Author's response to reviewer and editor comments

We are thankful for the constructive additional comments. Below we list all referee remarks and suggestions (in *italics)* along with our responses.

Editor Comments

Editor comment: *in line 43, it would be more correct to speak about NOx emissions, not NO2 emissions.*
Author's response: Thanks for pointing this out, we changed the text accordingly in the revised version of the manuscript.

Editor comment: *in the caption of Figure 5, it may be worthwhile to remind the readers that the 2019 data were part of the training data set.*
Author's response: We added this information to the caption of Figure 5.

Anonymous Referee #1

Reviewer comment: *The revised version has been substantially improved but I must say I am still a bit confused about the methodology implemented for training and evaluating the machine learning models.*
First, the authors are not performing any tuning of their models while this may substantially improve the performance of the predictions. Rather, they are using the default hyper-parameters. Why so? This does not follow the good practices of the field. Is this choice made for computational reasons?
Author's response: Performing hyperparameter tuning across all sites would indeed not be possible due to computational constraints. However, we did perform hyperparameter sensitivity tests at a handful of sites (grid search) and found only marginal improvements in performance. Due to this, we decided to stick with the default XGBoost model parameters. For clarification, we added the following sentence to section 2.3.1 of the revised version of the manuscript:

"The design of the XGBoost framework is determined by a set of hyperparameters, such as the learning rate, maximum tree depth, or minimum loss reduction. While a full hyperparameter optimization across all sites - e.g., by using a grid search approach – would be computationally prohibitive, we conducted hyperparameter sensitivity tests at few selected sites and found that the XGBoost performance only improved marginally at these sites when using other hyperparameter than the model defaults (less than 5% improvement). In addition, we found that the sites respond differently to the same change in hyperparameter setup, suggesting that there is no uniform hyperparameter design that is optimal across all sites. Based on this, we chose to use the default XGBoost model parameters at all locations, with a learning rate of 0.3, minimum loss reduction of 0, maximum tree depth of 6, and L1 and L2 regularization terms of 0 and 1, respectively."

Reviewer comment: *Secondly, cross-validation can be used for two different purposes : (1) for tuning the ML model, and/or (2) for estimating the performance of the final model. Given that no*

tuning is performed, I understand the authors are thus using cross-validation here only to estimate the performance of their models. Then, regarding their strategy, at each station, the authors are training 8 different models (model M[8] trained on X[1,2,3,4,5,6,7] and tested on X[8], model M[2] trained on X[1,3,4,5,6,7,9] and tested on X[2], etc.), which should give them 8 values of RMSE (computed on X[8], X[2], etc., respectively) or any other statistical metric they are interested in. A simple and relatively robust approach to estimate the (test) performance of their predictions would consist in computing the corresponding average RMSE (ideally providing also the standard deviation). Which average RMSEs are obtained following this simple approach? Eventually, another approach could be to first gather all the test subsets on which predictions are made (X[8], X[2], etc.) and compute the overall RMSE. Any of these approaches would provide an estimate of the performance of their predictions. Then, in a second step, in order to get the best possible final ML model, a last ML model (to be used to make predictions in 2020) could be trained using the entire 2018-2019 dataset in order to take benefit from the largest possible dataset during the training phase. The performance previously estimated could be used as a conservative estimate of the performance of this final model ("conservative" because this final model may perform slightly better than the 8 models previously evaluated given that it has been trained on a slightly larger dataset).
Rather, for a reason I don't really understand, the authors are finally considering a new model that is the average of the 8 models initially trained ("Once trained, the final model prediction at each location consists of the average prediction of the eight models."), which sounds strange to me. Then, in order to estimate the performance of this final model, the authors are "[omitting] the center week of each training segment from the 8-fold cross validation and use it for testing only". Why one week? All this part of the methodology seems a bit "baroque" to me, both for evaluating a ML model and for taking into account the auto-correlation. Regarding the auto-correlation, considering a 8-fold cross-validation is already an improvement compared to the random splitting proposed in the first version of the manuscript. I do not really understand why the authors then need to left apart only one week for testing.
These different aspects of the methodology should be clarified and eventually corrected. The choices made need to be comprehensively described and justified, ideally following the good practices in the field of machine learning.
Author's response: Following the reviewer's suggestion, we recalculated the model skill scores using the left out segment from the 8-fold cross validation as test segment. The methodology description in section 2.3.1 and 2.3.3 has been updated accordingly. The updated skill score values are almost identical to the previous ones.

Reviewer comment: *About the estimation of the uncertainties (Section 2.3.4), the authors are computing the uncertainties as the standard deviation of the model-observation residuals. One potential issue I see here is that they are assuming implicitly that individual ML models do not have any bias, which is roughly true when averaging all models at all stations, but not at individual stations where NMB ranges between -20 and +10% roughly (Fig. 3). As an illustration, if we consider an hypothetical ML model that would represent perfectly the observations but with a 1 ppbv (systematic) bias. In this case, the residuals all worth 1 ppbv, and the corresponding standard deviation is thus zero. So this model would be considered as perfect while it is not.*
Then, another aspect is how to translate uncertainties estimated for hourly predictions at a given individual station to uncertainties over a longer period (7 days for instance) and entire country.

While it is likely reasonable to consider that predictions on longer time scales are reduced due to error compensations, it might not be always and fully the case on the spatial dimension on which model-observation errors might be at least partly correlated to each other. Consider for instance a set of 2 stations located close to each other. The concentrations observed at these stations might be quite well correlated to each other given the short distance separating them, as well as the ML predictions given the fact that the input variables used are taken from a geophysical model at 25x25 km resolution. Therefore, the way I understand it, the model-observation residuals at these 2 stations might not fully compensate each other, while the authors implicitly assume so. As a consequence, the uncertainties affecting the combination of these two stations would be reduced by a factor of 1.4 (=2^0.5), which might be overly optimistic, as might also be the uncertainties close to zero mechanically obtained in countries with numerous stations, as shown in Fig. 5. I think this should be further discussed, and the assumptions used to estimate the uncertainties should be clarified.

Author's response: In the revised version of the manuscript, we recalculated the uncertainties based on the model-observation comparisons from the 8 test segments obtained from the 8 fold cross validation. Also, to clarify the assumptions that go into our uncertainty estimation, we added the following paragraph to section 2.3.4 of the revised version of the manuscript:

"This assumes that the errors across individual sites are uncorrelated, which they often are given the very local nature of the bias correction models. In addition, our uncertainty calculation also implies that the aggregated mean error approaches zero. Given that the average mean biases of the machine learning models are clustered around zero (Fig. 2 and Fig. 3), this is a valid general assumption - especially when aggregating across multiple sites. However, it might lead to overly optimistic uncertainty estimates for sites with a relatively large mean bias of 10% or higher."