Dear Dr Herckes & Editorial Support

Thank you for your feedback (dated 22 Mar 2022) on the latest draft of our manuscript (AMT-2021-282).

Firstly, may I extend my apologies for the extra resource this issue has required. I hope the additional context below and commentary attached helps to move it along in a positive way.

First identifying the discrepancies

We first found discrepancies in the manuscript whilst responding to reviewer#2 comments posted on 4 Mar 2022. The discrepancies arose as a direct result of re-running code to reproduce colour-blind accessible figures for the manuscript. After they were identified, we requested to extend the deadline for responding to reviewer #2's comments, as the MS records will show. This allowed us time to fully consider the implications of the discrepancies prior to resubmission.

Diagnosing the issue

Despite our code base being segregated from other research code base(s) to protect it from inadvertent changes, it is not part of a formal subversion system. We have identified that a proportion of the random forest (RF) model code was changed in error resulting in the training parameters for the RF models being overwritten. The timing of the overwrite appears to have been late on in the code base development – sometime after our preferred RF model configurations had already been agreed and documented, but before the results were compiled. As a result, the training parameters presented in Table 3 are correct, but do not deliver the results presented in Tables 4, 5, 6, & 7 in a reproducible way. We believe the change occurred during the preparation of follow-up research - the timing and type of changes reflect this follow-on work, and we anticipate that that code base was copied for re-use / development on the downstream project. After copying of the code to a new development environment for modifications, we believe that changes were subsequently saved back to the original location in error.

Safeguarding for the future

We will use code repositories e.g. GitHub, to further protect against similar mistakes in the future. At the time of this current analysis we did not have this procedure in place.

Progressing the manuscript

We have since reconstructed the RF models according to the parameters set out in Table 3 and achieved reproducible results. These are now reflected in the most recent version of the manuscript uploaded on to AMT on 21 Mar 2022 and available in track-changes at
https://editor.copernicus.org/index.php?_mdl=msover_md&_jrl=400&_lcm=oc3lcm4w&_acm=get_file&_ms=97838&id=1905110&salt=6825354511725275278.

General comments on changes identified

We have carefully considered the implications of the changes prior to re-submitting the manuscript. A point-by-point review of the changes required, is provided in the attached

document. From our review we consider that; (i) there are several groups of small changes mainly restricted to the MAE / R-squared values, which do not change the message(s) conveyed by the study; (ii) there are larger changes in the expanded uncertainty estimates, including one in particular which has resulted in a 20% swing in the expanded uncertainty estimate for the NO_2 correction model, based on the unseen data (not used RF model training & validation). We note too, that this latter change, results in this model exceeding the data quality criteria used in Europe to identify suitable methods for 'supplementary assessment' by 5%, where supplementary techniques are defined as techniques used to impart additional spatial context to high-quality reference measurements taken at a single location e.g. Palmes type diffusion tubes etc.

As a result of the review, and despite the changes, we consider that the key messages of our research remain unchanged. Our reasoning is set out below;

1. A relatively simple, effective, and flexible method for improving the quality of AQ sensor data is presented and demonstrated
2. We present evidence on the scale of improvements that the research has achieved and what others might expect, broadly >90% reduction in MAE
3. Despite the model for NO_2 not achieving the data quality objectives (by 5%), the scale of improvement uncorrected vs corrected is significant and worth sharing with the AQ and sensor communities
4. Our intention in using expanded uncertainty and associated European criteria / thresholds, was to provide real-world context on the efficacy of the models developed
5. We do not imply and have refrained from recommending a particular sensor or correction method that can / can't, should / should not be used. The correction models we present will require retraining for each application, and it is expected that there will be a small variation in results achieved because of this. However, our evidence suggests that significant improvements can be achieved which may approach or exceed the European criteria. We feel this is valuable to make this research finding available by publication.
6. Co-author, Brian Stacey (BS), is the convenor of CEN TC 264 WG15 (Measurement of PM_{10} and $PM_{2.5}$). This group is responsible for constructing the method for expanded uncertainty calculation and the spreadsheet tool used in this study. BS has indicated that there are legitimate statistical reasons set out by the working group that could be used to improve the expanded uncertainty of the NO_2 correction model. These relate to the application of a random error term of ~2.85 in the calculations as they currently stand. If this term were to be removed (which could be justified), the expanded uncertainty estimate for the NO_2 correction model would reduce to ~10% which meets the target expanded uncertainty.

Our position

Based on this reasoning, our team feels that the changes do not alter sufficiently the overall message of the paper. We have not re-worked the uncertainty calculations (6) because of the extra work involved and again this does not significantly change our findings. We are, however, confident that the research presented as-is, is of high quality,

reproducible, transparent and will be useful to readers in developing a solution to sensor data quality for their own applications.

Also note that, the studies code base and data will be made publicly available, a condition of our Natural Environment Research Council funding [NE/V010360/1]. As a result, traceability and repeatability of code is of utmost importance to the study investigators.

Yours sincerely,

Dr Tony Bush,
Department of Engineering Science, University of Oxford & Apertum (Co-Investigator)

Dr Felix Leach,
Department of Engineering Science, University of Oxford (Co-Principal Investigator and Corresponding Author)

Dr Suzanne Bartington
Institute of Applied Health Research University of Birmingham (Co-Principal Investigator)

Comments on changes to the manuscript text

These comments apply to modifications in the associate Tables.

Line 19

We demonstrate improvements of between 37% and 94% in the mean absolute error term of fully corrected sensor datasets; equivalent to performance within ±2.6 ppb of the reference method for NO_2, ±4.4 µg/m3 for PM_{10} and ±2.7 µg/m3 for $PM_{2.5}$. Expanded uncertainty estimates for PM_{10} and $PM_{2.5}$ correction models are shown to meet performance criteria recommended by European air quality legislation, whilst that of the NO_2 correction model was found to be narrowly (~5%) outside of its acceptance envelope. Expanded uncertainty estimates for corrected sensor datasets not used in model training were 29%, 21% and 27% for NO_2, PM_{10} and $PM_{2.5}$ respectively.~~A mean absolute error of 2.6 ppb, 5.1 µg/m3 and 2.9 µg/m3 for NO2, PM10, and PM2.5 respectively, was achieved for the full and final corrected field-deployed sensors compared to a reference method. When used to correct data collected under environmental conditions outside model training, results meet European data quality objectives, albeit with lower accuracy than data from within the trained range.~~

These changes now ensure the text and numbers quoted in the abstract are consistent with the revised content of manuscript and main messages

Line 337

Clearly, the $PM_{2.5}$ model performs excellently in this respect with an R-squared value of 0.~~96~~ 91 and OLS slope and intercept terms approaching unity.

A 5% change in coefficient of determination, we believe this does not change message - with the help of the baseline & RF model correction, the (corrected) sensor observations explain the majority (~90%) of the variability observed in the reference method.

Line 338

The respective R-squared value for both PM_{10} and NO_2 RF models (0.~~82~~ 79 and 0.86) also indicate good model performance.

As above, but a 3% reduction in coefficient of determination. The message is unchanged, corrected sensor observations explain the good proportion of the variability observed in the reference method (~80%).

Line 349

In concentration units this equates to fully ~~-~~corrected NO_2 sensor observations within approximately ±~~1~~ 1.2 ppb of the reference observation. Similar comparisons for PM_{10} and $PM_{2.5}$ indicate ~~corrected~~fully corrected concentrations within ±0.9 µg/m^3 (PM_{10}) and 1.9 µg/m3 ($PM_{2.5}$)~~1-2~~ µg/m^3 of the reference method.

Two modifications being seen here.

Firstly, Reviewer#2 queried the use of integer values in the manuscript text, indicating it raised issues with cross referencing to table values. We have re-introduced the values at 1 d.p. to reflect the tables.

Second, we have ~0.1-0.2 (ppb / ugm-3, depending on pollutant) changes in MAE, arising from the model re-runs. From an AQ perspective this is negligible. Manuscript message is unchanged.

Line 361

The data shown are, as expected, less favourable compared with the validation set, returning higher values for the MAE metric, but for air quality context, within ~~1~~1.4 ppb (NO_2) and ~~2-3~~2.5 µg/m^3 (PM_{10}) and ~~2.9~~1.8 µg/m^3 ($PM_{2.5}$) ~~-~~of the MAE returned by the model validation set (Tables 4 and 5).

As above, we observe the effect of d.p. change & a small model performance change. The difference between the validation & unseen MAE metrics, which is presented here, is consistent with that of the previous version of the manuscript. Message unchanged.

Line 377

Improvements in MAE attributable to the RF model in the range of 37-94% are shown; equivalent to ~~corrected~~fully corrected observation within, on average approximately ±~~3~~ 2.6 ppb of the reference method for NO_2, ±~~5~~ 4.4 µg/m^3 for PM_{10} and ±~~3~~ 2.7 µg/m^3 for $PM_{2.5}$.

As above, we observe the effect of d.p. change & an small MAE change, which is consistent with that presented at integer level in the previous version of the manuscript. Message unchanged.

Line 382

Section 3.2.4 has been thoroughly recast to respond to reviewer#2 comments.

Line 388

Table 6 presents expanded uncertainty estimates associated with fully corrected sensor data from the validation dataset, (data not used in the RFR model training) and shows that these data for all pollutants perform well against the target expanded uncertainty criteria recommended by European legislation, (expanded uncertainties of 21%, 40% and 19% respectively for NO_2, PM_{10} and $PM_{2.5}$).

6% & 1% increase in expanded uncertainty of the validation set for PM10 & PM2.5 respectively, even so the values are within the data quality objective thresholds.

Line 391

The result of this further correction is presented in Table 6 as the 'full and final correction'. Expanded uncertainty estimates for the validation set with full and final corrections applied were 17%, 15% and 12% for NO_2, PM_{10} and $PM_{2.5}$ respectively.

2-3% increase in coefficient of determination for PM, NO2 increased by 13% but still within data quality objectives. Messaging unchanged.

Line 401

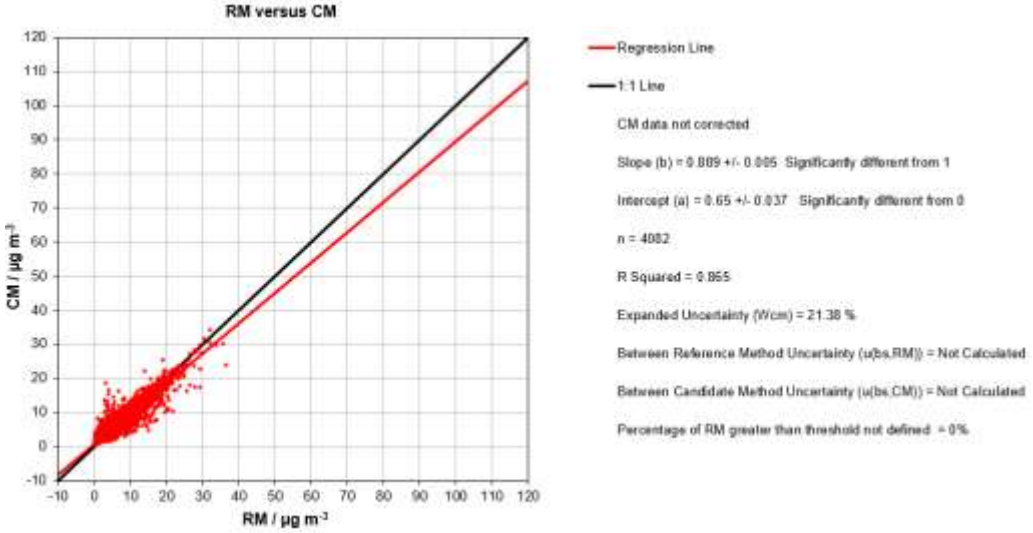
Table 7 presents these data for fully corrected sensor observations from December 2020. Table 7 shows the expanded uncertainty estimates for fully corrected unseen sensor data of 29%, 21% and 27% respectively for NO_2, PM_{10} and $PM_{2.5}$ are returned.

8% increase in expanded uncertainty (NO2) & reduction for PM10 & PM2.5 of 13% & 2% respectively relative to previous version. PM values remain within the data quality objectives. The expanded uncertainty associated with the NO2 correction model we agree is now outside data quality objectives we quote. However, we feel that the message we wish to convey remains the same - the scale of improvement relative to uncorrected sensor is good & the utility of this relatively simple approach to reducing sensor data uncertainty to approximately acceptable levels is of benefit to AQ sensor community. Our position is supported by a regression analysis present as track

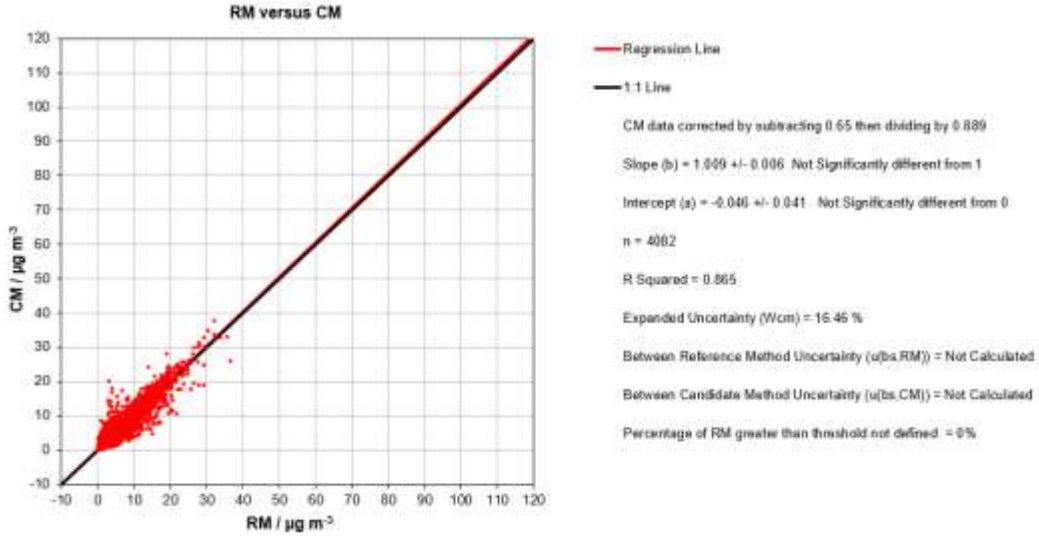
changes (in this document only). The 4 additional plot show for context the relationships between corrected sensor & reference methods for validation & unseen data sets under uncorrected & corrected slope & intercept conditions. In these plots all of which fall outside of the target thresholds we see that the relationship is generally really quite good,

We also maintain that with continued training the expanded uncertainty would improve - the electrochemical sensors used for NO2 are very (more) sensitive to T & RH interference than the OPCs used for PM. The ability of the models to cope with variation its model features to

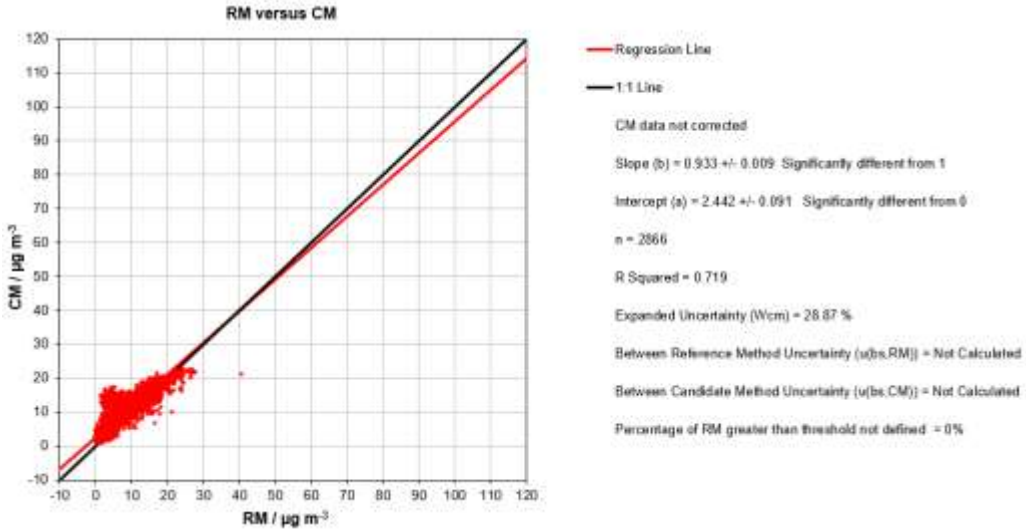
Relationship between corrected sensor & reference method- validation set, no slope & intercept correction



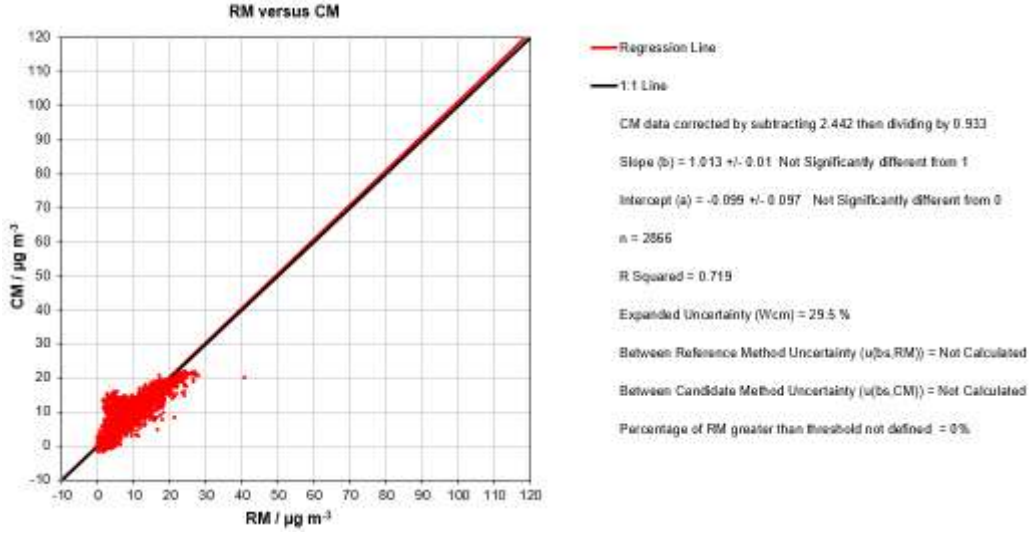
Relationship between corrected sensor & reference method- validation set, with slope & intercept correction



Relationship between corrected sensor & reference method – unseen dataset, no slope & intercept correction



Relationship between corrected sensor & reference method – unseen dataset, with slope & intercept correction



Line 402

Further corrections, for slope and intercept terms, had negligible change on these estimates, (30%, 25% and 28% expanded uncertainty respectively for NO_2, PM_{10} and $PM_{2.5}$).

As you indicate, increases in expanded uncertainty across the board. Increases are (now) marginal relative to the validation set.