Responses to the referee comments: Assessing the climate and air quality effects of future aerosol mitigation in India using a global climate model combined with statistical downscaling

Miinalainen, T., Kokkola, H., Lipponen, A., Hyvärinen, A.-P., Soni, V. K., Lehtinen, K. E. J., and Kühn, T.: Assessing the climate and air quality effects of future aerosol mitigation in India using a global climate model combined with statistical downscaling, Atmos. Chem. Phys. Discuss. [preprint], https://doi.org/10.5194/acp-2022-513, in review, 2022.

We thank Anonymous Referees #2 and #3 and the Editor for the good comments that helped us to improve our manuscript. Our responses are written below each comment separately. The referee comments are marked with *yellow color and italic*, and the author replies are marked with gray color. The original manuscript text is marked with pink color, and updated text with dark magenta. The line numbers refer to the 1st revised, submitted version of the manuscript which was peer-reviewed.

Replies to the comments made by the Editor:

"[...] In addition to these, I found some very small technical questions. Firstly, what are Aprc and Aprl in Figure 1 - I couldn't quite see which input variable from Table 3. Are they the precipitation variables? Perhaps putting that abbreviation in Table 3 could be helpful. The others were clear to me. Also, in Figure 1a there isn't a dark blue line in the legend, rather it shows up gray to me. Finally, winter and summer months are defined in line 473, but you talk about winter and summer before that. It may help to define the months the first time you discuss the seasons."

We thank the Editor for carefully reading our manuscript and for the suggestions how to improve the readability of our manuscript. We have made the following changes:

*Table 1 is now updated as suggested to describe the abbreviations that will be used later in Figures 1 and 2.

*Figure 1a has been updated by changing the shade of blue for the lines that represent the individual RF-corrected PM2.5 values.

*We have modified the manuscript text to mention explicitly the months that we are referring to when discussing summer and winter seasons.

Furthermore, we noticed that there was some missing information about the modeling setup used for the global model simulations. We added the following sentences after line 116:

"In addition, we used an additional setup where the emitted BC was assumed to get directly internally mixed with sulfate (Holopainen et al., 2020). Therefore, BC-containing particles were modeled to be more soluble already at the time of their emission."

****

Replies to the comments made by the Anonymous Referee #2:

The manuscript has improved substantially after the revision and the effort of the authors in addressing all the aspects of the review is appreciable. The second part of the work which examines the impact of future emission scenarios in terms of radiative forcing is now well-discussed with proper references. While all seem perfect, a few concerns/queries regarding the ML technique based on the responses to the reviewer comments are still pending which are briefed below.

We thank Anonymous Referee #2 for the comprehensive review and for the valuable comments. The more detailed replies are listed under each comment separately

• *Why this model is 'not at all sensitive or unaffected' to the input parameters (Fig./Table1 for the 2nd reviewer) as well as hyper-parameters (Fig. 2 for the 2nd reviewer)? The correlation coefficient values are expected to change if the parameters have an association with the target values, but there is no change. According to the 2nd figure, the model appears insensitive to the max depth feature. The model can acquire all the necessary information by the initial split itself, which seems unrealistic.*

We repeated the test that was conducted for the previous paper revision and where we altered the maximum depth parameter. This time we changed the max_depth parameter from 1 to 5, and used again the year 2020 data (i.e., outside of training and testing data). The error statistics for the repeated test are listed in the table below.

max_depth	1	2	3	4	5
RMSE (µg/m^3)	58.05	55.03	53.75	53.31	53.01
MRE (%)	71.81	66.58	63.49	62.81	61.86
MAE (µg/m^3)	45.91	42.97	41.55	41.08	40.64
R^2	0.46	0.51	0.53	0.54	0.55
Pearson correlation	0.80	0.80	0.80	0.80	0.80

The results clearly show that the model improves with max_depth being increased. At some point the model's maximum capacity is reached and accuracy metrics do not improve anymore. This is the case in the earlier reply to referee and the accuracy metric was not changing notably

regadless of e.g. different max_depths. In many machine learning models, such as neural networks, overfitting may be a significant problem. If the capacity of the machine learning model is increased too much it may lead to overfitting and the accuracy metrics of the test data decreases. Here we have shown that our Random Forest model does not suffer from overfitting and produces the best possible results based on our training data given the capacity of the model is large enough. In our case, the capacity leading to optimal performance is achieved already with a relatively low number of max_depth.

Note that here the error statistics differ from the error statistics presented in the manuscript. This is because we used here data from the year 2020, whereas in the manuscript the testing data is from the years 2018 and 2019.

Furthermore, error statistics are calculated based on the average over all corrections and daily average values of the measurement stations. This means that slight changes in one RF-correction do not affect the outcome as heavily as it would be the case if analyzing the outcome of one single RF model. All 31 RF models use the same hyperparameters, and there is no individual tuning for the RF model parameters.

• *L467-469 in the track-changed version: These two sentences are mutually contradicting. As per the first sentence, 'the feature importance value indicates the contribution of a feature to the total reduction in the error criterion', then, what is meant by the second sentence- 'Feature importance values do not reveal the sensitivity of the RF model to specific input feature'? Please clarify/modify.*

Thank you for bringing to our knowledge that this part was not clear in the manuscript text. The error criteria used in our model is the mean squared error, and the feature importance values indicate the contribution of a feature to the total reduction in the error. However, the reduction in error does not necessarily describe the total effect of a feature on the model prediction. Some features might affect, for instance, summertime values by increasing the magnitude of the estimate, but still the net error might be of the same magnitude as without the feature.
We have updated in the lines 413-414 the sentence from:
"However, importance values do not reveal the sensitivity of the RF model to specific input features." to
"However, importance values do not reveal the sensitivity of the RF model to specific input features as the reduction in error does not indicate directly how much a feature affects the RF model output trends and magnitude."

• *How important/necessary is normalization in Random Forest which is a tree-based model and not a 'Neural Network Model'?*

Input normalization is not important in training a Random Forest model. The splitting of the data in the construction of the regression trees is based on the ordering of the input variable.

The input variable ordering is not affected by the normalization and thus the input normalization is not important in Random Forests. Therefore, we did not normalize the input data used in our RF corrections. However, the feature importance values, which are an output of the RF training procedure, were normalized to be in a scale from 0 to 1.

• *When the authors state that bagging is not used, does it mean that the entire dataset is fed into the model rather than in short batches? Not able to understand the term 'Bootstrap bagging'.*

Since the bagging is not used, the whole data set is used for building each tree. However, the parameter "max_features" controls the number of features used in each split. This way, there are differences in the trees and variation in the output of different trees. Bagging refers to conducting model training with bootstrap samples many times. As mentioned in the previous revision comments, bootstrap sampling was not used in our analysis.

****

Replies to the comments made by the #Anonymous referee 3:

General remarks: The manuscript deals with aerosol near surface concentrations from a GCM, downscaled with a random forest approach over Dehli (India). The downscaling correction substantially improves the GCM performance in much better agreement with the observations. Therefore, the authors show the potential of the method. After a few minor corrections, the paper should be accepted for publication in my opinion.

We thank Anonymous Referee #3 for the excellent suggestions and for dedicating time to go through our manuscript. The detailed answers are listed below under each comment.

Content:
It is unclear how the station values are obtained from the GCM simulations, i.e. is it nearest neighbour or linear or cubic interpolation. How are potential altitude misrepresentations considered in determining the training values from the GCM at the station locations?

Thank you for making us aware that this aspect was not mentioned in the manuscript text. The data for New Delhi from ECHAM-HAMMOZ was retrieved by using nearest neighbor mapping. In practice, this was done by using the CDO program method "remapnn" and using the lowest model level output data. It is true that for some of the stations, the altitude of a station might be slightly higher than what the lowest model level represents. However, we estimated that in New Delhi the stations are located mostly near ground level, and this would not cause a significant error in our modeling studies. Furthermore, the aim of this study was to obtain one average PM2.5 value for the New Delhi urban region, which would be representative of whole

area. Therefore, in order to minimize the order of complexity of the method, all RF input feature data was retrieved from the same vertical model layer, and for the same latitude and longitude coordinates.

We have modified the sentence in line 272 to be from:

"For some input features, we used values representing one grid box surrounding the New Delhi region (point)." to
"For some input features, we used values representing one grid box surrounding the New Delhi region (point). These were retrieved from ECHAM-HAMMOZ data by using nearest neighbor interpolation."

How can the R^2 value in table 4 be negative for the uncorrected output?
The R-squared value is computed as $R^2 = 1 - SS_{res} / SS_{tot}$, where SS_{res} is the sum of squared residual errors between the modelled and measured values, and SS_{tot} is the sum of variance in the modelled data. If the modelled values do not follow the trends of measured data, the R-squared values can be negative. The R-squared value does not represent the squared Pearson correlation value, though this might be understood from Table 4 as we had Pearson correlation also marked with the letter R. Therefore, we have modified Table 4 to explicitly mention Pearson correlation and removed the abbreviation "R" to avoid confusion with the R-squared.

Can it be estimated, how large is the impact of the ML based correction on the surface concentrations on the total forcing? Even though there is the difficulty of estimating the effect of near surface concentrations on TOA forcing, the question whether the surface values have a large impact on the total forcing and therefore the correction would be also beneficial for the radiative forcing corrections. A short discussion on this topic should be added to the manuscript.

This was a good question. In principle, what the referee suggests could be achieved, but there are several to keep in mind: The corrected PM2.5 values represent very local PM2.5 values for New Delhi urban region, and do not represent the whole grid box. Therefore, to make better radiative forcing predictions for one grid box, calculations would have to be done on a sub-grid scale, including much more observational data spanning the entire grid box (for ECHAM-HAMMOZ the mentioned 2°x2°). However, TOA radiative fluxes are analyzed for larger areas, and not only one grid box as it would not be very representative if considering the energy fluxes over longer term periods. Therefore, the estimation of changes in radiative fluxes for larger areas would require even more station PM2.5 data, from various locations in India, and this would require separate RF models for each of the stations. In addition, as radiative forcing calculations are performed online, the RF correction would need to be conducted dynamically as well, i.e., during the ECHAM-HAMMOZ simulation, in order for the surface concentrations to affect TOA radiative fluxes. Furthermore, we here only correct for PM2.5, which is the integral over the aerosol size distribution up to 2.5 µm. For radiation calculations, on the other hand, the aerosol size information must be preserved, as ARI strongly varies with aerosol particle size.

This exercise, though very interesting, would be computationally quite demanding, and therefore is out of the scope of this study.

We have added after Line 596 the following:
"As a continuation to this study, one could apply the downscaling method described here during the ECHAM-HAMMOZ simulation (instead of after, as it was done here) and for larger areas. With some extensions which address additional aspects like, e.g., information about aerosol size, this may allow the bias correction to also affect the computation of other aerosol impacts, like radiative fluxes. However, such a dynamical approach for larger areas would also require a much larger spatial coverage of observational data, which would make the model computationally more demanding. Furthermore, without proper evaluation, such a model extension might introduce further uncertainties to the radiative forcing estimation."

line 148: specie -> species
Thanks for the suggestion, we have corrected this typing error. We have changed in line 148:
"the values for each grid box and emission specie from year 2015 to year 2020." to
"the values for each grid box and emission species from year 2015 to year 2020."