/expected in independent (real-world) measurements. The values are too optimistic. The authors should make this more clear and present and discuss the results/sensor performance based on the numbers in Table 7 (the MAE's from Table 7 are mentioned in the abstract, which is good and correct!)
I think it is generally fine to leave Table 6 in the paper as it provides information on the effect of the performed data correction method. Corresponding numbers are presented in many other papers and they are useful for comparison. However, the authors should make clear that these numbers should not be interpreted as the uncertainties to be expected in subsequent atmospheric measurements as they are too optimistic and not representative for sensor applications. Indeed, I think the comparison of numbers in Tables 6 and 7 is useful for readers who intend using sensors for atmospheric measurements. The authors should make clear that the special situations during the collection of unseen data in December is not an exception but the rule, when using sensors in a real-world setting and therefore should be included in the performance assessment.

AC. Thank you for this comment. We have revised the manuscript to clearly state uncertainties in validation estimates

Other comments:

Page 5, section 2.2, fourth line. Typo, should be "were" instead of "work".

AC. Corrected, thank you.

In section 3.2.3, page 13 the authors now have a sentence "… within about 1ppb (NO2) and 2-3ug/m3 (PM) of the MAE returned by the model validation set.". Where do these numbers come from? – Prob-ably from Table 4, but the reader does not know as numbers for PM are in Table 4 different. Same in the Conclusions section.

AC. Thank you. Revised with the inclusion of 1 decimal place & cross-reference to source Table 6 for clarity.

Page 11, second paragraph, first line, "Fig. ", number is missing.

AC. Page 11, line 1 references "Figs. 6-8" the version I am editing. Perhaps a pdf issue. We will review to ensure the issue does not perpetuate.

Figure 5. Should be indicated what is shown here. Also rolling 3h means as in Fig. 4, 15-minute mean values, or different?

AC. Thank you. We have revised the title of Fig 5 to indicate 15-minute mean averaging period of the data shown.