We would like to thank the reviewer for the comments. Below, we address the comments by the reviewer. The reviewer comments are typed in bold font and our replies to them in regular font. To help the reviewer, we also list some parts of the revised manuscript in our replies and these parts are typed in italic font or with quotation marks for small comments.

In addition to corrections according to referee comments, we have also made some minor (e.g. typos, grammar) related changes to the manuscript.

Referee 2

This article applied a previous developed concept of using machine learning (ML) to bias-correct aerosol optical depth (AOD) and other aerosol data from conventional aerosol product. Original concept of ML post-processing of satellite data against ground truth is introduced in author's previous journal articles. This time a feed forward neural network is used on Sentinel-3 data to produce two aerosol products: machine learning generated aerosol data and bias-corrected Level-2 Synergy Product. The article claims that the post-process corrected the Sentinel-3 synergy product is a high resolution, better accuracy data products than the original aerosol product and the aerosol product generated from pure FFNN model. Within resent decade, machine learning has been rapidly applied to Earth Science field. One of doubtfulness of relying on ML is that the approach is not based on physics. The idea of machine learning post-process include both the state of art machine learning technique and traditional algorithm-based approach, which maintain the physics within the retrieval process. It is a conservative way of using ML and if successful, can be applied to many fields. However, the statement of the post-process corrected aerosol data has higher accuracy than full ML predicted aerosol data is not convincing, especially in terms of AOD. Figure 4, 5, and 6 all show comparisons between these two products. There is no significant improvement from post-process corrected product to full machine learning output. Although the error statistics against AERONET are slightly better in post-process corrected data, when investigate details in Figure 4 we can see that the overestimation of AOD especially at AOD < 0.2, is amplified in post-process corrected data than fully learned regressor model output. The smaller bias statistics in post-processed product is balanced by the overestimation in low AOD regime (AOD < 0.2) and underestimation in high AOD regime (AOD > 0.5). If we look at other evaluation plots, such as error histogram or error diagnostic plot. We may have much better look at the error distribution of two data sets. Similarly for AE comparisons, it is hard to say that the accuracy of AE prediction is improved between the two ML-involved products.

We thank the referee for the careful evaluation of our manuscript and the comments.

We kindly disagree with the referee's statement "There is no significant improvement from post-process corrected product to full machine learning output." At first, the absolute improvements may not seem significant. However, the relative improvement, for example, in AOD at 550 nm is significant (R^2 improves by about 9%, RMSE is about 8% smaller, and BIAS decreases by 20% in post-process corrected model when compared to fully learned model). In some applications, such as data assimilation, these relative improvements may be significant for the accuracy of the data assimilation model.

The referee also claims that "…when investigate details in Figure 4 we can see that the overestimation of AOD especially at AOD < 0.2, is amplified in post-process corrected data than fully learned regressor model output.". This claim is not true. The biases for AERONET AOD smaller than 0.2 and larger than 0.5 are shown in the tables 1 and 2 below. The post-process corrected AOD has the best bias metric for all wavelengths (best model shared with the fully learned model in 3 cases) and thus the data does not support the referee's claim.

Table 1. AOD biases corresponding to data points with AERONET AOD smaller than 0.2. The gray background indicates the best-performing model.

Wavelength	Synergy AOD bias	Fully learned AOD bias	Post-process corrected AOD bias
440 nm	0.380	0.011	0.011
500 nm	0.333	0.010	0.010
550 nm	0.303	0.010	0.009
675 nm	0.249	0.008	0.008
870 nm	0.188	0.007	0.006

Table 2. AOD biases corresponding to data points with AERONET AOD larger than 0.5. The gray background indicates the best-performing model.

Wavelength	Synergy AOD bias	Fully learned AOD bias	Post-process corrected AOD bias
440 nm	0.484	-0.294	-0.271
500 nm	0.417	-0.267	-0.245
550 nm	0.379	-0.243	-0.222
675 nm	0.299	-0.194	-0.175
870 nm	0.247	-0.137	-0.122

We added the following paragraph to the results section:

> To evaluate the models' performance in low and high AOD conditions, we evaluated the results corresponding to AERONET AOD at 550 nm smaller than 0.2 and larger than 0.5. The results are shown in Table 1 [of the manuscript]. The post-process corrected model results in the best bias metric in both low and high AOD conditions. In addition, the post-process corrected model results in the best R^2 in low AOD and the best RMSE in high AOD conditions. The fully learned model results in about 4 % lower RMSE than the post-process corrected model in small AOD. The Synergy R^2 is the best for the high AOD cases but there are only 163 samples in the high AOD cases so more data would be needed for more reliable evaluation of the models in high AOD conditions.

We also added the following table of the results for low and high AOD in the manuscript:

Table 1. Error metrics for the satellite data product AOD at 550 nm corresponding to small (<0.2) and large (>0.5) AERONET AOD. The bold font indicates the best performing model.

AOD 550 nm < 0.2 (N=4708)			
Metric	Synergy	Fully learned	Post-process corrected
R^2	0.113	0.270	**0.310**
RMSE	0.412	**0.050**	0.052
Bias	0.303	0.010	**0.009**

AOD 550 nm > 0.5 (N=163)			
Metric	Synergy	Fully learned	Post-process corrected
R^2	**0.497**	0.273	0.377
RMSE	0.433	0.313	**0.279**
Bias	0.379	-0.243	**-0.222**

Other specific comments are:

Line 27, atmospheric spelled wrong.

Corrected.

Line 67 remove "accurate"

Removed.

Line 107 In section ? missing a number.

Corrected. "In section 2,…"

Line 190 please specific list the time/spatial criteria for collocation.

The temporal and spatial collocation is now better described. The sentence citing Petrenko et al. (2012) was revised to: "We use the same ±30 minutes temporal thresholds for the collocation procedure as in Petrenko et al. (2012) and spatial collocation radius of 5 km."

Line 197-198 Can random split for each region result in data from a few sites dominate the results for one region?

We have tested how the random split affects the results by running the analyses with multiple different random splits. As there are quite many stations in each region of interest there are no single station that would dominate the results and therefore different random splits do not significantly change the results. To show this result to the readers we have added the following sentences to the manuscript: "To study the effect of randomness on the splits of AERONET stations, we tested our approach with multiple random splits. We did not observe significant differences in the results between different random splits of the AERONET stations."

Line 211-212 Regarding normalization method. If we use all data mean/std to do the z-score standardization, all the data is converted equally still within the same scale as they are originally. What is the point of normalization? For fill data, what average is used? and how much missing data is there?

The normalization is often used in machine learning to ensure we do not run into numerical problems due to input values of different orders of magnitudes. Large differences in the values of the data may cause numerical problems in the training or evaluation of the neural network. This is the reason we carried out the normalization.

The missing values were filled with the average value of the corresponding variable in the training data set (in the manuscript: "In case some of the inputs contains a missing value, it is filled with the average value of the training dataset.")

Most of the missing values were due to different swath widths of OLCI, SLSTR nadir and oblique views. On average, there were about 8 % and 6 % missing values in the fully learned model and post-process correction model datasets, respectively. We added the following sentence to the manuscript: "On average the input data of the fully learned and post-process correction models contained about 8 % and 6 % of missing values, respectively."

Section 3.5 What is the accuracy for the two-folds testing results for training/testing/validation datasets?

As mentioned in the manuscript we have split the data into two sets by random selection of AERONET stations. In the evaluation, the models are always trained using the other set and evaluated using the other. In the training of the neural network models, we use, according to proper machine learning practices, early truncation based on monitoring of the validation loss to avoid overfitting. The accuracy metrics for the data computed using the models trained on the same data are significantly better than the ones computed for the stations not included in the training data. This is expected and well-known behavior in machine learning and should be avoided. We think it is not informative to present the overoptimistic results that contain evaluation data corresponding to models trained with same data. To get an idea of this type of evaluation results for AOD at 550 nm obtained with models trained on the same data see the figure below.

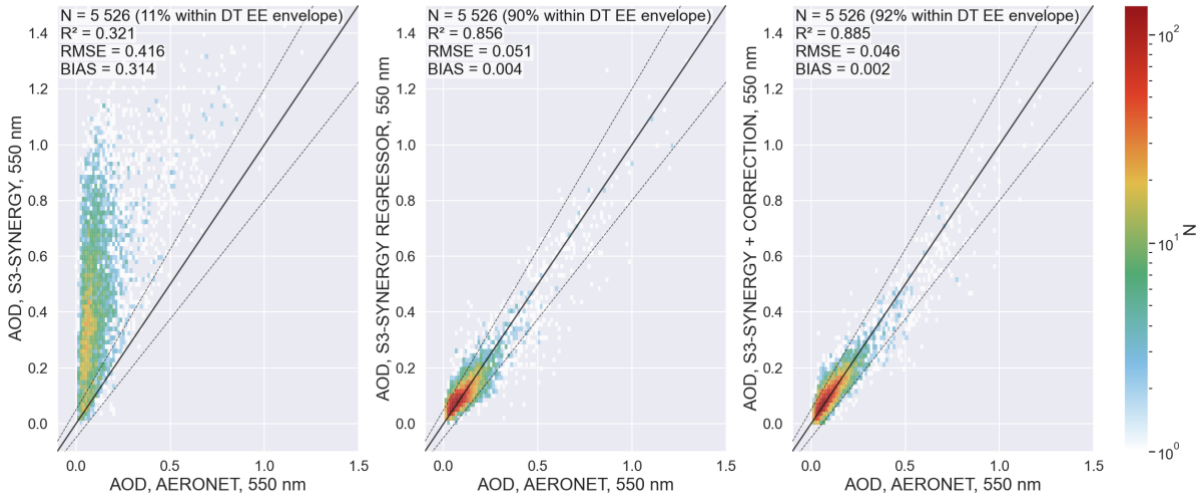

Figure 2. AOD (550nm). Left: Sentinel-3 level-2 Synergy product. Middle: Fully learned regressor model trained. Right: Post-process correction model. Please note the models have been evaluated using the training datasets and thus do not represent the true error metric values.